UAB
Universitat Autònoma de Barcelona

Métodos Cuantitativos para la Investigación de Mercado y el Marketing

Universitat Autònoma de Barcelona

Giuseppe Lamberti

Contents

Estadística descriptiva

1. Introducción

La estadística descriptiva es una herramienta clave en la investigación de mercado y marketing, ya que permite a las empresas entender los datos obtenidos de encuestas, estudios de clientes, ventas, entre otros. A través de la estadística descriptiva, podemos resumir y presentar datos de forma clara y sencilla, lo que facilita la toma de decisiones basadas en evidencia.

Por ejemplo, si una empresa desea lanzar un nuevo producto, la estadística descriptiva puede ayudar a identificar las preferencias de los consumidores a través de encuestas. Asimismo, puede identificar tendencias y patrones en los datos de ventas históricos, como la estacionalidad o el comportamiento del consumidor.

2. ¿A que sirve?

La **estadística descriptiva** ofrece una serie de beneficios clave en el análisis de datos, especialmente en la investigación de mercados y marketing. Algunas de sus principales **ventajas y usos** incluyen:

- **Resumir grandes volúmenes de datos**: Facilita la comprensión de conjuntos de datos extensos, al sintetizar la información en medidas simples como medias, medianas y desviaciones estándar.
- **Identificar tendencias y patrones**: Ayuda a detectar comportamientos recurrentes, estacionalidades o tendencias de consumo, proporcionando información valiosa para la toma de decisiones estratégicas.
- **Facilitar la comparación de grupos**: Mediante gráficos y tablas, permite comparar fácilmente diferentes segmentos de mercado o subgrupos de datos.
- **Visualización clara**: Los gráficos como histogramas, boxplots y diagramas de barras simplifican la interpretación de los datos, haciendo que los resultados sean más accesibles para stakeholders no técnicos.
- **Identificación y comparación de perfiles**: Facilita la caracterización de diferentes perfiles de clientes, como consumidores frecuentes o compradores esporádicos, lo que permite adaptar estrategias comerciales a sus necesidades específicas.
- **Soporte para la toma de decisiones**: Al proporcionar una visión clara y objetiva de los datos, la estadística descriptiva permite a las empresas fundamentar sus decisiones en evidencia, reduciendo la incertidumbre y el riesgo.

Estos beneficios hacen que la estadística descriptiva sea una herramienta esencial para analizar y gestionar información de manera eficiente.

3. Diferencia entre las variables numericas y categoricas.

Al realizar análisis estadístico descriptivo, es crucial **diferenciar entre variables numéricas y categóricas**, ya que el tipo de variable determina las técnicas y herramientas apropiadas a emplear. Las **variables numéricas** (o cuantitativas), como los ingresos o las calificaciones, permiten el uso

de medidas de tendencia central (media, mediana) y dispersión (desviación estándar) para resumir los datos. Además, se pueden aplicar gráficos como histogramas y boxplots para visualizar su distribución.

Por otro lado, las **variables categóricas** (o cualitativas), como el género o las preferencias de color, requieren el uso de tablas de frecuencia y gráficos como diagramas de barras para representar el número de ocurrencias en cada categoría. No es posible aplicar operaciones aritméticas directas en este tipo de variables, por lo que es fundamental utilizar las técnicas adecuadas para evitar interpretaciones erróneas.

Distinguir correctamente entre estos dos tipos de variables garantiza que los análisis estadísticos sean precisos y que las conclusiones extraídas sean relevantes y aplicables.

4. Estadistica descriptiva univariante

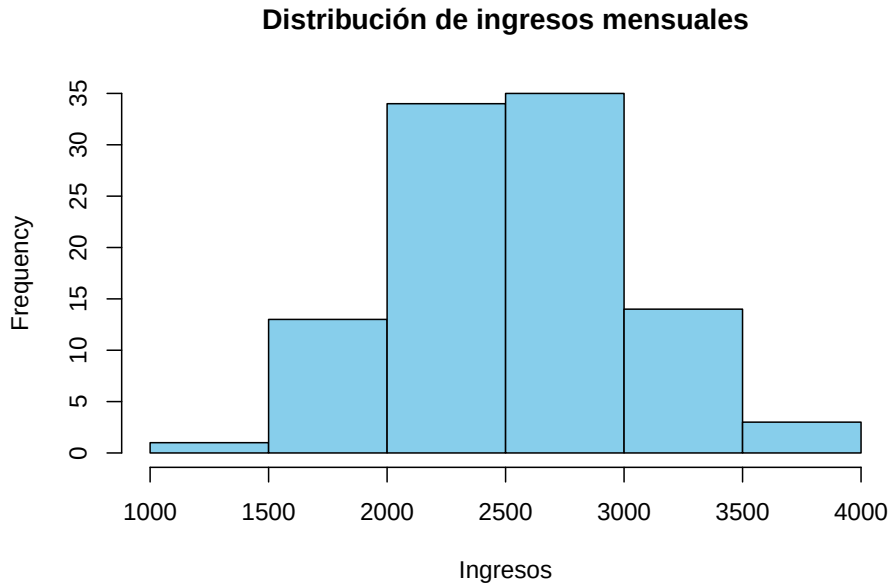
La **estadística descriptiva univariante** es un tipo de análisis que se enfoca en describir y resumir las características de una sola variable dentro de un conjunto de datos. Su objetivo es proporcionar una comprensión clara de cómo se distribuyen los valores de esa variable, sin tener en cuenta relaciones con otras variables. Para lograr esto, se utilizan diversas herramientas como las **medidas de tendencia central** (media, mediana, moda), que permiten identificar el valor promedio o típico de los datos, y las **medidas de dispersión** (rango, desviación estándar, varianza), que indican el grado de variabilidad en los datos. Además, se emplean gráficos como histogramas y boxplots para visualizar la distribución de la variable. La estadística descriptiva univariante es fundamental para obtener una primera impresión de los datos antes de realizar análisis más complejos o inferenciales.

4.1 Principales gráficos y medidas para variables numericas (analisis univariante)

Histograma

El histograma es un gráfico que representa la distribución de una variable cuantitativa continua. Permite visualizar la frecuencia de los diferentes intervalos de valores.

Supongamos que una empresa ha recogido datos sobre los ingresos mensuales de sus clientes. Un histograma podría ayudar a visualizar cómo se distribuyen esos ingresos.

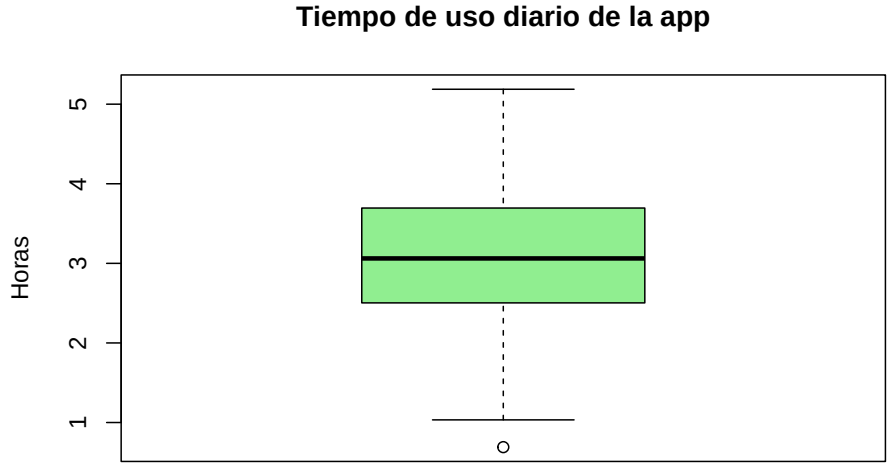


Interpretación: El histograma muestra la distribución de los ingresos mensuales de los clientes. En este caso, podemos ver que la mayoría de los clientes tienen ingresos en el rango de 2000 a 3000 euros.

Gráfico de caja (Boxplot)

El gráfico de caja es una representación gráfica que resume la distribución de una variable cuantitativa mostrando la mediana, los cuartiles y los valores atípicos. Se podría considerar como una fotografía de la distribución de una variable mediante histograma realizada desde arriba.

Una empresa de tecnología quiere analizar el tiempo de uso diario de su app por parte de sus usuarios.



Interpretación: El boxplot muestra que la mediana de tiempo de uso es de 3 horas, indicando que el 50% de usuarios pasan 3 horas o menos utilizando la app. La distribución e simentrica. El gráfico también indica que hay un usuario que usa la app menos de un hora al dia, hecho que representa una valor anómalo respecto al uso del resto de usuarios.

Medidas de tendencia central y dispersión

El promedio (media) es una medida que indica el valor medio de una distribución. El mínimo y el máximo son los valores extremos de los datos, mientras que la desviación estándar mide la dispersión de los datos respecto a la media.

Supongamos que una tienda en línea quiere analizar las calificaciones de sus productos.

promedio	min	max	sd
4.35	3.80	5.00	0.47

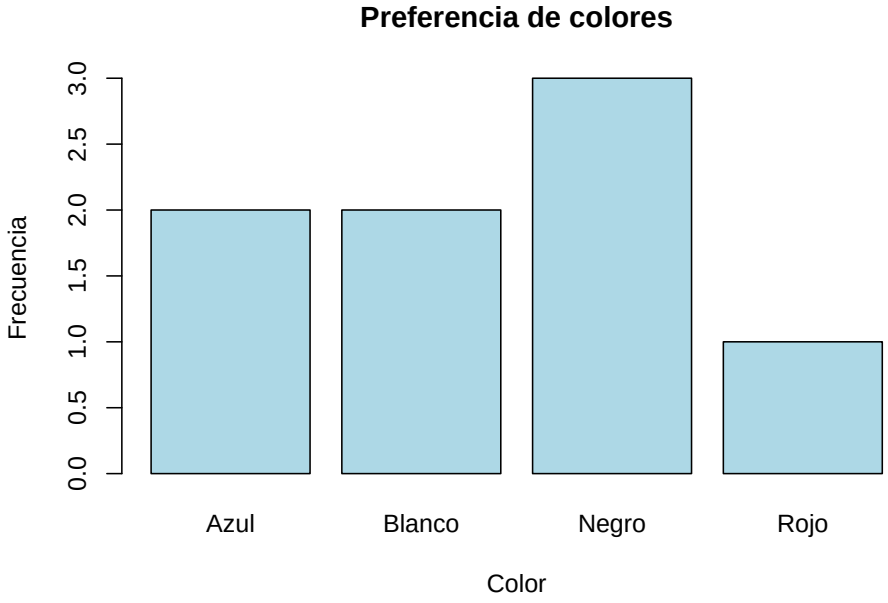
Interpretación: La media de las calificaciones es 4.35, lo que indica que, en promedio, los clientes están bastante satisfechos con los productos. La desviación estándar es 0.47, lo que significa que las calificaciones no están muy dispersas en torno a la media. La poca dispersión está confirmada por los valores minimos (3.80) y maximo (5). En particular, el valor mínimo indica que no hay valoraciones muy bajas o negativas.

4.2 Principales gráficos y medidas para variables categóricas

Diagrama de barras

Un diagrama de barras es una representación gráfica utilizada para variables categóricas, donde cada barra representa la frecuencia de una categoría (numeros de occurencias).

Supongamos que una tienda de ropa está interesada en analizar la preferencia de los clientes por diferentes colores.



Interpretación: El gráfico muestra que el color más preferido es el negro, lo que puede influir en las decisiones de stock para la tienda.

Frecuencias y frecuencias relativas

Las frecuencias son el conteo de veces que una categoría aparece en un conjunto de datos. Las frecuencias relativas expresan ese conteo como un porcentaje del total.

Supongamos que una encuesta ha recogido las respuestas a una pregunta sobre la satisfacción del cliente (satisfecho, neutral, insatisfecho).

satisfacción	frec.	frec. relativas
Insatisfecho	1	0.17
Neutral	2	0.33
Satisfecho	3	0.50

Interpretación: El 50% de los encuestados están satisfechos, lo que indica una opinión positiva de la mayoría de los clientes.

5. Estadistica descriptiva bivariante

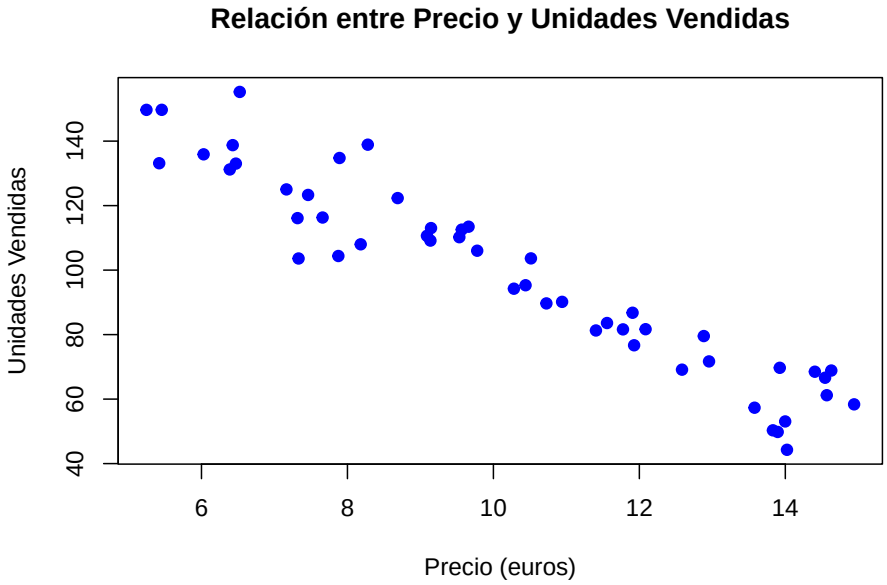
La **estadística descriptiva bivariante** es un tipo de análisis que se enfoca en la relación entre dos variables. Su objetivo es explorar cómo se comportan estas variables en conjunto y si existe algún tipo de asociación entre ellas. Para ello, se utilizan diversas herramientas como los **diagramas de dispersión** (scatter plots), que permiten visualizar la relación entre dos variables numéricas, y las **tablas de contingencia**, útiles para analizar la relación entre variables categóricas. Además, se emplean medidas estadísticas como la **covarianza** y el **coeficiente de correlación** para cuantificar el grado de asociación entre variables. La estadística bivariante es fundamental para identificar patrones de comportamiento conjuntos, y permite inferir si existe alguna relación lineal o no lineal entre las dos variables analizadas.

5.1 Principales gráficos y medidas para variables numericas (analisis bivariante)

Diagrama de dispersión

El **diagrama de dispersión** es un gráfico utilizado para visualizar la relación entre dos variables numéricas. Cada punto en el gráfico representa un par de valores correspondientes a las dos variables, lo que permite identificar patrones, tendencias o relaciones entre ellas. Es común en la investigación de mercado para analizar cómo variables como precio y ventas se relacionan entre sí.

Supongamos que una empresa quiere analizar la relación entre el precio de un producto y las unidades vendidas.



Interpretación: En el gráfico, se observa una relación negativa entre el precio y las unidades vendidas: a medida que el precio aumenta, las ventas disminuyen. Este tipo de información es clave para tomar decisiones sobre estrategias de precios.

Coeficiente de correlación

El **coeficiente de correlación** es una medida estadística que cuantifica la fuerza y la dirección de la relación entre dos variables numéricas. Su valor oscila entre -1 y 1: un valor de 1 indica una correlación perfecta y positiva, 0 indica que no hay correlación, y -1 indica una correlación perfecta y negativa.

En el mismo ejemplo anterior, podemos calcular el coeficiente de correlación entre el precio y las unidades vendidas.

Interpretación: El coeficiente de correlación obtenido cuantifica la relación entre precio y ventas. El valor es cercano a -1 (-0.95), esto confirma que existe una fuerte relación negativa: cuando el precio aumenta, las ventas disminuyen.

5.2 Principales gráficos y medidas para variables categoricas (analisis bivariante)

Tabla de contingencia

La **tabla de contingencia** es una tabla que muestra la distribución conjunta de dos variables categóricas. Es útil para analizar la relación entre categorías, como en el caso de encuestas donde se desea observar la preferencia de los clientes según su grupo de edad.

Supongamos que una empresa quiere analizar la relación entre la **preferencia de productos** (Producto A o Producto B) y el **grupo de edad** de los clientes (Jóvenes, Adultos). A continuación, se presenta una tabla de contingencia con las frecuencias absolutas, relativas, relativas por fila y relativas por columna.

Frecuencias

Grupo de Edad	Producto A (Absoluta)	Producto B (Absoluta)	Total
Jóvenes	30	20	50
Adultos	25	25	50
Total	55	45	100

Frecuencias Relativas (Total)

Grupo de Edad	Producto A (Relativa)	Producto B (Relativa)	Total
Jóvenes	0.30	0.20	0.50
Adultos	0.25	0.25	0.50
Total	0.55	0.45	1.00

Frecuencias Relativas por Fila

Grupo de Edad	Producto A (Rel. Fila)	Producto B (Rel. Fila)	Total
Jóvenes	0.60	0.40	1.00
Adultos	0.50	0.50	1.00

Frecuencias Relativas por Columna

Grupo de Edad	Producto A (Rel. Columna)	Producto B (Rel. Columna)	Total
Jóvenes	0.55	0.44	
Adultos	0.45	0.56	

Las frecuencias absolutas nos muestran el número total de clientes en cada categoría. Por ejemplo, de los **50 jóvenes encuestados**, **30** prefieren el Producto A y **20** prefieren el Producto B. Del mismo modo, entre los **50 adultos**, **25** prefieren el Producto A y **25** el Producto B. Esto indica que, en términos absolutos, la preferencia por los productos está distribuida de manera más equitativa entre los adultos que entre los jóvenes.

Las frecuencias relativas con respecto al total de la muestra nos indican las proporciones de cada combinación de grupo de edad y preferencia de producto. Por ejemplo, el **30% de los encuestados** son jóvenes que prefieren el Producto A, mientras que el **20%** son jóvenes que prefieren el Producto B. Estas frecuencias son útiles para entender el peso que tiene cada combinación en el conjunto total de encuestados.

Las frecuencias relativas por fila nos permiten ver la distribución de las preferencias dentro de cada grupo de edad. Por ejemplo, el **60% de los jóvenes** prefieren el Producto A, mientras que el **40%** prefieren el Producto B. Entre los adultos, las preferencias están equilibradas, con un **50%** de preferencia para cada producto. Esto indica que el Producto A es más popular entre los jóvenes en comparación con los adultos.

Las frecuencias relativas por columna muestran la distribución de los grupos de edad dentro de cada categoría de producto. Por ejemplo, el **55% de los clientes que prefieren el Producto A** son jóvenes, mientras que el **45%** son adultos. Para el Producto B, el **44%** de los clientes son jóvenes y el **56%** son adultos. Esto sugiere que, en términos de preferencia por producto, los jóvenes dominan ligeramente en la preferencia del Producto A, mientras que los adultos tienen una mayor proporción en la preferencia por el Producto B.

5.2.1 Interpretación como Probabilidades

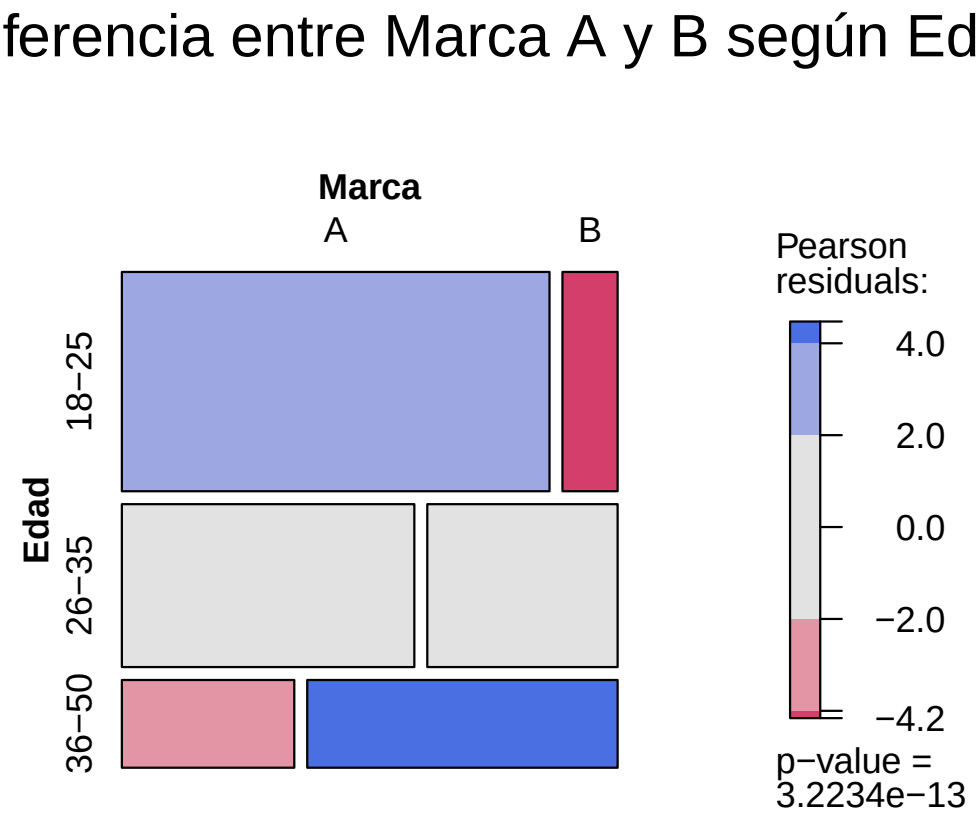
Si interpretamos estas frecuencias como probabilidades:

- La **probabilidad de que un cliente sea joven y prefiera el Producto A** es del 30%.
- La **probabilidad de que un cliente sea adulto y prefiera el Producto B** es del 25%.
- La **probabilidad de que un cliente prefiera el Producto A**, sin importar su edad, es del 55%.
- La **probabilidad de que un cliente prefiera el Producto A siendo joven**, es del 60%.
- La **probabilidad de que un cliente sea joven habiendo preferido el producto B**, es del 44%.

Estas probabilidades permiten a la empresa prever el comportamiento de un cliente promedio en función de su edad y sus preferencias de producto, lo que puede ser útil para diseñar estrategias de marketing personalizadas.

Mosaic Plot

El **mosaic plot** representa gráficamente tablas de contingencia. Cada rectángulo refleja la frecuencia observada y permite detectar asociaciones entre categorías.



Interpretación: El mosaic plot revela preferencias diferenciadas por edad. El grupo **18–25** aparece en azul intenso hacia **Marca A** indicando fuerte preferencia. El grupo **36–50** se inclina claramente por **Marca B**, con sombreado azul en esa celda. El grupo **26–35** muestra un patrón intermedio.

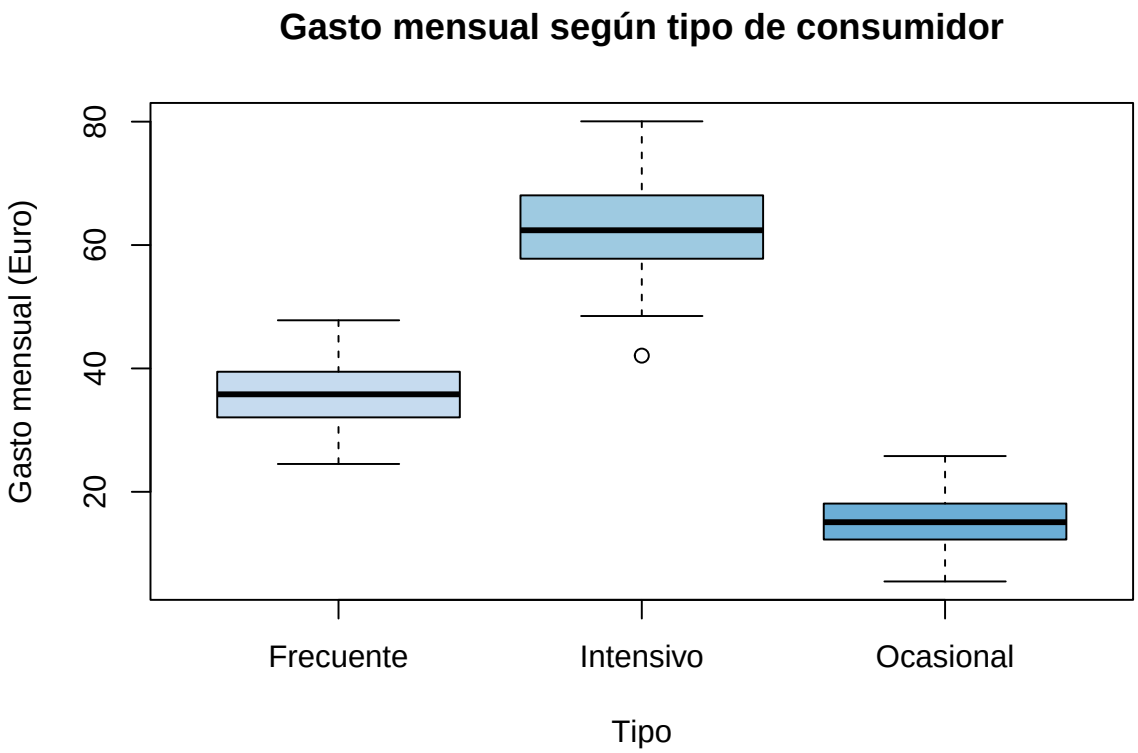
5.3 Principales gráficos para una variable numerica y unacategorica (analisis bivariante)

Cuando se estudia conjuntamente una variable numérica y una categórica, en investigación de mercado es habitual analizar la variable numérica en función de la variable categórica. Por ejemplo, comparar los ingresos de los consumidores según sus tipologías.

El gráfico estándar para este tipo de análisis es el diagrama de caja, representado uno por cada categoría. Este permite comparar de forma visual las distribuciones: medianas, dispersión, asimetrías y posibles valores atípicos.

La interpretación suele centrarse en detectar diferencias claras entre grupos. Con frecuencia, esta comparación se complementa con un test estadístico o con intervalos de confianza para determinar si las diferencias observadas son realmente significativas y no producto del azar.

Supungamos que quieremos comparar el **gasto mensual Euro** en bebidas energéticas según el **tipo de consumidor**.



Interpretación: El grupo **Intensivo** presenta **el gasto más alto**, claramente por encima del resto. Su caja es más alta y puede mostrar mayor dispersión, típico de consumidores muy activos. Los consumidores **Frecuentes** gastan una cantidad **intermedia**, estable y claramente diferenciada. El grupo **Ocasional** muestra **gasto bajo y menos variable**, acorde a un uso esporádico del producto.

6. Software estadísticos R, STATA, y JMP (SAS)

En este apartado, se describen las principales funciones y herramientas para calcular estadísticas univariantes y bivariantes, tanto en **R**, como en **Stata** y **JMP**. Se incluyen medidas y gráficos, así como una explicación de las funciones utilizadas en cada software.

6.1. R

6.1.1 Estadísticas univariantes en R

Las estadísticas univariantes permiten describir una sola variable. Las principales medidas y gráficos que se utilizan son:

- **Medidas de tendencia central**: media, mediana, moda.
- **Medidas de dispersión**: varianza, desviación estándar, rango.

- **Gráficos univariantes**: histograma, boxplot.

Funciones en R para variables numéricas:

```r
# Cálculo de la media, mediana y desviación estándar
mean(x)  # Media
median(x)  # Mediana
sd(x)  # Desviación estándar

# Cálculo del rango
range(x)

# Gráficos
hist(x)  # Histograma
boxplot(x)  # Boxplot
```

Funciones en R para variables categóricas:

```r
# Frecuencias
table(x)  # Tabla de frecuencias absolutas

# Gráfico de barras
barplot(table(x), main="Diagrama de Barras", col="lightblue")
```

Descripción de las funciones: - `mean(x)`, `median(x)`, `sd(x)`: Estas funciones calculan la media, mediana y desviación estándar de las variables numéricas. - `range(x)`: Proporciona el valor mínimo y máximo de `x`. - `hist(x)`, `boxplot(x)`: Generan gráficos univariantes para variables numéricas. - `table(x)`: Crea una tabla de frecuencias absolutas para variables categóricas. - `barplot(table(x))`: Crea un diagrama de barras a partir de una tabla de frecuencias.

6.1.2 Estadísticas bivariantes en R

Para analizar la relación entre dos variables, se utilizan estadísticas bivariantes como la **correlación** y los **diagramas de dispersión**.

Funciones en R:

```r
# Cálculo del coeficiente de correlación
cor(x, y)

# Gráfico de dispersión
plot(x, y)

# Tabla de contingencia para variables categóricas
```

```r
table(x, y)

# Gráfico de barras apiladas para variables categóricas
barplot(table(x, y), beside=TRUE, legend=TRUE, col=c("lightblue", "lightgreen"))

# Gráfico mosaicplot
mosaic(~ x + y, data = datos, shade = TRUE, main = " ")

# Gráfico boxplot (numerica y categorica)
boxplot (y~x)
```

Descripción de las funciones: - `cor(x, y)`: Calcula el coeficiente de correlación entre las variables `x` y `y`. - `plot(x, y)`: Genera un diagrama de dispersión para visualizar la relación entre dos variables numéricas. - `table(x, y)`: Crea una tabla de contingencia entre dos variables categóricas. - `barplot(table(x, y))`: Crea un gráfico de barras para visualizar las frecuencias conjuntas de dos variables categóricas. - `mosaic(~ x + y)`: Crea un gráfico mosaico para visualizar graficament elos resultados de una tabla de contigencia - `boxplot (y~x)`: Crea un boxplot de una variable numerica en función de la categorica.

6.2. Stata

6.2.1. Estadísticas univariantes en Stata

En Stata, las estadísticas univariantes se calculan utilizando las siguientes funciones y comandos:

Comandos para variables numéricas:

```
summarize varname  # Resumen estadístico (media, desviación estándar, etc.)
tabstat varname, statistics(mean median sd range)  # Tabla con medidas específicas
```

Comandos para variables categóricas:

```
tabulate varname  # Tabla de frecuencias absolutas

# Gráfico de barras
graph bar (count), over(varname)
```

Descripción de las funciones: - `summarize varname`, `tabstat varname`: Comandos para generar medidas estadísticas univariantes de variables numéricas. - `tabulate varname`: Genera una tabla de frecuencias absolutas para variables categóricas. - `graph bar (count), over(varname)`: Crea un diagrama de barras para representar las frecuencias de una variable categórica.

6.2.2. Estadísticas bivariantes en Stata

Comandos para variables numéricas:

```
correlate var1 var2  # Correlación entre dos variables
graph twoway scatter var1 var2  # Diagrama de dispersión
```

Comandos para variables categóricas:

```
tabulate var1 var2  # Tabla de contingencia para dos variables categóricas

# Gráfico de barras para variables categóricas
graph bar (count), over(var1) over(var2)
```

Descripción de las funciones: - `correlate var1 var2`: Calcula la correlación entre dos variables numéricas. - `graph twoway scatter var1 var2`: Genera un gráfico de dispersión para variables numéricas. - `tabulate var1 var2`: Crea una tabla de contingencia entre dos variables categóricas. - `graph bar (count), over(var1) over(var2)`: Crea un gráfico de barras para visualizar la relación entre dos variables categóricas.

6.3. Estadísticas Univariantes y Bivariantes en JMP

En **JMP**, las estadísticas univariantes y bivariantes se calculan a través de la interfaz gráfica. Para realizar estos análisis, sigue estos pasos:

Estadísticas Univariantes:

1. Importa tus datos a JMP.
2. Selecciona **Analyze** > **Distribution**.
3. En la ventana emergente, selecciona la variable que deseas analizar y haz clic en **OK**.
4. JMP generará un resumen de las estadísticas univariantes, incluyendo gráficos como histogramas y boxplots.
5. Para variables categóricas, elige la variable en la sección **Y, Columns** y JMP generará tablas de frecuencia y gráficos de barras.

Estadísticas Bivariantes:

1. Ve a **Analyze** > **Fit Y by X**.
2. Selecciona las dos variables que deseas analizar: una como X (independiente) y otra como Y (dependiente).
3. Haz clic en **OK** para generar un gráfico de dispersión y obtener el coeficiente de correlación para variables numéricas.
4. Para variables categóricas, selecciona ambas variables en la sección **X, Factor** y **Y, Response**, y JMP generará tablas de contingencia y gráficos de barras.

Estadística inferencial

1. Introducción a la Inferencia Estadística y Muestras

La *inferencia estadística* es una rama de la estadística que nos permite hacer afirmaciones o generalizaciones sobre una población, basándonos en una muestra (subconjunto) de datos extraída de esa población. En el contexto de la investigación de mercado, la inferencia es una herramienta muy valiosa, ya que, en los análisis, rara vez se puede obtener información de todos los consumidores (población), por lo que se suele recurrir a muestras representativas.

A nivel teórico una **población** y una **muestra** se pueden definir como:

- **Población**: Conjunto total de elementos que queremos estudiar (por ejemplo, todos los clientes de una tienda).
- **Muestra**: Subconjunto de la población del cual se recopilan datos (por ejemplo, 200 clientes de la tienda seleccionados para una encuesta).

Por ejemplo, en lugar de encuestar a todos los clientes de una cadena de supermercados, una empresa podría seleccionar una muestra de 500 clientes y después generalizar los resultados a todos los clientes usando la inferencia.

2. Límites para trabajar con datos poblacionales

En la investigación de mercado, a menudo se trabaja con muestras de la población en lugar de utilizar datos poblacionales completos. Esto se debe a varias razones que incluyen factores logísticos, económicos y de precisión en la recolección de datos. A continuación, se explican los principales límites para trabajar con datos poblacionales y se ofrecen ejemplos teóricos de cada uno.

1. Costos elevados

Trabajar con toda la población implica altos costos en términos de tiempo y recursos. La recolección de datos de cada miembro de una población puede ser financieramente prohibitiva, especialmente en mercados grandes o en estudios a nivel global.

Imaginemos que una empresa quiere hacer una encuesta para medir la satisfacción del cliente en un país con 50 millones de habitantes. Si cada encuesta cuesta $2, el costo total de una encuesta a la población completa sería de $100 millones, lo cual es inalcanzable para la mayoría de las empresas.

2. Dificultades logísticas

Reunir información de toda una población implica retos logísticos importantes. Algunas personas pueden vivir en áreas remotas, otras pueden ser difíciles de contactar o simplemente no estar dispuestas a participar.

Supongamos que una empresa quiere encuestar a todos los usuarios de un servicio de internet en un país. Hay personas que viven en áreas rurales donde no llega el correo o internet de forma confiable. Coordinar encuestas con estas personas sería un desafío significativo, especialmente si la población es dispersa geográficamente.

3. Problemas de actualización de los datos

Incluso si logramos recolectar datos de toda la población, esos datos pueden quedar desactualizados rápidamente. Las poblaciones cambian continuamente debido a factores como migración, cambios en el estilo de vida y demografía.

Si una empresa de productos de consumo realiza un censo de toda la población para analizar patrones de compra, pero los gustos y preferencias de las personas cambian rápidamente, los datos pueden volverse irrelevantes antes de que se pueda actuar sobre ellos.

4. Tiempo de procesamiento y análisis

Procesar datos poblacionales requiere un tiempo considerable. A mayor cantidad de datos, mayor será la complejidad de la tarea de análisis, lo que puede ralentizar el proceso de toma de decisiones.

Una empresa quiere lanzar un nuevo producto basado en las preferencias de consumo de toda la población. Sin embargo, procesar y analizar estos datos a nivel poblacional podría tomar meses, lo que retrasaría el lanzamiento del producto. Con una muestra, este análisis se podría completar en semanas.

En resumen, las principales razones por las que no se suele trabajar con datos poblacionales en la investigación de mercado son:

1. **Costos elevados**: recolectar datos de toda la población es demasiado caro.
2. **Dificultades logísticas**: es difícil acceder a toda la población.
3. **Datos desactualizados**: los datos pueden volverse obsoletos antes de ser utilizados.
4. **Tiempos largos de análisis**: analizar datos poblacionales toma mucho tiempo.

Trabajar con una muestra representativa y bien diseñada ofrece una solución más eficiente y rentable para las investigaciones de mercado. Las muestras son más fáciles de recolectar y analizar, y permiten obtener resultados rápidos que pueden ser generalizados a la población total si están bien diseñadas.

3. Cambios en la Interpretación

Cuando realizamos una investigación de mercado utilizando datos muestrales, es importante tener en cuenta que no estamos trabajando con toda la población, sino con una pequeña parte de ella. Esto introduce incertidumbre en nuestras conclusiones, ya que los resultados de la muestra pueden no reflejar perfectamente las características de la población completa. Para abordar esta incertidumbre, usamos herramientas estadísticas como los **intervalos de confianza** y los **test estadísticos**, que nos permiten hacer afirmaciones sobre la población con un grado de confianza medible.

3.1. ¿Por qué existe incertidumbre en los datos muestrales?

Cuando recolectamos datos de una muestra en lugar de hacerlo de toda la población, estamos asumiendo que esta muestra es representativa. Sin embargo, debido a que no hemos medido a todos los individuos de la población, siempre existe la posibilidad de que la muestra no refleje con precisión el comportamiento o las características de la población general. Esta diferencia entre la

muestra y la población se denomina **error muestral**, y es la fuente principal de incertidumbre en la investigación basada en muestras.

Imaginemos que una empresa de bebidas quiere conocer el porcentaje de personas que prefieren su nuevo refresco en una ciudad de un millón de habitantes. Es imposible encuestar a todas las personas de la ciudad, por lo que deciden encuestar a 1,000 personas. El resultado de la encuesta muestra que el 60% de las personas encuestadas prefieren el nuevo refresco. Pero, ¿cómo podemos estar seguros de que este 60% refleja el comportamiento del millón de habitantes?

Aquí es donde entra la incertidumbre. Debido a que solo hemos encuestado una muestra, existe la posibilidad de que si encuestáramos a otras 1,000 personas, obtendríamos un resultado ligeramente diferente, como 58% o 62%.

3.2. ¿Cómo manejamos esta incertidumbre?

Para manejar esta incertidumbre, utilizamos herramientas estadísticas que nos permiten cuantificar la confianza que podemos tener en nuestras estimaciones basadas en la muestra. Dos de las herramientas más importantes son los **intervalos de confianza** y los **test estadísticos**.

El **intervalo de confianza** es un rango de valores dentro del cual esperamos que caiga el verdadero valor poblacional con un cierto nivel de confianza. El nivel de confianza más común es del 95%, lo que significa que si repitiéramos el experimento 100 veces, el 95% de las veces, el intervalo incluiría el verdadero valor poblacional.

Volviendo al ejemplo del refresco, supongamos que obtenemos un intervalo de confianza del 95% que va del 57% al 63%. Esto significa que, aunque nuestra muestra arrojó un 60% de preferencia, podemos decir con un 95% de confianza que el porcentaje real de personas en la población que prefieren el refresco está entre el 57% y el 63%. En otras palabras, aunque no estamos completamente seguros, podemos estar razonablemente confiados en que el verdadero valor está dentro de este rango.

Los **test estadísticos** nos permiten comprobar hipótesis sobre la población a partir de los datos de la muestra. Estos tests nos ayudan a determinar si una diferencia observada en nuestra muestra es estadísticamente significativa, es decir, si es poco probable que haya ocurrido simplemente por azar.

Supongamos que la empresa de bebidas quiere saber si su nuevo refresco es preferido por una mayor proporción de personas que el refresco de la competencia, que históricamente ha tenido una preferencia del 55% en la población. Un test estadístico como una **prueba de proporciones** nos permitiría evaluar si la diferencia entre el 60% de nuestra muestra y el 55% de la competencia es estadísticamente significativa, o si esa diferencia podría haber ocurrido simplemente por el azar de la muestra.

Si el test muestra una diferencia significativa, podríamos concluir que el nuevo refresco es preferido por una mayor proporción de la población. Si no es significativo, la diferencia podría deberse al azar y no a una verdadera preferencia en la población.

4. Intervalos de Confianza

Un intervalo de confianza (IC) es un rango de valores dentro del cual, con un cierto nivel de confianza (generalmente 95%), se espera que esté el valor verdadero del parámetro poblacional. Es una herramienta clave en la inferencia estadística.

El intervalo de confianza mas usado en la investigación de mercado es el que se calcula para el valor promedio y se determina a partir de la formula:

$$IC = \bar{X} \pm Z\alpha/2\frac{s}{\sqrt{n}}$$

Donde:
- $\bar{X}$ es la media muestral
- $Z\alpha/2$ es el valor de la distribución normal para el nivel de significancia α deseado
- s es la desviación estándar de la muestra
- n es el tamaño de la muestra

Supongamos que una empresa quiere estimar el gasto promedio de sus clientes. Se toma una muestra de 100 clientes, y se encuentra que el gasto promedio es de 50 euro con una desviación estándar de 5 euro. Asi que tendriamos tendriamos:

- $\bar{X} = 50$
- $Z(\alpha = 5 \rightarrow 1-\alpha = 95\%) = 1.96$ es el valor de la distribución normal para el nivel de confianza deseado
- $s = 5$
- $n = 100$

Si aplicamos la formula tendríamos:

$$IC = 50 \pm 1.96\frac{5}{\sqrt{100}} = IC[49.02 - 50.97]$$

El intervalo de confianza nos indica que, con un 95% de confianza, el gasto promedio de los clientes está entre 49.02E y 50.98E. Esto proporciona a la empresa una base sólida para estimar el comportamiento de gasto de su base de clientes.

Cuando trabajamos con **frecuencias relativas**, por ejemplo, el porcentaje de consumidores que prefieren una marca, tambien podriamos estar inteersado en generalizar el resultados a nivel de la población. Normalmente en este caso se suele usar la aproximación normal cuando la muestra es suficientemente grande (np 5 y n(1−p) 5).

El IC para una proporción se calcula así:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Si por ejemplo, una empresa encuesta a **n = 400** personas para saber si prefieren el **Producto A**. Un total de **260** personas dicen que sí y disponemeos de la proporción muestral $\hat{p} = 260/400 = 0.65$, y quremeos deterinar el intervalo con nivel de confianza del 95%, podriamos calcular el IC en tres pasos:

1. Error estándar:

$$SE = \sqrt{\frac{0.65(1-0.65)}{400}} = 0.0238$$

2. Margen de error:

$$ME = 1.96 \times 0.0238 = 0.0467$$

3. Intervalo de confianza:

$$0.65 \pm 0.0467 \Rightarrow (0.6033,\ 0.6967)$$

Interpretación: podemos afirmar, con un 95% de confianza, que **entre el 60.3% y el 69.7%** de los consumidores de la población prefieren el Producto A.

4.1. Factores que Afectan la Amplitud del IC

El amplitud de un intervalo de confianza puede variar en función de tres parámetros:

1. **Tamaño de la muestra**: A mayor tamaño de muestra, menor amplitud del intervalo. Esto es debido a que si la muestra será mas grande, será mas parecida a la población, proporcionando estimaciones mas precisas.

2. **Desviación estándar**: A mayor dispersión en los datos, mayor será la amplitud del intervalo. Claramente si hay una mayor dispersión que se puede traducir en mayor heterogeneidad en la composición de la población, será mas difícil proporcionar estimaciones robustas y en consecuencia los resultados obtenidos seran menos precisos, algo que se traduce en intervalos mas amplios.

3. **Nivel de confianza**: A mayor nivel de confianza, mayor será la amplitud del intervalo. Esto depende de la precisión que se quiere alcanzar, si no somos dispuestos a asumir errores, para estar mas seguros que el intervalo incluya el valor, tendremos que ampliar sus limites.

4.2. Aplicón de los intervalos de confianza en la investigación de mercado

A continuación, veremos dos casos en los que se aplican los intervalos de confianza: para generalizar la estimación de un promedio muestral y para comparar promedios entre diferentes grupos.

Generalización de un Promedio Muestral

En investigación de mercado, a menudo recolectamos datos de una muestra de consumidores y queremos hacer una inferencia sobre el promedio poblacional. El intervalo de confianza nos ayuda a estimar el rango probable en el que se encuentra el verdadero promedio de la población.

Supongamos que una empresa realiza una encuesta de satisfacción entre 500 clientes y obtiene una puntuación media de satisfacción de 7.5 sobre 10. Si calculamos un intervalo de confianza del 95%, obtenemos que el verdadero promedio de satisfacción en toda la población de clientes podría estar entre 7.3 y 7.7.

Esto significa que, aunque el promedio muestral es 7.5, podemos decir con un 95% de confianza que el verdadero promedio de satisfacción de todos los clientes se encuentra en algún punto entre 7.3 y

7.7. Este rango nos da una idea de la precisión de nuestra estimación y nos ayuda a entender mejor el comportamiento de la población general, sin necesidad de encuestar a todos los clientes.

En este caso, el intervalo de confianza nos permite no solo reportar un valor puntual (7.5), sino también reconocer la incertidumbre asociada a la muestra. Nos ofrece una forma más confiable de comunicar los resultados, ya que en lugar de afirmar que el promedio es exactamente 7.5, indicamos que es probable que esté dentro del rango estimado.

Comparación de Promedios entre Diferentes Grupos

Otra aplicación común de los intervalos de confianza en investigación de mercado es la comparación de promedios entre diferentes grupos. Por ejemplo, podríamos querer comparar la satisfacción de dos grupos de clientes: aquellos que compran online y aquellos que compran en tiendas físicas. En este contexto, los intervalos de confianza nos ayudan a evaluar si las diferencias entre los promedios de ambos grupos son significativas.

Una regla práctica para la comparación de intervalos de confianza es observar si los intervalos **se solapan**. Si los intervalos de confianza de dos grupos se solapan, no podemos concluir que haya una diferencia significativa entre los grupos. Por el contrario, si los intervalos **no se solapan**, es un indicio de que las medias de los grupos probablemente son diferentes.

Volviendo al ejemplo anterior, supongamos de de encuestar a 400 clientes online y a 400 clientes en la tienda, y que se obtienen los siguientes resultados:

- Promedio de satisfacción para clientes online: 7.8, con un intervalo de confianza del 95% entre 7.6 y 8.0.
- Promedio de satisfacción para clientes en tienda física: 7.2, con un intervalo de confianza del 95% entre 7.0 y 7.4.

En este caso, los intervalos de confianza **no se solapan** (el intervalo de los clientes online va de 7.6 a 8.0 y el de los clientes de tienda física va de 7.0 a 7.4). Esto sugiere que hay una **diferencia significativa**, considerando un nivel de confianza del 95% ,en los niveles de satisfacción entre los clientes online y los de tienda física. Podemos inferir que los clientes online están significativamente más satisfechos que los de la tienda física.

Por el contrario, si los intervalos de confianza hubieran sido los siguientes:

- Clientes online: 7.6 a 8.0.
- Clientes en tienda física: 7.4 a 7.8.

En este caso, los intervalos de confianza **se solapan** (ambos incluyen el rango de valores desde 7.6 a 7.8), lo que indicaría que no podemos afirmar con seguridad que haya una diferencia significativa, cnsiderando un nivel de confianza del 95%, entre la satisfacción de ambos grupos. Es posible que la diferencia observada sea atribuible al azar en la muestra.

5. Test Estadísticos

Los test estadísticos permiten comprobar si las diferencias observadas en los datos son estadísticamente significativas o se deben al azar. En la investigación de mercado, se utilizan para tomar

decisiones basadas en datos. Por ejemplo, supongamos que una empresa quiere saber si una nueva campaña de marketing ha aumentado las ventas. Para comparar las ventas antes y después de la campaña y ver si las ventas se han incrementados se utilizaría un test estadístico.

Para construir un test estadístico necesitamos identificar:

1. **Hipótesis nula (H_0)**: No hay efecto o diferencia (por ejemplo, "la campaña no ha afectado las ventas").

2. **Hipótesis alternativa (H_1)**: Existe un efecto o diferencia (por ejemplo, "la campaña ha aumentado las ventas").

3. **Estadístico de prueba**: Valor calculado a partir de los datos que se compara con una distribución teórica.

4. **Error de tipo I (alpha)**: Probabilidad de rechazar la hipótesis nula cuando es cierta.

5. **Valor p**: Probabilidad de obtener un resultado tan extremo como el observado si la hipótesis nula fuera verdadera.

Hay que precisar que cada test estadístico tiene un propósito específico, y, por lo tanto, sus **hipótesis** y **estadísticos** varían. Dependiendo de lo que queramos estudiar o analizar en el contexto de la investigación de mercado, debemos elegir el test adecuado.

5.1. ¿Qué son las Hipótesis y los Estadísticos?

Una **hipótesis** es una afirmación que hacemos sobre la población que estamos estudiando, y la finalidad de un test estadístico es determinar si hay suficiente evidencia en los datos muestrales para aceptar o rechazar esta hipótesis. Normalmente se plantean dos hipótesis:

- **Hipótesis Nula (H_0):** Es una afirmación que indica que no hay efecto o diferencia en la población. Por ejemplo, que no hay diferencia entre las medias de dos grupos.
- **Hipótesis Alternativa (H_1):** Es la afirmación contraria a la hipótesis nula, es decir, que existe un efecto o diferencia en la población.

El **estadístico** es una medida calculada a partir de los datos muestrales, y su valor se utiliza para tomar la decisión de si se rechaza o no la hipótesis nula. Dependiendo del tipo de test que estemos realizando, el estadístico puede ser diferente (por ejemplo, el estadístico t para una prueba t). Hay que recordar que los estadístico están definidos de tal forma que siempre van asociados a una distribución estadística teórica conocidas.

5.2. Los Errores en los Test Estadistico

Cuando realizamos un test estadístico en una investigación de mercado, siempre existe la posibilidad de cometer errores debido a la naturaleza probabilística de los datos muestrales. En particular, hay dos tipos de errores que pueden ocurrir: el **error de tipo 1** y el **error de tipo 2**. Cada uno tiene implicaciones importantes para la interpretación de los resultados del test. En este apartado nos enfocaremos en entender ambos errores, con ejemplos concretos, y en particular, discutiremos por qué es crucial evitar cometer el **error de tipo 1**.

Error de Tipo 1 (Falso Positivo)

El **error de tipo 1** ocurre cuando rechazamos incorrectamente la hipótesis nula (H_0) cuando en realidad es verdadera. En otras palabras, concluimos que hay una diferencia o un efecto en la población cuando en realidad no lo hay. Este tipo de error se conoce también como **falso positivo**.

El **nivel de significancia** (α) que fijamos antes de realizar el test estadístico controla la probabilidad de cometer un error de tipo 1. Por ejemplo, si establecemos un nivel de significancia de $= 0.05$, estamos aceptando un 5% de riesgo de rechazar la hipótesis nula cuando en realidad es verdadera.

Supongamos que una empresa lanza una nueva campaña publicitaria y quiere evaluar si ha aumentado significativamente las ventas en comparación con las campañas anteriores. Se realiza un test estadístico con un nivel de significancia del 5% ($\alpha = 0.05$), y el resultado del test indica que la nueva campaña ha aumentado las ventas de forma significativa, por lo que se rechaza la hipótesis nula (H_0: "la campaña no tiene efecto en las ventas").

Sin embargo, la realidad es que la nueva campaña no ha afectado realmente las ventas, pero debido al error de tipo 1, el test ha indicado erróneamente que sí hubo un aumento. La empresa ahora podría invertir más recursos en una campaña que, en realidad, no es más efectiva que las anteriores.

¿Por qué no queremos cometer un error de tipo 1?

El error de tipo 1 es particularmente grave porque puede llevar a **conclusiones falsas** que afecten la toma de decisiones. En el ejemplo anterior, la empresa podría desperdiciar grandes cantidades de recursos en una estrategia de marketing basada en un resultado incorrecto. En otras palabras, el error de tipo 1 nos lleva a actuar como si algo fuera cierto cuando no lo es, lo que podría generar consecuencias costosas.

Evitar el error de tipo 1 es crucial, especialmente en decisiones de alto impacto en marketing, ya que actuar sobre una diferencia que no existe puede llevar a decisiones erróneas que afecten negativamente al negocio.

¿Cómo Controlamos el Error de Tipo 1?

El error de tipo 1 se controla fijando el **nivel de significancia** (α) antes de realizar el test estadístico. El nivel de significancia comúnmente usado es 0.05, lo que significa que estamos aceptando una probabilidad del 5% de cometer un error de tipo 1. Esto implica que, en promedio, en 5 de cada 100 tests, podríamos rechazar la hipótesis nula cuando es verdadera.

Si queremos reducir la probabilidad de cometer un error de tipo 1, podemos elegir un nivel de significancia más bajo, como $\alpha = 0.01$.

Supongamos que una empresa está evaluando la efectividad de dos anuncios publicitarios, y fija un nivel de significancia de 0.05 para comparar el impacto en las ventas. Si el resultado del test

estadístico muestra que las ventas han aumentado significativamente con uno de los anuncios ($p < 0.05$), la empresa puede decidir invertir más en ese anuncio.

Error de Tipo 2 (Falso Negativo)

El **error de tipo 2** ocurre cuando no rechazamos la hipótesis nula (H_0) cuando en realidad es falsa. Es decir, concluimos que no hay una diferencia o efecto en la población cuando, de hecho, sí lo hay. Este error se conoce también como **falso negativo**.

La **potencia** del test (1 - β) determina la probabilidad de evitar un error de tipo 2. Una potencia elevada (generalmente se busca que sea al menos del 80%) reduce la probabilidad de cometer este error.

Continuando con el ejemplo anterior, supongamos que la empresa realiza el test estadístico y, en esta ocasión, no se rechaza la hipótesis nula (H_0: "la campaña no tiene efecto en las ventas"). Sin embargo, en realidad la nueva campaña sí ha tenido un impacto significativo en las ventas, pero debido a un error de tipo 2, no se detectó esta diferencia.

En este caso, la empresa perdería la oportunidad de invertir en una campaña que realmente está generando resultados positivos.

5.3. Los Principales Test Estadisticos Usados en Investigación de Mercado

En investigación de mercado se aplican diferentes test estadísticos dependiendo del problema que estamos analizando. A continuación se presentan los test mas comunes.

5.3.1. Test t: Comparación de medias

El **t-test** es una de las pruebas estadísticas más comunes para comparar medias. Puede usarse en varios contextos, pero aquí nos enfocaremos en dos casos específicos:

1. **Prueba t para una muestra**: Se utiliza para comparar la media de una muestra con un valor conocido o esperado (por ejemplo, una media poblacional).

2. **Prueba t para dos muestras independientes**: Se usa para comparar las medias de dos grupos independientes entre sí.

T-test para una muestra

Se utiliza cuando queremos comparar la media de una muestra con un valor de referencia conocido (por ejemplo, una media poblacional) para determinar si la media de la muestra es significativamente diferente de ese valor.

En este caso las hipótesis serán:

- **Hipótesis nula (H_0):** La media de la muestra es igual al valor de referencia.

- **Hipótesis alternativa (H_1):** La media de la muestra es diferente del valor de referencia.

El estadístico t asociado se calcula con la siguiente fórmula:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Donde:
- $\bar{x}$ es la media de la muestra.
- μ_0 es el valor de referencia o media poblacional.
- s es la desviación estándar de la muestra.
- n es el tamaño de la muestra.

Ejemplo. Supongamos que queremos evaluar si la satisfacción media de los clientes es diferente de 7 en una escala de 1 a 10. Recolectamos una muestra de 10 clientes con las siguientes puntuaciones: 6, 7, 8, 6, 7, 8, 9, 7, 6, 8.

Paso 1. Calculo de media de la muestra ($\bar{x}$):

$$\bar{x} = \frac{6+7+8+6+7+8+9+7+6+8}{10} = 7.2$$

Paso 2. Calculo la desviación estándar de la muestra ($\bar{\sigma}$):

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = 1.03$$

Paso 3. Cálculo del estadístico *t*:

$$t = \frac{7.2 - 7}{\frac{1.03}{\sqrt{10}}} = 0.61$$

Paso 4. Determinación del p-valor:

Con un t de 0.61 y 9 grados de libertad, usando una tabla de t o un software estadístico, encontramos que el p-valor es aproximadamente 0.56 (> 0.05). Esto indica que no hay evidencia suficiente para rechazar la hipótesis nula. Considerendo un nivel de confianza del 95%, el promedio meustral y el teorico coinciden.

Cálculo en R:

```r
# Datos
muestra <- c(6, 7, 8, 6, 7, 8, 9, 7, 6, 8)

# Prueba t
t.test(muestra, mu = 7)
```

Cálculo en STATA:

```
* Crear los datos de la muestra
clear
input muestra
6
7
8
6
7
8
9
7
6
8
end

* Prueba t
ttest muestra == 7
```

Cálculo en JMP:

1. Ve a **Analyze > Distribution**.
2. Introduce los datos de la muestra.
3. Selecciona la columna de datos, y en el menú contextual selecciona **Test Mean**.
4. Introduce el valor de referencia (en este caso 7) y ejecuta la prueba.

T-test para dos muestras independientes

Se usa para comparar las medias de dos grupos independientes y determinar si las diferencias entre ellas son estadísticamente significativas.

En este caso las hipótesis serán:

- **Hipótesis nula (H_0):** Las medias de los dos grupos son iguales.
- **Hipótesis alternativa (H_1):** Las medias de los dos grupos son diferentes.

El estadístico t para dos muestras independientes se calcula como:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Donde:
- $\bar{x}_1$ y $\bar{x}_2$ son las medias de los dos grupos.
- s_1 y s_2 son las desviaciones estándar de los dos grupos.
- n_1 y n_2 son los tamaños de las muestras.

Ejemplo. Supongamos que una empresa quiere comparar la satisfacción de clientes que compran en tienda física con los que compran online. Se recolectan dos muestras:

- Tienda física: 6, 4, 5, 4, 6
- Tienda online: 8, 9, 7, 8, 9

Paso 1. Calculo Media y desviación estándar:
- $\bar{x}_1 = 5.0$, $s_1 = 1$ (Tienda física).
- $\bar{x}_2 = 8.2$, $s_2 = 0.84$ (Tienda online).

Paso 2. Cálculo del estadístico t:

$$t = \frac{5.0 - 8.2}{\sqrt{\frac{1^2}{5} + \frac{0.84^2}{5}}} = -5.488$$

Paso 3. Determinación del p-valor:

Con un t de -5.488 y 8 grados de libertad, el p-valor es aproximadamente 0.0006. Este valor es mas bajo de 0.05, por lo que podemos deccir que, considerando in nivel de confianza del 95%, hay una diferencia significativa entre los dos grupos.

Cálculo en R:

```r
# Datos
tienda_fisica <- c(6, 4, 5, 4, 6)
tienda_online <- c(8, 9, 7, 8, 9)

# Prueba t para dos muestras
t.test(tienda_fisica, tienda_online)
```

Cálculo en STATA:

```
* Crear los datos de las dos muestras
clear
input tienda_fisica tienda_online
6 8
4 9
5 7
4 8
6 9
end

* Prueba t para dos muestras independientes
ttest tienda_fisica == tienda_online
```

Cálculo en JMP:

1. Ve a **Analyze > Fit Y by X**.
2. Selecciona la variable de grupo (tienda física vs tienda online).
3. En el menú contextual, selecciona **t-test** para comparar las medias de los dos grupos.
4. El software realizará la prueba y te dará el valor de t y el p-valor.

5.3.2. ANOVA

El **ANOVA** (Análisis de Varianza) es una prueba estadística utilizada para comparar las medias de tres o más grupos. A diferencia del **t-test**, que solo compara dos grupos, el ANOVA nos permite evaluar si existe una diferencia significativa entre las medias de múltiples grupos simultáneamente. Por ejemplo, en marketing, podríamos querer comparar las preferencias de los clientes entre tres o más estrategias publicitarias diferentes. El ANOVA responde a la pregunta: ¿hay alguna diferencia significativa entre las medias de estos grupos?

En este caso las hipótesis serán:

- **Hipótesis nula (H_0):** Las medias de los grupos son iguales.
- **Hipótesis alternativa (H_1):** Al menos una de las medias es diferente.

El estadístico **F** utilizado en ANOVA se calcula como la razón de dos varianzas:

$$F = \frac{\text{Variabilidad entre los grupos}}{\text{Variabilidad dentro de los grupos}}$$

La fórmula para el estadístico F es la siguiente:

$$F = \frac{MSB}{MSW}$$

Donde:
- MSB es el **cuadrado medio entre los grupos** (Mean Square Between), que mide la variabilidad entre las medias de los grupos.
- MSW es el **cuadrado medio dentro de los grupos** (Mean Square Within), que mide la variabilidad dentro de cada grupo.

Fórmulas:

- **Variabilidad entre los grupos (SSB)**:

$$SSB = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_G)^2$$

Donde:
- n_i es el tamaño de la muestra del grupo i.
- $\bar{x}_i$ es la media del grupo i.
- $\bar{x}_G$ es la media general de todos los grupos.

- **Variabilidad dentro de los grupos (SSW)**:

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

Donde:
- x_{ij} es el valor de la j-ésima observación en el grupo i.

Finalmente, los cuadrados medios se obtienen dividiendo cada suma de cuadrados por los grados de libertad correspondientes:

- **Cuadrado medio entre los grupos (MSB)**:

$$MSB = \frac{SSB}{k-1}$$

- **Cuadrado medio dentro de los grupos (MSW)**:

$$MSW = \frac{SSW}{N-k}$$

Donde:
- k es el número de grupos.
- N es el número total de observaciones.

Ejemplo. Supongamos que una empresa de marketing desea comparar la efectividad de tres campañas publicitarias diferentes (A, B y C) midiendo la satisfacción de los clientes (en una escala de 1 a 10). Se seleccionan muestras de clientes para cada campaña con los siguientes datos:

- Campaña A: 6, 8, 7, 9, 6
- Campaña B: 7, 6, 8, 7, 6
- Campaña C: 8, 9, 8, 8, 7

Paso 1: Cálculo de las medias de los grupos y la media general:

- Media campaña A: $\bar{x}_A = 7.2$
- Media campaña B: $\bar{x}_B = 6.8$
- Media campaña C: $\bar{x}_C = 8.0$
- Media general: $\bar{x}_G = 7.33$

Paso 2: Cálculo de SSB (Suma de cuadrados entre los grupos):

$$SSB = 5(7.2 - 7.33)^2 + 5(6.8 - 7.33)^2 + 5(8.0 - 7.33)^2 = 2.93$$

Paso 3: Cálculo de SSW (Suma de cuadrados dentro de los grupos):

$$SSW = (6-7.2)^2+(8-7.2)^2+(7-7.2)^2+(9-7.2)^2+(6-7.2)^2+(7-6.8)^2+(6-6.8)^2+(8-6.8)^2+$$
$$(7-6.8)^2+(6-6.8)^2+(8-8.0)^2+(9-8.0)^2+(8-8.0)^2+(8-8.0)^2+(7-8.0)^2 = 10.4$$

Paso 4: Cálculo de MSB y MSW:

- $MSB = \frac{2.93}{2} = 1.465$
- $MSW = \frac{10.4}{12} = 0.87$

Paso 5: Cálculo del estadístico F:

$$F = \frac{1.465}{0.87} = 1.68$$

Paso 6: Determinación del p-valor:

Con un valor de F = 1.68 y grados de libertad (2, 12), el p-valor asociado es aproximadamente 0.23. Esto indica que no hay evidencia suficiente para rechazar la hipótesis nula; por lo tanto, no podemos concluir que las medias de las tres campañas sean significativamente diferentes.

Comparaciones múltiples en ANOVA

Cuando se realiza un ANOVA y se rechaza la hipótesis nula, sabemos que al menos una de las medias es diferente, pero no sabemos cuál. Para identificar qué grupos son significativamente diferentes entre sí, se pueden realizar **comparaciones múltiples** (post-hoc tests). Un método común es la **prueba de Tukey** o **t-test**, que compara todas las combinaciones de medias de los grupos y ajusta el nivel de significancia para múltiples comparaciones.

Cálculo en R:

```r
# Datos
grupo_A <- c(6, 8, 7, 9, 6)
grupo_B <- c(7, 6, 8, 7, 6)
grupo_C <- c(8, 9, 8, 8, 7)

# Crear data frame
datos <- data.frame(
  valor = c(grupo_A, grupo_B, grupo_C),
  grupo = factor(rep(c("A", "B", "C"), each = 5))
)

# Realizar ANOVA
anova_resultado <- aov(valor ~ grupo, data = datos)

# Resumen del ANOVA
summary(anova_resultado)

# Comparaciones múltiples (Tukey)
TukeyHSD(anova_resultado)
```

Cálculo en STATA:

```
* Crear los datos para cada grupo
clear
input valor grupo
6 1
8 1
7 1
9 1
6 1
7 2
6 2
8 2
7 2
6 2
8 3
9 3
8 3
8 3
7 3
end

* Etiquetar los grupos
label define grupo_label 1 "A" 2 "B" 3 "C"
label values grupo grupo_label

* Realizar ANOVA
anova valor grupo

* Comparaciones múltiples (Tukey)
* En Stata, usamos el comando `pwcompare` con ajuste de Tukey
para realizar comparaciones múltiples.
pwcompare grupo, mcompare(tukey)
```

Cálculo en JMP:

1. Ve a **Analyze > Fit Y by X**.
2. Introduce la variable dependiente (la satisfacción) y la variable categórica (la campaña).
3. En el menú contextual, selecciona **ANOVA**.
4. Para realizar comparaciones múltiples, selecciona **Compare Means** y luego **Tukey HSD** o **t-test**.

5.3.2. Test Chi-cuadrado (X^2)

El **test Chi-cuadrado (X^2)** es una prueba estadística utilizada para determinar si existe una relación significativa entre dos variables categóricas.

En este caso las hipótesis serán:

- **Hipótesis nula (H_0):** Las dos variables son independientes.

- **Hipótesis alternativa (H_1):** Las dos variables están relacionadas.

El estadístico $\mathbf{X^2}$ se calcula con la siguiente fórmula:

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Donde:
- O_i son las frecuencias observadas.
- E_i son las frecuencias esperadas bajo la hipótesis nula.

Los **grados de libertad (d.f.)** seran:

$$d.f. = (r-1)(c-1)$$

, donde r es el número de filas y c es el número de columnas en la tabla de contingencia.

Ejemplo. Supongamos que una tienda quiere evaluar si la preferencia por un tipo de producto (A, B o C) está relacionada con el género del cliente. Se realiza una encuesta y se obtienen los siguientes datos:

	Producto A	Producto B	Producto C	Total
Hombres	30	10	20	60
Mujeres	20	25	15	60
Total	50	35	35	120

Paso 1. Cálculo de las frecuencias esperadas

Primero, calculamos las frecuencias esperadas. El total de hombres es 60, el total de mujeres es 60 y el total general es 120. Las frecuencias esperadas para cada celda se calculan como:

$$E_{ij} = \frac{\text{Fila total} \times \text{Columna total}}{\text{Total general}}$$

Por ejemplo, la frecuencia esperada para hombres que prefieren el Producto A es:

$$E_{11} = \frac{60 \times 50}{120} = 25$$

El mismo procedimiento se aplica para las demás celdas:

	Producto A	Producto B	Producto C
Hombres	25	17.5	17.5
Mujeres	25	17.5	17.5

Paso 2. Cálculo del estadístico X²

Para cada celda, aplicamos la fórmula:

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Aplicando esta fórmula para cada celda:

$$X^2 = \frac{(30-25)^2}{25} + \frac{(10-17.5)^2}{17.5} + \frac{(20-17.5)^2}{17.5} + \frac{(20-25)^2}{25} + \frac{(25-17.5)^2}{17.5} + \frac{(15-17.5)^2}{17.5}$$

$$X^2 = 1 + 3.21 + 0.36 + 1 + 3.21 + 0.36 = 9.14$$

Paso 3. Determinación del p-valor

Con un estadístico Chi-cuadrado de 9.14 y 2 grados de libertad ($d.f. = (2-1)(3-1) = 2$), usando una tabla de Chi-cuadrado o software estadístico, encontramos que el p-valor es aproximadamente 0.01. Dado que este valor es menor que 0.05, rechazamos la hipótesis nula y concluimos que la preferencia por el producto está relacionada con el género del cliente.

Cálculo en R

```r
# Crear tabla de contingencia
tabla <- matrix(c(30, 10, 20, 20, 25, 15), nrow = 2, byrow = TRUE)

# Realizar prueba Chi-cuadrado
chisq.test(tabla)
```

Cálculo en STATA

```
* Crear tabla de contingencia
clear
input row column count
1 1 30
1 2 10
1 3 20
2 1 20
2 2 25
2 3 15
end
```

```
* Convertir a formato de tabla de contingencia
table row column, c(sum count) chi2

* También se puede usar el siguiente comando para una tabla
bidimensional con Chi-cuadrado
tabulate row column, chi2
```

*#### **Cálculo en JMP**

1. Ve a **Analyze > Fit Y by X**.
2. Introduce las dos variables categóricas (por ejemplo, género y preferencia de producto).
3. En el menú contextual, selecciona **Chi-Square Test** para realizar la prueba de independencia.

6. Tecnicas de Muestreo

En la investigación de mercado, elegir una buena técnica de muestreo es crucial para obtener resultados representativos y confiables. Las técnicas de muestreo se dividen en dos categorías principales: **muestreo probabilístico** y **muestreo no probabilístico**. Cada técnica tiene ventajas, desventajas y aplicaciones específicas, y la elección de una depende de la naturaleza del estudio.

6.1. Muestreo Probabilístico

En el **muestreo probabilístico**, cada individuo de la población tiene una probabilidad conocida y no nula de ser seleccionado. Esto permite hacer inferencias válidas sobre la población.

6.1.1. Muestreo aleatorio simple

El **muestreo aleatorio simple** es una técnica en la que cada miembro de la población tiene la misma probabilidad de ser seleccionado. Se utiliza cuando tenemos acceso a una lista completa de todos los miembros de la población.

Sus Ventajas son:
- Es sencillo de entender y aplicar.
- Garantiza que la muestra sea representativa de la población (si es lo suficientemente grande).

Mientras que su desventajas incluyen:
- Requiere un listado completo de la población.
- Puede ser ineficiente si la población es muy dispersa.

Ejemplo. Supongamos que queremos seleccionar una muestra de 50 clientes al azar de una base de datos de 1,000 clientes. Utilizamos una técnica de muestreo aleatorio simple para elegir a los clientes.

6.1.2. Muestreo sistemático

El **muestreo sistemático** consiste en seleccionar cada k-ésimo individuo de la población después de elegir un punto de partida al azar. Se usa cuando hay una lista ordenada de la población y es difícil realizar un muestreo aleatorio simple.

Sus Ventajas son:
- Más sencillo y rápido que el muestreo aleatorio simple.
- Garantiza una distribución uniforme de la muestra.

Mientras que si desventajas incluyen:
- Si los datos siguen un patrón, puede sesgar los resultados.

Ejemplo. Seleccionamos cada 10° cliente de una lista de 1,000 clientes para crear una muestra de 100 personas.

6.1.3. Muestreo por conglomerados (cluster)

En el **muestreo por conglomerados**, la población se divide en grupos (conglomerados), y se seleccionan algunos de estos conglomerados al azar. Luego, se toman muestras dentro de esos conglomerados. Se utiliza cuando la población está dividida naturalmente en grupos (por ejemplo, tiendas, barrios, etc.) que presenten carteristicas parecidas, y es costoso o difícil realizar un muestreo aleatorio simple.

Sus Ventajas son:
- Reduce costes y tiempo de muestreo.
- No requiere una lista completa de toda la población.

Mientras que si desventajas incluyen:
- Menos preciso que otros métodos si los conglomerados no son homogéneos.

Ejemplo. Supongamos que queremos estudiar la satisfacción del cliente en una cadena de tiendas. Seleccionamos al azar 10 tiendas y luego encuestamos a todos los clientes de esas tiendas.

6.2. Muestreo No Probabilístico

En el **muestreo no probabilístico**, algunos individuos de la población tienen una mayor probabilidad de ser seleccionados que otros. Las inferencias que se obtienen son menos robustas de las obtenidas con el muestreo probabilístico.

6.2.1. Muestreo de conveniencia

El **muestreo de conveniencia** consiste en seleccionar a los individuos más accesibles o fáciles de encontrar. Se usa cuando el investigador busca rapidez o está limitado por los recursos.

Sus ventajas son:
- Rápido y fácil de implementar.

- Bajo coste.

Sus desventajas incluyen:
- Los resultados pueden no ser representativos.
- Existe un alto riesgo de sesgo.

Ejemplo. Un investigador que hace encuestas en un centro comercial está utilizando un muestreo de conveniencia al seleccionar a las personas que estén disponibles en ese momento.

6.2.2. Muestreo por respuesta voluntaria

En el **muestreo por respuesta voluntaria**, los participantes eligen si desean o no participar en el estudio. Se aplica cuando se busca recolectar respuestas de un gran número de personas en poco tiempo, como encuestas en línea.

Sus ventajas son:
- Muy fácil de implementar.
- Puede atraer a personas que están muy interesadas en el tema.

Sus desventajas incluyen:
- Los participantes voluntarios pueden no ser representativos.
- Suele haber sesgo hacia opiniones extremas.

Ejemplo. Una encuesta en un sitio web donde los visitantes pueden elegir participar o no es un ejemplo de muestreo por respuesta voluntaria.

6.2.3. Muestreo en bola de nieve (snowball)

El **muestreo en bola de nieve** consiste en que los participantes iniciales reclutan a otros participantes, formando una "bola de nieve". Se utiliza para llegar a poblaciones difíciles de acceder o cuando no hay una lista clara de los miembros de la población.

Sus ventajas son:
- Útil para llegar a poblaciones ocultas o pequeñas.
- Bajo costo de implementación.

Sus desventajas incluyen:
- Puede haber un sesgo hacia los grupos que los participantes conocen.
- Los resultados no suelen ser representativos de la población general.

Ejemplo. Si un investigador quiere estudiar consumidores de un producto de nicho, puede pedir a los primeros encuestados que le presenten la encuesta a otros usuarios del producto.

6.2.4. Muestreo por juicio o propósito (purposive)

El **muestreo por juicio** se basa en la selección de individuos que el investigador considera más representativos o adecuados para el estudio. Se utiliza cuando el investigador quiere incluir a participantes con características específicas relevantes para el estudio.

Sus ventajas son:
- Permite seleccionar a personas clave o expertos en un tema.
- Útil en estudios exploratorios.

Sus desventajas incluyen:
- Existe un alto riesgo de sesgo.
- La representatividad depende del juicio del investigador.

Ejemplo. Un investigador que selecciona a 10 ejecutivos de alto nivel en una empresa para estudiar las estrategias de marketing estaría utilizando un muestreo por juicio.

6.3. Muestreo Estratificado o por cuotas

El **muestreo estratificado** es una técnica en la que la población se divide en subgrupos homogéneos (estratos) y luego se toma una muestra de cada estrato. Esto asegura que cada subgrupo esté representado proporcionalmente en la muestra. Se usa cuando la población se puede dividir en grupos con características relevantes para el estudio y es importante que todos los subgrupos estén representados.

Sus ventajas son:
- Aumenta la precisión y representatividad de la muestra.
- Asegura que todos los subgrupos estén representados en la muestra.

Sus desventajas incluyen:
- Requiere una lista completa y detallada de la población.
- Puede ser costoso y laborioso.

Ejemplo 1.Una empresa quiere medir la satisfacción de clientes de diferentes rangos de edad. Se divide la población en tres grupos: menores de 30 años, entre 30 y 50 años, y mayores de 50 años. Luego, se seleccionan participantes de cada grupo de manera proporcional a su tamaño en la población general.

Ejemplo 2.Un estudio de mercado busca evaluar el uso de un producto en diferentes niveles de ingreso. La población se divide en estratos según los ingresos (bajo, medio, alto) y se seleccionan muestras proporcionales de cada grupo.

7. Tamaño Muestral

El cálculo del **tamaño de la muestra** es crucial para garantizar que los resultados de un estudio sean representativos y precisos. El tamaño de la muestra depende de varios parámetros clave, como

el **nivel de confianza**, el **error de muestreo** y la **proporción esperada**. En este apartado, explicaremos cómo se calculan los tamaños de muestra para poblaciones finitas e infinitas, y daremos ejemplos de su cálculo paso a paso.

7.1.Parámetros para determinar el tamaño de la muestra

El calculo del tamaño de la muestra se define a partir de los siguientes paramentros.

Nivel de confianza $(1-\alpha)$

El nivel de confianza refleja la certeza de que el valor real de la población está dentro del intervalo de confianza calculado. Los niveles de confianza comunes son 95% y 90%.

El nivel de confianza está relacionado con el **valor crítico z**, que corresponde al número de desviaciones estándar que abarca el nivel de confianza:

- Para un nivel de confianza del 95%, el valor de **z** es 1.96.
- Para un nivel de confianza del 90%, el valor de **z** es 1.645.

Error de muestreo o margen de error (e)

El margen de error (e) es la cantidad de error que estamos dispuestos a aceptar en nuestras estimaciones. Este error determina qué tan lejos puede estar la estimación de la media poblacional verdadera. Margen de error comunes son 5%, 2.5% y 1.5%.

7.1.3. Proporción esperada (p)

Es la proporción esperada de la población que posee la característica de interés. Si no se tiene información previa, se utiliza un valor conservador de 0.5 para maximizar el tamaño de la muestra.

7.2. Fórmulas para calcular el tamaño de la muestra

Poblaciones infinitas

Cuando la población es lo suficientemente grande, podemos asumir que es infinita. La fórmula para calcular el tamaño de la muestra es:

$$n = \frac{z^2 \cdot p \cdot (1-p)}{e^2}$$

Donde:
- n es el tamaño de la muestra.
- z es el valor crítico z asociado al nivel de confianza.
- p es la proporción esperada de la población (si se desconoce, se usa 0.5).
- e es el margen de error.

Poblaciones finitas

Si conocemos el tamaño de la población (N), el tamaño de la muestra ajustada para una población finita se calcula con la siguiente fórmula:

$$n_f = \frac{N \cdot n}{N + n - 1}$$

Donde: - N es el tamaño de la población. - n es el tamaño de la muestra calculado para una población infinita (con la fórmula anterior).

7.3. Ejemplo: Cálculo del tamaño de la muestra

Supongamos que queremos calcular el tamaño de una muestra para una población infinita y una población finita de 5,000 personas. Queremos un margen de error del 5%, y un nivel de confianza del 95%. Asumimos que la proporción esperada es 0.5.

Paso 1. Cálculo para población infinita

Usamos la fórmula para poblaciones infinitas:

$$n = \frac{(1.96)^2 \cdot 0.5 \cdot (1 - 0.5)}{0.05^2} = 384.16$$

El tamaño de la muestra es aproximadamente 385.

Paso 2. Cálculo para población finita

Para una población finita de 5,000 personas, usamos la fórmula ajustada:

$$n_f = \frac{5000 \cdot 384.16}{5000 + 384.16 - 1} = 357.98$$

El tamaño ajustado de la muestra es aproximadamente 358.

7.4. Tabla: Tamaños de muestra para diferentes márgenes de error y niveles de confianza

A continuación se muestra una tabla que compara los tamaños de muestra para diferentes márgenes de error y niveles de confianza para una población infinita y una finita de 5,000 personas.

Tabla de Tamaños de Muestra

A continuación, se presentan los tamaños de muestra correspondientes para poblaciones finitas e infinitas:

Nivel de Confianza	Error Muestral	Tamaño Muestra Infinita	Tamaño Muestra Finita (N=10000)
95%	5%	385	370
95%	2.5%	1,537	1,340
95%	1.5%	4,267	2,770
90%	5%	271	261
90%	2.5%	1,084	964

Nivel de Confianza	Error Muestral	Tamaño Muestra Infinita	Tamaño Muestra Finita (N=10000)
90%	1.5%	3,018	2,162

Regresión múltiple

1. Introducción

La **regresión lineal múltiple** extiende la regresión lineal simple a escenarios donde existen múltiples variables predictoras. Ya que se trata de una extensión, hipótesis, estimación de los parametros, significancia, validación, análisis de los residuos, siguen las mismas pautas y criterios que los de la **regresión lineal simple** bajo la consideración que ahora tenemos mas de una variable predictora.

2. Regresión Lineal Múltiple: Definición y Fórmula

La **regresión lineal múltiple** modela la relación entre una variable dependiente Y y múltiples variables independientes (i.e., predictoras) $X_1, X_2, ..., X_n$. La fórmula general es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Donde:
- Y es la variable dependiente.
- $X_1, X_2, \dots, X_n$ son las variables independientes.
- β_0 es la costante.
- $\beta_1, \beta_2, ..., \beta_n$ son los coeficientes que representan el efecto de cada predictor sobre Y.
- ϵ es el término de error.

En el contexto del modelo de regresión lineal múltiple se suela utilizar la notación matricial para explicar la relación entre variable dependiente y variables predictoras. La formula de la **regresión lineal múltiple** expresada en notación matricial es:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Donde:
- $\mathbf{Y}$ es un vector de $n \times 1$ con las variables dependientes (o respuesta).
- $\mathbf{X}$ es una matriz de $n \times (k+1)$ que contiene las variables explicativas (o predictoras), siendo k el número de variables independientes.
- β es un vector de $(k+1) \times 1$ que contiene los coeficientes del modelo, incluyendo el intercepto.
- ϵ es un vector de $n \times 1$ de los términos de error, que se asume tienen media cero.

2. Hipótesis del Modelo

Para que el modelo de **regresión lineal múltiple** sea válido, se deben cumplir las mismas hipótesis que se valen para el modelo de **regresión lineal simple** recordando de nuevo que en este caso tenemos varias variables predictoras. A parte las 4 hipótesis clásicas en este caso también hay una nueva hipótesis sobre la multicolinealidad. Recordando la notación matricial tendremos:

1. **Linealidad**. Se asume que la relación entre las variables independientes ($\mathbf{X}$) y la variable dependiente ($\mathbf{Y}$) es lineal. Esto significa que la variable dependiente se puede modelar como una combinación lineal de las variables explicativas y sus coeficientes.

La hipótesis es:

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$$

2. **Independencia de los residuos**. Los residuos (ϵ) son independientes entre sí. Esto implica que no hay correlación entre los términos de error de diferentes observaciones. Matemáticamente, esto se expresa como:

$$\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \forall\, i \neq j$$

3. **Homoschedasticidad**. Los residuos tienen varianza constante, es decir, la varianza de los términos de error es la misma para todas las observaciones. En notación matricial, la hipótesis de homocedasticidad se expresa como:

$$\text{Var}(\epsilon) = \sigma^2 \mathbf{I}_n$$

Donde $\mathbf{I}_n$ es la matriz identidad de tamaño $n \times n$.

4.**Normalidad de los Errores**. Los residuos se distribuyen normalmente con media cero y varianza constante.

$$\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)$$

5. **No Colinealidad Perfecta**. No debe haber colinealidad perfecta entre las variables independientes, es decir, ninguna columna de $\mathbf{X}$ puede ser una combinación lineal exacta de otras columnas.

3. Estimación de los parametros del modelo (Metodo de los minimos cuadrado OLS)

El objetivo de la regresión lineal múltiple es ajustar un modelo lineal de la forma:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Donde:
- $\mathbf{Y}$ es un vector de $n \times 1$ con las variables dependientes.
- $\mathbf{X}$ es una matriz de $n \times (k+1)$ que contiene las variables independientes (incluyendo un vector de unos para el término constante).
- β es un vector de $(k+1) \times 1$ con los coeficientes del modelo.
- ϵ es un vector de $n \times 1$ de los términos de error.

Nuestro objetivo es encontrar los coeficientes β que minimicen el error cuadrático entre los valores observados y los valores predichos por el modelo. Formalmente, buscamos minimizar la suma de los cuadrados de los errores, es decir, minimizar:

$$\epsilon^T \epsilon \Rightarrow (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta)$$

Esta expresión representa la suma de los cuadrados de los residuos. Para encontrar el valor óptimo de β, derivamos esta función respecto a β y la igualamos a cero.

$$\epsilon^T \epsilon \Rightarrow S(\beta) = \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

Ahora derivamos $S(\beta)$ con respecto a β:

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta$$

Para minimizar $S(\beta)$, igualamos la derivada a cero:

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\beta$$

Finalmente, despejamos β:

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Este es el estimador de los mínimos cuadrados ordinarios (MCO) para β.

4. Significancia de los Parámetros

La significatividad de los coeficientes en la regresión lineal múltiple se calcula de la misma manera que en la regresión lineal simple. En ambos casos, se utiliza el **test t** para evaluar si cada coeficiente es significativamente diferente de cero. Las hipótesis que se contrastan son las mismas:

- Hipótesis nula (H_0): El coeficiente del predictor es igual a cero ($\beta_i = 0$).
- Hipótesis alternativa (H_1): El coeficiente del predictor es diferente de cero ($\beta_i \neq 0$).

El valor p resultante del test indica la probabilidad de que los coeficientes observados se deban al azar, bajo la suposición de que la hipótesis nula es cierta. La interpretación del valor p sigue el mismo principio: si es menor que un nivel de significancia establecido (por ejemplo, 0.05), se rechaza la hipótesis nula.

La diferencia principal en la regresión lineal múltiple es que este proceso se realiza para varios predictores en lugar de uno solo. Es decir, se evalúa la significatividad de cada predictor de manera individual, pero el concepto y el proceso estadístico no cambian respecto a la regresión simple.

Como en el caso de la regresión lineal simple, también podemos usar los intervalos de confianza. La interpretación será la misma.

5. Validación del Modelo

La validación de un modelo de regresión lineal múltiple es crucial para evaluar su calidad y la precisión de las predicciones. Como en el caso de la regresión lineal simple, existen varios indicadores y pruebas estadísticas para este propósito (incluyendo el **R²** y el **test ANOVA**). Ademas se suele utilizar el el **R² ajustado** que se mas indicado en el caso de varias variables predictoras.

R²

El **R²** mide la proporción de la variabilidad total de la variable dependiente que es explicada por las variables independientes. Se define como:

$$R^2 = 1 - \frac{SSE}{SST}$$

Donde: - SSE es la suma de los cuadrados de los errores (sum of squared errors). - SST es la suma total de cuadrados (total sum of squares), que mide la variabilidad total de los datos.

Interpretación: Un valor de R^2 cercano a 1 indica que el modelo explica una gran parte de la variabilidad de la variable dependiente. Sin embargo, un R^2 alto no garantiza que el modelo sea adecuado, ya que no penaliza el uso de variables adicionales que puedan no ser útiles.

R² ajustado

El **R² ajustado** es una versión modificada de R^2 que tiene en cuenta el número de predictores en el modelo. A diferencia del R^2, que siempre aumenta al añadir más variables, el R^2 ajustado solo aumenta si las nuevas variables mejoran significativamente el modelo. Se calcula como:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

Donde:
- n es el número de observaciones.
- k es el número de predictores.

Interpretación: El R^2_{adj} penaliza el uso de predictores innecesarios. Un valor más alto indica un mejor ajuste, teniendo en cuenta la complejidad del modelo.

Análisis de Varianza

El **Análisis de Varianza** (ANOVA) es una prueba estadística que evalúa si el modelo general es significativo. Se basa en la comparación de dos fuentes de variabilidad: - Variabilidad explicada por el modelo (SSR, sum of squares regression). - Variabilidad no explicada por el modelo (SSE, sum of squares error).

La **prueba F** se utiliza para determinar si la proporción de la variabilidad explicada respecto a la no explicada es significativa. La fórmula es:

$$F = \frac{SSR/k}{SSE/(n - k - 1)}$$

Donde: - k es el número de predictores. - n es el número de observaciones.

Interpretación: Un valor de F alto y un valor p bajo indican que el modelo en su conjunto es significativo.

6. Validación de los Residuos en la Regresión Lineal Múltiple

La validación de los residuos es una etapa crucial en el análisis de regresión lineal, ya que nos permite evaluar si los supuestos del modelo se cumplen. En un modelo de regresión lineal múltiple, los residuos deben cumplir con las siguientes condiciones (mismas de regresion lineal simple):

1. **Independencia**: Los residuos deben ser independientes entre sí.
2. **Normalidad**: Los residuos deben seguir una distribución normal.
3. **Homocedasticidad**: Los residuos deben tener una varianza constante.
4. **Linealidad**: La relación entre los residuos y los predictores debe ser lineal.

6.1. Residuos

Los **residuos** son las diferencias entre los valores observados y los valores predichos por el modelo:

$$e_i = Y_i - \hat{Y}_i$$

Donde: - Y_i son los valores observados. - $\hat{Y}_i$ son los valores predichos por el modelo.

El análisis de residuos busca verificar si los residuos se comportan como ruido aleatorio, lo que indicaría que el modelo ajusta bien los datos.

6.1. Gráficos de Validación de los Residuos

Gráfico de Residuos vs Valores Ajustados

Este gráfico ayuda a identificar patrones no lineales o heterocedasticidad (variabilidad no constante de los residuos). En este gráfico, los residuos deben estar distribuidos aleatoriamente alrededor de cero sin mostrar ningún patrón definido. Si los residuos están distribuidos aleatoriamente alrededor de la línea cero sin patrones, indica que el modelo ajusta bien y que no hay problemas de especificación o heterocedasticidad.

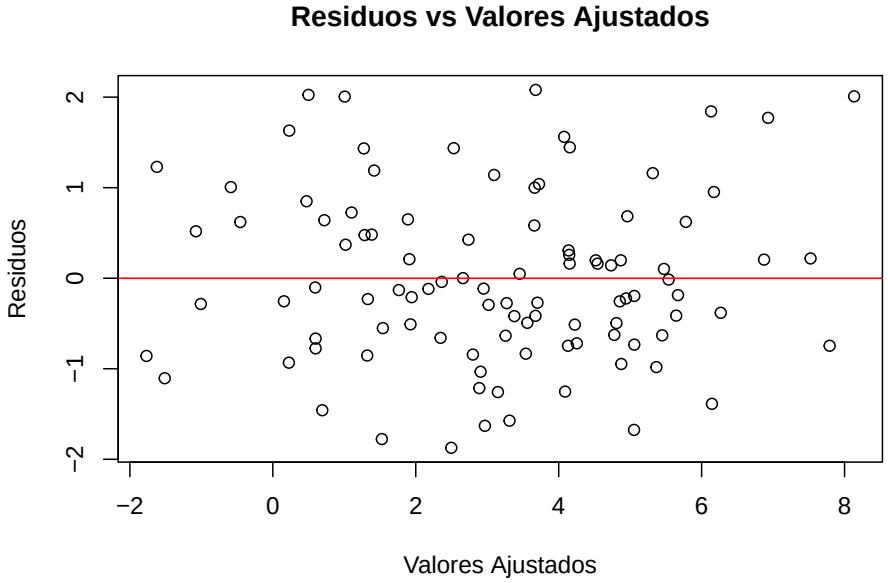


Gráfico Q-Qnorm (Quantile-Quantile)

Este gráfico compara los residuos estandarizados con una distribución normal teórica. Es útil para verificar si los residuos se distribuyen normalmente. Si los puntos siguen aproximadamente la línea roja, indica que los residuos se distribuyen normalmente, lo cual es una de las suposiciones clave del modelo de regresión.

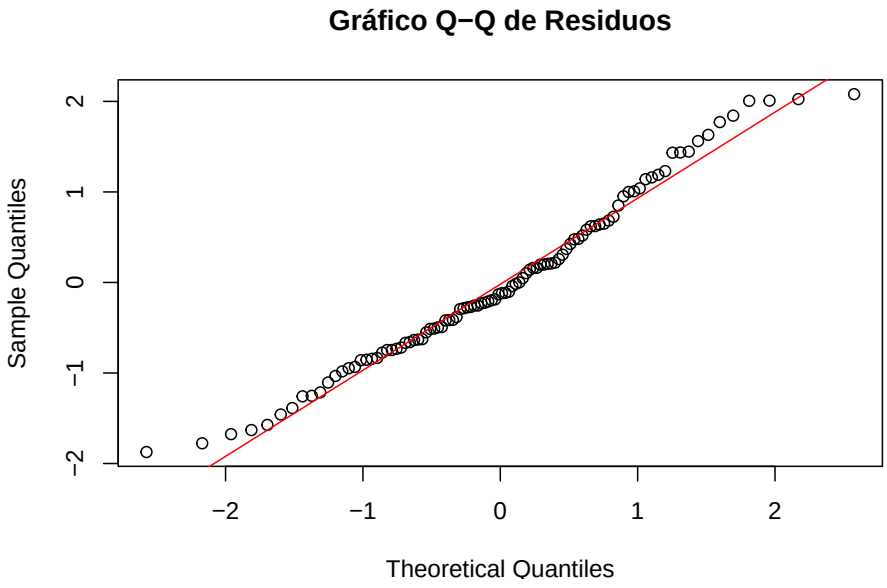
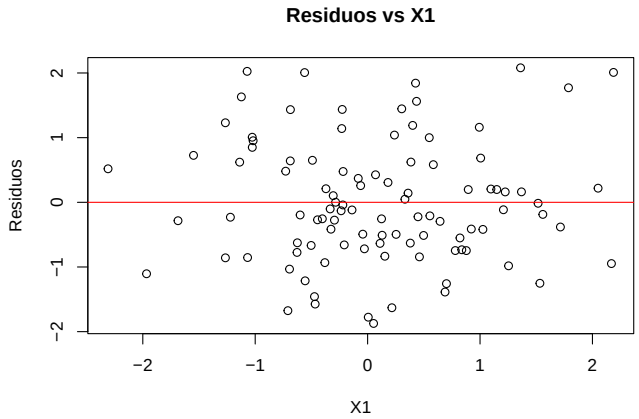
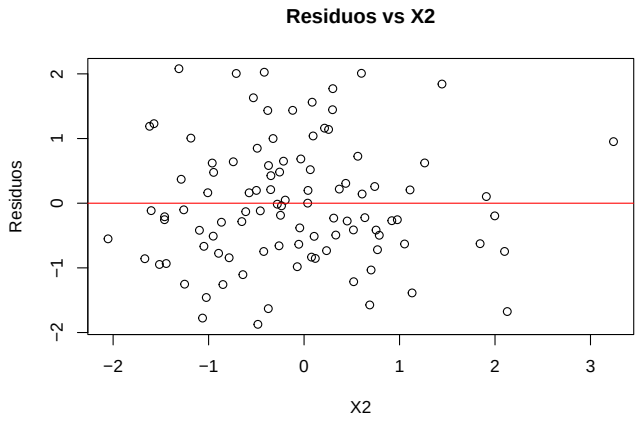


Gráfico de Residuos vs Predictores

Este gráfico verifica si hay relaciones no capturadas entre los predictores y los residuos. En este gráfico, los residuos deben estar distribuidos aleatoriamente. Al igual que en el gráfico de residuos vs valores ajustados, no debe haber patrones visibles. Si los residuos están distribuidos de manera aleatoria, el modelo ajusta bien.





6.2. Pruebas Estadísticas para los Residuos

Test de Normalidad de Shapiro-Wilk

El **test de Shapiro-Wilk** se utiliza para verificar si los residuos siguen una distribución normal. La hipótesis nula del test es que los residuos se distribuyen normalmente. Si el valor p es mayor que un nivel de significancia (generalmente 0.05), no se rechaza la hipótesis nula, lo que sugiere que los residuos se distribuyen normalmente.

$$H_0 : \text{Los residuos se distribuyen normalmente.}$$

Test de Durbin-Watson para Autocorrelación

El **test de Durbin-Watson** verifica la autocorrelación de los residuos. La hipótesis nula es que no hay autocorrelación. Un valor p mayor que 0.05 indica que no hay autocorrelación entre los residuos, lo que es deseable para un buen ajuste del modelo.

$$H_0 : \text{No hay autocorrelación entre los residuos.}$$

7. Multicolinealidad en la Regresión Lineal Múltiple

La **multicolinealidad** ocurre cuando dos o más variables predictoras en un modelo de regresión lineal múltiple están altamente correlacionadas entre sí. Esto significa que una de las variables predictoras puede ser predicha linealmente a partir de las otras con un alto grado de precisión. La multicolinealidad puede tener efectos negativos en la interpretación de los coeficientes del modelo.

Los principales problemas asociados con la multicolinealidad son:

1. **Inestabilidad en las estimaciones de los coeficientes**: Los coeficientes pueden volverse sensibles a pequeños cambios en los datos.
2. **Altos errores estándar**: Debido a la colinealidad, los coeficientes pueden tener valores elevado de variablidiad, lo que dificulta determinar si un predictor es significativo.
3. **Dificultad para interpretar los coeficientes**: Cuando hay multicolinealidad, los efectos de las variables individuales pueden ser difíciles de interpretar, ya que sus impactos están confusamente compartidos con otras variables predictoras.

Existen varias maneras de detectar la multicolinealidad:

- **Matriz de correlaciones**: Correlaciones muy altas entre los predictores pueden indicar multicolinealidad.
- **Factor de Inflación de la Varianza (VIF)**: El VIF es una medida cuantitativa que evalúa cuánto la varianza de un coeficiente estimado aumenta debido a la colinealidad con las otras variables predictoras.

Factor de Inflación de la Varianza (VIF)

El **VIF** mide cuántas veces la varianza de un coeficiente aumenta debido a la multicolinealidad en comparación con un escenario donde las variables son independientes. Se calcula de la siguiente manera. Para cada predictor X_j, se ajusta un modelo de regresión en el cual X_j es la variable dependiente y las demás variables son las predictoras. El R_j^2 es el coeficiente de determinación de este modelo. Luego, el VIF se calcula como:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Donde:
- R_j^2 es el coeficiente de determinación de la regresión de X_j sobre las demás variables.

Un valor de $VIF > 5$ se considera indicativo de una multicolinealidad problemática.

El VIF puede ser derivado de la matriz de la variable predictoras. Consideremos el modelo de regresión múltiple estándar:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

Donde $\mathbf{X}$ es la matriz que incluye todas las variables predictoras. Si calculamos la varianza de los estimadores β_j, uno de los componentes de la varianza es:

$$\text{Var}(\beta_j) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}$$

La multicolinealidad aumenta la magnitud de este término debido a la estructura de correlación entre las columnas de $\mathbf{X}$.

7.1. Como se soluciona

Para poder solucionar la multicolinealidad hay diversoso métodos:

1. **Eliminar variables altamente correlacionadas**: Una solución es eliminar una o más de las variables que están fuertemente correlacionadas.
2. **Combinar variables**: Crear nuevas variables combinando las variables correlacionadas en una sola, como con la suma o el promedio.
3. **Usar Regularización (Ridge o Lasso)**: Estos métodos penalizan el tamaño de los coeficientes, reduciendo el impacto de la multicolinealidad.

4. **Aplicar Análisis de Componentes Principales (ACP)**: El ACP transforma las variables correlacionadas en un nuevo conjunto de variables no correlacionadas llamadas componentes principales, que se pueden utilizar en la regresión.

Variables	Precio	CalidadServicio	Cobertura	AtencionCliente
Precio	1.00	0.88	-0.14	0.76
CalidadServicio	0.88	1.00	-0.09	0.85
Cobertura	-0.14	-0.09	1.00	-0.10
AtencionCliente	0.76	0.85	-0.10	1.00
Satisfaccion	0.78	0.79	0.21	0.78

Tras analizar las correlacione, resulta evidente que hay correlaciones altas entre algunas variables de la encuesta en particular: **CalidadServicio**, **AtencionCliente** (0.84) y entre **Precio** y **CalidadServicio**, **AtencionCliente** (0.88,0.75). Esto puede ser indicador de presencia de multicolinealidad y de que tengamos problemas con la estimación de los coeficientes.

Realizaremos una regresión lineal múltiple para predecir la satisfacción a partir de las percepciones del cliente.

Variables	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.6498	3.4569	-1.35	0.1818
Precio	0.4616	0.1044	4.42	0.0000
CalidadServicio	0.1020	0.1393	0.73	0.4661
Cobertura	0.2300	0.0337	6.83	0.0000
AtencionCliente	0.5139	0.1160	4.43	0.0000

Observando los resultados, podemos ver que la variable **CalidadServicio** no es significativa ($p - value = 0.46$). Las otras variables si lo son.

Antes de seguir vamos a verificar si realmente hay multicolinealidad. A tal fin usaremos el Factor de Inflación de la Varianza (VIF).

Precio	CalidadServicio	Cobertura	AtencionCliente
4.602718	6.972337	1.028793	3.588558

Observando el VIF, notamos que en el caso de la variable **CalidadServicio** el indicador es mas alto del nivel de referencia (VIF<5). En concreto el valor es igual a 6.97. Para solucionar este problema, aplicaremos un Análisis en Componentes Principales (ACP),y, en lugar de usar las variables originales, usaremos las componentes. Esto nos permitirá eliminar las correlaciones.

Variables	Dim.1	Dim.2	Dim.3	Dim.4
Precio	-0.93	-0.02	-0.32	-0.16
CalidadServicio	-0.96	-0.09	-0.05	0.24
Cobertura	0.19	-0.98	-0.02	-0.01
AtencionCliente	-0.92	-0.08	0.38	-0.10
Cumulative Proportion	0.6705	0.9154	0.97667	1.00000

El analisis indica que con dos componentes **PC1** y **PC2** podemos explicar el 91.54% de la variabilidad contenida en la nubes de puntos. La componente **PC1** está asociada a **Precio** (-0.93), **CalidadServicio** (-0.96), y **AtencionCliente** (-0.91); la componente **PC2** está asociada a **Cobertura** (-0.98).

Teniendo en cuenta este aspecto nombramos la componente **PC1**, **Valor Percibido del Servicio**, mientras que la componente **PC2**, **Cobertura**.

En lugar de las variables predictoras originales correlacionadas, usaremos las componentes principales en la regresión.

Variables	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
costante	50.8185	0.5084	99.96	0.0000
ValorPercibido	5.3022	0.3120	16.99	0.0000
Cobertura	4.0297	0.5162	7.81	0.0000
R^2=0.78				

Después de aplicar el ACP, las nuevas variables (**PC1**, **PC1**) no están correlacionadas, lo que mejora la interpretación de los coeficientes. **ValorPercibido** es el coeficiente mas importante seguido por **Cobertura**. El valor del $R^2 = 0.78$ se mantiene. Ya que la **CalidadServicio** define el **ValorPercibido**, también es relevante para estimar la percepción de los consumidores. Estos resultados difieren del primer modelo que daba una interpretación sesgada del efecto de los coeficientes.

8. Ejemplos

8.1. Caso de Estudio: Presupuesto publicitario

En este caso de estudio, analizaremos los factores que influyen en las ventas de un producto en diferentes tiendas. Consideraremos las siguientes variables: - **Ventas (Y)**: La variable dependiente que queremos explicar. - **Publicidad en TV (X1)**: Gasto en publicidad en televisión. - **Publicidad en Radio (X2)**: Gasto en publicidad en radio. - **Publicidad en Redes Sociales (X3)**: Gasto en publicidad en redes sociales.

Nuestro objetivo es ajustar un modelo de regresión lineal múltiple para predecir las ventas en función de las inversiones en publicidad, evaluar la significatividad de los predictores y realizar un análisis completo de validación del modelo.

Para este ejemplo, generaremos un conjunto de datos simulado. A continuación, ajustamos el modelo de regresión lineal múltiple para predecir las ventas en función de la publicidad en TV, radio y redes sociales.

Variables	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**VIF**
costante	51.8558	5.5666	9.32	0.0000	
TV	0.4861	0.0292	16.64	0.0000	1.0191
Radio	0.3231	0.0547	5.90	0.0000	1.0030
SocialMedia	0.1426	0.1122	1.27	0.2069	1.0175
$R^2 adjusted$	**ANOVA**	**Shapiro**			
0.752	0.001	0.939			

En el resumen del modelo, podemos observar los coeficientes estimados para cada variable, el valor de R^2 y el valor p para cada predictor, lo que nos permite evaluar la significatividad de los coeficientes, el resultado del ANOVA el VIF y taben el test de normalidad de los residuos.

En relación a los coeficientes podemos observar:
1. La **costante** $\beta_0 = 51.85$ representa el valor de las **ventas** si todas los gastos en publicidad (**TV**, **radio** y **redes sociales**) fueron iguales a 0. En concreto las **ventas** seria 51 unidades.

2. El coeficiente $\beta_{TV} = 0.48$. Esto quiere decir que al aumentar una unidad (1000 euro) los gastos de **publicidad** relacionados con la **TV**, produciría un aumento de $0.48 \times 1000 = 480$ unidades, sin tener en cuenta los otros gastos.

3. El coeficiente $\beta_{Radio} = 0.32$. Esto quiere decir que al aumentar una unidad (1000 euro) los gastos de **publicidad** relacionados con la **radio**, produciría un aumento de $0.32 \times 1000 = 320$ unidades, sin tener en cuenta los otros gastos.

4. El coeficiente $\beta_{SocialMedia} = 0.14$. Esto quiere decir que al aumentar una unidad (1000 euro) los gastos de **publicidad** relacionados con la **redes sociales** produciría un aumento de $0.32 \times 1000 = 142$ unidades, sin tener en cuenta los otros gastos.

Comparando las diferentes tipologías de gastos, aquel relacionado con la **TV** es el que produce un aumento en las **ventas** mas importante (i.e., es el factor mas importante a la hora de estimar las ventas).

El **test t** evalúa si los coeficientes son significativamente diferentes de cero. Si el valor p es menor que 0.05, se considera que el predictor tiene un efecto significativo sobre las **ventas**. En este caso todos los coeficientes son significativos a excepción de **SocialMedia** $p - valor = 0.207 > 0.05$.

El R^2 mide qué proporción de la variabilidad de las ventas está explicada por el modelo. Un R^2 alto indica que el modelo explica una gran parte de la variabilidad de las ventas. El R^2 ajustado es una versión corregida que penaliza la inclusión de variables no significativas. En este caso los dos son muy parecidos e altos ($R^2 = 0.759$ y elR^2ajustado $= 0.752$). Esto nos permite concluir que el modelo es un buen modelo.

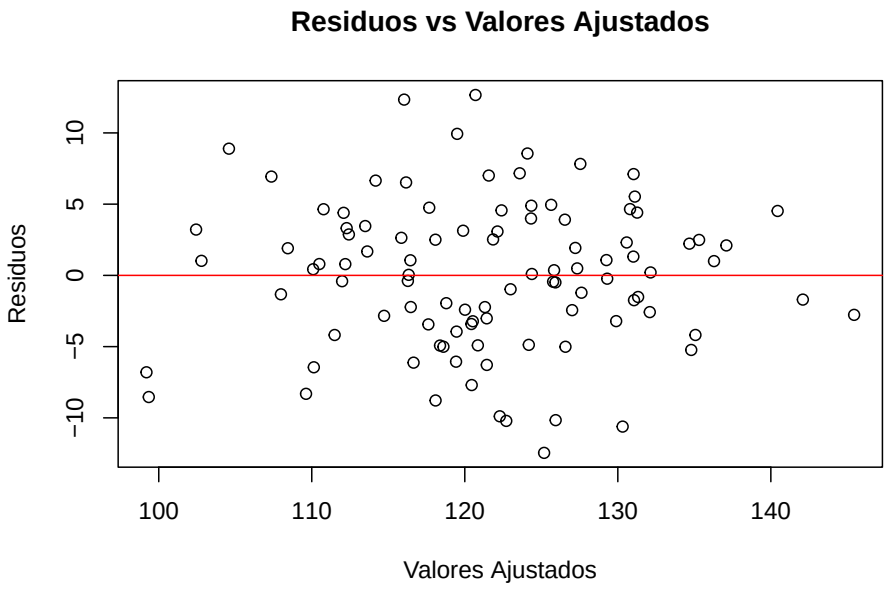
El **ANOVA** nos permite evaluar si el modelo en su conjunto es significativo. Si el valor p es menor que 0.05, el modelo es estadísticamente significativo. En este caso, el $p - valor < 0.001 < 0.05$, así que podemos rechazar la H_0 el modelo explica mas del modelo nulo que solo tiene en cuenta el intercepto.

Utilizamos el **VIF** para evaluar la multicolinealidad entre los predictores.Un valor de VIF mayor que 5 indica un problema de multicolinealidad. Si los VIF son inferiores a 5, podemos concluir que no hay problemas graves de multicolinealidad.En este estudio ya que no hay ningun valor mayor de 5 podemos concluir que no hay multicolinealidad.

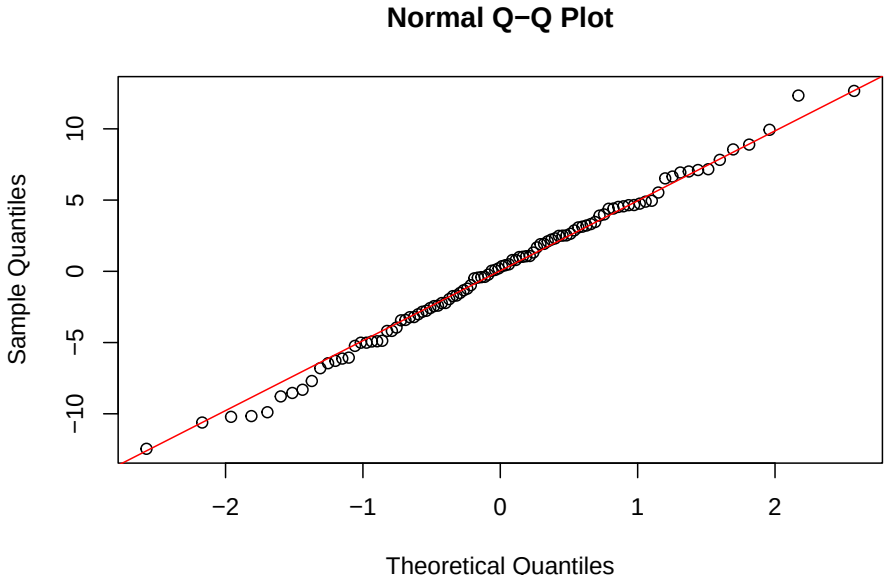
Análisis de los Residuos

Verificamos si los residuos cumplen con las suposiciones de normalidad y homocedasticidad.

1. **Gráfico de Residuos vs Valores Ajustados**. El gráfico de residuos vs valores ajustados nos permite verificar si se cumplen las condiciones de homocedasticidad y linealidad. Podemos observar que no hay ningún patron. Los valores se distribuyen al azar. Las hipótesis de homocedasticidad y linealidad se cumplen.



2. **Gráfico Q-Qnorm para Normalidad de Residuos**. El gráfico Q-Qnorm nos permite evaluar si los residuos siguen una distribución normal.Si los puntos siguen la línea diagonal, los residuos pueden considerarse normalmente distribuidos. En este caso los puntos siguen la linea diagonal y podemos concluir que la hipótesis de normalidad se cumple.



3. **Test de Shapiro-Wilk para Normalidad de los Residuos.** Este resultado es conformado por el test de Shapiro-Wilk para verificar si los residuos siguen una distribución normal. En este caso la hipótesis nula H_0 es que los residuos siguen una distribución normal. Si el p-valor del test de Shapiro-Wilk es mayor que 0.05, no rechazamos la hipótesis nula de que los residuos siguen una distribución normal. En este caso, siendo p-valor asociado $0.939 > 0.05$ no rechazamos H_0 y concluimos que los residuos siguen una distribución normal.

8.2. Caso de Estudio: Producto Interno Bruto

En este caso de estudio, analizaremos los factores que influyen en el **Producto Interno Bruto (PIB)** de un país, utilizando un modelo de regresión lineal múltiple. Este análisis incluirá variables macroeconómicas clave como:

- **PIB (Y)**: La variable dependiente que queremos explicar.
- **Inversión (X1)**: Nivel de inversión en la economía.
- **Consumo (X2)**: Gasto en consumo.
- **Exportaciones (X3)**: Exportaciones netas.

El objetivo es ajustar un modelo para explicar el PIB en función de estas variables, verificar la significatividad de los coeficientes, y realizar un análisis completo de validación del modelo.

Paso 1. Estimación del Modelo

Ajustamos el modelo de regresión lineal múltiple para predecir el PIB en función de la inversión, el consumo y las exportaciones.

Variables	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**VIF**
costante	10116.96	233.81	43.27	0.00	
Inversion	0.87	0.07	12.36	0.00	1.019
Consumo	0.71	0.03	21.74	0.00	1.00
Exportaciones	0.03	1.12	0.02	0.98	1.01
$R^2 adjusted$	**ANOVA**	Shapiro test			
0.858	0.001	0.939			

En el resumen del modelo, observamos los coeficientes estimados para cada variable, el valor de R^2 y los valores p para cada predictor. Esto nos permite evaluar la significatividad de los coeficientes.

En relación a los coeficientes podemos observar:

1. La **costante** $\beta_0 = 10323.43$ representa el valor del **PIB** si todos los indicadores (**Inversion**, **Consumo** y **Exportaciones**) fueron iguales a 0. En concreto el **PIB** seria 10116.96 M.

2. El coeficiente $\beta_{Inversion} = 0.87$. Esto quiere decir que al aumentar una unidad (1 M euro) la **inversion**, produciría un aumento de $0.87 \times 1M = 870.000$ Euros en el **PIB**, sin tener en cuenta los otros indicadores.

3. El coeficiente $\beta_{Consumo} = 0.71$. Esto quiere decir que al aumentar una unidad (1 M euro) el **consumo**, produciría un aumento de $0.71 \times 1M = 710.000$ Euros en el **PIB**, sin tener en cuenta los otros indicadores.

4. El coeficiente $\beta_{Exportaciones} = 0.03$. Esto quiere decir que al aumentar una unidad (1 M euro) las **exportaciones** produciría una disminución de $0.03 \times 1M = 30.000$ Euros en el **PIB**, sin tener en cuenta los otros indicadores.

Comparando las diferentes indicadores, aquel relacionado con las **Inversion** es el mas importante (i.e., es el factor mas importante a la hora de estimar el **PIB**).

El **test t** evalúa si los coeficientes son significativamente diferentes de cero. Si el valor p es menor que 0.05, se considera que el predictor tiene un efecto significativo sobre las ventas. En este caso todos los coeficientes son significativos a excepción de las **exportaciones** $p - valor = 0.98 > 0.05$.

El R^2 mide qué proporción de la variabilidad de las ventas está explicada por el modelo. Un R^2 alto indica que el modelo explica una gran parte de la variabilidad de las ventas. El R^2 ajustado es una versión corregida que penaliza la inclusión de variables no significativas. En este caso los dos son muy parecidos y altos ($R^2 = 0.868$ y elR^2ajustado $= 0.858$). Esto nos permite concluir que el modelo es un buen modelo para estimar el **PIB**.

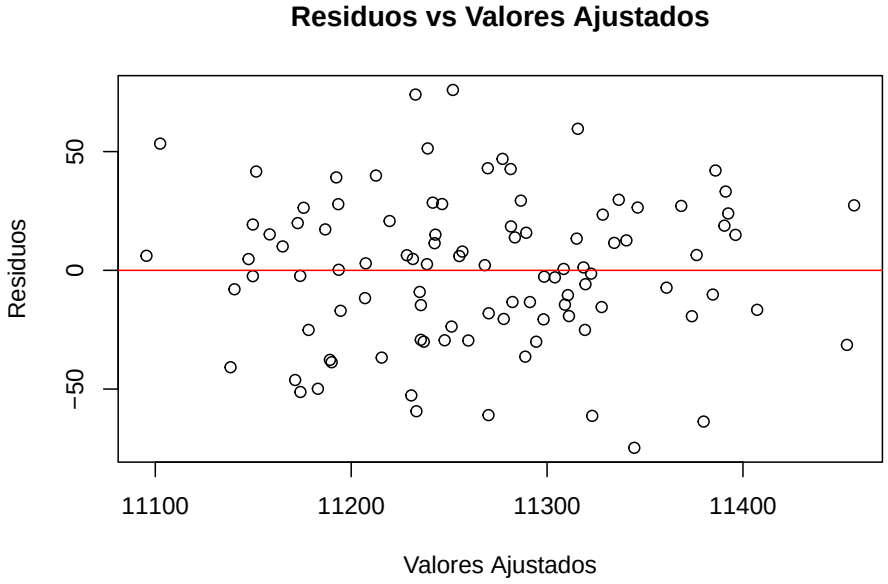
El **ANOVA** nos permite evaluar si el modelo en su conjunto es significativo. Si el valor p es menor que 0.05, el modelo es estadísticamente significativo. En este caso, el $p - valor =< 0.001 < 0.05$, así que podemos rechazar la H_0 el modelo explica mas del modelo nulo que solo tiene en cuenta la costante.

Utilizamos el **VIF** para evaluar la multicolinealidad entre los predictores. Un valor de VIF mayor que 5 indica un problema de multicolinealidad. Si los VIF son inferiores a 5, podemos concluir que no hay problemas graves de multicolinealidad. En este estudio no hay ningun valor mayor de 5.

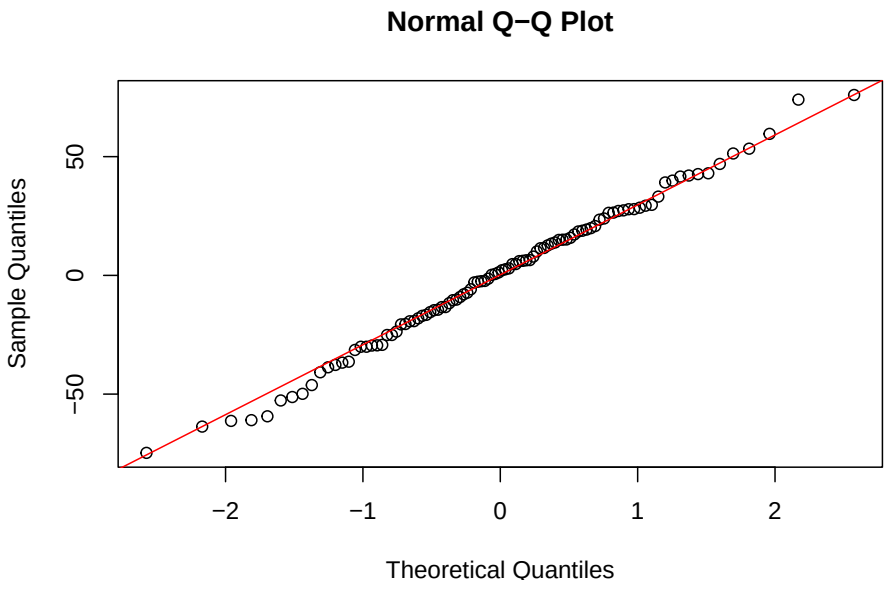
Paso 2. Análisis de los Residuos

Verificamos si los residuos cumplen con las suposiciones de normalidad y homocedasticidad.

1. **Gráfico de Residuos vs Valores Ajustados**. El gráfico de residuos vs valores ajustados nos permite verificar si se cumplen las condiciones de homocedasticidad y linealidad. Podemos observar que no hay ningún patron. Los valores se distribuyen al azar. Las hipótesis de homocedasticidad y linealidad se cumplen.



2. **Gráfico Q-Qnorm para Normalidad de Residuos**. El gráfico Q-Qnorm nos permite evaluar si los residuos siguen una distribución normal. Si los puntos siguen la línea diagonal, los residuos pueden considerarse normalmente distribuidos. En este caso los puntos siguen la linea diagonal y podemos concluir que la hipótesis de normalidad se cumple.



3. **Test de Shapiro-Wilk para Normalidad de los Residuos.** Este resultado es confirmado por el test de Shapiro-Wilk para verificar si los residuos siguen una distribución normal. Si el p-valor del test de Shapiro-Wilk es mayor que 0.05, no rechazamos la hipótesis nula de que los residuos siguen una distribución normal. En este caso, siendo p-valor asociado $0.939 > 0.05$ no rechazamos H_0 y concluimos que los residuos siguen una distribución normal.

8.3. Caso de Estudio: U invertida

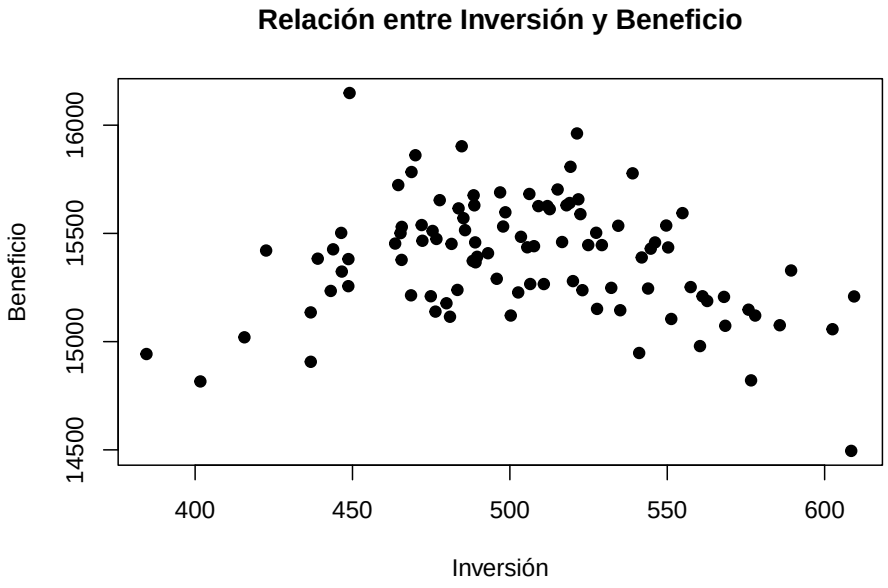
En este caso de estudio, analizaremos la relación en forma de **U invertida** entre la inversión en una empresa y los beneficios generados. A medida que aumenta la inversión, los beneficios también aumentan hasta cierto punto, pero luego comienzan a disminuir debido a rendimientos decrecientes. Este tipo de relación no puede ser capturada por un modelo lineal clásico, por lo que agregaremos un término cuadrático para capturar la relación.

Las variables que utilizaremos son:

- **Beneficio (Y)**: Variable dependiente, que representa los beneficios generados por la empresa.
- **Inversión (X1)**: La cantidad de inversión realizada por la empresa.
- **OtrosGastos (X2)**: Otros gastos en la empresa que pueden influir en los beneficios.

Paso 1. Visualización de la Relación U Invertida

Para detectar la posible forma de U invertida entre inversión y beneficio, graficamos los datos.



El gráfico muestra una relación en forma de **U invertida** entre la inversión y el beneficio, lo que sugiere que los beneficios aumentan con la inversión hasta un punto máximo, después del cual comienzan a disminuir.

Paso 2. Estimación del Modelo Lineal Clásico

A continuación, ajustamos un modelo de regresión lineal clásico sin el término cuadrático para observar los problemas que surgen al omitir la no linealidad.

Variables	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**	**VIF**
costante	13658.9447	246.7012	55.37	0.0000	
Inversion	-0.8399	0.3496	-2.40	0.0182	1.002
OtrosGastos	10.8730	0.8251	13.18	0.0000	1.002

$R^2 adjusted$	**ANOVA**
0.646	0.001

El coeficiente de inversión podría estar subestimado o sobreestimado, ya que no se tiene en cuenta la relación cuadrática. El ajuste del modelo puede ser deficiente y podría no capturar la verdadera relación entre inversión y beneficio. En particular notamos que el coeficiente es (-0.83) lo que se interpreta como: a mas inversión menos beneficios. Esta intepretacion es contradictoria y solo parece reflejar una parte de la relación entre las dos variables.

Paso 3. Incorporar el Término Cuadrático

Para resolver el problema de la no linealidad, agregamos un término cuadrático de la inversión al modelo de regresión.

Variables	Estimate	Std. Error	t value	Pr(>|t|)
costante	2321.5474	866.6940	2.68	0.0087
Inversion	44.8979	3.4519	13.01	0.0000
$Inversion^2$	-0.0452	0.0034	-13.27	0.0000
OtrosGastos	10.1911	0.4952	20.58	0.0000
$R^2 adjusted$	**ANOVA**			
0.874	0.001			

Observamos que: - El coeficiente de **Inversion** es positivo, indicando que los beneficios aumentan con la inversión inicialmente. - El coeficiente de $Inversion^2$ es ser negativo, confirmando la forma de U invertida. - Ambos coeficientes son significativos para que podamos concluir que la relación en forma de U invertida está presente.

Paso 5: Comparación de Modelos

Vamos a comparar el ajuste del modelo clásico frente al modelo cuadrático utilizando el R^2 y el **ANOVA** para evaluar cuál de los dos modelos es mejor.

Interpretación:
- El R^2 del modelo cuadrático resulta mayor, indicando un mejor ajuste del modelo a los datos. - El **test F** del ANOVA nos permite comparar si el modelo cuadrático mejora significativamente el ajuste en comparación con el modelo lineal clásico. Ya que el p-valor es <0.001, concluimos que el modelo cuadrático explica mas del modelo clasico.

9. Software estadísticos R, STATA, y JMP (SAS)

En este documento, explicaremos cómo realizar una regresión lineal múltiple en tres herramientas: **R**, **STATA**, y **JMP (SAS)**.

9.1. Regresión Lineal Múltiple en R

En **R**, la función principal para ajustar un modelo de regresión lineal múltiple es `lm()`

```
lm(formula, data, subset, weights, na.action)
```

- **formula**: La fórmula que describe el modelo, por ejemplo, `Y ~ X1 + X2`.
- **data**: El conjunto de datos que contiene las variables.
- **subset**: Un subconjunto opcional de los datos a usar en el ajuste.
- **weights**: Pesos opcionales a aplicar a las observaciones.
- **na.action**: Cómo manejar los valores ausentes.

Vamos a generar un ejemplo donde estimamos el impacto de la educación y la experiencia sobre el ingreso de los individuos.

```
##
## Call:
```

```
## lm(formula = Ingreso ~ Educacion + Experiencia, data = data_r)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3651 -3.3037 -0.6222  3.1068 10.3991
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54.43236    3.40564   15.98   <2e-16 ***
## Educacion    9.66707    0.26217   36.87   <2e-16 ***
## Experiencia  5.02381    0.09899   50.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.756 on 97 degrees of freedom
## Multiple R-squared:  0.9748, Adjusted R-squared:  0.9743
## F-statistic:  1879 on 2 and 97 DF,  p-value: < 2.2e-16
```

Interpretación de los Parámetros:
- **Estimate**: Los coeficientes estimados para cada predictor. En este caso, son los efectos de la educación y la experiencia sobre el ingreso. - **Std. Error**: Los errores estándar asociados a cada coeficiente. - **t value**: El valor de la estadística t para cada predictor. - **Pr(>|t|)**: El valor p asociado a cada predictor, que nos indica si los coeficientes son significativos (usualmente si es menor que 0.05).

9.2. Regresión Lineal Múltiple en STATA

En **STATA**, la regresión lineal múltiple se realiza con el comando `regress`. Este comando sigue la estructura:

```
regress dependent_var independent_vars
```

Supongamos que estamos trabajando con los mismos datos de educación, experiencia e ingreso.

```
* Generar los datos
set obs 100
gen Educacion = rnormal(12, 2)
gen Experiencia = rnormal(10, 5)
gen Ingreso = 50 + 10*Educacion + 5*Experiencia + rnormal(0, 5)

* Ajustar el modelo de regresión
regress Ingreso Educacion Experiencia
```

Parámetros y Resultados en STATA:
- **Coeficiente**: Indica la magnitud del efecto de cada variable independiente sobre la variable dependiente. - **Std. Err.**: El error estándar de cada coeficiente. - **t**: El valor de la prueba t. -

P>|t|: El valor p asociado a la prueba t. - **R-squared**: El coeficiente de determinación, que mide qué proporción de la variabilidad de la variable dependiente está explicada por el modelo.

Otras Opciones Útiles:
- `robust`: Realiza estimaciones robustas de errores estándar. - `vce(cluster varname)`: Calcula errores estándar agrupados por una variable. - `outreg2`: Permite exportar los resultados a tablas de forma fácil.

9.3. Regresión Lineal Múltiple en JMP (SAS)

En **JMP (SAS)**, la regresión lineal múltiple se puede realizar a través de la interface gráfica. Los pasos son los siguientes:

Pasos para Realizar la Regresión en JMP:

1. **Cargar los Datos**:
   - Abrir JMP e importar el archivo de datos que contiene las variables `Ingreso`, `Educacion`, y `Experiencia`.
2. **Seleccionar la Herramienta de Regresión**:
   - En el menú de JMP, ir a **Analyze** > **Fit Model**.
3. **Configurar el Modelo**:
   - En la ventana de **Fit Model**:
     - Arrastra la variable dependiente (`Ingreso`) al campo **Y**.
     - Arrastra las variables independientes (`Educacion` y `Experiencia`) al campo **Construct Model Effects**.
4. **Ajustar el Modelo**:
   - Haz clic en **Run** para ajustar el modelo.

Interpretación en JMP:

- **Estimates**: Los coeficientes estimados para cada variable predictora.
- **Std Error**: El error estándar asociado a cada estimación.
- **t Ratio**: El valor de la prueba t.
- **Prob>|t|**: El valor p asociado a la significatividad de cada predictor.

Funciones Avanzadas en JMP:
- **Diagnostics**: JMP ofrece herramientas gráficas avanzadas para diagnosticar problemas de ajuste del modelo, como gráficos de residuos, gráficas de leverage, etc. - **Transformations**: JMP permite fácilmente transformar las variables (por ejemplo, al cuadrado) para capturar relaciones no lineales.

ANEXO 1. Estimar una regression manualmente.

Ahora, vamos a desarrollar un ejemplo utilizando un conjunto de datos simulado y calcular manualmente los parámetros del modelo utilizando la fórmula matricial anterior.

Paso 1. Generar Datos Simulados

Vamos a generar datos para la variable dependiente Y y dos variables independientes X_1 y X_2:

```
# Generar datos simulados
set.seed(123)
n <- 100
X1 <- rnorm(n)
X2 <- rnorm(n)
epsilon <- rnorm(n, mean = 0, sd = 1)
Y <- 3 + 2*X1 + 1.5*X2 + epsilon
```

Paso 2. Crear la Matriz de Diseño

Creamos la matriz de diseño $\mathbf{X}$ que incluye un vector de unos para el intercepto y las variables $X1$ y $X2$:

```
# Crear la matriz de diseño
X <- cbind(1, X1, X2)  # Matriz con el intercepto y las variables independientes
```

Paso 3. Calcular β

Usamos la fórmula $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ para calcular los coeficientes manualmente:

```
# Calcular beta manualmente
XtX_inv <- solve(t(X) %*% X)   # (X^T * X)^-1
XtY <- t(X) %*% Y              # X^T * Y
beta <- XtX_inv %*% XtY        # beta = (X^T * X)^-1 * X^T * Y
beta  # Mostrar los coeficientes estimados
```

```
##        [,1]
##    3.135065
## X1 1.866828
## X2 1.523811
```

Paso 4. Comparación con `lm()`

Para verificar los resultados, usamos la función `lm()` de R para ajustar el mismo modelo:

```
# Ajustar el modelo con lm
modelo <- lm(Y ~ X1 + X2)
summary(modelo)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8730 -0.6607 -0.1245  0.6214  2.0798
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.13507    0.09614   32.61   <2e-16 ***
## X1           1.86683    0.10487   17.80   <2e-16 ***
## X2           1.52381    0.09899   15.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9513 on 97 degrees of freedom
## Multiple R-squared:  0.8448, Adjusted R-squared:  0.8416
## F-statistic:   264 on 2 and 97 DF,  p-value: < 2.2e-16
```

Al comparar los resultados obtenidos manualmente con la fórmula matricial y los resultados de la función `lm()`, podemos ver que son idénticos, confirmando que hemos aplicado correctamente la teoría de los mínimos cuadrados.

Regresión simple

1. Introducción

1.1. ¿Qué es la regresión lineal?

La **regresión lineal** es una técnica estadística que, mediante una equacion linenal, modela la relación entre una variable dependiente (i.e., variable target, variable objectivo) Y y una o más variables independientes (i.e., variables predictoras) X.

La regresión lineal se considera un método supervisado porque, durante el proceso de entrenamiento, el algoritmo utiliza un conjunto de datos que incluye tanto las variables explicativas (independientes) como la variable objetivo (dependiente) cuyos valores ya son conocidos. En otras palabras, el modelo "aprende" a predecir los valores de la variable dependiente usando los valores de las variables dependientes y indipendietes que ya son conocidos.

En su forma más simple, la regresión lineal se refiere a la **regresión lineal simple**, que involucra solo una variable independiente. Sin embargo, cuando se utilizan varias variables predictoras, hablamos de **regresión lineal múltiple**.

La forma general de un modelo de regresión lineal simple es:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Donde:

- Y es la variable dependiente o respuesta.
- X es la variable independiente o predictor.
- β_0 es (costante), que representa el valor esperado (i.e., valor promedio estimado) de Y cuando $X = 0$.
- β_1 es el coeficiente de regresión, que mide el cambio en Y por cada unidad de cambio en X (i.e., la relación que existe entre las dos variables: predictora y dependiente)
- ϵ es el término de error o residuo, que captura la variabilidad no explicada por el modelo.

1.2. ¿Para qué sirve la regresión lineal?

En investigación de mercado, la regresión lineal, es uno de los metodos mas utilizado ya que permite:

1. **Predecir valores**: hacer predicciones sobre la variable dependiente con base en los valores de la(s) variable(s) independiente(s). Esto quiere decir que a partir de unos valores conocidos de la(s) variable(s) independiente(s) y de las variable dependiente, una vez estimado el modelo y comprendida la relación que existe entre las variables podemos predecir nuevo valores de la variables dependiente. Por ejemplo, determinar el valor de las ventas en función de un nuevo budjet, predecir indicadores económicos o patrones de consumo.

2. **Entender relaciones**: facilita la comprensión de la relación entre las variables y la magnitud del efecto que una variable independiente tiene sobre la dependiente. Los coeficientes de la regresión lineal, pueden ser utilizados para determinar hasta que punto una variable predictora

tiene efecto sobre la variable dependiente. Por ejemplo, podemos determinar cuales son los aspectos que más influyen en la satisfacción de un consumidor, o identificar los indicadores que mas importantes en la calidad de vida de una persona.

3. **Evaluar la significancia**: Nos permite evaluar la significancia estadística de las relaciones entre las variables (i.e., si una variable predictora tiene un efecto significativo o no segun un determinado nivel de confianza).

4. **Explicar variabilidad**: A través del coeficiente R^2, podemos saber qué proporción de la variabilidad (i.e., información contenida en la variable) en Y es explicada por el modelo de regresión.

5. **Identificar posibles outliers**: Analizando los residuos, se pueden identificar observaciones que se comportan de manera anómala.

1.3. Aplicaciones de la regresión lineal

En la investigación de mercado, la regresión lineal se usa para:

- **Predecir el comportamiento del consumidor**: Se puede analizar cómo factores como el precio, la publicidad o la satisfacción del cliente afectan las decisiones de compra.
- **Estimación de demanda**: Permite modelar la relación entre el precio de un producto y su demanda.
- **Segmentación de mercado**: Ayuda a identificar qué características de los clientes influyen más en su comportamiento de compra.

Ejemplo en investigación de mercado:

Supongamos que una empresa desea predecir el nivel de ventas en función del presupuesto publicitario en TV. Un modelo de regresión lineal simple podría tener la forma:

$$Ventas = \beta_0 + \beta_1 \times Presupuesto\_TV$$

Este modelo nos permitiría estimar cuántas unidades adicionales se venderán por cada mil dólares adicionales en publicidad en televisión.

En marketing, la regresión lineal tiene aplicaciones directas, tales como:

- **Efectividad de campañas publicitarias**: Determina qué porcentaje de variación en las ventas se puede explicar por las campañas publicitarias.
- **Precios óptimos**: Modela la relación entre el precio de un producto y las ventas para determinar cuál es el precio óptimo.
- **Retorno de inversión (ROI)**: Analiza el impacto de diferentes iniciativas de marketing en los ingresos de una empresa.

2. Hipótesis del modelo de regresión lineal simple

Para establecer una relación entre una variable dependiente (respuesta) y una o más variables independientes (predictoras) utilizando la **regresión lineal** hay que verificar algunas hipótesis.

La hipótesis de modelo son:
1. Linealidad
2. Independencia de los errores
3. Homocedasticidad
4. Normalidad de los errores

El cumplimento de la hipótesis es importante para garantizar la validez del modelo.

Linealidad

La relación entre la variable dependiente Y y la variables independiente X debe ser lineal. Esto quiere decir que su relación puede ser explicada y formalizada mediante una ecuación lineal:

$$Y_i = \beta_0 + \beta X_i + \epsilon_i$$

Donde:

- Y_i es el valor de la variable dependiente para el i-ésimo observación.
- β_0, β son los coeficientes del modelo.
- X_i es el valor de la variable predictora para el i-ésimo observación.
- ϵ_i es el error aleatorio o residuo para el i-ésimo observación.

Ejemplo. Supongamos que queremos predecir las ventas (Y) de una empresa en función del presupuesto publicitario en TV (X_1). Para poder aplicar la **regresión lineal** es necesario que la relación entre el presupuesto publicitario y ventas se tiene que poder explicar/formular mediante la ecuación lineal:

$$\text{Ventas} = \beta_0 + \beta \times \text{TV} + \epsilon$$

Independencia de los errores

Los errores ϵ_i deben ser independientes entre sí. Es decir, no debe haber correlación entre los residuos.

Hipótesis nula (H_0):

$$\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j$$

El incumplimiento de H_0 puede suponer que el modelo no sea capaz de representar y estimar la relación entre la variable dependiente y predictora o que esta relación no sea lineal.

Homocedasticidad

La varianza de los errores ϵ_i debe ser constante para todos los valores de X.

Hipótesis nula (H_0):

$$\text{Var}(\epsilon_i) = \sigma^2$$

Si la varianza de los errores cambia a lo largo de los valores de X, se produce heterocedasticidad, lo que puede invalidar los resultados del modelo. De nuevo, el incumplimiento de H_0, puede suponer que el modelo no sea capaz de representar y estimar la relación entre la variable dependiente y predictora o que esta relación no sea lineal.

Normalidad de los errores

Los errores ϵ_i deben seguir una distribución normal con media cero y varianza constante.

Hipótesis nula (H_0):

$$\epsilon_i \sim N(0, \sigma^2)$$

El incumplimiento de H_0, puede suponer una menor capacidad predictiva del modelo e estimaciones menos eficientes, es decir, podrían no ser las mejores o más precisas posibles. La normalidad de los errores, como veremos más adelante, se puede verificar visualmente con un gráfico Q-Qnorm o mediante pruebas estadísticas de normalidad (Shapiro-Wilk, Kolmogorov-Smirnov).

3. Como se estiman los parametros del modelo (metodo de los minimos cuadrado OLS)

El objetivo de la regresión lineal es encontrar una ecuación lineal que represente de manera óptima la relación entre las variables independientes y la variable dependiente. Esto se consigue ajustando una línea a los datos utilizando el método de mínimos cuadrados. Si definimos el modelo como:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Donde:

- Y es la variable dependiente.
- X es la variable independiente.
- β_0 es la ordenada en el origen (costante).
- β_1 es la pendiente de la recta.
- ϵ es el error aleatorio, que captura la desviación entre los valores observados y los predichos.

El objetivo es encontrar los valores de β_0 y β_1 que minimicen el error.

3.1. Cálculo de los parámetros mediante mínimos cuadrados

El método de mínimos cuadrados se usa para estimar los parámetros β_0 y β_1. Este método minimiza la suma de los errores cuadrados (RSS: Residual Sum of Squares) entre los valores observados Y_i y los valores predichos por el modelo $\hat{Y}_i$.

Paso 1. Definir el error cuadrático residual

El error residual o residuo se define como la diferencia entre el valor observado Y_i y el valor predicho $\hat{Y}_i$. El valor predicho está dado por la ecuación de la recta:

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

El residuo para cada observación i es:

$$e_i = Y_i - \hat{Y}_i$$

La suma de los errores al cuadrado es:

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Paso 2. Minimizar la función objetivo

Para encontrar los valores de β_0 y β_1 que minimicen RSS, necesitamos derivar esta función con respecto a β_0 y β_1, y luego igualar estas derivadas a cero.

Primero, derivamos RSS con respecto a β_0:

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

Igualamos a cero para minimizar:

$$-2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

Dividimos por -2:

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

Esto se puede reescribir como:

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$

Si dividimos todos los valores por n y llamemos $\bar{Y}$ y $\bar{X}$ las medias de Y y X, respectivamente, obtenemos:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Paso 3. Derivar con respecto a β_1

Ahora derivamos RSS con respecto a β_1:

$$\frac{\partial RSS}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

Igualamos a cero para minimizar:

$$-2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

Dividimos por -2:

$$\sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

Sustituimos β_0 y simplificamos:

$$\sum_{i=1}^{n} X_i Y_i = (\bar{Y} - \beta_1 \bar{X}) \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

Despejamos β_1:

$$\beta_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Esta es la fórmula para la pendiente β_1.

Paso 4. Cálculo de β_0

Una vez que tenemos β_1, podemos calcular β_0 usando la fórmula:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Resumen

Los parámetros de la regresión lineal simple se calculan con las siguientes fórmulas:

- **Pendiente**:

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- **Costante**:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

3.2. Ejemplo práctico

Supongamos que tenemos los siguientes datos de ventas y presupuesto publicitario en TV:

TV ($1000)	Ventas (unidades)
230	22
44	10
17	5
151	14
180	17

Paso 1. Calcular las medias $\bar{X}$ y $\bar{Y}$

Las medias de las variables TV y Ventas son:

$$\bar{X} = \frac{230 + 44 + 17 + 151 + 180}{5} = \frac{622}{5} = 124.4$$

$$\bar{Y} = \frac{22 + 10 + 5 + 14 + 17}{5} = \frac{68}{5} = 13.6$$

Paso 2: Calcular β_1

Usamos la fórmula para β_1:

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

4. Significancia de los parámetros

Una vez estimados los parámetros del modelo, es muy importante determinar su validez. En otras palabras, tenemos que determinar si son significantivos (i.e., diversos de 0) a un determinado nivel de confianza.

La significancia de $\beta_0 \quad y \quad \beta_1$ se puede verificar utilizando dos métodos principales:

1. **Prueba de hipótesis con el t-test**.

2. **Intervalos de confianza**.

t-test

Queremos verificar si los coeficientes de regresión β_1 y β_0 son significativamente diferentes de cero. Las hipótesis para la pendiente β_1 son:

- **Hipótesis nula (H_0)**: El coeficiente de pendiente es igual a cero (no hay relación lineal entre X y Y):

$$H_0 : \beta_1 = 0$$

- **Hipótesis alternativa (H_1)**: El coeficiente de pendiente es diferente de cero (hay una relación lineal entre X y Y):

$$H_1 : \beta_1 \neq 0$$

Para la costante β_0, el procedimiento es análogo.

Para cada parámetro, podemos realizar una prueba de hipótesis utilizando el estadístico t, que se calcula como:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Donde:

- $\hat{\beta}_1$ es el valor estimado del parámetro de la pendiente.
- $SE(\hat{\beta}_1)$ es el error estándar de la estimación de β_1.

El error estándar de $\hat{\beta}_1$ se calcula como:

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Donde s es la desviación estándar de los residuos y se calcula como:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

Calculado el estadístico, el valor de t calculado se compara con una distribución t de Student con $n-2$ grados de libertad para determinar si rechazamos la hipótesis nula.

Si el p-valor asociado es menor que el nivel de significancia (comúnmente $\alpha = 0.05$), rechazamos H_0 y concluimos que el parámetro es significativamente diferente de cero.

Notar que en los principales software estadísticos, el el valor de t calculado, los grados de libertad, y p-valor asociado ya vienen calculado. Así que lo unico que tendremos que hacer es comparar el p-valor asociado es menor que el nivel de significancia (comúnmente $\alpha = 0.05$).

Intervalos de confianza

Otra forma de verificar si los parámetros β_0 y β_1 son significativos es construyendo intervalos de confianza y verificando que el intervalo no incluya el 0.

El intervalo de confianza para un parámetro β_1 se calcula como:

$$IC(\beta_1) = \hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot SE(\hat{\beta}_1)$$

Donde:

- $\hat{\beta}_1$ es la estimación del parámetro.
- $t_{\alpha/2,n-2}$ es el valor crítico de la distribución t de Student para un nivel de confianza $(1-\alpha)$ y $n-2$ grados de libertad.
- $SE(\hat{\beta}_1)$ es el error estándar de $\hat{\beta}_1$.

Si el intervalo de confianza no incluye el valor 0, podemos concluir que el parámetro es significativamente diferente de cero. Esto es equivalente a rechazar la hipótesis nula en la prueba de t-test.

Por ejemplo, si calculamos el intervalo de confianza para β_1 y obtenemos $IC(\beta_1) = [0.042, 0.052]$, dado que 0 no está en este intervalo, concluimos que el coeficiente es significativo.

5. Validación del Modelo

Después de estimar un modelo de regresión lineal simple, y verificar las significancia de los parametros, es fundamental validar el modelo. Esto implica verificar qué tan bien se ajusta el modelo a los datos y si es significativo. A continuación, veremos dos formas de validar un modelo de regresión lineal simple:

1. El coeficiente de determinación R^2.
2. El test ANOVA (análisis de varianza).

Coeficiente de determinación R^2

El coeficiente de determinación R^2 mide la proporción de la variabilidad en la variable dependiente Y que puede ser explicada por la variable independiente X. Su valor está comprendido entre 0 y 1:

- $R^2 = 1$: El modelo explica toda la variabilidad de Y.
- $R^2 = 0$: El modelo no explica nada de la variabilidad de Y.

El coeficiente R^2 se calcula como:

$$R^2 = \frac{\text{Suma de los cuadrados explicados (SSR)}}{\text{Suma total de los cuadrados (SST)}}$$

También podemos expresarlo como:

$$R^2 = 1 - \frac{\text{Suma de los cuadrados residuales (SSE)}}{\text{Suma total de los cuadrados (SST)}}$$

Donde:

- **SST**: Suma total de los cuadrados.

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- **SSE**: Suma de los cuadrados residuales.

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- **SSR**: Suma de los cuadrados explicados.

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

Consideremos los datos del siguiente ejemplo:

TV ($1000)	Ventas (unidades)
230	22
44	10
17	5
151	14
180	17

Primero, estimamos el modelo de regresión lineal simple para obtener:

$$\hat{Y}_i = 7.032 + 0.047X_i$$

Paso 1. Calcular $\bar{Y}$ La media de Y es:

$$\bar{Y} = \frac{22 + 10 + 5 + 14 + 17}{5} = 13.6$$

Paso 2: Calcular SST (Suma total de los cuadrados)

$$SST = (22 - 13.6)^2 + (10 - 13.6)^2 + (5 - 13.6)^2 + (14 - 13.6)^2 + (17 - 13.6)^2$$

$$SST = 70.56$$

Paso 3. Calcular SSE (Suma de los cuadrados residuales)

Primero, obtenemos los valores predichos $\hat{Y}_i$:

- Para $X_1 = 230$, $\hat{Y}_1 = 7.032 + 0.047 \times 230 = 17.842$
- Para $X_2 = 44$, $\hat{Y}_2 = 7.032 + 0.047 \times 44 = 9.100$
- Para $X_3 = 17$, $\hat{Y}_3 = 7.032 + 0.047 \times 17 = 7.832$
- Para $X_4 = 151$, $\hat{Y}_4 = 7.032 + 0.047 \times 151 = 14.109$
- Para $X_5 = 180$, $\hat{Y}_5 = 7.032 + 0.047 \times 180 = 15.472$

Luego, calculamos SSE:

$$SSE = (22 - 17.842)^2 + (10 - 9.100)^2 + (5 - 7.832)^2 + (14 - 14.109)^2 + (17 - 15.472)^2$$

$$SSE = 17.88$$

Paso 4. Calcular SSR (Suma de los cuadrados explicados)

$$SSR = SST - SSE = 70.56 - 17.88 = 52.68$$

Paso 5. Calcular R^2

$$R^2 = \frac{SSR}{SST} = \frac{52.68}{70.56} = 0.746$$

El valor de $R^2 = 0.746$ indica que el 74.6% de la variabilidad de las ventas puede ser explicada por el presupuesto en TV. Esto sugiere que el modelo se ajusta razonablemente bien a los datos.

Test ANOVA

El análisis de varianza (ANOVA) nos permite probar la significancia general del modelo de regresión. En particular, nos permite probar si al menos una de las variables predictoras tiene un efecto significativo sobre la variable dependiente.

- **Hipótesis nula (H_0)**: El coeficiente de regresión es igual cero (el modelo no tiene poder explicativo).

$$H_0 : \beta_1 = 0$$

- **Hipótesis alternativa (H_1)**: El coeficiente de regresión es diferente de cero (el modelo tiene poder explicativo).

$$H_1 : \beta_1 \neq 0$$

El estadístico F se calcula como:

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

Donde:

- k es el número de predictores en el modelo (en la regresión simple, $k = 1$).
- n es el número de observaciones.

Usamos los valores de SSR y SSE que calculamos anteriormente:

- $SSR = 52.68$
- $SSE = 17.88$
- $n = 5$ (número de observaciones)
- $k = 1$ (una variable independiente)

Entonces:

$$F = \frac{\frac{52.68}{1}}{\frac{17.88}{5-1-1}} = \frac{52.68}{5.96} = 8.84$$

El valor $F = 8.84$ se compara con el valor crítico de F para 1 y 3 grados de libertad. Si el valor F es mayor que el valor crítico, o si el p-valor es menor que 0.05, rechazamos H_0, indicando que el modelo explica mas del modelo nulo (que solo considera la intercepta)

6. Validación de los Residuos en un Modelo de Regresión Lineal Simple

La validación de los residuos es una etapa crucial en el análisis de regresión lineal, ya que nos permite evaluar si los supuestos del modelo se cumplen. En un modelo de regresión lineal simple, los residuos deben cumplir con las siguientes condiciones:

1. **Independencia**: Los residuos deben ser independientes entre sí.
2. **Normalidad**: Los residuos deben seguir una distribución normal.
3. **Homocedasticidad**: Los residuos deben tener una varianza constante.
4. **Linealidad**: La relación entre los residuos y los predictores debe ser lineal.

El análisis de los residuos se realiza principalmente mediante gráficos y test estadísticos.

6.1. Gráficos

a. Gráfico de Residuos vs. Valores Ajustados

Este gráfico nos permite verificar la **homocedasticidad** y la **linealidad**. En un modelo bien ajustado, los residuos deben distribuirse de manera aleatoria alrededor de 0, sin mostrar patrones claros.

- **Cómo se construye**: En el eje X se colocan los valores ajustados $\hat{Y}_i$, y en el eje Y los residuos e_i.
- **Interpretación**:
  - Si los residuos están distribuidos de forma aleatoria, se cumple la homocedasticidad y la linealidad.
  - Si los residuos presentan un patrón (curvado o cónico), esto indica una violación de estos supuestos.

b. Gráfico Q-Q norm (Quantile-Quantile)

El gráfico Q-Qnorm se utiliza para evaluar la **normalidad** de los residuos. Compara los cuantiles observados de los residuos con los cuantiles teóricos de una distribución normal.

- **Cómo se construye**: Se representan los residuos en el eje Y y los cuantiles teóricos de una distribución normal en el eje X.
- **Interpretación**:
  - Si los puntos caen sobre la línea diagonal, los residuos son aproximadamente normales.
  - Si los puntos se desvían de la línea, hay una violación del supuesto de normalidad.

c. Gráfico de Residuos vs. Predictores

Este gráfico nos ayuda a verificar la **independencia** y la **linealidad** de los residuos en relación con las variables predictoras.

- **Cómo se construye**: En el eje X se coloca el valor de los predictores X_i y en el eje Y, los residuos e_i.
- **Interpretación**:
  - La independencia se cumple si no se observa un patrón claro.
  - La linealidad se cumple si los residuos están distribuidos de manera aleatoria alrededor de 0.

6.2. Pruebas Estadísticas para Validar Residuos

Además de los gráficos, existen pruebas estadísticas que nos permiten verificar cuantitativamente si los residuos cumplen con los supuestos del modelo.

a. Prueba de Normalidad: Test de Shapiro-Wilk

El **test de Shapiro-Wilk** es una prueba estadística que evalúa si los residuos siguen una distribución normal.

- **Hipótesis**:
  - H_0: Los residuos siguen una distribución normal.
  - H_1: Los residuos no siguen una distribución normal.
- **Fórmula**:

$$W = \frac{\left(\sum_{i=1}^{n} a_i r_{(i)}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Donde a_i son los coeficientes tabulados y $r_{(i)}$ son los residuos ordenados.

- **Interpretación**:
  - Si el p-valor es menor que 0.05, rechazamos H_0, indicando que los residuos no son normales.

c. Prueba de Independencia: Test de Durbin-Watson

El **test de Durbin-Watson** se utiliza para verificar la **independencia** de los residuos

- **Hipótesis**:
  - H_0: No hay correlación entre los residuos.

- H_1: Hay correlación entre los residuos.

- **Fórmula**:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

- **Interpretación**:

  - Un valor de DW cercano a 2 indica independencia.
  - Valores cercanos a 0 sugieren autocorrelación positiva, y valores cercanos a 4 sugieren autocorrelación negativa.

7. Ejemplo: Analisis de la relación entre el ingreso personal y el nivel de consumo

Este caso de estudio se centra en analizar la relación entre el **ingreso personal** y el **nivel de consumo** en una muestra de individuos. El objetivo es determinar si el ingreso tiene un impacto significativo en el nivel de consumo y evaluar la validez del modelo estimado. Supongamos que tenemos datos sobre el ingreso personal (en miles de dólares) y el nivel de consumo (también en miles de dólares) de 10 individuos:

	Ingreso	**Consumo**
1	25.00	20.00
2	35.00	30.00
3	45.00	35.00
4	30.00	25.00
5	55.00	45.00
6	40.00	33.00
7	60.00	50.00
8	50.00	42.00
9	70.00	55.00
10	65.00	52.00

Paso 2: Estimación del modelo

Ajustamos un modelo de regresión lineal para determinar si el ingreso personal afecta significativamente el nivel de consumo.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
costante	1.4485	1.3151	1.10	0.3027
Ingreso	0.7842	0.0265	29.59	0.0000

R^2	**ANOVA**	**Shapiro**	**DW**
0.99	0.001	0.6302	2.232

A partir de la salida podemos realizar la siguiente interpretación:

1. El valor de la costante β_0 representa el nivel de **consumo** cuando el ingreso personal es cero. Sin tener en cuenta los **ingresos** (i.e., si los ingresos fueron 0) el valor esperado del consumo seria $1.4485 = 1448.5 \quad dolares$. El coeficiente de la variable β_1 indica cuánto aumenta el **consumo** por cada mil dólares adicionales en ingreso personal. El aumento de una unidad en los **ingresos** (1000 dólares) produce un aumento en el **consumo** de $0.7842 = 784.2 \quad dolares$. Notar que multiplicamos por 1000 ya que las dos variables están expresadas en 1000 de dolares.

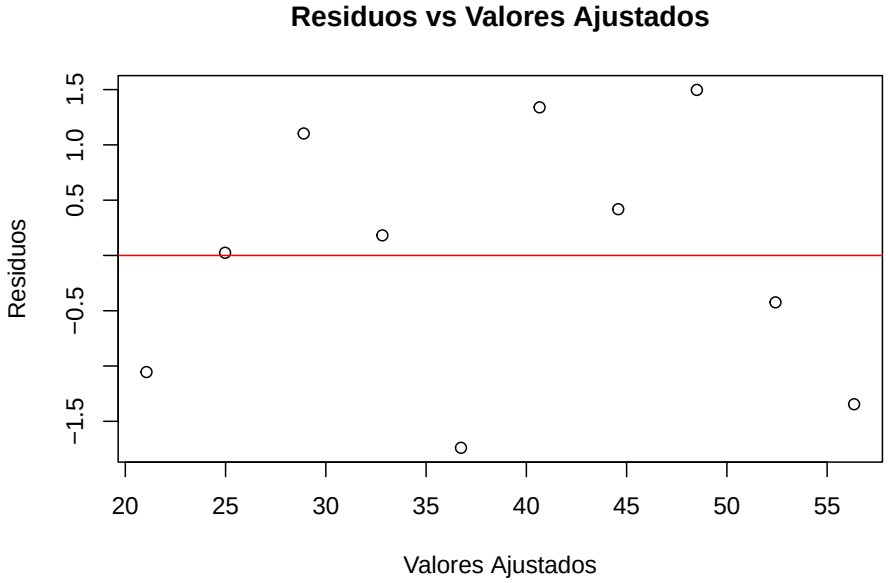
2. La significatividad de los parámetros se evalúa utilizando el t-test. En este caso queremos verificar si el ingreso tiene un efecto significativo en el nivel de consumo. Observamos el p-valor para verificar si $\beta_0 \quad$ y $\quad \beta_1$ son significativamente diferente de cero. Podemos concluir que la costante no es significativa mientras que el **consumo** si lo es, considerando un nivel de confianza de $95\% = p - valor 0.05$. Notar que la costante no es determinante en el análisis pero si lo es el coeficiente β_1 ya que es este que permite entender la relación entre **consumo** y **ingresos**.

3. El coeficiente R^2 nos indica qué proporción de la variabilidad del consumo es explicada por el ingreso personal. Si el valor de R^2 es cercano a 1, esto significa que una gran parte de la variabilidad en el consumo es explicada por el modelo. Si es cercano a 0, el modelo no es bueno. En este caso el valor es muy alto 0.99 (tratándose de valores simulados). Esto quiere decir que el modelo explica casi perfectamente el **consumo**. La componente residual será muy pequeña.

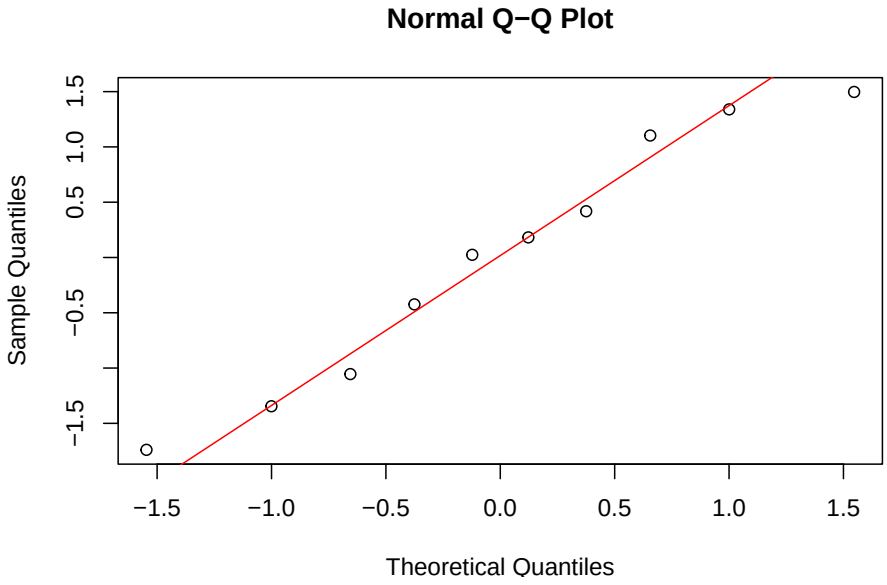
4. El análisis de varianza (ANOVA) nos permite probar la significancia general del modelo. El estadístico F y su p-valor asociado nos indican si el modelo en su conjunto es significativo. Si el p-valor es menor que 0.05, podemos concluir que el modelo tiene poder explicativo. En este caso, el p-valor asociado es $< 0.001 < 0.05$. Esto nos permite rechazar la hipótesis H_0, el modelo explica mas de el modelo nulo (el que simplemente incluye la intercepta).

5. Análisis de los residuos permiten terminar de validar el modelo.

- El gráfico de residuos vs valores ajustados nos permite verificar si se cumplen las condiciones de homocedasticidad y linealidad. Podemos observar que no hay ningún patron. Los valores se distribuyen al azar. Las hipótesis de homocedasticidad y linealidad se cumplen.



El gráfico Q-Qnorm nos permite evaluar si los residuos siguen una distribución normal. Si los puntos siguen la línea diagonal, los residuos pueden considerarse normalmente distribuidos. En este caso los puntos siguen la linea diagonal y podemos concluir que la hipótesis de normalidad se cumple.



Este resultado es conformado por el test de Shapiro-Wilk. En este caso la hipótesis nula H_0 es que los residuos siguen una distribución normal. Si el p-valor del test de Shapiro-Wilk es mayor que 0.05, no rechazamos la hipótesis nula de que los residuos siguen una distribución normal. En este caso siendo p-valor asociado $0.6302 > 0.05$ no rechazamos H_0 y concluimos que los residuos siguen una distribución normal.

Por ultimo podemos aplicar el test de Durbin-Watson para verificar la independencia de los residuos. Un valor de DW cercano a 2 indica que los residuos son independientes. Valores cercanos a 0 sugieren autocorrelación positiva, mientras que valores cercanos a 4 sugieren autocorrelación negativa. En este caso $DW = 2.2323$. La independencia de los residuos se cumple.

8. Software estadísticos R, STATA, y JMP (SAS)

En este apartado, explicaremos cómo calcular la regresión lineal en tres software estadísticos ampliamente utilizados: **R**, **STATA**, y **JMP (SAS)**. A lo largo de la explicación, abordaremos las funciones principales, los parámetros más importantes y proporcionaremos ejemplos prácticos.

8.1. Regresión Lineal en R

En R, el cálculo de la regresión lineal se realiza principalmente a través de la función `lm()`. Esta función permite ajustar modelos lineales de manera simple. La sintaxis general es:

```
lm(formula, data)
```

- **formula**: Especifica la relación entre la variable dependiente y las independientes. Se escribe en formato `y ~ x1 + x2`, donde `y` es la variable dependiente y `x1, x2` son las variables independientes.
- **data**: El conjunto de datos donde se encuentran las variables.

La función `lm()` devuelve:

1. **formula**: La relación entre las variables dependientes e independientes. Ejemplo: `Ventas ~ Presupuesto_TV`.
2. **data**: El conjunto de datos en forma de data frame.
3. **coefficients**: Nos devuelve los coeficientes estimados del modelo, es decir, los valores de β_0, β_1, etc.
4. **fitted.values**: Devuelve los valores ajustados $\hat{Y}_i$, es decir, los valores predichos por el modelo.
5. **residuals**: Devuelve los residuos, es decir, la diferencia entre los valores observados Y_i y los valores predichos $\hat{Y}_i$.

Ejemplo en R

```r
# Cargar datos simulados
TV <- c(250, 200, 180, 300, 350)
Ventas <- c(25, 20, 18, 30, 35)
datos <- data.frame(TV, Ventas)

# Ajustar el modelo de regresión lineal
modelo <- lm(Ventas ~ TV, data = datos)

# Resumen del modelo
summary(modelo)

# Coeficientes estimados
modelo$coefficients
```

8.2. Regresión Lineal en STATA

En STATA, la regresión lineal se realiza utilizando el comando `regress`. La sintaxis es bastante directa y clara. La sintaxis general es:

```
regress y x1 x2
```

Donde: - `y` es la variable dependiente. - `x1`, `x2` son las variables independientes.

La función `regress` devuelve:

1. **y**: Variable dependiente.
2. **x1, x2, ...**: Variables independientes.
3. **_b[]**: Devuelve los coeficientes de las variables independientes y del intercepto.
4. **_se[]**: Devuelve los errores estándar de los coeficientes estimados.
5. **r2**: Devuelve el coeficiente de determinación R^2.

Ejemplo en STATA

Supongamos que tenemos un conjunto de datos donde queremos analizar la relación entre el presupuesto en TV y las ventas.

```
regress Ventas TV
```

8.3. Regresión Lineal en JMP (SAS)

En JMP, la regresión lineal se realiza utilizando las opciones de menú o mediante **scripts**. El entorno JMP es altamente visual y facilita la interpretación de los resultados.

Realizar la regresión en JMP

1. **Cargar los datos**: Importar los datos a JMP en formato de tabla.
2. **Análisis de regresión**:
    - Ir a **Analyze** > **Fit Model**.
    - En la ventana emergente, seleccionar la variable dependiente bajo **Y, Response**.
    - Seleccionar las variables independientes bajo **Construct Model Effects**.
    - Hacer clic en **Run** para generar el modelo.

Parámetros y Salidas principales en JMP

1. **Coeficientes**: Estimaciones de los coeficientes de regresión β_0 y β_1.
2. **Summary of Fit**:
    - **R^2**: Proporción de la variabilidad en Y explicada por las variables X.
    - **Root Mean Square Error (RMSE)**: Medida del error estándar de los residuos.
3. **Analysis of Variance (ANOVA)**:
    - **F-ratio**: Estadístico que nos ayuda a evaluar si el modelo es significativo.
    - **p-value**: Para determinar si los coeficientes son significativos.

Ejemplo en JMP

Supongamos que tenemos un conjunto de datos que contiene las columnas **Presupuesto_TV** y **Ventas**. Para realizar una regresión en JMP:

1. Importamos los datos en JMP.
2. Vamos a **Analyze** > **Fit Model**.
3. Seleccionamos **Ventas** como la variable dependiente y **Presupuesto_TV** como la variable independiente.
4. Hacemos clic en **Run**.

Regresión logística

1. Introducción

La **regresión logística** (también conocida como regresión logit o modelo logit) fue desarrollada por el estadístico David Cox en 1958. Es un modelo de regresión donde la variable de respuesta, Y, es categórica. La regresión logística permite alcanzar dos objetivos principales:

1. Estimar la probabilidad de una respuesta categórica en función de una o más variables predictoras (X).
2. Medir en qué grado una variable predictora incrementa o disminuye la probabilidad de un resultado específico, expresado como un porcentaje.

En este curso, nos enfocaremos en el caso en que la variable de respuesta es **binaria**, es decir, cuando la variable toma dos valores, "0" y "1", que representan resultados como aprobar/fallar, ganar/perder, o comprar/no comprar. Introduciremos la **regresión logística simple** cuando solo hay una variable predictora, y la **regresión logística múltiple** cuando hay más de un predictor.

La regresión logística tiene diversas aplicaciones en el ámbito del Marketing, entre ellas:

- **Previsión**: La regresión logística se puede utilizar para predecir oportunidades y riesgos futuros. Por ejemplo, podemos emplearla para estimar la cantidad de artículos que un consumidor probablemente comprará o predecir cuántos compradores pasarán frente a una cartelera específica, lo que puede ayudar a calcular el valor máximo de oferta por un anuncio. También es útil para medir las tasas de éxito de campañas de marketing. Las compañías de seguros utilizan frecuentemente la regresión para estimar la solvencia crediticia de sus asegurados y prever el número de reclamaciones en un período determinado.
- **Toma de decisiones**: Hoy en día, las empresas se enfrentan a una sobrecarga de datos sobre finanzas, operaciones y comportamiento del cliente. El análisis de regresión aporta un enfoque científico a la gestión empresarial. Al transformar grandes cantidades de datos en información procesable, la regresión facilita la toma de decisiones más inteligentes y precisas.
- **Corrección de errores**: La regresión no solo se utiliza para identificar errores en decisiones pasadas. Por ejemplo, un gerente de una tienda podría creer que ampliar el horario de atención aumentará significativamente las ventas. Sin embargo, el análisis de regresión puede mostrar que el incremento en los ingresos no compensaría los costos adicionales de operación, proporcionando así un apoyo cuantitativo para evitar decisiones basadas únicamente en la intuición.
- **Identificación de patrones**: Las técnicas de análisis de regresión pueden descubrir relaciones entre diferentes variables al identificar patrones previamente no detectados. Por ejemplo, el análisis de datos de los sistemas de punto de venta puede revelar patrones de demanda estacional, como el aumento de ventas en ciertos días del año.

2. ¿Por qué usar regresión logística en lugar de regresión lineal para variables categóricas?

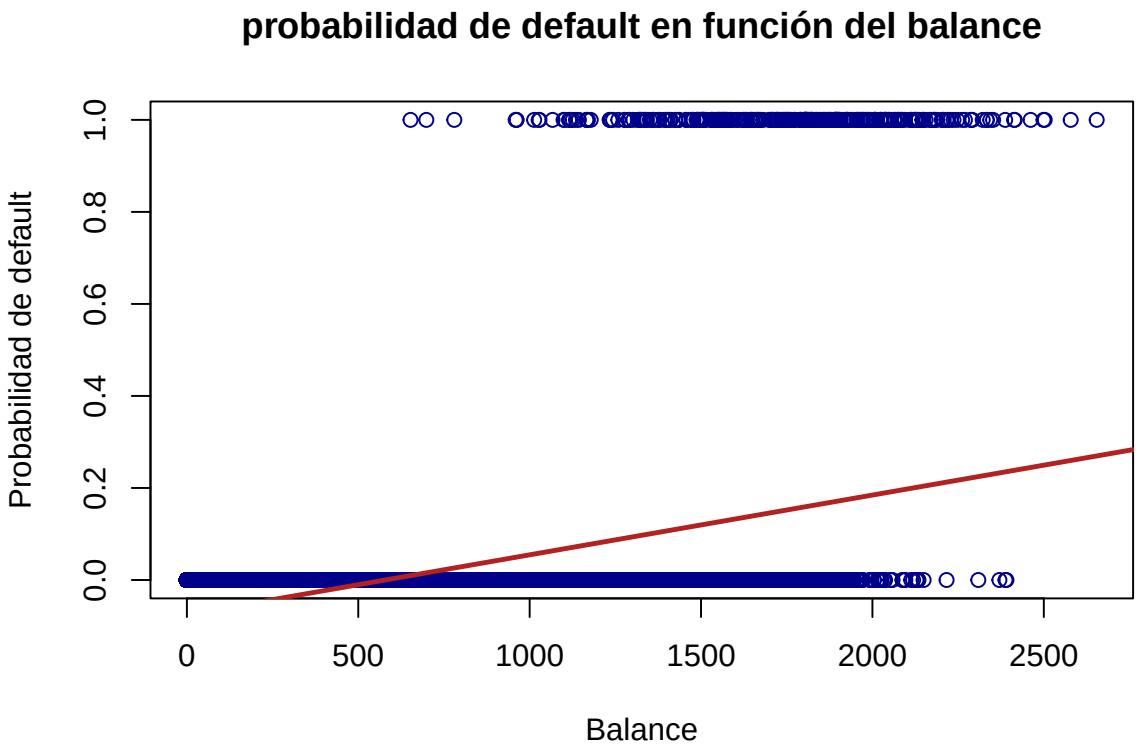
Cuando una variable categórica con dos niveles se codifica como 1 y 0, es técnicamente posible ajustar un modelo de regresión lineal. En este caso, el modelo intentaría predecir la probabilidad de que la variable dependiente Y pertenezca al nivel de referencia asignado, en función de los valores de la variable predictora X.

Sin embargo, la regresión lineal tiene una limitación importante en este tipo de situaciones: como se ajusta una línea recta, para valores extremos del predictor X, el modelo puede generar predicciones de Y menores que 0 o mayores que 1. Esto es problemático, ya que las probabilidades, por definición, siempre deben estar dentro del rango [0,1]. Este comportamiento no es adecuado para modelar probabilidades.

Por esta razón, se prefiere la **regresión logística**, que emplea la función logística para ajustar la relación entre X y Y. La regresión logística garantiza que las predicciones de probabilidades estén siempre dentro del rango [0,1], lo que la convierte en una mejor opción para problemas donde la variable dependiente es binaria o categórica.

A continuación, se presenta un ejemplo en el que se modela la probabilidad de fraude por impago (*default*) en función del balance de la cuenta bancaria (*balance*).

```
library(ISLR)
levels(Default$default) <- c("0", "1")
Default$default <- as.character(Default$default)
Default$default <- as.numeric(Default$default)
modelo_lineal <- lm(default ~ balance, data = Default)
plot(x = Default$balance, y = Default$default, col = "darkblue",
     main = "probabilidad de default en función del balance",
     xlab = "Balance", ylab = "Probabilidad de default")
abline(modelo_lineal, lwd = 2.5, col = "firebrick")
```



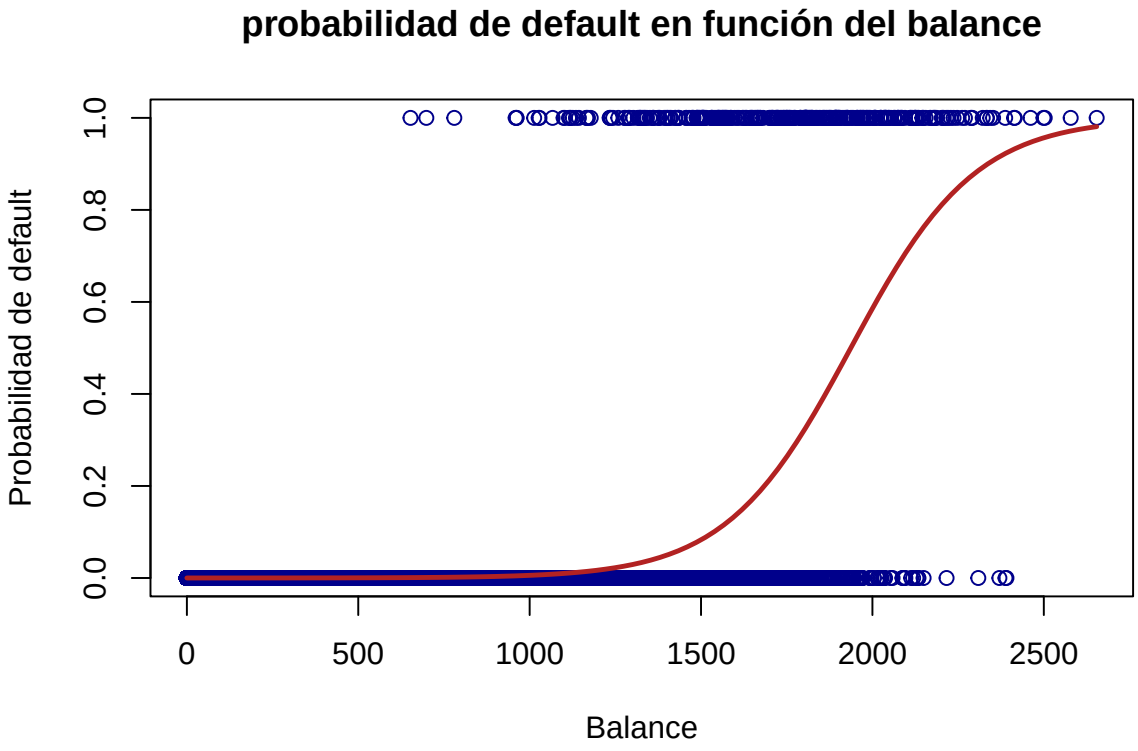
Cuando se modela la probabilidad de una variable dependiente binaria Y, la regresión lineal puede generar problemas al producir valores fuera del rango de las probabilidades (0 y 1). Para evitar estos problemas, la **regresión logística** modela la probabilidad de Y usando una función que garantiza que el resultado siempre esté comprendido entre 0 y 1, sin importar los valores del predictor X.

Existen varias funciones que cumplen esta propiedad, pero la más comúnmente utilizada es la **función logística**, que se define de la siguiente manera:

$$\pi = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

La función exponencial, representada por $e^{\beta_0+\beta_1 X}$, asegura que el valor de la probabilidad sea siempre positivo. Al dividir el numerador por el denominador, se garantiza que la probabilidad nunca exceda el valor de 1. De esta manera, la función logística, combinando estos dos aspectos, asegura que la probabilidad esté siempre entre 0 y 1, resolviendo los problemas que surgen con la regresión lineal.

```
modelo_logistico <- glm(default ~ balance, data = Default, family = "binomial")
plot(x = Default$balance, y = Default$default, col = "darkblue",
main = "probabilidad de default en función del balance",
xlab = "Balance", ylab = "Probabilidad de default")
curve(predict(modelo_logistico, data.frame(balance = x), type = "response"),
      add = TRUE, col = "firebrick", lwd = 2.5)
```



Para transformar la ecuación logística en una forma lineal, se aplica el logaritmo natural a los **odds** (razón de probabilidades), lo que da lugar a lo que se conoce como el **logaritmo de los odds** o **log-odds**. Esta transformación linealiza la relación, permitiendo que las técnicas de regresión lineal se apliquen de manera efectiva.

$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X$

Donde: - π es la probabilidad de que un evento ocurra. - $1 - \pi$ es la probabilidad de que el evento no ocurra. - $\frac{\pi}{1-\pi}$ es llamado **odds ratio** (razón de probabilidades). - $\log\left(\frac{\pi}{1-\pi}\right)$ es conocido como el **logit**.

2.1. Demostración

$\pi = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$, $\frac{\pi}{1-\pi}$ implica que $\frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}$

haciendo algunos cálculos se obtiene:

$$\frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{\frac{1+e^{\beta_0+\beta_1 X}-e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}$$

que se puede simplificar:

$$\frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{\frac{\cancel{1+e^{\beta_0+\beta_1 X}}-\cancel{e^{\beta_0+\beta_1 X}}}{1+e^{\beta_0+\beta_1 X}}} = \frac{\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}}{\frac{1}{1+e^{\beta_0+\beta_1 X}}}$$

realizando algunos cálculos mas obtenemos:

$$\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} * 1 + e^{\beta_0+\beta_1 X} = \frac{e^{\beta_0+\beta_1 X}}{\cancel{1+e^{\beta_0+\beta_1 X}}} * \cancel{1+e^{\beta_0+\beta_1 X}}$$

al considerar el logaritmo el exponencial desaparece devolviendo una función lineal:

$log(\frac{\pi}{1-\pi}) = log(e^{\beta_0+\beta_1 X} = \cancel{log}(\cancel{e}^{\beta_0+\beta_1 X} = \beta_0 + \beta_1 X$

3. Concepto de ODDS, razón de probabilidad (ODDS ratio) y logaritmo de ODDS

En la regresión lineal, se modela el valor de la variable dependiente Y en función del valor de la variable independiente X. Sin embargo, en la **regresión logística**, se modela la probabilidad de que la variable respuesta Y pertenezca al nivel de referencia seleccionado, utilizando el **logaritmo de los odds (log-odds)**.

Por ejemplo, supongamos que la probabilidad de que un evento ocurra es 0.8, por lo tanto, la probabilidad de que el evento no ocurra es $1 - 0.8 = 0.2$. Los **odds** o razón de probabilidad de que el evento sea verdadero se definen como el cociente entre la probabilidad de que ocurra el evento y la probabilidad de que no ocurra. En este caso, los odds de que el evento sea verdadero son:

$$\text{odds} = \frac{0.8}{0.2} = 4$$

Esto significa que se esperan 4 eventos verdaderos por cada evento falso. La transformación de probabilidades a odds es **monotónica**: si la probabilidad aumenta, los odds también lo hacen, y si la probabilidad disminuye, los odds disminuyen.

Propiedades de los ODDS y el Logaritmo de los ODDS:

- Si $p(\text{verdadero}) = p(\text{falso})$, entonces $\text{odds}(\text{verdadero}) = 1$.
- Si $p(\text{verdadero}) < p(\text{falso})$, entonces $\text{odds}(\text{verdadero}) < 1$.
- Si $p(\text{verdadero}) > p(\text{falso})$, entonces $\text{odds}(\text{verdadero}) > 1$.
- A diferencia de la probabilidad, que está acotada entre 0 y 1, los odds no tienen límite superior.
- Si $\text{odds}(\text{verdadero}) = 1$, entonces $logit(p) = 0$.
- Si $\text{odds}(\text{verdadero}) < 1$, entonces $logit(p) < 0$.
- Si $\text{odds}(\text{verdadero}) > 1$, entonces $logit(p) > 0$.
- La transformación logit no existe para $p = 0$.

En resumen, el logaritmo de los odds (logit) es una transformación que convierte las probabilidades, limitadas entre 0 y 1, en un valor que puede tomar cualquier valor real, lo que facilita el modelado de la relación entre los predictores y la variable dependiente en regresión logística.

4. Estimación de los parámetros: el método de máxima verosimilitud

Una vez establecida la relación lineal entre el logaritmo de los odds y la variable predictora X, es necesario estimar los parámetros β_0 y β_1. Para esto, se utiliza el método de **máxima verosimilitud** (*maximum likelihood*), mientras que en la regresión lineal se emplea el método de **mínimos cuadrados**.

El método de **máxima verosimilitud** (ML) es un proceso computacional y matemático que encuentra los valores de los parámetros β_0 y β_1 que maximizan la probabilidad de observar los datos tal como han sido registrados. Es decir, el ML busca los parámetros que hacen más probable que las observaciones se ajusten al modelo.

A diferencia de los mínimos cuadrados en regresión lineal, que minimizan la suma de los errores cuadrados, el ML se enfoca en maximizar la función de verosimilitud, que mide la probabilidad de obtener los datos observados dados ciertos valores de los parámetros.

5. Evaluación del modelo

Existen diferentes técnicas estadísticas para evaluar la significancia de un modelo logístico en su conjunto (p-valor del modelo). Estas técnicas consideran que el modelo es útil si muestra una mejora respecto al **modelo nulo**, que es un modelo sin predictores (solo contiene el parámetro β_0). Dos de las pruebas más utilizadas para evaluar la significancia del modelo son:

1. **Wald test**: Similar al t-test en regresión lineal, se utiliza para evaluar la significancia de los coeficientes del modelo. Este test analiza si el coeficiente de cada predictor es significativamente diferente de cero.
2. **Likelihood ratio test**: Esta prueba compara la probabilidad de obtener los valores observados bajo el modelo logístico con predictores, frente a un modelo sin relación entre las variables (modelo nulo). Calcula la significancia de la diferencia de residuos entre el modelo con predictores y el modelo nulo. El estadístico resultante sigue una distribución chi-cuadrado, con grados de libertad que equivalen a la diferencia en los grados de libertad entre los dos modelos comparados. Al comparar con el modelo nulo, los grados de libertad son iguales al número de predictores en el modelo.

Para determinar la significancia individual de cada predictor en un modelo de regresión logística, se utiliza el **estadístico Z** y el **test de Wald**. En R, este método es el que se emplea para calcular los p-valores que aparecen al realizar el *summary* del modelo.

6. Interpretación del modelo

En la regresión logística, la interpretación de los coeficientes β_1 es diferente a la de la regresión lineal. Mientras que en la regresión lineal β_1 representa el cambio promedio en la variable dependiente Y debido a un incremento de una unidad en el predictor X, en la **regresión logística** β_1 representa el cambio en el **logaritmo de los odds** por cada incremento de una unidad en X.

Dado que la relación entre $p(Y)$ (la probabilidad de que ocurra el evento) y X no es lineal, β_1 no indica directamente el cambio en la probabilidad de Y por unidad de cambio en X. En cambio, β_1 refleja cómo cambia el **log-odds**.

La cantidad en que se incrementa la probabilidad de Y por cada unidad de cambio en X depende del valor de X, es decir, de la posición en la **curva logística** en la que se encuentre. En otras palabras, el impacto de un cambio en X sobre la probabilidad de Y varía según el punto de la curva logística en el que nos encontremos, siendo mayor en los tramos medios de la curva y menor en los extremos.

6.1. Algunas reglas básicas

Al igual que en la regresión lineal, en la regresión logística podemos interpretar los coeficientes basándonos en las siguientes reglas básicas:

1. **Signo**: Si el coeficiente es positivo, significa que el efecto de la variable predictora sobre la variable dependiente es positivo, es decir, un aumento en la variable predictora X incrementará los odds de que ocurra el evento (y viceversa si el signo es negativo).
2. **Significatividad**: La significancia de los coeficientes se determina utilizando el **test de Wald**. Dependiendo del p-valor de este test, se puede concluir si los parámetros de la regresión son significativos o no. Un coeficiente significativo implica que la variable predictora tiene un efecto relevante en el modelo.

Interpretación de los odds ratio

El **odds ratio (OR)** es una medida que nos ayuda a entender el impacto de un predictor sobre las probabilidades de que ocurra un evento en un modelo de regresión logística.

En R para obtener **odds ratio (OR)** hay que calcular los exponenciales de los coeficientes $\beta \Rightarrow \exp(\beta)$.

Como dicho anteriormente: - Un **odds ratio** de 1 significa que no hay ningún efecto. El evento tiene la misma probabilidad de ocurrir con o sin el predictor. - Un **odds ratio** mayor que 1 indica que el predictor aumenta las probabilidades del evento. - Un **odds ratio** menor que 1 indica que el predictor disminuye las probabilidades del evento.

Para facilitar la interpretación de los **odds ratios**, podemos utilizar dos fórmulas comunes:

Fórmula 1: $(\textbf{odds} - 1) \times 100$

Esta fórmula se utiliza cuando el **odds ratio** es mayor que 1, para calcular el **porcentaje de incremento** en las probabilidades del evento debido al predictor.

Ejemplo: - Si el **odds ratio** es 1.5, podemos calcular:

$$(1.5 - 1) \times 100 = 0.5 \times 100 = 50\%$$

Esto significa que un aumento de una unidad en la variable predictora incrementa las probabilidades del evento en un **50%**.

Fórmula 2: $(1 - \textbf{odds}) \times 100$

Esta fórmula se utiliza cuando el **odds ratio** es menor que 1, para calcular el **porcentaje de disminución** en las probabilidades del evento.

Ejemplo: - Si el **odds ratio** es 0.7, podemos calcular:

$$(1 - 0.7) \times 100 = 0.3 \times 100 = 30\%$$

Esto significa que un aumento de una unidad en la variable predictora disminuye las probabilidades del evento en un **30%**.

En resumen:

- Si el **odds ratio** es mayor que 1, usamos la fórmula $(\text{odds} - 1) \times 100$ para interpretar el incremento en las probabilidades.
- Si el **odds ratio** es menor que 1, usamos la fórmula $(1 - \text{odds}) \times 100$ para interpretar la disminución en las probabilidades.

Estas fórmulas nos permiten convertir los **odds ratios** en porcentajes de aumento o disminución, lo que facilita su interpretación en contextos prácticos.

A parte la interpretación de los **odds ratios**. Si queremos identificar el efecto mas importante entre todos los predictores lo que tenemos que verificar es si el **OR** es menor de 1. En este caso tendremos que calcular el valor inverso: $\frac{1}{OD}$. Para que el effecto sea siempre positivo en todos los coeficientes.

Hecha esta trasformación el efecto mas importante será aquel asociado al **OR** mas elevado.

6.2. Predicciones

Una vez estimados los coeficientes del modelo logístico, es posible calcular la probabilidad de que la variable dependiente pertenezca al nivel de referencia, dado un determinado valor del predictor. Para ello, se utiliza la siguiente ecuación del modelo:

$$\pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Donde π representa la probabilidad de que el evento ocurra, β_0 es el intercepto, y β_1 es el coeficiente asociado al predictor X.

6.3. Comparación de clasificación predicha y observaciones

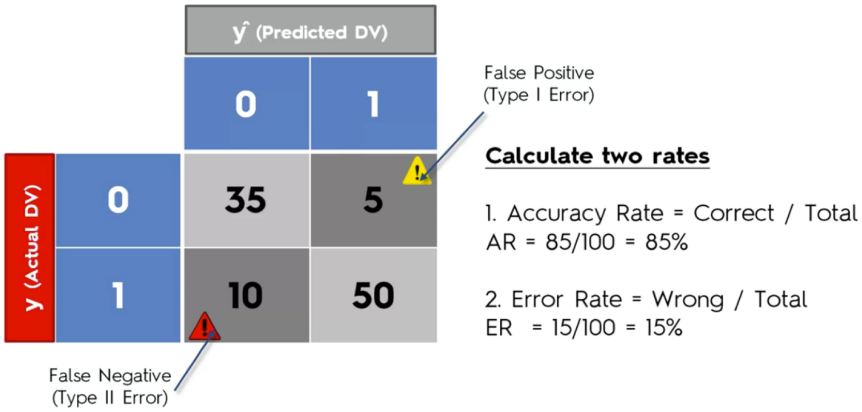
Una forma común de evaluar la capacidad de un modelo logístico es utilizando una **matriz de confusión**, la cual muestra el número de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Estos valores indican el rendimiento de las predicciones del modelo:

1. **Verdaderos positivos (cuadrante inferior derecho)**: Predijimos que el cliente incumpliría y efectivamente lo hizo.

2. **Verdaderos negativos (cuadrante superior izquierdo)**: Predijimos que no habría incumplimiento y el cliente efectivamente no incumplió.

3. **Falsos positivos (cuadrante superior derecho)**: Predijimos que el cliente incumpliría, pero en realidad no lo hizo (también conocido como "error tipo I").

4. **Falsos negativos (cuadrante inferior izquierdo)**: Predijimos que no habría incumplimiento, pero el cliente sí incumplió (también conocido como "error tipo II").



La matriz de confusión es una tabla cruzada que describe el rendimiento de clasificación del modelo. En la tabla, las filas representan el valor observado (si el cliente incumplió o no), y las columnas representan la predicción del modelo.

Cálculo de métricas de evaluación

A partir de la matriz de confusión, se pueden calcular varios índices útiles para evaluar el rendimiento del modelo, como:

$$\text{Accuracy} = \frac{A + D}{A + B + C + D}$$

$$\text{Error} = \frac{B + C}{A + B + C + D}$$

Donde: - A y D representan los valores correctos (verdaderos positivos y verdaderos negativos). - B y C representan los errores (falsos positivos y falsos negativos).

7. La regresión logística múltiple

La **regresión logística múltiple** es una extensión de la regresión logística simple, que permite incluir múltiples predictores (continuos o categóricos) en el modelo. La ecuación del modelo logístico múltiple es:

$$\pi = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}}$$

Donde cada β_i representa el coeficiente asociado al predictor X_i.

Para evaluar la validez y la calidad del modelo de regresión logística múltiple, se analiza tanto el modelo completo como cada uno de los predictores. Un modelo es útil si muestra una mejora respecto al modelo nulo (sin predictores). Existen tres pruebas estadísticas que cuantifican esta mejora:

1. **Likelihood ratio**: Compara la probabilidad de los datos observados con el modelo ajustado frente al modelo nulo.
2. **Score test**: Evalúa la capacidad predictiva del modelo en función de los residuos.
3. **Wald test**: Evalúa la significancia individual de los coeficientes.

Aunque las tres pruebas no siempre coinciden en sus conclusiones, en muchos casos se recomienda confiar en el **likelihood ratio test** como la medida más robusta para evaluar el modelo.

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/nested_tests.html

Ejemplos

8.1 Vinoteca

Una bodega de vino desea lanzar una nueva marca al mercado y está interesada en entender si las ventas tendrían éxito entre sus clientes. Para determinar la probabilidad de compra, la bodega cuenta con tres variables clave. En particular, le gustaría analizar qué sucedería si el precio fuera 80 euros y la calidad del vino fuera 30.

Las variables disponibles son:

- **Quality**: Calidad del vino (escala de 1 a 100).
- **Price**: Precio del vino (en euros).
- **Purchased**: Variable binaria que indica la probabilidad de compra (1 = Sí, 0 = No).

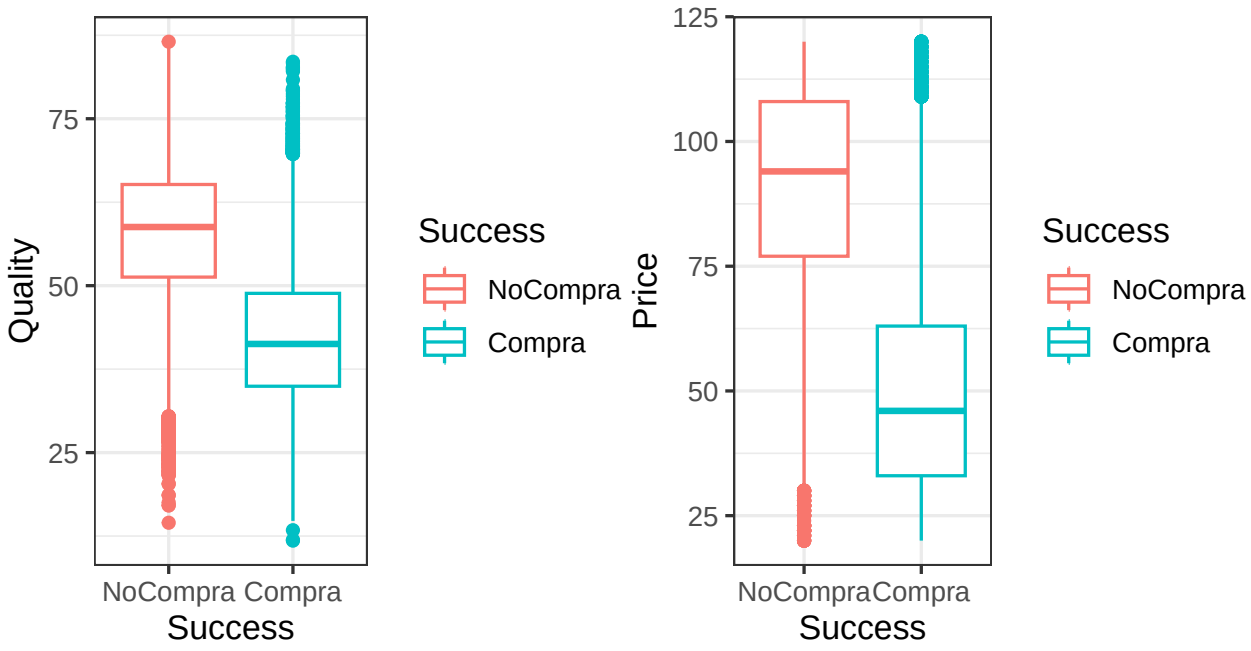
El objetivo es predecir la probabilidad de que un cliente compre la nueva marca de vino, dado un precio de 80 euros y una calidad de 30, utilizando estas variables en un modelo de regresión logística.

```
# cargo las liberáis necesarias
library(ggplot2)
library(gridExtra)

wine <- read.csv("wine.csv", header = TRUE, sep = ",")

wine$Success = factor(wine$Purchased, labels=c("NoCompra","Compra"))
```

8.1.1. Análisis descriptiva

El primer paso del análisis consiste en examinar cómo se relacionan las variables predictoras con la variable de respuesta (*Purchased*). Esto se puede realizar mediante gráficos **box-plot**, los cuales permiten visualizar la distribución de las variables **Quality** y **Price** en función de la probabilidad de compra.

```
ggplot(data = wine, mapping = aes(x = Success , y = Quality, colour = Success)) +
geom_boxplot() + theme_bw()
ggplot(data = wine, mapping = aes(x = Success , y = Price, colour = Success )) +
geom_boxplot() + theme_bw()
```



Al observar los gráficos **box-plot**, podemos identificar algunos patrones interesantes en la relación entre las variables predictoras (*Price* y *Quality*) y la variable de respuesta (*Purchased*):

- A medida que el **precio** aumenta, el número de personas que deciden comprar el nuevo vino disminuye. Esto sugiere que un precio elevado podría estar afectando negativamente las decisiones de compra.
- Un patrón similar se observa con la variable **calidad**. Aunque la **calidad** del vino es, en teoría, un aspecto positivo, en este caso parece tener un efecto negativo sobre la decisión de compra. Esto puede deberse a que la calidad está fuertemente correlacionada con el precio del vino. Es decir, a mayor calidad, también mayor precio, lo que podría disuadir a algunos clientes de comprar.

Además, en ambos gráficos, es posible observar un número considerable de **outliers**. Estos puntos representan un pequeño grupo de consumidores que, a pesar de que el precio o la calidad sean altos, siguen interesados en comprar el vino. Aunque no disponemos de otras variables que expliquen

este comportamiento, podemos intuir que estos consumidores podrían tener un poder adquisitivo elevado, lo que los hace menos sensibles al precio.

Por otro lado, también se detecta un grupo de clientes que, aunque el precio y la calidad del vino sean bajos, no comprarían el vino. Esto sugiere que existen otros factores no incluidos en este análisis que podrían influir en la decisión de compra, como las preferencias personales o la lealtad a otras marcas.

8.1.2. Estimación modelo Purchased ~ Price

En R, podemos estimar un modelo de **regresión logística** utilizando la función `glm()` (Generalized Linear Models). Para especificar que estamos ajustando un modelo logístico, usamos el parámetro `family = "binomial"`. A continuación, se muestra cómo hacerlo:

```
modelo1 <- glm(Purchased ~ Price, data = wine, family = "binomial")
summary(modelo1)
```

```
##
## Call:
## glm(formula = Purchased ~ Price, family = "binomial", data = wine)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.4676556  0.0468286   116.8   <2e-16 ***
## Price       -0.0779729  0.0006425  -121.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69315  on 49999  degrees of freedom
## Residual deviance: 38757  on 49998  degrees of freedom
## AIC: 38761
##
## Number of Fisher Scoring iterations: 5
```

Al observar los resultados del modelo de regresión logística, podemos interpretar los coeficientes estimados de la siguiente manera:

- El coeficiente estimado para la **intercepta** es **5.46**. Este valor corresponde al logaritmo de los **odds** de que un consumidor compre el vino cuando el precio es 0 euros. Como podemos esperar, los **odds** son muy altos en este caso:

$$e^{5.46} = 231.09$$

esto significa que las probabilidades de compra del vino cuando el precio es 0 euros son extremadamente altas. Aplicando la fórmula $(\text{odds} - 1) \times 100$ se obtiene:

$$(231.09 - 1) \times 100 = 230.09 \times 100 = 23.009\%$$

Esto indica que, cuando el precio del vino es 0 euros, las probabilidades de que un consumidor compre el vino son **23.009%** más altas en comparación con una situación en la que los **odds** son 1 (lo que corresponde a una probabilidad del 50%). En otras palabras, los **odds** de compra son extremadamente favorables en esta situación.

- El coeficiente asociado al **precio** es **-0.07**, lo que indica que a medida que aumenta el precio, disminuye la probabilidad de compra. Si calculamos el exponencial de este coeficiente:

$$e^{-0.07} = 0.93$$

El **odds ratio** asociado al precio del vino es **0.93**, lo que indica que a medida que el precio aumenta, la probabilidad de que un consumidor compre el vino disminuye.

Para interpretar este odds ratio utilizando la fórmula $(1 - \text{odds}) \times 100$, calculamos el porcentaje de disminución en las probabilidades de compra por cada incremento unitario en el precio del vino:

$$(1 - 0.93) \times 100 = 0.07 \times 100 = 7\%$$

Este resultado significa que por cada incremento de 1 euro en el precio del vino, las probabilidades de que un consumidor compre el vino disminuyen en un **7%**.

- Todos los coeficientes son significativos según el **Wald test**, ya que ambos tienen un p-valor inferior a 0.001, lo que indica que estos coeficientes tienen un impacto estadísticamente significativo en el modelo.

Finalmente, podríamos calcular la probabilidad de que un consumidor compre el vino cuando el precio es 80 euros, uno de los valores sugeridos por el vendedor. Para hacer esto, simplemente sustituimos el valor del precio en la ecuación del modelo:

$$\pi = \frac{e^{5.46 - 0.07 \times 80}}{1 + e^{5.46 - 0.07 \times 80}}$$

Esto nos dará la probabilidad estimada de compra para un precio de 80 euros.

```
y_HAT = 5.467-(0.077*80)
p=exp(y_HAT)/(1+exp(y_HAT))
p
```

```
## [1] 0.333366
```

8.1.3. Estimación modelo Purchased ~ Price+Quality

```r
modelo2 <- glm(Purchased ~ Price+Quality, data = wine, family = "binomial")
summary(modelo2)
```

```
##
## Call:
## glm(formula = Purchased ~ Price + Quality, family = "binomial",
##     data = wine)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.494772   0.067066   52.11   <2e-16 ***
## Price       -0.121824   0.001432  -85.06   <2e-16 ***
## Quality      0.100903   0.002732   36.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69315  on 49999  degrees of freedom
## Residual deviance: 37307  on 49997  degrees of freedom
## AIC: 37313
##
## Number of Fisher Scoring iterations: 5
```

Podemos observar los siguientes coeficientes y sus interpretaciones en el modelo de regresión logística:

El coeficiente estimado para la **intercepta** es **3.49**, lo que corresponde al valor esperado del logaritmo de los **odds** de que un consumidor compre el vino cuando tanto el precio como la calidad son 0. El exponencial de este valor nos da los **odds**:

$$e^{3.49} = 32.78$$

Esto indica que, en estas condiciones, los odds de que un consumidor compre el vino son muy altos. Para interpretarlo con la fórmula $(\text{odds} - 1) \times 100$, calculamos:

$$(32.78 - 1) \times 100 = 31.78 \times 100 = 3178\%$$

Cuando el precio y la calidad son 0, la probabilidad de que un consumidor compre el vino es **3178%** más alta. Esto refleja una probabilidad extremadamente alta de compra en esta situación.

El coeficiente asociado al **precio** es **-0.12**, lo que sugiere que, a medida que el precio aumenta, la probabilidad de compra disminuye. El exponencial de este coeficiente nos da el **odds ratio**:

$$e^{-0.12} = 0.89$$

Usando la fórmula $(1 - \text{odds}) \times 100$, calculamos:

$$(1 - 0.89) \times 100 = 0.11 \times 100 = 11\%$$

Esto significa que por cada aumento de 1 euro en el precio del vino, la probabilidad de compra disminuye en un **11%**.

El coeficiente asociado a la **calidad** es **+0.10**, lo que sugiere que a medida que aumenta la calidad, la probabilidad de compra también aumenta. El exponencial de este coeficiente nos da el **odds ratio**:

$$e^{0.10} = 1.10$$

Para interpretarlo con la fórmula $(\text{odds} - 1) \times 100$, calculamos:

$$(1.10 - 1) \times 100 = 0.10 \times 100 = 10\%$$

Esto indica que por cada incremento de una unidad en la calidad del vino, la probabilidad de compra aumenta en un **10%**.

En resumen:

- Un **odds** de 32.78 para la intercepta indica que, cuando el precio y la calidad son 0, la probabilidad de compra es extremadamente alta (**3178%** más alta que un odds de 1).
- El **odds ratio** de 0.89 asociado al precio indica que por cada euro adicional en el precio, la probabilidad de compra disminuye en un **11%**.
- El **odds ratio** de 1.10 asociado a la calidad indica que por cada incremento de una unidad en la calidad, la probabilidad de compra aumenta en un **10%**.

Por ultimo podríamos decidir calcular la probabilidad de que el vino sea comprado siendo el precio 80 Euro y la calidad fuera 30: .

```
y_HAT = 3.49+(+0.10*30)+(-0.12*80)
p=exp(y_HAT)/(1+exp(y_HAT))
p
```

```
## [1] 0.04269664
```

8.1.4. Evaluación del modelo

El **Likelihood ratio test** evalúa la significancia de la diferencia de los residuos entre el modelo de interés y el modelo nulo (modelo sin predictores). El estadístico del test sigue una distribución **chi-cuadrado** con grados de libertad equivalentes a la diferencia de grados de libertad entre los dos modelos comparados.

Podemos calcular este test en R mediante el siguiente comando:

```r
# Comparación de dos modelos con el test de Chi-cuadrado
anova(modelo1, modelo2, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: Purchased ~ Price
## Model 2: Purchased ~ Price + Quality
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1     49998      38757
## 2     49997      37307  1   1449.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al observar los resultados del test, podemos ver que el segundo modelo es significativamente mejor que el primero (p-value < 0.05). Esto indica que la inclusión de más variables predictoras en el segundo modelo aporta información significativa para mejorar las predicciones.

8.1.5. Matriz de confusión

Una **matriz de confusión** nos permite evaluar el rendimiento del modelo al comparar las predicciones con las observaciones reales. A continuación, calculamos las matrices de confusión para el **Modelo 1** y el **Modelo 2**:

```r
library(vcd)

# Matriz de confusión para el Modelo 1
predicciones1 <- ifelse(test = modelo1$fitted.values > 0.5, yes = 1, no = 0)
matriz_confusion1 <- table(wine$Success, predicciones1, dnn = c("observaciones", "predicciones"

# Matriz de confusión para el Modelo 2
predicciones2 <- ifelse(test = modelo2$fitted.values > 0.5, yes = 1, no = 0)
matriz_confusion2 <- table(wine$Success, predicciones2, dnn = c("observaciones", "predicciones"

# Visualizar las matrices de confusión
matriz_confusion1
```

```
##              predicciones
## observaciones     0     1
##      NoCompra 20646  4318
##      Compra    4179 20857
```

```r
matriz_confusion2
```

```
##              predicciones
## observaciones     0     1
##      NoCompra 20808  4156
##      Compra    4075 20961
```

Cálculo de métricas: Accuracy y Error

A partir de las matrices de confusión, podemos calcular las métricas de evaluación **accuracy** y **error** para ambos modelos:

1. **Accuracy**: mide el porcentaje de predicciones correctas.

```
# Cálculo de accuracy para Modelo 1 y Modelo 2
accuracy_modelo1 <- (20646 + 20857) / (20646 + 20857 + 4318 + 4179)
accuracy_modelo2 <- (20808 + 20961) / (20808 + 20961 + 4156 + 4075)

accuracy_modelo1
```

```
## [1] 0.83006
```

```
accuracy_modelo2
```

```
## [1] 0.83538
```

- **Modelo 1**:
$$\frac{20646 + 20857}{20646 + 20857 + 4318 + 4179} = 0.83$$
- **Modelo 2**:
$$\frac{20808 + 20961}{20808 + 20961 + 4156 + 4075} = 0.84$$

2. **Error**: mide el porcentaje de predicciones incorrectas.

```
# Cálculo de error para Modelo 1 y Modelo 2
error_modelo1 <- (4318 + 4179) / (20646 + 20857 + 4318 + 4179)
error_modelo2 <- (4156 + 4075) / (20808 + 20961 + 4156 + 4075)

error_modelo1
```

```
## [1] 0.16994
```

```
error_modelo2
```

```
## [1] 0.16462
```

- **Modelo 1**:
$$\frac{4318 + 4179}{20646 + 20857 + 4318 + 4179} = 0.17$$
- **Modelo 2**:
$$\frac{4156 + 4075}{20808 + 20961 + 4156 + 4075} = 0.16$$

Aunque ambos modelos son similares, podemos observar que el **Modelo 2** tiene un rendimiento ligeramente mejor, con una mayor **accuracy** y un menor **error**.

Cálculo de pérdidas en la campaña

Si imputamos un coste de 5 euros por cliente para la campaña, podemos determinar las pérdidas asociadas tanto a incluir clientes que no realizarán compras como a excluir clientes que sí lo harían.

Las pérdidas totales se pueden calcular como:

$$5 \times (4156 + 4075) = 41,155 \text{ euros}$$

Este valor representa las pérdidas estimadas por errores en las predicciones de compra.

8.2 Caso de estudio: Campaña de préstamos juveniles

Supongamos que somos el director de marketing del banco **CyndiCat**, encargado de evaluar la viabilidad de realizar una nueva campaña en 2019 para conceder **préstamos juveniles** destinados a la compra de un auto nuevo.

El año anterior se llevó a cabo una campaña similar, en la que, con la ayuda del departamento de riesgos, se identificaron **10.000 clientes potenciales** a quienes se les concedió el préstamo de forma automática, teniendo en cuenta algunas características clave, como:

- Si el cliente era estudiante.
- El saldo promedio en su tarjeta de crédito.
- Sus ingresos anuales.

Después de un año, el director dispone de datos sobre los clientes que, tras recibir el préstamo y gastar el dinero en la compra de un auto, no devolvieron la deuda. De los 10.000 clientes, un **3.33%** no devolvieron el préstamo. Aunque este porcentaje es pequeño, representa una pérdida significativa para el banco, ya que cada préstamo no devuelto supone una pérdida estimada de **6.000 euros**.

Evaluación previa a la nueva campaña

Antes de lanzar la nueva campaña, el director decide aprovechar la información recopilada durante la campaña anterior para desarrollar un modelo que permita:

1. **Identificar con antelación** a los clientes que podrían no devolver el préstamo.
2. Determinar **cuáles son las características más relevantes** para identificar a los buenos clientes.
3. Probar los filtros utilizados en la campaña anterior calculando la probabilidad de **default** en un cliente típico (por ejemplo, un estudiante con un saldo de 2.000 euros y hasta 40.000 euros de ingresos).

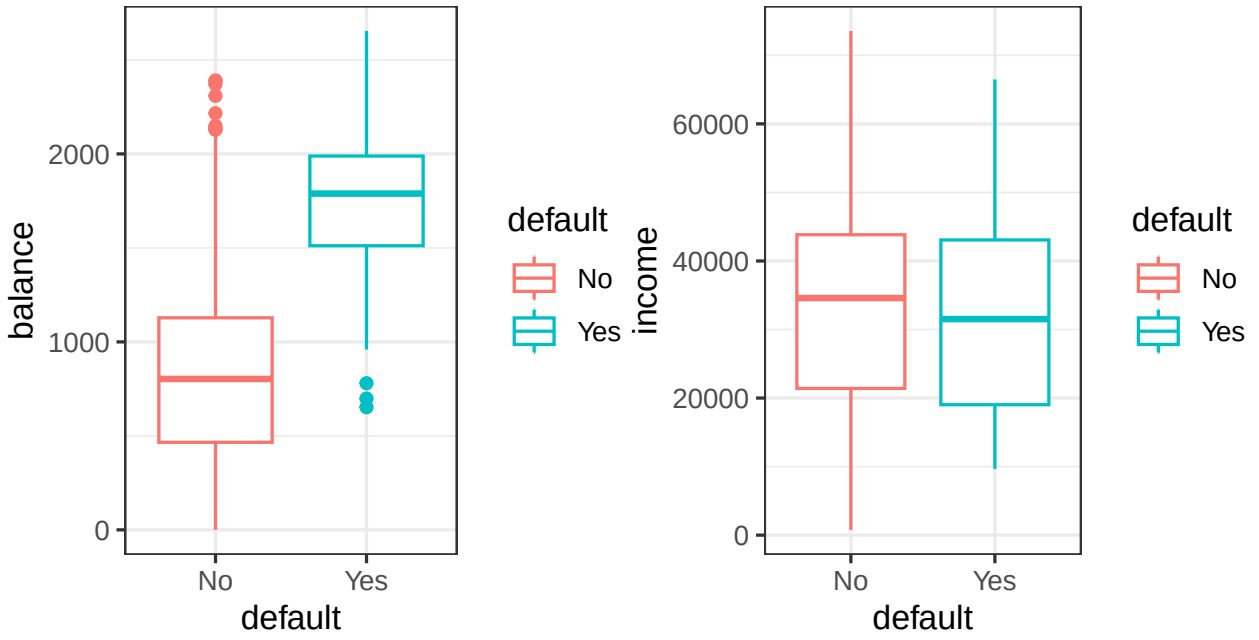
El objetivo es construir un modelo predictivo que ayude a minimizar el riesgo y a identificar los factores clave que influencian la probabilidad de que un cliente no devuelva el préstamo.

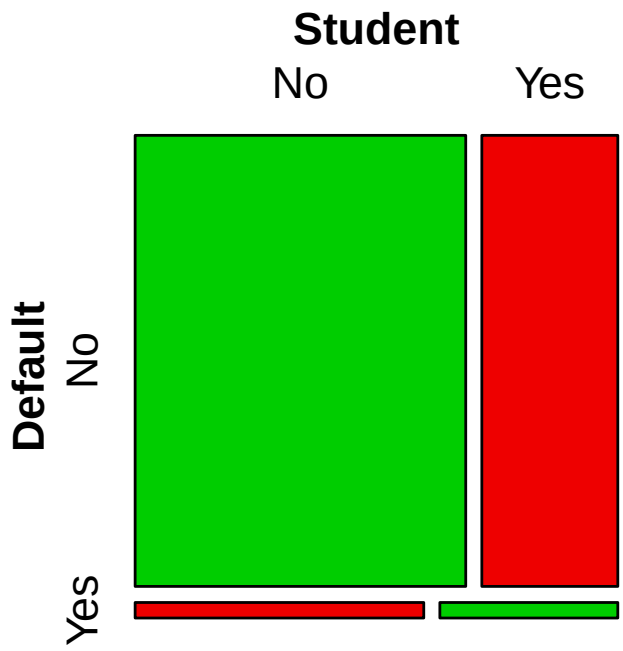
```r
# cargo las liberias necesarias
library(ggplot2)
library(gridExtra)

# leo mis datos
credit = read.csv(file="credit.csv", header = TRUE,sep=",")

credit$default = as.factor(credit$default)
credit$student = as.factor(credit$student)
```

8.2.1. Analisis descriptiva

El primer paso del análisis consiste en examinar cómo se relacionan las variables predictoras con la variable de respuesta. Esto se puede hacer mediante gráficos **box-plot**, los cuales permiten visualizar la distribución de las variables predictoras en función de la variable de respuesta.

```r
ggplot(data = credit, mapping = aes(x = default , y = balance, colour = default)) +
geom_boxplot() + theme_bw()
ggplot(data = credit, mapping = aes(x = default , y = income, colour = default )) +
geom_boxplot() + theme_bw()
mosaic(table(credit$default, credit$student,
             dnn = c("Default", "Student")), shade = T, colorize = T,
gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))
```





A partir de los gráficos **box-plot**, podemos hacer las siguientes observaciones:

1. **Saldo de la cuenta**:
    - El saldo de la cuenta parece ser una variable que permite diferenciar claramente a los clientes. Los clientes con **saldos más bajos** presentan un menor número de impagos, mientras que la mayoría de los impagos se concentran entre aquellos con **saldos más elevados**. Esto sugiere que el saldo promedio de la tarjeta de crédito es un factor clave para predecir el riesgo de impago.
2. **Ingresos**:
    - La variable ingresos, por otro lado, no parece ser un buen diferenciador entre los clientes que incumplen y los que no. Los **grupos de clientes que devuelven el préstamo y los que no lo hacen son bastante similares en términos de ingresos**, lo que sugiere que esta variable no tiene un impacto significativo en la probabilidad de impago.
3. **Estudiante**:
    - En la mayoría de los casos de impago, los clientes son **estudiantes**. Esto indica que ser estudiante podría estar relacionado con un mayor riesgo de incumplimiento, posiblemente debido a factores como la inestabilidad financiera o ingresos futuros inciertos.

8.2.2. Estimación modelo impago ~ saldo

En R, podemos estimar un modelo de **regresión logística** utilizando la función `glm()` (Generalized Linear Models). Para especificar que estamos ajustando un modelo logístico, usamos el argumento `family = "binomial"`.

```r
modelo1 <- glm(default ~ balance, data = credit, family = "binomial")
summary(modelo1)
```

```
##
## Call:
## glm(formula = default ~ balance, family = "binomial", data = credit)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

La siguiente tabla muestra las estimaciones de los coeficientes y la información obtenida ajustando un modelo de **regresión logística** para predecir la **probabilidad de incumplimiento = Sí** utilizando la variable **saldo** como predictor.

- **Intercepta**:
  - El coeficiente de la **intercepta** es negativo ($\beta_0 = -11$), lo que indica el logaritmo de los **odds** de que un cliente incumpla con el pago cuando su saldo es 0. Dado que el valor es muy bajo, el exponencial de -11 nos da los **odds**:

$$e^{-11} \approx 0$$

Esto implica que las probabilidades de incumplimiento cuando el saldo es 0 son prácticamente nulas.

- **Saldo**:
  - El coeficiente estimado para el **saldo** es positivo ($\beta_1 = 0.0057$), lo que indica que, a medida que aumenta el saldo de un cliente, también aumenta la probabilidad de que incumpla con el pago. El exponencial de este coeficiente es:

$$e^{0.0057} \approx 1.0057$$

Para interpretar el **odds ratio** utilizando la fórmula $(\text{odds} - 1) \times 100$:

$$(1.0057 - 1) \times 100 = 0.57\%$$

Esto significa que por cada euro adicional en el saldo del cliente, los **odds** de no pagar aumentan en un **0.57%**. Aunque el incremento es pequeño, muestra que el saldo tiene un impacto positivo en la probabilidad de no pagar.

Por ultimo podríamos decidir calcular la probabilidad de que el cliente impaga siendo el saldo 30.000 Euro:

```
y_HAT = -11+(0.0057*1500)
p=exp(y_HAT)/(1+exp(y_HAT))
p
```

```
## [1] 0.07943855
```

8.2.3. Estimación modelo impago ~ saldo+ingresos+estudiante

Ajustemos un modelo que prediga la **probabilidad de impago** en función de tres variables: **saldo**, **ingresos** (en miles de dólares) y la variable de estado de **estudiante** (si es estudiante o no). Hay un resultado interesante en los coeficientes:

1. Los **p-valores** asociados con las variables **saldo** y **estado de estudiante = Sí** son muy pequeños, lo que indica que estas variables están fuertemente asociadas con la probabilidad de incumplimiento.
2. El coeficiente para la variable **estudiante** es **negativo**, lo que sugiere que los estudiantes tienen **menos probabilidades** de incumplimiento en comparación con los no estudiantes, cuando controlamos por el saldo y los ingresos.

Podemos ajustar el modelo en R de la siguiente manera:

```
modelo2 <- glm(default ~ balance + income + student, data = credit, family = "binomial")

summary(modelo2)
```

```
##
## Call:
## glm(formula = default ~ balance + income + student, family = "binomial",
##     data = credit)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
##
## Number of Fisher Scoring iterations: 8
```

En relacioón a los coeficientes:

- **Saldo**:
  - El coeficiente asociado al **saldo** es positivo. Esto significa que, a medida que aumenta el saldo de un cliente, también lo hace la probabilidad de incumplimiento. El **odds ratio** para esta variable es mayor que 1, lo que implica un **incremento** en la probabilidad de impago. Supongamos que el **odds ratio** es $e^{\beta_1} = 1.007$, lo interpretamos de la siguiente manera:

$$(1.007 - 1) \times 100 = 0.7\%$$

Por cada incremento unitario en el saldo del cliente, los **odds** de incumplimiento aumentan en un **0.7%**. Esto muestra una relación directa entre saldo elevado y mayor riesgo de impago.

- **Ingresos**:
  - El coeficiente asociado a **ingresos** no es estadísticamente significativo, lo que sugiere que los ingresos no tienen un efecto claro en la probabilidad de incumplimiento cuando controlamos por el saldo y el estado de estudiante.
- **Estudiante**:
  - El coeficiente asociado a la variable **estudiante** es negativo. Esto indica que, manteniendo constantes otras variables, los estudiantes tienen una **menor probabilidad** de incumplimiento en comparación con los no estudiantes. Supongamos que el **odds ratio** para esta variable es $e^{\beta_3} = 0.85$. La interpretación con la fórmula $(1 - extodds)imes100$ sería:

$$(1 - 0.85) \times 100 = 15\%$$

Esto significa que los estudiantes tienen un **15%** menos de probabilidades de incumplir en comparación con los no estudiantes, cuando controlamos por saldo y otros factores.

Las variables **saldo** y **estado de estudiante** están correlacionadas. Los estudiantes tienden a tener niveles más altos de deuda, lo cual está asociado con una mayor probabilidad de incumplimiento. En otras palabras, es más probable que los estudiantes tengan grandes saldos en sus tarjetas de crédito, lo que a su vez está vinculado a tasas más altas de impagos.

Aunque un estudiante individual con un saldo determinado tiene una **probabilidad de incumplimiento más baja** en comparación con un no estudiante con el mismo saldo, el hecho de que los estudiantes, en promedio, tengan saldos más altos en sus tarjetas de crédito significa que, en general, los estudiantes tienen una **tasa de impago más alta** que los no estudiantes.

Como antes, podemos hacer predicciones fácilmente con este modelo. Por ejemplo, para un **estudiante** con un saldo de tarjeta de crédito de **1.500 euros** y un ingreso de **40.000 euros**, podemos estimar la probabilidad de incumplimiento de la siguiente manera:

```
# Estimación de la probabilidad de incumplimiento para un estudiante con saldo de 1500 y 40k de
new_data <- data.frame(balance = 1500, income = 40, student = "Yes")
pred_prob <- predict(modelo2, newdata = new_data, type = "response")
pred_prob
```

```
##          1
## 0.05161531
```

La probabilidad estimada de incumplimiento para este estudiante será el valor calculado por el modelo.

```
y_HAT = -10.90+(0.0057*1500)+(0.00001*40)-(0.809*1)
p=exp(y_HAT)/(1+exp(y_HAT))
p
```

```
## [1] 0.04075375
```

Esta probabilidad aumenta sinsiblemente al aumnetar el saldo (2.000,2.500)

```
y_HAT = -10.90+(0.0057*2000)+(0.00001*40)-(0.809*1)
p=exp(y_HAT)/(1+exp(y_HAT))
p
```

```
## [1] 0.4234565
```

```
y_HAT = -10.90+(0.0057*2500)+(0.00001*40)-(0.809*1)
p=exp(y_HAT)/(1+exp(y_HAT))
p
```

```
## [1] 0.9269936
```

También podemos identificar el coeficiente mas relevante para determinar la probabilidad de impago. utilizando la función R *varImp()* de la libreria *caret*

```r
library(caret)
varImp(modelo2)
```

```
##               Overall
## balance    24.737563
## income      0.369815
## studentYes  2.737646
```

Podemos ver que la variable más importante es el **saldo de la tarjeta**, seguida del **estatus de estudiante**, y por último los **ingresos**, que, como vimos, no eran significativos.

El **Likelihood ratio test** evalúa la significancia de la diferencia de los residuos entre el modelo de interés y el modelo nulo. El estadístico sigue una distribución **chi-cuadrado** con grados de libertad equivalentes a la diferencia de grados de libertad entre los dos modelos.

Podemos calcular este test en R mediante el siguiente comando:

```r
anova(modelo1, modelo2, test ='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: default ~ balance
## Model 2: default ~ balance + income + student
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      9998     1596.5
## 2      9996     1571.5  2   24.907 3.904e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observamos que el **segundo modelo** es mejor que el primero (p-value < 0.05). Esto indica que la introducción de más variables predictoras aporta **información significativa**.

La **matriz de confusión** es útil para evaluar el rendimiento de las predicciones del modelo. A continuación, mostramos el código para calcular la matriz de confusión para el **Modelo 2**:

```r
library(vcd)

# Modelo 2
predicciones <- ifelse(test = modelo2$fitted.values > 0.5, yes = 1, no = 0)
matriz_confusion2 <- table(credit$default, predicciones, dnn = c("observaciones", "predicciones

# Visualización de la matriz de confusión
matriz_confusion2
```

```
##              predicciones
## observaciones    0    1
##           No  9627   40
##           Yes  228  105
```

A partir de la matriz de confusión, podemos calcular las métricas de **accuracy** y **error**:

1. **Accuracy**:

$$\text{Accuracy} = \frac{A + D}{A + B + C + D}$$

   Para el **Modelo 2**:

$$\text{Modelo 2} \Rightarrow \frac{9627 + 105}{9627 + 105 + 40 + 228} = 0.97$$

2. **Error**:

$$\text{Error} = \frac{B + C}{A + B + C + D}$$

   Para el **Modelo 2**:

$$\text{Modelo 2} \Rightarrow \frac{40 + 228}{9627 + 105 + 40 + 228} = 0.03$$

Si imputamos un coste para la campaña de **6.000 euros por cliente**, podemos calcular las pérdidas generadas por incluir en la campaña a clientes que no realizarán compras o no incluir a clientes que sí lo harían.

Las pérdidas totales serían:

$$228 \times 6.000 = 1.368.000 \text{ euros}$$

8.2.4.Conclusiones

El director del estudio comprendió que la variable más importante que afecta la probabilidad de impago es el **saldo de la tarjeta de crédito**. Para **limitar los riesgos de impago**, hubiera sido mejor poner un **límite máximo** en el saldo, concediendo préstamos solo a clientes con saldos inferiores a **1.500 euros**.

Además, al clasificar a los clientes utilizando el modelo de regresión, las pérdidas estimadas habrían sido **1.368.000 euros**, que es más bajo que las pérdidas de la campaña anterior, donde se perdieron:

$$333 \times 6.000 = 1.998.000 \text{ euros}$$

El director también se dio cuenta de que había una relación entre las variables **estudiante** y **saldo**, y que ambas debían considerarse en el modelo para obtener estimaciones más precisas.

Anexo Demostración Teórica: Verosimilitud en Regresión Logística"

En la regresión logística, el objetivo es modelar la probabilidad de que ocurra un evento binario (1 o 0) utilizando un conjunto de variables predictoras X. La función que modela esta probabilidad es la función logística:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

donde: - $\pi(x)$ es la probabilidad de que el evento ocurra ($Y = 1$). - β_0 es el intercepto o término independiente. - β_1 es el coeficiente asociado a la variable X.

El objetivo es **maximizar la función de verosimilitud** para obtener las mejores estimaciones de β_0 y β_1.

A1. Paso 1: Definir la función de verosimilitud

Supongamos que tenemos un conjunto de datos con n observaciones. Si asumimos que las observaciones son independientes, la **verosimilitud** del conjunto de datos es el producto de las probabilidades de cada observación:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \left(\pi(x_i)^{y_i} \times (1 - \pi(x_i))^{1-y_i}\right)$$

donde: - $y_i = 1$ si el evento ocurre, $y_i = 0$ si no ocurre. - $\pi(x_i)$ es la probabilidad de que el evento ocurra dado el valor de X_i.

A2. Paso 2: Tomar el logaritmo de la función de verosimilitud

Para simplificar la maximización, tomamos el logaritmo natural de la función de verosimilitud. Esto convierte el producto en una suma:

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[y_i \log(\pi(x_i)) + (1 - y_i) \log(1 - \pi(x_i))\right]$$

Sustituyendo la expresión de $\pi(x_i)$, obtenemos:

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}\right) + (1 - y_i) \log\left(\frac{1}{1 + e^{\beta_0 + \beta_1 X_i}}\right)\right]$$

A3. Paso 3: Derivar la función log-verosimilitud

Para obtener los coeficientes β_0 y β_1, necesitamos **maximizar** la función de log-verosimilitud. Esto se hace derivando la función respecto a β_0 y β_1 y luego igualando a cero.

Las derivadas para los coeficientes son:

$$\frac{\partial \log L}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \pi(x_i))$$

$$\frac{\partial \log L}{\partial \beta_1} = \sum_{i=1}^{n} (y_i - \pi(x_i)) X_i$$

A4. Paso 4: Resolver las ecuaciones para obtener los coeficientes

No hay una solución analítica sencilla para estas ecuaciones, por lo que normalmente se utiliza un **algoritmo numérico**, como el método de Newton-Raphson, para resolverlas y obtener los valores de β_0 y β_1.

A5. Ejemplo calculado a mano

Consideremos un ejemplo con 3 observaciones. Supongamos que tenemos la siguiente tabla de datos:

X	Y
1	1
2	0
3	1

Queremos ajustar un modelo de regresión logística para predecir la probabilidad de $Y = 1$ en función de X.

La función de verosimilitud para estos datos es:

$$L(\beta_0, \beta_1) = \pi(1) \times (1 - \pi(2)) \times \pi(3)$$

donde:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Sustituyendo en la fórmula de log-verosimilitud:

$$\log L(\beta_0, \beta_1) = \log\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) + \log\left(\frac{1}{1+e^{\beta_0+2\beta_1}}\right) + \log\left(\frac{e^{\beta_0+3\beta_1}}{1+e^{\beta_0+3\beta_1}}\right)$$

Podemos hacer una suposición inicial de $\beta_0 = 0$ y $\beta_1 = 0$ para calcular los valores iniciales y luego ajustar mediante iteraciones numéricas. En este caso, resolveríamos manualmente para obtener aproximaciones de β_0 y β_1.

Análisis cluster

1. Introducción al Clustering

El término *clustering* se refiere a un amplio conjunto de técnicas no supervisadas (*unsupervised*) cuyo objetivo es encontrar patrones o grupos (*clusters*) dentro de un conjunto de observaciones. Las particiones se definen de manera que las observaciones dentro de un mismo grupo sean similares entre sí, mientras que las observaciones en grupos distintos sean diferentes. Se trata de un método no supervisado, ya que el proceso no toma en cuenta ninguna variable de respuesta que indique a qué grupo pertenece cada observación (si tal variable existiera).

En Marketing, diferenciamos principalmente dos tipos de *clusters*:

- **Partitioning Clustering**: Estos algoritmos requieren que el usuario especifique de antemano el número de *clusters* que se van a crear (por ejemplo, *K-means*, *K-medoids*, *CLARA*).
- **Hierarchical Clustering**: En este tipo de algoritmos, no es necesario que el usuario especifique previamente el número de *clusters* (por ejemplo, *agglomerative clustering* y *divisive clustering*).

2. El Concepto de Distancias

Todos los métodos de clustering tienen un aspecto en común: para poder llevar a cabo las agrupaciones, es necesario definir y cuantificar la similitud entre las observaciones.

El término "distancia" se utiliza en el contexto del clustering para referirse a la cuantificación de la similitud o diferencia entre observaciones. Si representamos las observaciones en un espacio de p dimensiones, donde p es el número de variables asociadas a cada observación, cuanto más se asemejen dos observaciones, más cercanas estarán entre sí. De ahí que se utilice el término "distancia". Una de las características que hace del clustering un método adaptable a diversos escenarios es la capacidad de emplear distintos tipos de distancia, lo que permite al investigador elegir la más adecuada para el estudio en cuestión. A continuación, se describen algunas de las distancias más utilizadas.

2.1. Distancia Euclidiana

La **distancia euclidiana** es la más comúnmente utilizada. Se define como la longitud del segmento que une dos puntos p y q. En coordenadas cartesianas, esta distancia se calcula aplicando el teorema de Pitágoras. Por ejemplo, en un espacio bidimensional donde cada punto está definido por las coordenadas (x, y), la distancia euclidiana entre los puntos p y q se expresa con la siguiente ecuación:

$$d_{euc}(p,q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$$

Esta fórmula puede generalizarse para un espacio euclidiano de n dimensiones, donde cada punto está definido por un vector de n coordenadas: $p = (p_1, p_2, ..., p_n)$ y $q = (q_1, q_2, ..., q_n)$:

$$d_{euc}(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

2.2. Otras Distancias

2.2.1. Distancia de Manhattan

La **distancia de Manhattan**, también conocida como *taxicab metric*, *rectilinear distance* o *L1 distance*, se define como la suma de las diferencias absolutas entre las coordenadas de dos puntos p y q. A diferencia de la distancia euclidiana, la distancia de Manhattan es más robusta frente a valores atípicos (*outliers*) porque no eleva las diferencias al cuadrado.

2.2.2. Distancia Basada en Correlación

La **correlación** es otra medida de distancia útil cuando la similitud entre observaciones se define en términos de patrones o formas, en lugar de desplazamientos o magnitudes. El **coeficiente de correlación de Pearson** es particularmente efectivo en una amplia variedad de contextos. Sin embargo, este coeficiente no es robusto frente a *outliers*, incluso si se cumple la condición de normalidad.

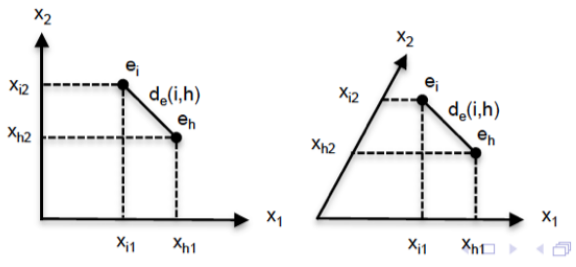
Para mitigar este efecto, se puede utilizar la **Jackknife correlation**, que calcula todos los posibles coeficientes de correlación excluyendo una observación en cada iteración. El promedio de todas las correlaciones calculadas reduce el impacto de los *outliers*.

2.2.3. Distancias para Variables Binarias

Cuando las variables utilizadas para determinar la similitud entre observaciones son binarias, no es apropiado utilizar medidas de distancia como la euclidiana o de Manhattan, ya que estas se basan en operaciones aritméticas que no tienen sentido en este contexto. Por ejemplo, si codificamos la variable "sexo" como 1 para mujeres y 0 para hombres, no tiene sentido decir que la media de la variable sexo en un conjunto de datos es 0.5. En estos casos, es necesario emplear otras medidas de similitud adecuadas para variables categóricas o binarias.

2.3. El problema de la corrección entre variables.

Unos de los principales problemas que podemos encontrar cuando calculamos las distancias es que este se ve afectada por la correlacionó entre variables.



Si observamos la figura podemos ver claramente que las distancias son distintas los los dos casos. De hecho, cuanto mas alta sea la correlación entre las variables menos será la distancia. Por esto es buena practica realizar un análisis en componentes principales o (como vendremos mas

adelante para las variables categóricas) un análisis de las correspondencias, y utilizar las componentes como substitutas de las variables originales que presentan la importante propiedad de ser incorrellacionadas.

3. K-means

El método de *k-means clustering* (MacQueen, 1967) agrupa las observaciones en k clusters distintos, donde el número k es determinado por el analista. K-means clustering busca los k mejores clusters, entendiendo como "mejor" aquel cuya variación interna (*intra-cluster variation*) sea lo más pequeña posible. Por lo tanto, se trata de un problema de optimización en el que las observaciones se distribuyen en k clusters de manera que la suma de las varianzas internas de todos ellos sea mínima.

Para resolver este problema, se utiliza un algoritmo basado en los siguientes pasos:

1. Especificar el número k de clusters que se desean crear.
2. Seleccionar de manera aleatoria k observaciones del conjunto de datos como centroides iniciales.
3. Asignar cada observación al centroide más cercano.
4. Recalcular el centroide de cada uno de los k clusters.
5. Repetir los pasos 3 y 4 hasta que las asignaciones no cambien o se alcance el número máximo de iteraciones establecido.

Dado que el algoritmo de *k-means* no evalúa todas las posibles distribuciones de las observaciones, sino solo una parte de ellas, los resultados obtenidos dependen de la asignación aleatoria inicial (paso 2). Por esta razón, es importante ejecutar el algoritmo varias veces (entre 20 y 50), cada una con una asignación inicial diferente, y seleccionar el resultado que logre una menor varianza total.

3.1. Ventajas y Desventajas

K-means es uno de los métodos de clustering más utilizados debido a la simplicidad y rapidez de su algoritmo, pero también presenta una serie de limitaciones que es importante considerar.

1. Requiere que el número de clusters (k) se especifique de antemano. Esto puede ser complicado si no se dispone de información suficiente sobre los datos. Una posible solución es ejecutar el algoritmo para un rango de valores de k y evaluar cuál proporciona mejores resultados, por ejemplo, minimizando la suma total de varianza interna.
2. Los resultados del clustering pueden variar dependiendo de la asignación inicial aleatoria de los centroides. Para reducir este problema, se recomienda repetir el proceso entre 20 y 50 veces, y seleccionar el resultado con menor varianza interna. Sin embargo, no se garantiza que los resultados sean siempre idénticos para un mismo conjunto de datos.
3. *K-means* podria ser no robusto frente a la presencia de *outliers*.

3.2. La funcion *kmeans()*

En este estudio, consideramos un conjunto de **datos simulados** que contiene observaciones pertenecientes a cuatro grupos distintos. Nuestro objetivo es aplicar el algoritmo de **K-means clustering** para identificar correctamente estos grupos a partir de las características de las observaciones.

El análisis consiste en dividir las observaciones en k clusters y compararlas con los grupos reales a los que pertenecen para evaluar la efectividad del método.

Para llevar a cabo este proceso, seguiremos los pasos estándar del algoritmo de K-means:

1. Especificar el número de grupos (k) que deseamos encontrar, en este caso $k = 4$.
2. Seleccionar aleatoriamente cuatro observaciones del conjunto de datos como los centroides iniciales.
3. Asignar cada observación al centroide más cercano.
4. Recalcular los centroides de los clusters.
5. Repetir los pasos 3 y 4 hasta que las asignaciones de las observaciones no cambien o se alcance el número máximo de iteraciones establecido.

Al final del análisis, compararemos los resultados obtenidos por K-means con los grupos originales, para determinar el nivel de precisión del método en este conjunto de datos simulado.

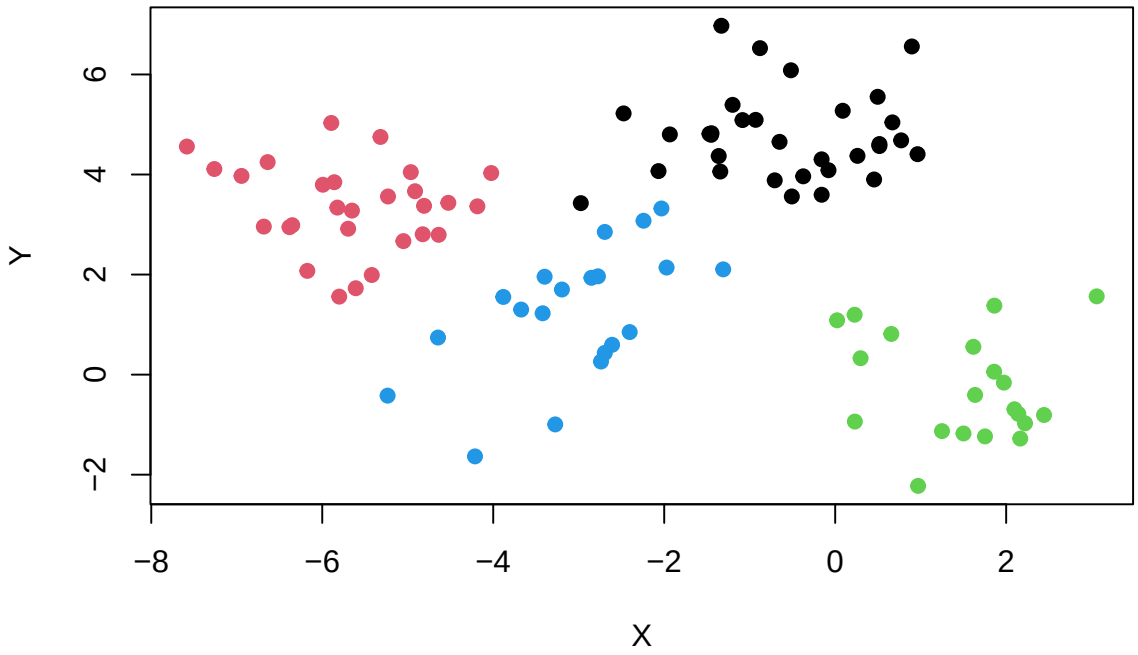
3.2.1 Resultados Esperados

Dado que los datos simulados contienen cuatro grupos claramente definidos, esperamos que el algoritmo de K-means sea capaz de identificar correctamente la mayoría de las observaciones y agruparlas de forma similar a los grupos originales. Sin embargo, como en todos los análisis basados en K-means, los resultados pueden variar ligeramente debido a la asignación inicial aleatoria de los centroides.

Para asegurar resultados más estables, ejecutaremos el algoritmo varias veces y seleccionaremos la ejecución que presente la menor suma total de varianza interna.

```
set.seed(101)
# Se simulan datos aleatorios con dos dimensiones
datos <- matrix(rnorm(n = 100*2), nrow = 100, ncol = 2,
                dimnames = list(NULL,c("x", "y")))

# Se determina la media que va a tener cada grupo en cada una de las dos
# dimensiones. En total 2*4 medias. Este valor se va a utilizar para
# separar cada grupo de los demás.

media_grupos <- matrix(rnorm(n = 8, mean = 0, sd = 4), nrow = 4, ncol = 2,
dimnames = list(NULL, c("media_x", "media_y")))
media_grupos <- cbind(grupo = 1:4, media_grupos)
```

```r
# Se genera un vector que asigne aleatoriamente cada observación a uno de
# los 4 grupos

grupo <- sample(x = 1:4, size = 100, replace = TRUE)
datos <- cbind(datos, grupo)

# Se incrementa el valor de cada observación con la media correspondiente al
# grupo asignado.

datos <- merge(datos, media_grupos, by = "grupo")
datos[,"x"] <- datos[, "x"] + datos[, "media_x"]
datos[,"y"] <- datos[, "y"] + datos[, "media_y"]

plot(x = datos[,"x"], y = datos[,"y"], col = datos[,"grupo"], pch = 19,
     xlab = "X", ylab = "Y")
```



3.2.2 Uso de la Función `kmeans()` en K-means Clustering

La función `kmeans()` de la librería **stats** en R es utilizada para realizar el **K-means clustering**. Entre sus argumentos más importantes se encuentran:

- `centers`: Define el número k de clusters que se desean generar.
- `nstart`: Determina el número de veces que se repetirá el proceso de clustering, cada vez con una asignación aleatoria inicial diferente.

Es recomendable establecer un valor alto para el argumento `nstart` (entre 20 y 50) para evitar obtener resultados subóptimos debido a una mala asignación inicial de los centroides. Un valor más alto asegura una mejor probabilidad de encontrar una solución óptima.

Dado que los datos simulados tienen aproximadamente la misma magnitud en todas las dimensiones, **no es necesario escalarlos ni centrarlos** antes de aplicar el algoritmo. Esto simplifica el preprocesamiento de los datos, permitiendo ejecutar el algoritmo directamente sobre el conjunto de datos original.

```
set.seed(101)
km_clusters <- kmeans(x = datos[, c("x", "y")], centers = 4, nstart = 50)
km_clusters
```

```
## K-means clustering with 4 clusters of sizes 20, 28, 32, 20
##
## Cluster means:
##            x          y
## 1  1.4989983 -0.2412154
## 2 -5.6518323  3.3513316
## 3 -0.5787702  4.7639233
## 4 -3.1104142  1.2535711
##
## Clustering vector:
##   [1] 3 3 3 3 3 3 3 4 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2
##  [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 3 4
##
## Within cluster sum of squares by cluster:
## [1] 34.95921 42.40322 53.04203 48.52107
##  (between_SS / total_SS =  85.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

El objeto devuelto por la función `kmeans()` contiene, entre otros elementos, los siguientes datos relevantes para la interpretación del clustering:

- La media de cada una de las variables para cada cluster.
- Un vector que indica a qué cluster ha sido asignada cada observación.
- La suma de cuadrados interna de cada cluster (*within-cluster sum of squares*).
- El ratio de la suma de cuadrados entre clusters y la suma de cuadrados totales ($\frac{between_{SS}}{total_{SS}}$).

Este último término es equivalente al R^2 en los modelos de regresión y representa el porcentaje de varianza explicada por el modelo en relación con la varianza total observada. Este valor puede utilizarse para evaluar la calidad del clustering obtenido. Sin embargo, al igual que ocurre con el R^2, este ratio ($\frac{between_{SS}}{total_{SS}}$) aumenta a medida que se incrementa el número de clusters. Por lo tanto, es importante tener en cuenta este fenómeno para evitar problemas de **overfitting**, es decir, ajustar el modelo en exceso a los datos.

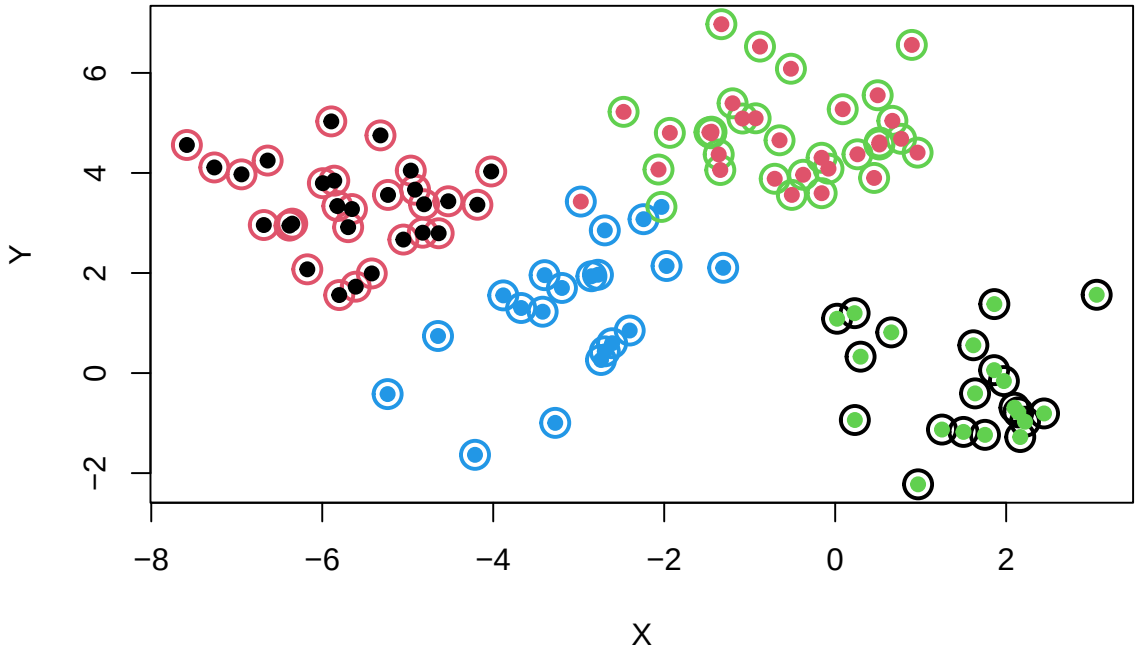
3.2.3 Evaluación del Clustering

En este caso particular, al tratarse de una simulación, conocemos el número real de grupos (4) y a qué grupo pertenece cada observación. Esta situación, aunque rara en la mayoría de los casos prácticos, es extremadamente útil para evaluar la efectividad del método de **K-means clustering** en la clasificación de las observaciones.

El conocimiento previo de los grupos proporciona una referencia objetiva que nos permite medir el rendimiento del algoritmo con mayor precisión. Además, esta información resulta valiosa para ajustar parámetros y mejorar el desempeño del método en situaciones reales, donde el número de grupos y sus asignaciones no son conocidos de antemano.

En contextos prácticos, donde los grupos no están definidos de forma explícita, la capacidad de validar y ajustar el clustering puede mejorar significativamente la utilidad del modelo.

```
# Se representan circunferencias con las asignaciones hechas por
# K-means-clustering
datos <- cbind(cluster = km_clusters$cluster, datos)
plot(x = datos[,"x"], y = datos[,"y"], col = km_clusters$cluster, pch = 1,
     cex = 2, lwd = 2, xlab = "X", ylab = "Y")

# Se rellenan las circunferencias con puntos del color real del grupo al
# que pertenecen las observaciones. Es necesario hacer coincidir los
# colores con el mismo orden que el devuelto por la función kmeans() ya
# que el clustering no asigna variable respuesta, solo agrupa las
# observaciones.

points(x = datos[,"x"], y = datos[,"y"],
col = c(2, 1, 3, 4)[datos[, "grupo"]], pch = 19)
```



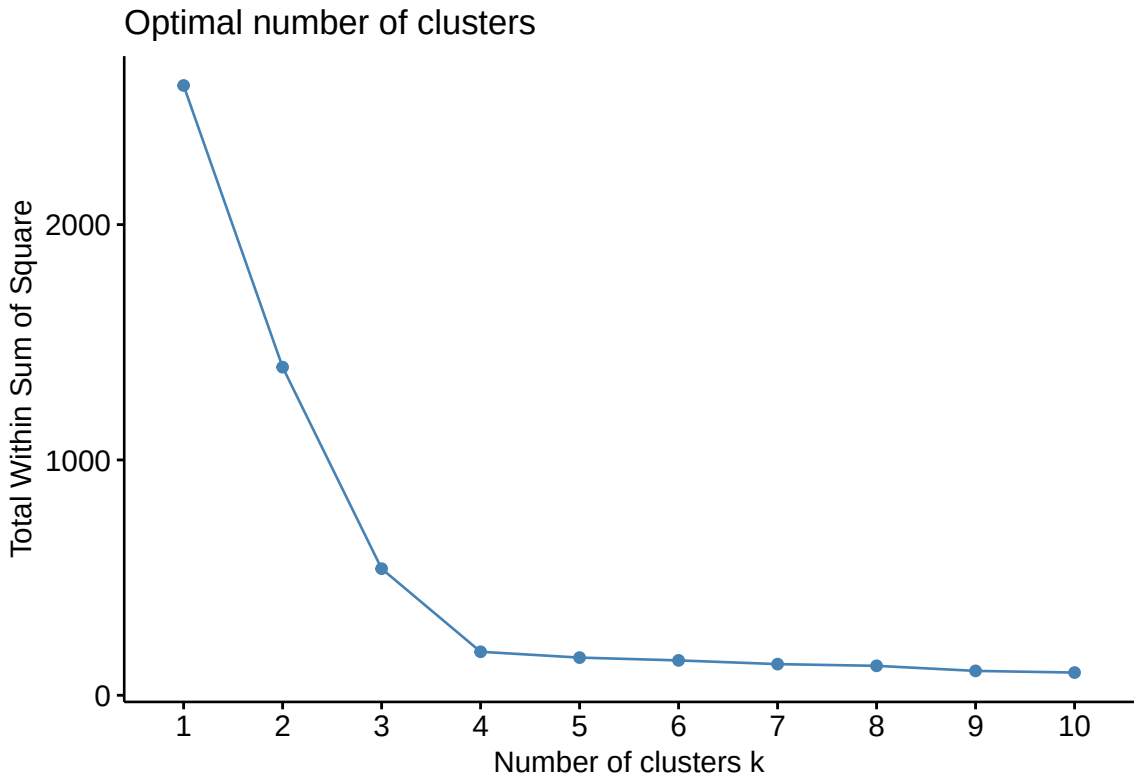
El gráfico muestra que solo dos observaciones han sido asignadas incorrectamente a sus clusters. Este tipo de visualización es muy útil e informativa, pero solo es posible cuando se trabaja con datos en dos dimensiones. Cuando los datos contienen más de dos variables (dimensiones), una solución posible es utilizar las **dos primeras componentes principales** obtenidas mediante un Análisis de Componentes Principales (PCA) para la visualización.

Otra opción sería aplicar primero un **clustering jerárquico** y, posteriormente, utilizar el gráfico de inercia (*scree plot*) para decidir el número óptimo de clusters a utilizar. Este enfoque es útil para identificar de manera visual cuántos grupos parecen estar presentes en los datos.

Por último, una manera sencilla de estimar el número óptimo de k clusters cuando no se dispone de información adicional es aplicar el algoritmo K-means para un rango de valores de k, identificando aquel a partir del cual la reducción en la suma total de la varianza intra-cluster deja de ser significativa. Este método es conocido como el "método del codo" (*elbow method*), y será detallado más adelante.

Para automatizar este proceso, la función `fviz_nbclust()` puede ser utilizada, lo que facilita la selección del número óptimo de clusters al analizar diferentes métricas y criterios.

```
suppressMessages(library(factoextra))
fviz_nbclust(x = datos, FUNcluster = kmeans, method = "wss",
             diss = dist(datos, method = "euclidean"))
```



4. Clúster jerárquico

El **Hierarchical Clustering** es una alternativa a los métodos de *Partitioning Clustering* que no requiere predefinir el número de clusters. Este método se divide en dos tipos principales, dependiendo

de la estrategia utilizada para formar los grupos:

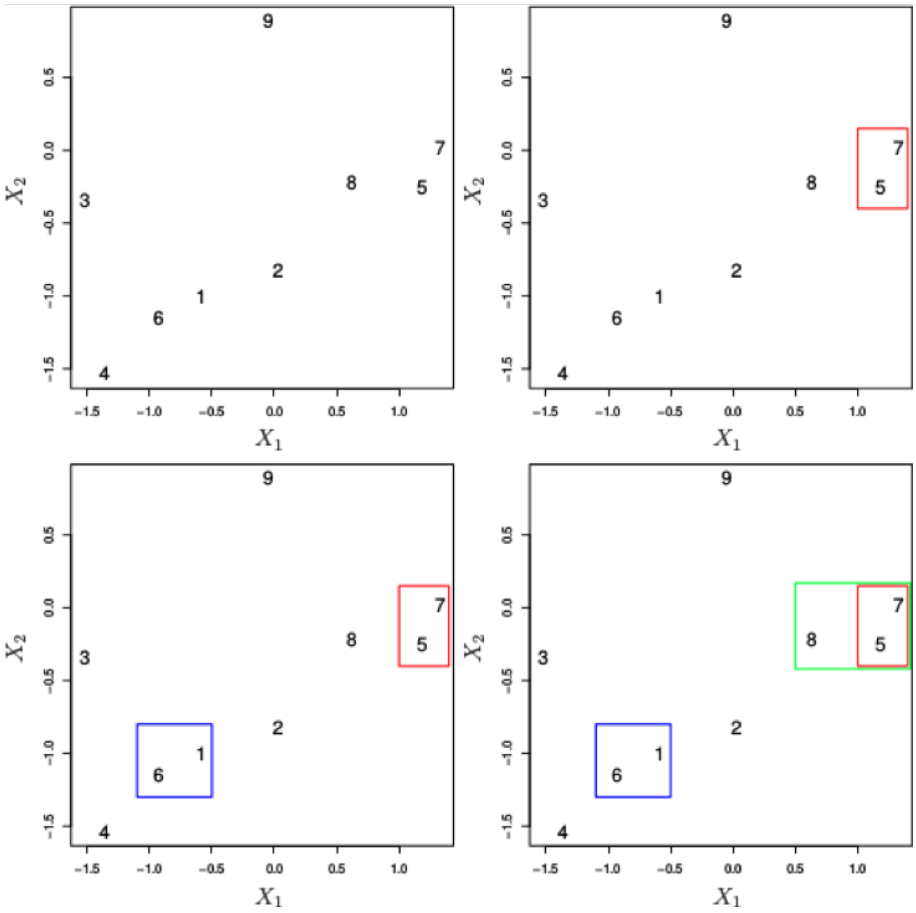
- **Agglomerative Clustering** (*bottom-up*): El agrupamiento comienza en la base del árbol, donde cada observación forma un clúster individual. A medida que el proceso avanza, los clusters se van combinando hasta que todos forman una única "rama" central.
- **Divisive Clustering** (*top-down*): Es el enfoque opuesto al *Agglomerative Clustering*. Comienza con todas las observaciones en un solo clúster, y se van realizando divisiones hasta que cada observación forma un clúster individual.

En ambos casos, los resultados pueden representarse de forma intuitiva mediante un árbol llamado **dendrograma**. En este capítulo, nos centramos exclusivamente en los métodos de *Hierarchical Clustering* de tipo aglomerativo.

4.1. El Algoritmo de Agglomerative Hierarchical Clustering

La estructura resultante de un **Agglomerative Hierarchical Clustering** se obtiene mediante un algoritmo sencillo:

1. El proceso comienza considerando cada observación como un clúster individual, formando así la base del dendrograma.
2. Se inicia un proceso iterativo hasta que todas las observaciones pertenecen a un único clúster:
    - Se calcula la distancia entre cada par de los n clusters. El investigador debe seleccionar la medida de distancia y el tipo de *linkage* (vinculación) que se empleará para cuantificar la similitud entre observaciones o grupos.
    - Los dos clusters más similares se fusionan, dejando un total de *n-1* clusters.
3. Se determina dónde cortar la estructura del árbol generada (dendrograma) para identificar los clusters.



Para que este proceso funcione correctamente, es fundamental definir cómo se mide la similitud entre clusters. Esto implica extender el concepto de distancia entre pares de observaciones para que sea aplicable a pares de grupos, cada uno formado por varias observaciones. Este proceso se conoce como **linkage**.

4.1.1. Tipos de Linkage

A continuación, se describen los tipos de *linkage* más comunes:

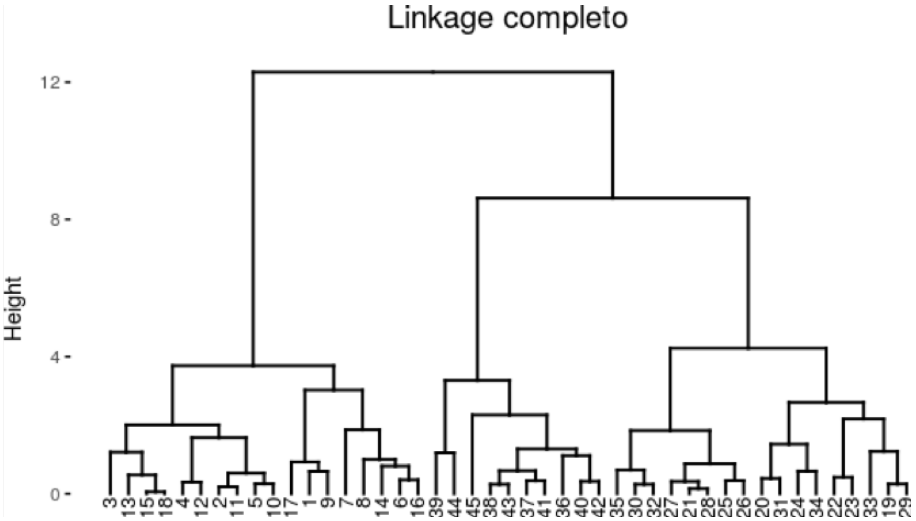
- **Complete (Maximum)**: Se calcula la distancia entre todos los pares posibles entre los clusters A y B, y se selecciona la mayor de todas como la distancia entre ambos clusters (*maximal intercluster dissimilarity*). Es la medida más conservadora.
- **Single (Minimum)**: Se calcula la distancia entre todos los pares posibles entre los clusters A y B, y se selecciona la menor de todas como la distancia entre los dos clusters (*minimal intercluster dissimilarity*). Es la medida menos conservadora.
- **Average**: Se calcula la distancia entre todos los pares posibles entre los clusters A y B, y se toma el promedio de todas como la distancia entre los dos clusters (*mean intercluster dissimilarity*).
- **Centroid**: Se calcula el centroide de cada clúster, y la distancia entre los centroides se toma como la distancia entre los clusters.
- **Ward's**: Es un método más general, donde se seleccionan los clusters a combinar en cada paso del proceso, minimizando el incremento de la varianza total intra-cluster. El método

Ward's minimum variance es un caso particular que busca minimizar la suma total de la varianza intra-cluster en cada paso.

Los métodos *Complete*, *Average* y *Ward's minimum variance* suelen ser los preferidos por los analistas porque generan dendrogramas más equilibrados. Sin embargo, la elección del mejor método depende del caso de estudio.

4.2. El Dendrograma

Supongamos que tenemos 45 observaciones en un espacio bidimensional que pertenecen a 3 grupos. Al aplicar *Hierarchical Clustering* utilizando la distancia euclidiana y el *Complete linkage*, se obtiene el siguiente dendrograma:

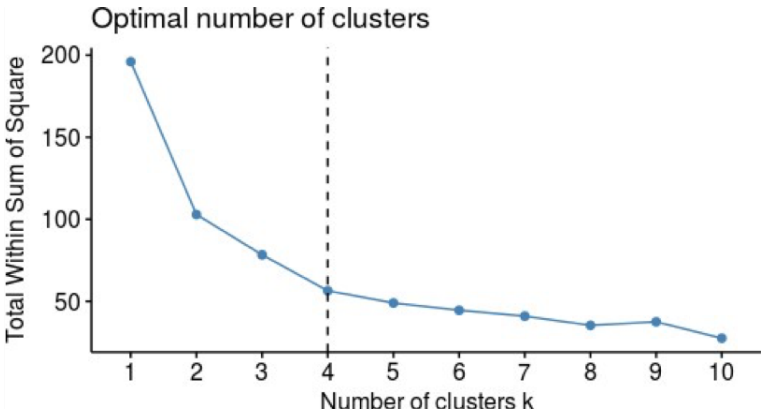


En la base del dendrograma, cada observación forma una hoja (*leaf*) individual. A medida que se asciende por la estructura, pares de hojas se combinan formando ramas. Estas uniones (nodos) representan las observaciones más similares. También puede ocurrir que ramas se fusionen con otras ramas o con hojas. Cuanto más cerca de la base ocurre una fusión, mayor es la similitud entre las observaciones o clusters.

Es importante destacar que los dendrogramas deben interpretarse únicamente basándose en el eje vertical (altura). Las posiciones horizontales de las observaciones no tienen un significado específico y pueden variar entre diferentes programas.

Por ejemplo, en el gráfico anterior, la observación 11 es la más similar a la observación 2, ya que es la primera fusión que recibe la observación 2. Sin embargo, observaciones más cercanas en el eje horizontal, como la 12, no necesariamente son más similares. De hecho, las observaciones 5 y 10 son más similares a la 2 que la observación 12, a pesar de estar más alejadas horizontalmente.

Además de representar la similitud entre observaciones, el dendrograma permite identificar el número de clusters. Esto se hace cortando el dendrograma a una determinada altura. El número de ramas que se extienden más allá de ese corte corresponde al número de clusters. Para determinar cuántos clusters formar, se suele utilizar el **inertia plot**, que muestra el grado de similitud al considerar más grupos. El número óptimo de clusters corresponde a un punto en el que se observa un cambio significativo en la inercia (elbow point).



4.3. La Función HCPC en Clustering Jerárquico

La función `HCPC` de la librería **FactoMineR** permite generar un **clustering jerárquico**. Esta función ofrece varios parámetros, entre los cuales destaca `nb.clust`, que permite definir el número de clusters. Es recomendable, en una primera etapa del análisis, establecer `nb.clust = -1` para que el software realice una primera agrupación automática. Posteriormente, se puede seleccionar el número óptimo de clusters utilizando el gráfico de inercia, que se obtiene mediante el comando `plot(hcpc, choice = "bar")`.

La función `HCPC` genera varios **outputs** importantes, entre ellos:

- `hcpc$desc.var`: Permite interpretar los grupos formados, proporcionando una descripción detallada de cada cluster.
- `hcpc$data.clust`: Devuelve el clúster al que pertenece cada observación.

Además, el **dendrograma** y el **mapa de las observaciones** se pueden visualizar con los comandos `plot(hcpc, choice = "tree")` y `plot(hcpc, choice = "map")`, respectivamente.

4.3.1. Descripción de los Grupos

Después de realizar el clustering, uno de los principales desafíos es la descripción de los grupos formados. Esto se puede lograr mediante el **output** `desc.var`. Todas las variables del conjunto de datos original, ya sean continuas, categóricas, activas o complementarias, se utilizan para describir los clusters. La metodología utilizada se describe en la sección 3.3 de **Le et al. (2008)** y en **Lebart, Morineau y Warwick (1984)**.

Para las **variables continuas**, el output proporciona:

- El promedio de la variable en cada grupo (media dentro del grupo).
- El promedio de la variable en todo el conjunto de datos (media general).
- Las desviaciones estándar asociadas.
- El valor p correspondiente a la prueba de hipótesis: “La media del grupo es igual a la media general”.

Un valor de `v.test` mayor que 1.96 corresponde a un valor p menor que 0.05, lo cual indica que la media del grupo difiere significativamente de la media general. El signo del `v.test` indica si la media del grupo es mayor o menor que la media general.

4.3.2. Individuos Representativos de los Clusters

Una forma interesante de ilustrar los clusters es utilizando **individuos específicos** de cada grupo. Para ello, se sugieren dos tipos de individuos:

- **Paragons**: Son los individuos que están más cerca del centro del clúster.
- **Individuos específicos**: Son los individuos que están más alejados de los centros de otros grupos.

El objeto `desc.ind` contiene, para cada grupo, los individuos ordenados por la distancia entre cada uno de ellos y el centro de su clúster, facilitando la identificación de los más representativos.

5. Ejemplos

5.1. Nissan case

Nissan, un fabricante líder en la industria automotriz, está preparando el lanzamiento de un nuevo coche deportivo para el mercado español. Para garantizar que las estrategias de marketing y posicionamiento sean efectivas, es crucial identificar a los competidores clave en el segmento y comprender cómo los consumidores perciben tanto el nuevo modelo como las marcas existentes.

El objetivo de este estudio es realizar un **análisis clúster** que permita a Nissan obtener una visión clara del panorama competitivo en el mercado español. A través de este análisis, se busca responder las siguientes preguntas:

- ¿Cómo perciben los consumidores las características del nuevo modelo en comparación con sus competidores?
- ¿Son adecuadas las estrategias de marketing actuales de Nissan?
- ¿Existe un competidor principal o varios competidores con características similares?

Data set El análisis se basará en un conjunto de datos que contiene información sobre varias marcas de automóviles, incluidas las percepciones de los consumidores en relación con las siguientes características:

- **Mecánica**
- **Estabilidad**
- **Habitabilidad**
- **Comodidad**
- **Equipamiento**
- **Prestaciones**
- **Consumo**

Cada variable ha sido evaluada en una escala de 1 a 5, donde 1 indica que el consumidor no asocia la característica con la marca y 5 indica una fuerte asociación. Las valoraciones son promedios obtenidos de una muestra de 500 entrevistados para cada marca.

Análisis ¿Es posible aplicar el análisis de clúster directamente a los datos disponibles?

Antes de aplicar el análisis clúster, es importante verificar si los datos son adecuados para este tipo de análisis. En este caso, dado que todas las variables están en la misma escala (1-5), no es necesario realizar un escalado previo. Sin embargo, sería recomendable realizar una revisión exploratoria de los datos para identificar posibles valores correlaciones entre las variables que afecten los resultados.

2. ¿Es necesario realizar un Análisis de Componentes Principales (ACP)?

Un Análisis de Componentes Principales (ACP) puede ser útil si queremos reducir la dimensionalidad de los datos y centrarnos en las variables que más contribuyen a la varianza en las percepciones de los consumidores. Esto podría simplificar la interpretación de los clústeres, pero no es estrictamente necesario para este análisis. Evaluaremos si el ACP aporta valor en este caso una vez que hayamos examinado la correlación entre las variables.

5.1.2 Cluster jerárquico

Para comenzar con el análisis, primero cargaremos el archivo de datos que contiene la información relativa a las percepciones promedio de los consumidores. Este archivo será leído utilizando la función `read.csv()` de R, que nos permitirá importar los datos en un formato adecuado para su análisis.

```
# leo mis datos
x = read.csv(file="marcas_coches.csv", header = TRUE,sep=";")
# uso los nombres de los estados como etiquetas de las líneas de mi data
rownames(x) = t(x[,1])
x=x[,-1]
x$Perfil = as.factor(x$Perfil)
```

También es recomendable verificar la presencia de correlaciones elevadas (mayores a 0.5) antes de decidir si conviene aplicar un Análisis de Componentes Principales (ACP).

```
# Analizo las correlaciones
round(cor(x[, -8]),2)
```

```
##                mecanica estabilidad habitabilidad comodidad equipamiento
## mecanica           1.00        0.06         -0.45     -0.60         0.39
## estabilidad        0.06        1.00         -0.18     -0.02         0.15
## habitabilidad     -0.45       -0.18          1.00      0.64        -0.59
## comodidad         -0.60       -0.02          0.64      1.00        -0.06
## equipamiento       0.39        0.15         -0.59     -0.06         1.00
## Prestaciones       0.46        0.20         -0.61     -0.60         0.39
## consumo           -0.23        0.00          0.00      0.21        -0.15
##                Prestaciones consumo
## mecanica               0.46   -0.23
## estabilidad            0.20    0.00
## habitabilidad         -0.61    0.00
```

```
## comodidad            -0.60    0.21
## equipamiento          0.39   -0.15
## Prestaciones          1.00   -0.46
## consumo              -0.46    1.00
```

Podemos observar que existen varias correlaciones significativas entre las variables. Por ejemplo:

- **Comodidad** muestra una correlación inversa considerable con **mecánica** (-0.60) y **prestaciones** (-0.60).
- En contraste, **comodidad** presenta una correlación positiva elevada con **habitabilidad** (0.64).
- También observamos que **habitabilidad** y **equipamiento** tienen una correlación negativa elevada (-0.59).

Debido a estas correlaciones notables entre las variables, hemos decidido calcular un **Análisis de Componentes Principales (ACP)** y centrarnos en las **dos primeras componentes principales** para el análisis.

```
library(FactoMineR)
pca = PCA(x,quali.sup=8, graph=FALSE)
```

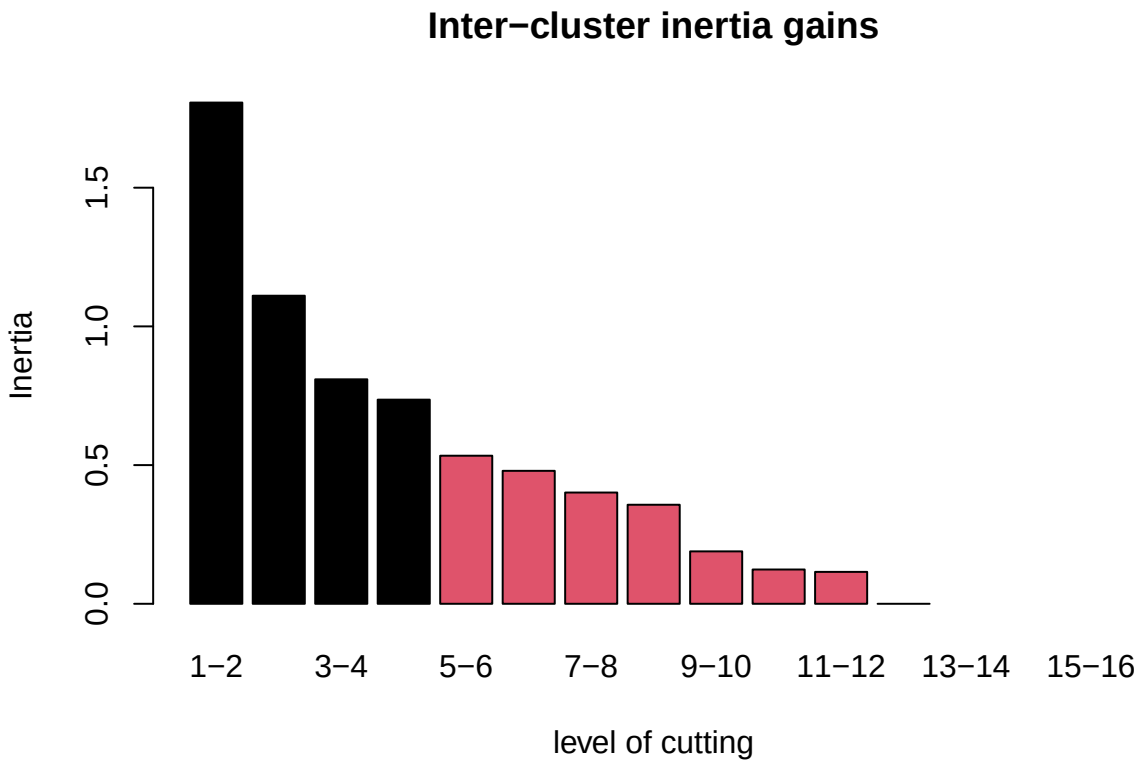
Step 1: Identifico el numero de clústers. Utilizo el gráfico de la **inercia intra clases** para identificar el número óptimo de clusters. Al aumentar el número de grupos, se observa cómo las barras en el gráfico se hacen más pequeñas, lo que indica una disminución gradual de la variabilidad dentro de cada clúster.

Es importante elegir un número limitado de grupos para que el análisis sea interpretable. Por lo tanto, la decisión debe ser un compromiso entre la calidad de los clusters y su interpretabilidad y usabilidad.

La mejor opción suele corresponder al punto del gráfico donde se produce el salto más significativo. En este caso, hemos decidido optar por 3 grupos, basándonos en el segundo salto significativo.

Es importante destacar que la decisión debe tener en cuenta también el número de variables y observaciones. En este caso, dado que ambos son limitados, no sería adecuado considerar un número elevado de grupos.

```
hcpc = HCPC(pca,nb.clust=-1,graph=FALSE)
plot(hcpc, choice="bar")
```




```
clust=3 # <-- modifica el valor numero con el numero que clusters deseado
hcpc = HCPC(pca,nb.clust=clust, graph=FALSE)

hcpc$desc.var
```

Step 2 y 3: Realizo en análisis y interpreto los grupos

```
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##                     Eta2     P-value
## habitabilidad 0.6944444 0.002663489
## Prestaciones  0.6196809 0.007956846
## comodidad     0.6163194 0.008314745
## mecanica      0.5873016 0.011971925
##
## Description of each cluster by quantitative variables
## ======================================================
## $`1`
##                 v.test Mean in category Overall mean sd in category Overall sd
## comodidad     2.583333             3.75     3.153846      0.4330127  0.5329387
## Prestaciones -2.441514             2.75     3.538462      0.4330127  0.7457969
## mecanica     -2.494438             3.00     3.538462      0.0000000  0.4985185
```

```
##                   p.value
## comodidad    0.009785073
## Prestaciones 0.014625802
## mecanica     0.012615667
##
## $`2`
## NULL
##
## $`3`
##                   v.test Mean in category Overall mean sd in category Overall sd
## equipamiento   2.087103         4.000000     3.153846      0.8164966  0.7692308
## Prestaciones   2.022217         4.333333     3.538462      0.4714045  0.7457969
## habitabilidad -2.792848         2.000000     3.000000      0.0000000  0.6793662
##                    p.value
## equipamiento  0.036878802
## Prestaciones  0.043153936
## habitabilidad 0.005224623
```

Step 4: visualizo el albor, la tabla con mis datos y el correspondiente clúster de cada observación Primero, observamos que las variables que más caracterizan a los clusters son, en orden: **habitabilidad**, **prestaciones**, **comodidad** y **mecánica**. El valor de η^2 puede interpretarse como una correlación: cuanto más elevado sea, más define la variable los grupos.

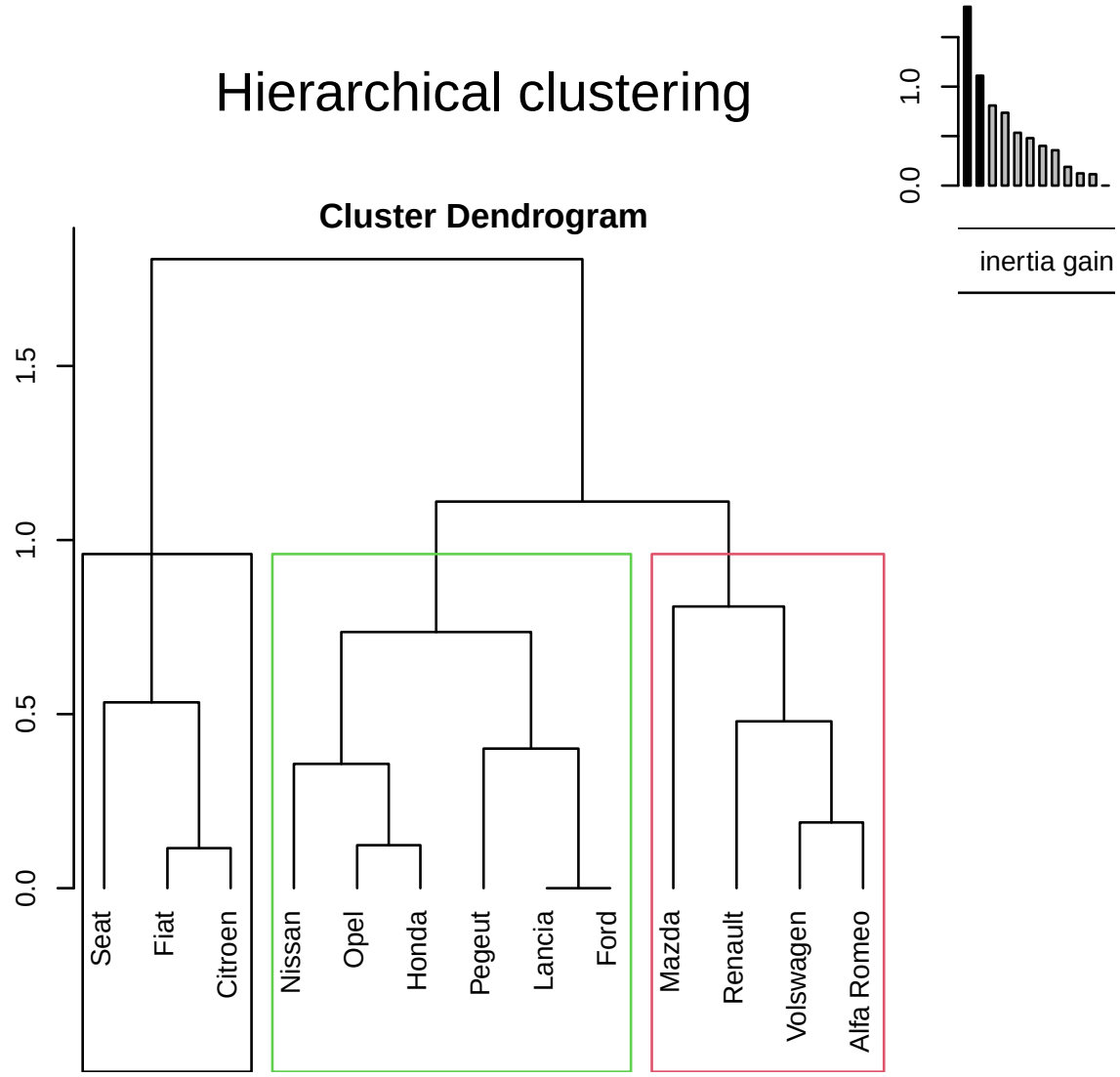
Posteriormente, procedemos a interpretar los grupos. El **v-test** nos indica si una variable es significativa para identificar un grupo. Un valor positivo del v-test sugiere que el grupo se caracteriza por tener un promedio más alto en esa variable, mientras que un valor negativo indica lo contrario. Además del v-test, la función `HCPC` proporciona el promedio de la variable en el grupo, el promedio general y sus desviaciones estándar (tanto intra-grupo como general).

Interpretación de los Grupos:

- **Grupo 1**: Este grupo se caracteriza principalmente por el atributo **comodidad** (v-test = 2.58, media = 3.75 frente a 3.1 de la media general). Sin embargo, los coches que forman este grupo no son percibidos como vehículos de **buenas prestaciones** (v-test = -2.48, media = 2.75 frente a 3.53 de la media general) ni de **buena mecánica** (v-test = -2.49, media = 3 frente a 3.53 de la media general). Este grupo podría interpretarse como el de **"los familiares"**.
- **Grupo 2**: Este grupo no se caracteriza por ningún atributo en particular, lo que indica que los coches que lo conforman son percibidos como vehículos con características promedio.
- **Grupo 3**: Este grupo se caracteriza por la variable **equipamiento** (v-test = 2.08, media = 4 frente a 3.15 de la media general) y por **prestaciones** (v-test = 2.02, media = 4.33 frente a 3.53 de la media general), mientras que se aleja de atributos como **habitabilidad** (v-test = -2.79, media = 2 frente a 3 de la media general). Este grupo podría identificarse como **"los deportivos"**.

Finalmente, podemos visualizar la clasificación de los clusters mediante el **dendrograma** y los datos obtenidos.

```
plot(hcpc,choice="tree")
```




```
hcpc$data.clust
```

```
##             mecanica estabilidad habitabilidad comodidad equipamiento
## Lancia             3           3              3         3             3
## Citroen            3           3              3         4             3
## Fiat               3           3              4         4             3
## Ford               3           3              3         3             3
## Honda              4           3              3         3             4
## Alfa Romeo         4           4              2         2             3
## Mazda              4           4              2         3             5
## Nissan             4           4              2         3             4
## Opel               4           3              3         3             3
## Pegeut             3           4              3         3             2
```

```
## Seat              3           5             4         4            3
## Renault           4           3             4         3            2
## Volswagen         4           4             3         3            3
##            Prestaciones consumo   Perfil clust
## Lancia                4       3    Mayor     2
## Citroen               3       3    Mayor     1
## Fiat                  2       3    Mayor     1
## Ford                  4       3    Mayor     2
## Honda                 3       3    Mayor     2
## Alfa Romeo            4       2    Joven     3
## Mazda                 5       2    Joven     3
## Nissan                4       4 Familiar     3
## Opel                  3       4 Familiar     2
## Pegeut                3       4 Familiar     1
## Seat                  3       3 Familiar     1
## Renault               4       2    Joven     2
## Volswagen             4       3    Joven     2
```

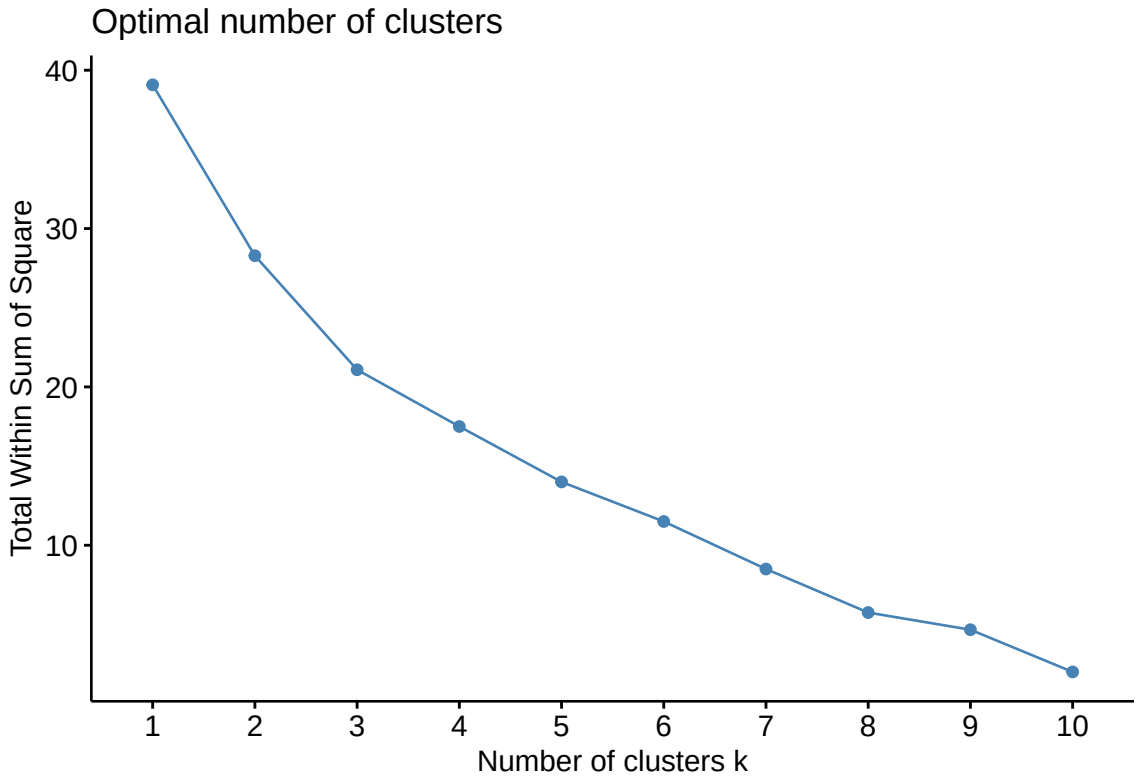
Por último, en respuesta a la pregunta de la empresa, podemos confirmar que **Nissan** pertenece al **clúster 2**, identificado como el de **los coches deportivos**. Dentro de este mismo clúster, **Mazda** y **Alfa Romeo** se destacan como los principales competidores de Nissan.

Las estrategias de marketing de la empresa son adecuadas, ya que Nissan tenía como objetivo posicionar su nuevo modelo como un coche deportivo, y los resultados del análisis confirman que este posicionamiento es coherente con la percepción del mercado.

5.1.3 K-menas

En el caso del análisis **K-means**, el primer paso es elegir el número óptimo de clusters. Para ello, utilizamos la función `fviz_nbclust()` de la librería **factoextra**, que nos ayuda a visualizar y determinar el número adecuado de clusters a utilizar en el análisis.

```
suppressMessages(library(factoextra))
fviz_nbclust(x[,-8], FUNcluster = kmeans, method = "wss",
             diss = dist(x[,-8], method = "euclidean"))
```



A partir de 3 clusters, la reducción en la suma total de cuadrados internos comienza a estabilizarse, lo que indica que $\mathbf{K = 3}$ es el número de grupos más adecuado para este análisis.

```
set.seed(123)
km_clusters <- kmeans(x = x[,-8], centers = 3, nstart = 25)
round(km_clusters$centers,3)
```

```
##   mecanica estabilidad habitabilidad comodidad equipamiento Prestaciones
## 1    3.667       3.167         3.167     3.000         3.00        3.667
## 2    4.000       4.000         2.000     2.667         4.00        4.333
## 3    3.000       3.750         3.500     3.750         2.75        2.750
##   consumo
## 1   3.000
## 2   2.667
## 3   3.250
```

Para interpretar las características de los grupos, es necesario analizar los **promedios** (más o menos elevados) dentro de cada grupo.

- El **primer grupo** presenta características promedio en casi todas las variables, con puntuaciones cercanas a 3.
- El **segundo grupo** se destaca por tener promedios más elevados en **mecánica**, **estabilidad** y **prestaciones**, atributos típicos de los coches deportivos.

- Finalmente, el **tercer grupo** se caracteriza por obtener mejores puntuaciones en atributos como **comodidad**, **habitabilidad** y **estabilidad**, lo que corresponde a las características típicas de los coches familiares.

Además, el paquete **factoextra** permite generar un gráfico de las agrupaciones. Si el número de variables (dimensionalidad) es mayor a 2, el paquete realiza automáticamente un **Análisis de Componentes Principales (PCA)** y representa las dos primeras componentes principales para facilitar la visualización.

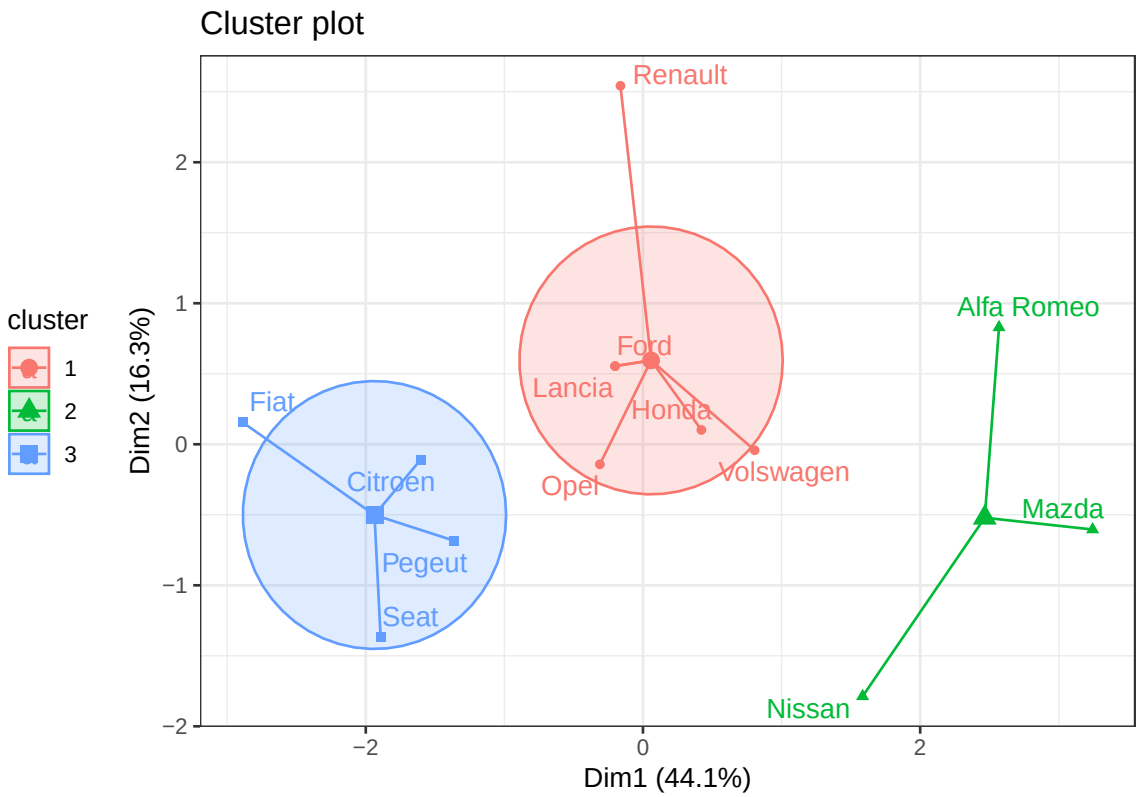
```
set.seed(123)
round(km_clusters$cluster,3)
```

```
##     Lancia    Citroen       Fiat       Ford      Honda Alfa Romeo      Mazda
##          1          3          3          1          1          2          2
##     Nissan       Opel     Pegeut       Seat    Renault  Volswagen
##          2          1          3          3          1          1
```

Para responder a la pregunta de la empresa utilizando el análisis **K-means**, los resultados muestran que **Nissan** pertenece al **clúster 2**, el cual agrupa a los **coches deportivos**. Dentro de este mismo clúster, **Mazda** y **Alfa Romeo** se identifican como los principales competidores de Nissan.

Las estrategias de marketing de la empresa son acertadas, ya que el objetivo era promocionar el nuevo modelo de Nissan como un coche deportivo, y los resultados del análisis confirman que este posicionamiento es coherente con la percepción del mercado.

```
fviz_cluster(object = km_clusters, data = x[,-8], show.clust.cent = TRUE,
             ellipse.type = "euclid", star.plot = TRUE, repel = TRUE) +
theme_bw() + theme(legend.position = "left")
```



5.2. Securitas case

Securitas Direct es una empresa dedicada a la seguridad, especializada en vigilancia, patrullaje móvil y consultoría, con sede en Estocolmo, Suecia. El grupo cuenta con más de 300,000 empleados distribuidos en 53 países.

Con más de 25 años de experiencia, Securitas Direct se originó en Suecia en 1988 como parte del grupo Securitas. Diez años después, la división de alarmas, Securitas Direct, comenzó a operar de forma independiente.

Desde sus inicios, Securitas Direct ha experimentado un crecimiento sostenido, expandiéndose constantemente en Europa. Actualmente, está presente en países como Bélgica, Dinamarca, Finlandia, Italia, Países Bajos, Noruega, Portugal, España, Suecia y Reino Unido. Además, ha ampliado su presencia en América del Sur, con oficinas en Chile, Brasil y Perú.

Líder en Europa, la empresa ha puesto su foco en el mercado estadounidense en los últimos años. Actualmente, está realizando varios estudios de mercado para identificar la mejor estrategia de entrada en este segmento. El equipo de marketing se encuentra trabajando en la identificación de los estados clave donde sería más conveniente establecer nuevas filiales del grupo.

Data set Para alcanzar este objetivo, se dispone de diversas fuentes de información sobre la criminalidad en los cincuenta estados de Estados Unidos. Los datos incluyen el número de arrestos por cada 100,000 habitantes en los siguientes delitos:

- Asalto (*Assault*)

- Asesinato (*Murder*)
- Violación (*Rape*)

Además, se ha registrado el porcentaje de la población que vive en áreas urbanas (*UrbanPop*), lo que permitirá analizar posibles correlaciones entre la densidad urbana y la criminalidad.

Análisis

1. Selección de mercados para estrategias de marketing Indicar a la empresa cuáles serían los grupos de países más adecuados para desarrollar sus estrategias de marketing. Además, especificar si el estado de Florida debería formar parte de este grupo y señalar cuál es el estado que presenta mayor potencial para la empresa.

2. Preguntas clave

1. ¿Es posible aplicar directamente un análisis de clúster a nuestros datos? ¿Qué paso previo sería necesario realizar antes del análisis?
2. Considerando la primera pregunta, ¿tendría sentido aplicar un Análisis de Componentes Principales (ACP)?

5.2.1. Cluster jerárquico

Se debe realizar un análisis de clúster tanto jerárquico como mediante el método de k-means. A continuación, se debe identificar el número óptimo de grupos y ofrecer una interpretación detallada de cada uno. Finalmente, responder a la pregunta: **¿Qué podríamos aconsejar a Securitas?**

```
# leo mis datos
securitas_USA = read.csv(file="securitasUSA.csv", header = TRUE)
x=securitas_USA[,-1]
# uso los nombre de los estados como etiquetas de las lineas de mi data
rownames(x) = t(securitas_USA[,1])
# Realizo un resumen de las variables
summary(x)
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

Para decidir si es necesario aplicar un **Análisis de Componentes Principales (ACP)** previo, es recomendable verificar la presencia de correlaciones elevadas (mayores a 0.5).

```r
# Analizo las correlaciones
round(cor(x),2)
```

```
##          Murder Assault UrbanPop Rape
## Murder     1.00    0.80     0.07 0.56
## Assault    0.80    1.00     0.26 0.67
## UrbanPop   0.07    0.26     1.00 0.41
## Rape       0.56    0.67     0.41 1.00
```

Podemos observar que existen diversas correlaciones significativas entre las variables. Por ejemplo, **Assault** presenta una correlación elevada con **Murder** (0.80) y con **Rape** (0.67). Asimismo, **Murder** y **Rape** también muestran una correlación elevada, aunque negativa (-0.56).

Debido a estas correlaciones, hemos decidido calcular un **Análisis de Componentes Principales (ACP)** y considerar las dos primeras componentes principales para el análisis.

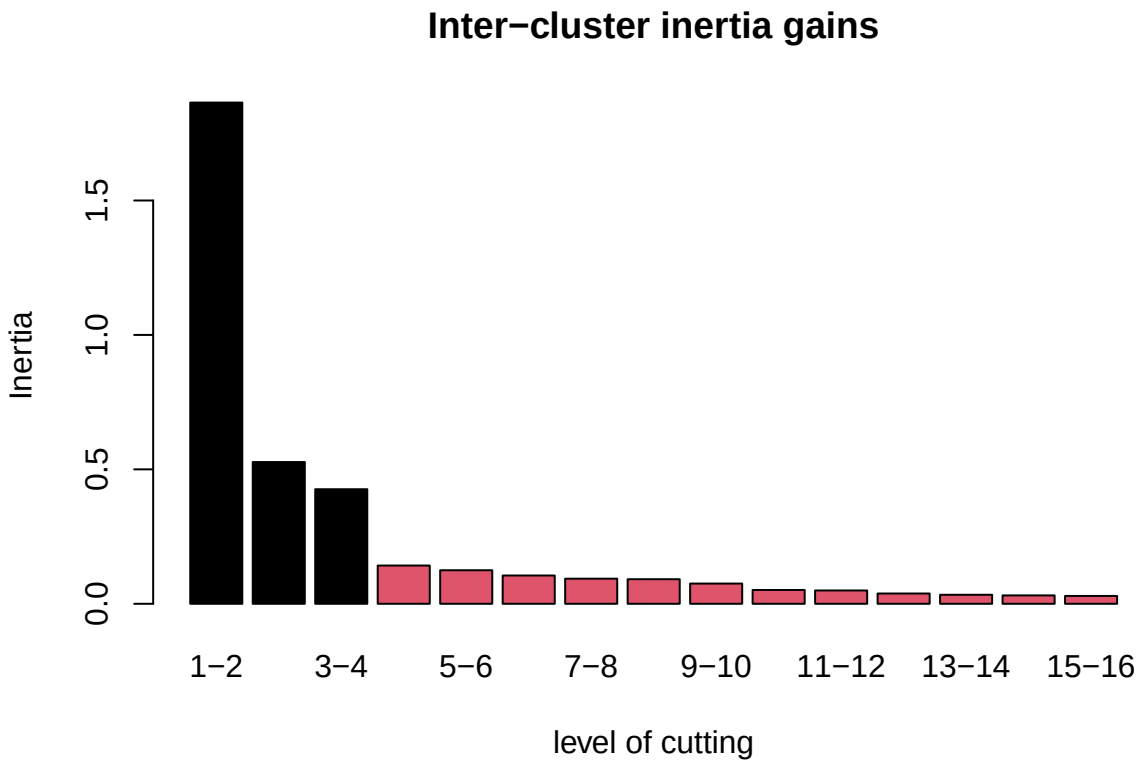
```r
library(FactoMineR)
pca = PCA(x, graph=FALSE)
```

Step 1: Identifico el numero de clusters. Para identificar el número óptimo de clusters, utilizamos el **gráfico de inercia intra-clases**. A medida que aumenta el número de grupos, podemos observar cómo las barras en el gráfico se hacen más pequeñas, lo que indica una disminución gradual de la variabilidad dentro de cada clúster.

Para que el análisis sea interpretable, es necesario seleccionar un número limitado de grupos. La mejor decisión suele corresponder al punto del gráfico donde ocurre el salto más significativo. En este caso, hemos decidido considerar **4 grupos**, basándonos en el primer salto significativo.

Es importante destacar que la decisión también debe tener en cuenta el número de variables y observaciones. En este caso, dado que ambos son limitados, no tendría sentido seleccionar un número elevado de grupos.

```r
hcpc = HCPC(pca,nb.clust=-1,graph=FALSE)
plot(hcpc, choice="bar")
```




```
clust=4 # <-- modifica el valor numero con el numero que clusters deseado
hcpc = HCPC(pca,nb.clust=clust, graph=FALSE)

hcpc$desc.var
```

Step 2 y 3: Realizo en análisis y interpreto los grupos

```
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##                Eta2      P-value
## Assault  0.7841402 2.376392e-15
## Murder   0.7771455 4.927378e-15
## Rape     0.7029807 3.480110e-12
## UrbanPop 0.5846485 7.138448e-09
##
## Description of each cluster by quantitative variables
## ======================================================
## $`1`
##             v.test Mean in category Overall mean sd in category Overall sd
## UrbanPop -3.898420         52.07692       65.540       9.691087  14.329285
## Murder   -4.030171          3.60000        7.788       2.269870   4.311735
## Rape     -4.052061         12.17692       21.232       3.130779   9.272248
```

```
## Assault  -4.638172         78.53846      170.760      24.700095  82.500075
##                p.value
## UrbanPop 9.682222e-05
## Murder   5.573624e-05
## Rape     5.076842e-05
## Assault  3.515038e-06
##
## $`2`
##             v.test Mean in category Overall mean sd in category Overall sd
## UrbanPop  2.793185         73.87500       65.540       8.652131  14.329285
## Murder   -2.374121          5.65625        7.788       1.594902   4.311735
##                p.value
## UrbanPop 0.005219187
## Murder   0.017590794
##
## $`3`
##             v.test Mean in category Overall mean sd in category Overall sd
## Murder    4.357187          13.9375        7.788       2.433587   4.311735
## Assault   2.698255         243.6250      170.760      46.540137  82.500075
## UrbanPop -2.513667          53.7500       65.540       7.529110  14.329285
##                p.value
## Murder   1.317449e-05
## Assault  6.970399e-03
## UrbanPop 1.194833e-02
##
## $`4`
##             v.test Mean in category Overall mean sd in category Overall sd
## Rape     5.352124         33.19231       21.232       6.996643   9.272248
## Assault  4.356682        257.38462      170.760      41.850537  82.500075
## UrbanPop 3.028838         76.00000       65.540      10.347798  14.329285
## Murder   2.913295         10.81538        7.788       2.001863   4.311735
##                p.value
## Rape     8.692769e-08
## Assault  1.320491e-05
## UrbanPop 2.454964e-03
## Murder   3.576369e-03
```

Step 4: visualizo el albor y la tabla con mis datos y el correspondiente clúster para cada observación Primero, podemos observar que las variables que más caracterizan los clusters son, en orden: **Assault**, **Murder**, **Rape**, y **UrbanPop**.

Interpretación de los Clusters

- **Clúster 1**: Se caracteriza por valores negativos del v-test en todas las variables: **UrbanPop** (-3.89), **Murder** (-4.03), **Rape** (-4.05), y **Assault** (-4.63). Estos valores indican que los estados pertenecientes a este clúster son zonas con baja criminalidad y poca densidad poblacional.

Podríamos identificar este grupo como **"Seguros y tranquilos"**. Este patrón se confirma al observar que los promedios intra-grupo son todos más bajos que los promedios generales.

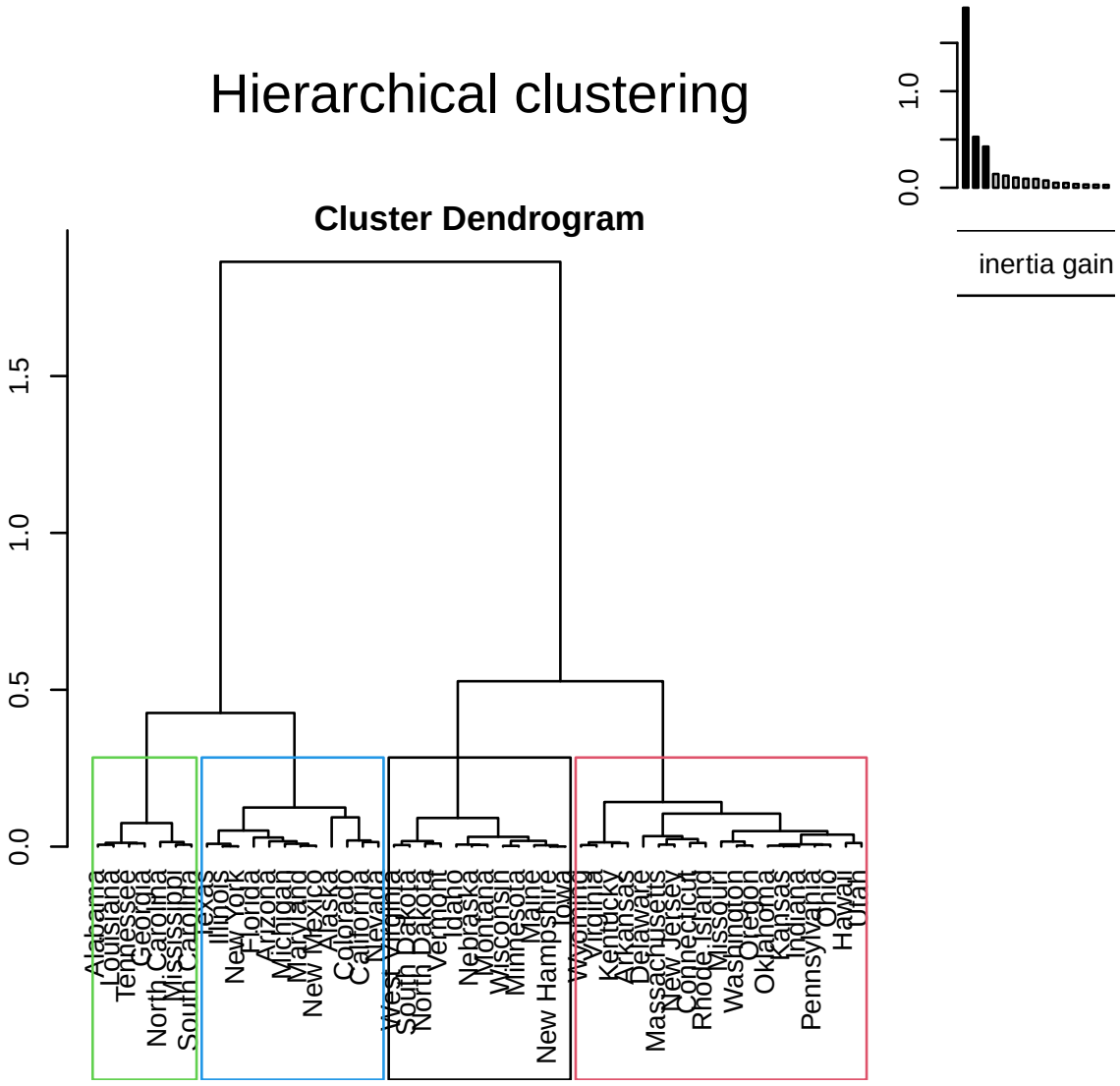
- **Clúster 2**: Este grupo se caracteriza por tener una baja tasa de **Murder** (v-test = 2.37, media = 5.65 frente a 7.78 de la media general) y una alta tasa de **UrbanPop** (v-test = 2.79, media = 73.87 frente a 65.54 de la media general). Podríamos identificar este grupo como **"No asesinados"**.

- **Clúster 3**: Se distingue por tener una tasa alta de **Murder** (v-test = 4.35, media = 13.93 frente a 7.78 de la media general) y de **Assault** (v-test = 2.69, media = 243.62 frente a 170.76 de la media general). Sin embargo, la tasa de **UrbanPop** es más baja que la media (v-test = -2.51, media = 53.75 frente a 65.54 de la media general). Podríamos identificar este grupo como **"Atracos y asesinatos en zonas rurales"**.

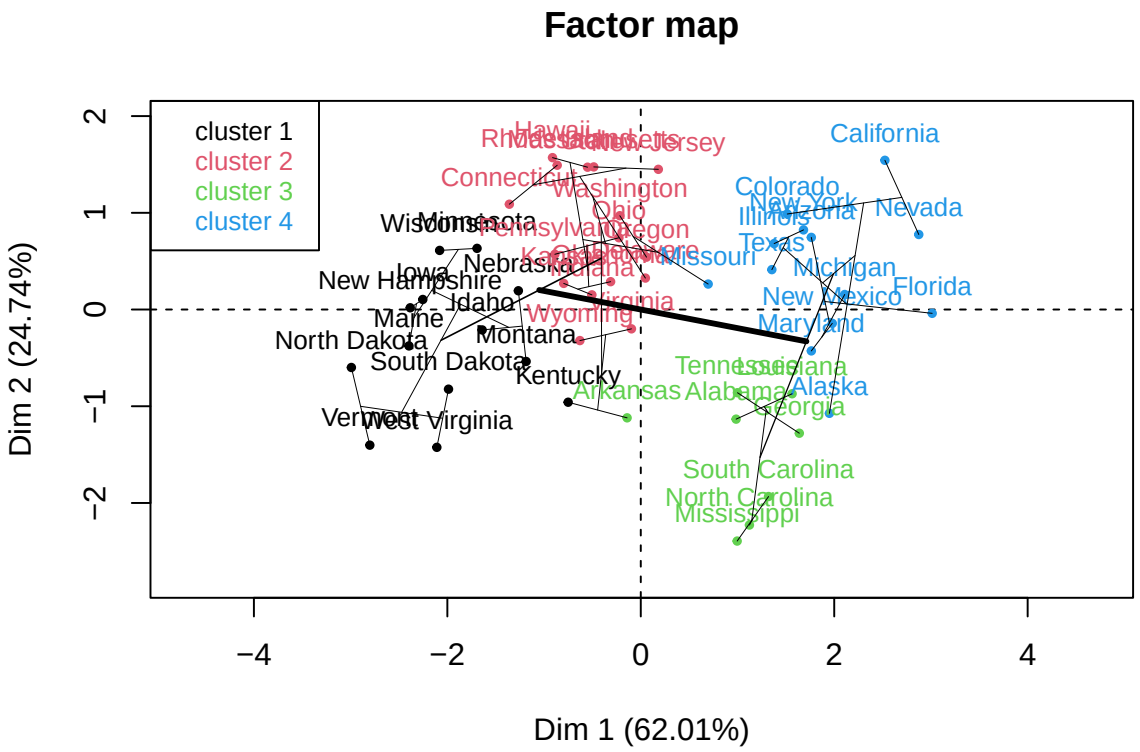
- **Clúster 4**: Se caracteriza por tener una tasa de criminalidad elevada. Todos los estados de este clúster presentan índices superiores a la media: **Rape** (33.19), **Assault** (257.38), y **Murder** (10.81). Además, la tasa de **UrbanPop** es también más alta (v-test = 3.02, media = 76.00 frente a 65.54 de la media general). Este grupo podría identificarse como **"Securitas Core Business"**.

Podemos visualizar la clasificación de los clusters utilizando un dendrograma y/o un scatter plot para entender mejor la distribución de las observaciones.

```
plot(hcpc,choice="tree")
```




```
plot(hcpc,choice="map")
```

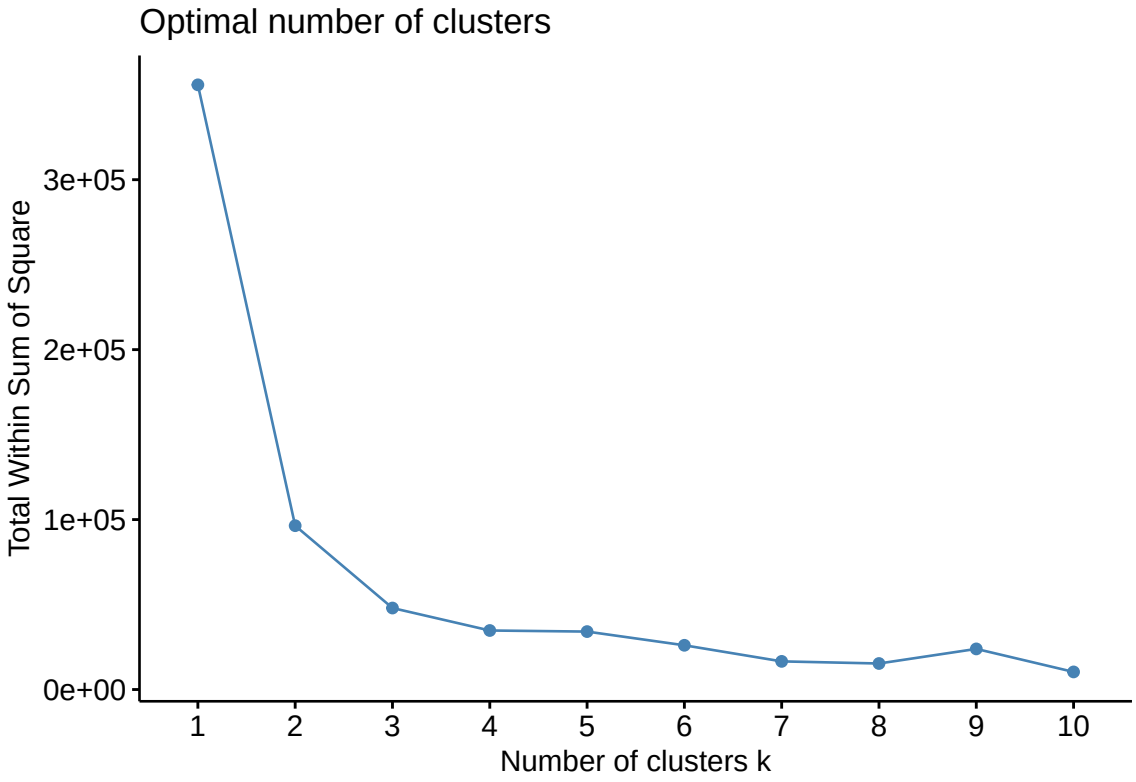


Por último, para responder a la pregunta de la empresa, podemos concluir que **Securitas** debería enfocar su campaña en los estados que pertenecen al **clúster 4**, como **Florida**, **California**, **Nevada**, y **New York**. Estos estados se caracterizan por tener índices elevados de criminalidad, lo que los convierte en mercados clave para los servicios de seguridad que ofrece Securitas.

5.2.3 K-menas

En el caso del análisis **K-means**, lo primero que debemos hacer es elegir el número óptimo de clusters. Esto lo logramos utilizando la función `fviz_nbclust()` de la librería **factoextra**.

```
suppressMessages(library(factoextra))
fviz_nbclust(x, FUNcluster = kmeans, method = "wss",
             diss = dist(x[,-8], method = "euclidean"))
```



A partir de 3 clusters, la reducción en la variabilidad entre países comienza a estabilizarse, lo que indica que $\mathbf{K = 3}$ es una opción adecuada para este análisis.

```
set.seed(124)
km_clusters <- kmeans(x , centers = 3, nstart = 25)
round(km_clusters$centers,3)
```

```
##   Murder Assault UrbanPop   Rape
## 1 11.812 272.562   68.312 28.375
## 2  4.270  87.550   59.750 14.390
## 3  8.214 173.286   70.643 22.843
```

Para interpretar las características de los grupos, observamos los **promedios** dentro de cada grupo.

- El **primer grupo** presenta valores más elevados en todos los índices de criminalidad: **Murder** (11.81), **Assault** (272.56), **Rape** (28.37), así como en la **densidad de población** (**UrbanPop**: 68.31). Este grupo representa el **"Core business"** de Securitas, donde los servicios de seguridad serían más demandados.
- El **segundo grupo** se caracteriza por tener valores por debajo del promedio en todos los índices de criminalidad: **Murder** (4.27), **Assault** (87.55), **Rape** (14.39), y una densidad de población relativamente baja (**UrbanPop**: 59.75). Podemos identificar este grupo como **"Seguros y tranquilos"**.

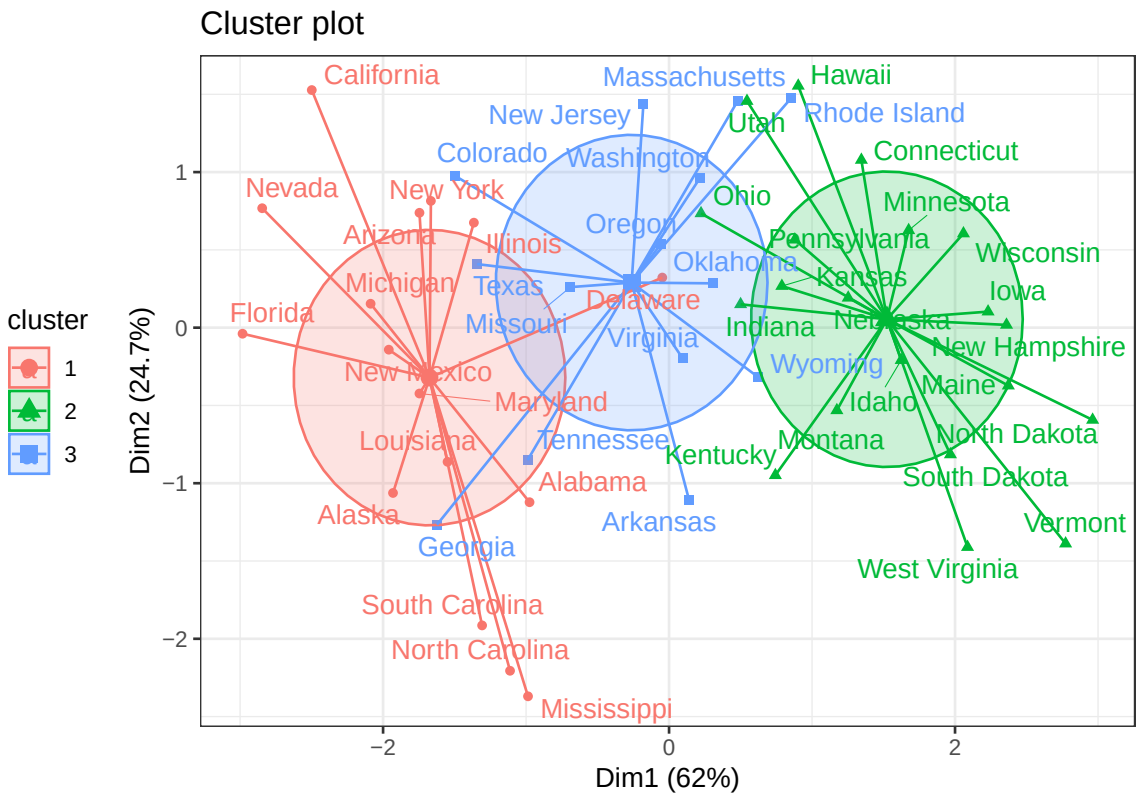
- El **tercer grupo** se distingue por tener un índice elevado de **densidad de población** (**UrbanPop**: 70.64), pero con índices de criminalidad más bajos. Este grupo podría identificarse como **"Poblado"**, ya que, aunque tiene alta densidad poblacional, los niveles de criminalidad son relativamente bajos.

El paquete **factoextra** también ofrece herramientas para obtener **visualizaciones** de las agrupaciones, lo que facilita la interpretación de los resultados.

```
set.seed(123)
round(km_clusters$cluster,3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              3              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              3              2              1              1              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              1              2              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              2              1              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              1              2              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              1              2              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              2              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              2              3              1
##   South Dakota      Tennessee          Texas          Utah        Vermont
##              2              3              3              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              2              2              3
```

```
fviz_cluster(object = km_clusters, data = x[,-8], show.clust.cent = TRUE,
             ellipse.type = "euclid", star.plot = TRUE, repel = TRUE) +
theme_bw() + theme(legend.position = "left")
```



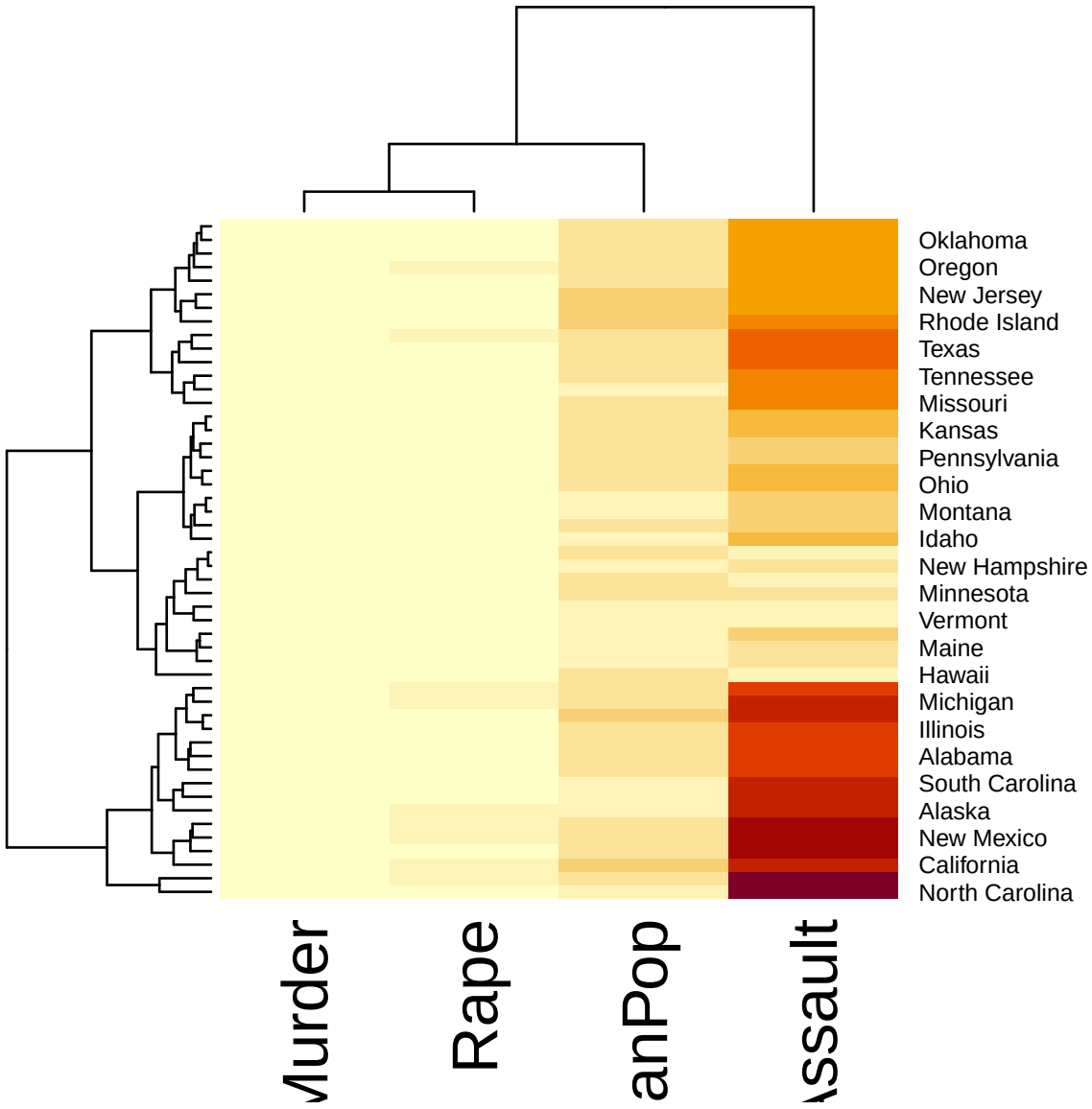
Anexo A. Heatmaps

Los **heatmaps** son una representación visual de una matriz de valores, en la que en lugar de números se utiliza un gradiente de color proporcional al valor de cada variable en cada posición de la matriz.

Combinar un **dendrograma** con un heatmap permite ordenar las filas y/o columnas de la matriz por semejanza, mostrando además el valor de las variables mediante un código de colores. Esto proporciona una representación más rica en información que un dendrograma simple, facilitando la identificación visual de posibles patrones característicos de cada clúster.

En R, existe una amplia variedad de funciones desarrolladas para la creación de heatmaps, entre las que se destaca la función `heatmap()`.

```
heatmap(as.matrix(x), scale = "none",
distfun = function(x){dist(x, method = "euclidean")},
hclustfun = function(x){hclust(x, method = "average")}, cexRow = 0.7)
```



Anexo A. Dimostraciones clúster jerárquico y kmeans

Clúster Jerárquico

En este documento realizaremos un clustering jerárquico aglomerativo utilizando la métrica de promedio (Average Linkage). Se trabajará con dos variables (`X` y `Y`) y cuatro observaciones. A continuación se describen los pasos detallados para llevar a cabo este proceso, incluyendo los cálculos matemáticos para la matriz de distancias, la agrupación de las observaciones y el cálculo del nuevo centroide de los clústeres formados.

Paso 1: Definir las observaciones Las cuatro observaciones son las siguientes:

$\text{Obs} = \{(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4)\}$

Los valores de las observaciones son:

Observación	X	Y
Obs1	2	3
Obs2	3	4
Obs3	6	7
Obs4	8	9

```r
# Definir los datos
data <- data.frame(
  obs = c("Obs1", "Obs2", "Obs3", "Obs4"),
  X = c(2, 3, 6, 8),
  Y = c(3, 4, 7, 9)
)
data
```

```
##    obs X Y
## 1 Obs1 2 3
## 2 Obs2 3 4
## 3 Obs3 6 7
## 4 Obs4 8 9
```

Paso 2: Cálculo de la matriz de distancias La distancia euclidiana entre dos puntos $A = (x_1, y_1)$ y $B = (x_2, y_2)$ se define como:

$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Ahora calculamos las distancias entre todas las parejas de observaciones:

1. **Distancia entre Obs1 y Obs2**:

$$d(Obs1, Obs2) = \sqrt{(3-2)^2 + (4-3)^2} = \sqrt{1^2 + 1^2} = \sqrt{1+1} = \sqrt{2} \approx 1.414$$

2. **Distancia entre Obs1 y Obs3**:

$$d(Obs1, Obs3) = \sqrt{(6-2)^2 + (7-3)^2} = \sqrt{4^2 + 4^2} = \sqrt{16+16} = \sqrt{32} \approx 5.657$$

3. **Distancia entre Obs1 y Obs4**:

$$d(Obs1, Obs4) = \sqrt{(8-2)^2 + (9-3)^2} = \sqrt{6^2 + 6^2} = \sqrt{36+36} = \sqrt{72} \approx 8.485$$

4. **Distancia entre Obs2 y Obs3**:

$$d(Obs2, Obs3) = \sqrt{(6-3)^2 + (7-4)^2} = \sqrt{3^2 + 3^2} = \sqrt{9+9} = \sqrt{18} \approx 4.243$$

5. **Distancia entre Obs2 y Obs4**:

$$d(Obs2, Obs4) = \sqrt{(8-3)^2 + (9-4)^2} = \sqrt{5^2 + 5^2} = \sqrt{25+25} = \sqrt{50} \approx 7.071$$

6. **Distancia entre Obs3 y Obs4**:

$$d(Obs3, Obs4) = \sqrt{(8-6)^2 + (9-7)^2} = \sqrt{2^2 + 2^2} = \sqrt{4+4} = \sqrt{8} \approx 2.828$$

Matriz de distancias La matriz de distancias resultante es la siguiente:

```r
# Crear la matriz de distancias manualmente
dist_matrix <- matrix(c(
  0, 1.414, 5.657, 8.485,
  1.414, 0, 4.243, 7.071,
  5.657, 4.243, 0, 2.828,
  8.485, 7.071, 2.828, 0
), nrow = 4, byrow = TRUE)
colnames(dist_matrix) <- rownames(dist_matrix) <- data$obs
dist_matrix
```

```
##       Obs1  Obs2  Obs3  Obs4
## Obs1 0.000 1.414 5.657 8.485
## Obs2 1.414 0.000 4.243 7.071
## Obs3 5.657 4.243 0.000 2.828
## Obs4 8.485 7.071 2.828 0.000
```

Paso 3: Agrupar las observaciones más cercanas Las observaciones más cercanas son `Obs1` y `Obs2`, con una distancia de 1.414. Formamos un nuevo clúster $C_1 = \{Obs1, Obs2\}$.

Paso 4: Cálculo del nuevo centroide El centroide del nuevo clúster C_1 se calcula promediando las coordenadas X y Y de las observaciones que lo componen:

$X_{C_1} = \frac{X_1+X_2}{2} = \frac{2+3}{2} = 2.5$ $Y_{C_1} = \frac{Y_1+Y_2}{2} = \frac{3+4}{2} = 3.5$

Por lo tanto, el nuevo centroide es $C_1 = (2.5, 3.5)$.

Paso 5: Actualizar la matriz de distancias Ahora recalculamos las distancias entre el nuevo clúster C_1 y las demás observaciones utilizando la métrica del promedio (Average Linkage). La distancia entre un clúster y una observación se calcula promediando las distancias individuales entre cada observación del clúster y la observación externa.

1. **Distancia entre C_1 y Obs3**:

$$d(C_1, Obs3) = \frac{d(Obs1,Obs3)+d(Obs2,Obs3)}{2} = \frac{5.657+4.243}{2} = 4.95$$

2. **Distancia entre C_1 y Obs4**:

$$d(C_1, Obs4) = \frac{d(Obs1,Obs4)+d(Obs2,Obs4)}{2} = \frac{8.485+7.071}{2} = 7.778$$

Nueva matriz de distancias La nueva matriz de distancias es:

```r
# Nueva matriz de distancias después de la primera unión
new_dist_matrix <- matrix(c(
  0, 4.95, 7.778,
  4.95, 0, 2.828,
  7.778, 2.828, 0
), nrow = 3, byrow = TRUE)
colnames(new_dist_matrix) <- rownames(new_dist_matrix) <- c("C1", "Obs3", "Obs4")
new_dist_matrix
```

```
##         C1  Obs3  Obs4
## C1   0.000 4.950 7.778
## Obs3 4.950 0.000 2.828
## Obs4 7.778 2.828 0.000
```

Paso 6: Formar nuevos clústeres La próxima pareja más cercana es `Obs3` y `Obs4`, con una distancia de 2.828. Agrupamos estas observaciones para formar el siguiente clúster $C_2 = \{Obs3, Obs4\}$. Calculamos el nuevo centroide y repetimos el proceso hasta formar un solo clúster.

K-means

En este documento realizaremos un clustering K-means utilizando dos variables (`X` y `Y`) y cuatro observaciones. A continuación se describen los pasos detallados para llevar a cabo este proceso, incluyendo los cálculos matemáticos para la inicialización de los centroides, la asignación de puntos a los clústeres y la actualización de los centroides hasta la convergencia.

Paso 1: Definir las observaciones Las cuatro observaciones son las siguientes:

$\text{Obs} = \{(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), (X_4, Y_4)\}$

Los valores de las observaciones son:

Observación	X	Y
Obs1	2	3
Obs2	3	4
Obs3	6	7
Obs4	8	9

```r
# Definir los datos
data <- data.frame(
  obs = c("Obs1", "Obs2", "Obs3", "Obs4"),
  X = c(2, 3, 6, 8),
  Y = c(3, 4, 7, 9)
)
data
```

```
##    obs X Y
## 1 Obs1 2 3
## 2 Obs2 3 4
## 3 Obs3 6 7
## 4 Obs4 8 9
```

Paso 2: Inicialización de los centroides Para el algoritmo K-means, primero debemos inicializar dos centroides aleatoriamente (en este caso para $k = 2$). Supongamos que los centroides iniciales son:

- Centroide 1: $C_1 = (2, 3)$
- Centroide 2: $C_2 = (6, 7)$

```
# Inicializar los centroides manualmente
centroids <- data.frame(
  X = c(2, 6),
  Y = c(3, 7)
)
rownames(centroids) <- c("C1", "C2")
centroids
```

```
##    X Y
## C1 2 3
## C2 6 7
```

Paso 3: Asignación de observaciones a los centroides Ahora, calculamos las distancias entre cada observación y cada centroide utilizando la distancia euclidiana, y asignamos cada observación al centroide más cercano.

1. **Distancia entre Obs1 y los centroides**:

$d(Obs1, C_1) = \sqrt{(2-2)^2 + (3-3)^2} = 0$

$d(Obs1, C_2) = \sqrt{(6-2)^2 + (7-3)^2} = \sqrt{16+16} = \sqrt{32} \approx 5.657$

Obs1 se asigna a C_1 porque está más cerca.

2. **Distancia entre Obs2 y los centroides**:

$d(Obs2, C_1) = \sqrt{(2-3)^2 + (3-4)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.414$

$d(Obs2, C_2) = \sqrt{(6-3)^2 + (7-4)^2} = \sqrt{9+9} = \sqrt{18} \approx 4.243$

Obs2 se asigna a C_1.

3. **Distancia entre Obs3 y los centroides**:

$d(Obs3, C_1) = \sqrt{(6-2)^2 + (7-3)^2} = \sqrt{16+16} = \sqrt{32} \approx 5.657$

$d(Obs3, C_2) = \sqrt{(6-6)^2 + (7-7)^2} = 0$

Obs3 se asigna a C_2.

4. **Distancia entre Obs4 y los centroides**:

$d(Obs4, C_1) = \sqrt{(8-2)^2 + (9-3)^2} = \sqrt{36+36} = \sqrt{72} \approx 8.485$

$d(Obs4, C_2) = \sqrt{(8-6)^2 + (9-7)^2} = \sqrt{4+4} = \sqrt{8} \approx 2.828$

Obs4 se asigna a C_2.

```r
# Asignación inicial de los puntos a los centroides
data$cluster <- c("C1", "C1", "C2", "C2")
data
```

Asignación inicial

```
##    obs X Y cluster
## 1 Obs1 2 3      C1
## 2 Obs2 3 4      C1
## 3 Obs3 6 7      C2
## 4 Obs4 8 9      C2
```

Paso 4: Actualización de los centroides Ahora calculamos los nuevos centroides promediando las coordenadas de las observaciones asignadas a cada clúster:

1. **Nuevo centroide** C_1 (Obs1 y Obs2):

$X_{C_1} = \frac{2+3}{2} = 2.5$

$Y_{C_1} = \frac{3+4}{2} = 3.5$

2. **Nuevo centroide** C_2 (Obs3 y Obs4):

$X_{C_2} = \frac{6+8}{2} = 7$

$Y_{C_2} = \frac{7+9}{2} = 8$

```r
# Actualización de los centroides
new_centroids <- data.frame(
  X = c(2.5, 7),
  Y = c(3.5, 8)
)
rownames(new_centroids) <- c("C1", "C2")
new_centroids
```

```
##      X   Y
## C1 2.5 3.5
## C2 7.0 8.0
```

Paso 5: Repetir hasta la convergencia El proceso se repite recalculando las distancias entre las observaciones y los nuevos centroides, reasignando las observaciones y actualizando los centroides hasta que no haya cambios en las asignaciones.

Componentes principales

1. Introducción

El **Análisis en Componentes Principales (ACP)**, es un método estadístico pensado para reducir el numero de variables, pero conservando al mismo tiempo la mayor parte de la información contenida en ellas.

Supongamos que tenemos una muestra con n individuos, cada uno de ellos medido en p variables (X_1, X_2, … , X_p), es decir, un espacio de p dimensiones. El ACP permite reducir este espacio encontrando un conjunto menor de factores subyacentes (z componentes principales, donde $z < p$) que explican de manera aproximada la misma información que las p variables originales. Esto significa que, en lugar de utilizar p valores para describir cada individuo, basta con usar z valores, lo que facilita el análisis y la interpretación de los datos. Cada uno de estos z valores se denomina **componente principal**.

El ACP pertenece a la familia de técnicas conocidas como **técnicas de aprendizaje no supervisado**. A diferencia de las **técnicas de aprendizaje supervisado**, cuyo objetivo es predecir una variable respuesta **Y** a partir de un conjunto de predictores (por ejemplo, regresión lineal), en el aprendizaje no supervisado no se tiene en cuenta una variable respuesta. En su lugar, se busca extraer patrones o estructuras latentes en los datos utilizando únicamente los predictores, por ejemplo, para identificar subgrupos o reducir dimensionalidad. Una de las principales dificultades de las técnicas no supervisadas es la validación de los resultados, ya que no existe una variable de respuesta que sirva como referencia para contrastar la calidad de las predicciones.

El ACP es una herramienta útil para **condensar** la información contenida en múltiples variables en un número reducido de componentes principales, lo que lo hace especialmente valioso como paso previo a la aplicación de otras técnicas estadísticas, como la **regresión** o el **clustering**.

2. Presentación intuitiva

Imaginemos que estamos participando en un concurso donde hay distintos equipos, y cada uno debe elegir un representante. El representante tiene la tarea de observar un objeto durante unos pocos segundos, recopilar información sobre él y luego transmitir a su equipo, de la manera más clara posible, de qué objeto se trata. Supongamos que el objeto es una tetera.

Dado que el tiempo es limitado, el representante decide tomar una fotografía del objeto para captar la mayor cantidad de información posible en un solo vistazo. Para que los otros participantes entiendan claramente que se trata de una tetera, el representante debe hacer **la mejor fotografía posible**, es decir, una imagen que transmita la mayor cantidad de información relevante sobre el objeto.

Para lograr esto, el representante busca la mejor posición de la tetera, rotándola hasta encontrar el ángulo óptimo que capture mejor su esencia.

De manera análoga, el **Análisis en Componentes Principales (ACP)** busca crear la mejor "fotografía" de nuestros datos: una **representación simplificada** que capte la mayor cantidad de información posible, pero en menos dimensiones.

En el contexto del marketing, los datos que analizaremos podrían ser indicadores económicos, preferencias de los consumidores o las percepciones relativas a un producto o empresa. El objetivo es identificar patrones subyacentes que permitan reducir la complejidad de la información y facilitar la toma de decisiones estratégicas.

3. Aplicaciones en Marketing

El **Análisis en Componentes Principales (ACP)** tiene múltiples aplicaciones en el campo del marketing. Algunas de las principales son:

1. Estudio del posicionamiento de una marca respecto a otras.
2. Comparación entre distintas marcas que operan en un mercado.
3. Identificación de segmentos de mercado potenciales.
4. Detección de posibles competidores.

El ACP ofrece diversas ventajas, entre ellas:

1. Generación de **mapas de posicionamiento**, que permiten visualizar cómo se sitúan las marcas o productos en relación con sus competidores.
2. **Reducción de la dimensionalidad**, lo que facilita trabajar con un número menor de variables no correlacionadas que pueden ser utilizadas en técnicas posteriores, como la **regresión lineal**.
3. Transformación de las variables antes de aplicar técnicas de segmentación, como el algoritmo de partición **k-means**, para identificar de manera más eficiente los diferentes grupos de consumidores.

Estas aplicaciones permiten optimizar el análisis de datos complejos en el ámbito del marketing, facilitando la toma de decisiones estratégicas.

4. El método

El **Análisis en Componentes Principales (ACP)** es una técnica que transforma un conjunto de variables cuantitativas en un nuevo conjunto reducido de dimensiones, denominadas **componentes principales**, con el objetivo de preservar la mayor cantidad de información posible.

El ACP genera estas nuevas dimensiones (componentes principales o variables sintéticas) a partir de las variables originales, permitiendo representar visualmente la información contenida en el conjunto de datos mediante un **gráfico de dispersión** (*scatter-plot*). Para lograr este objetivo, una vez definidas las componentes principales, los datos se proyectan sobre ellas, generando un gráfico donde los ejes corresponden a las componentes y los puntos representan nuestras observaciones. Las componentes principales se calculan como **combinaciones lineales** de las variables originales, garantizando que:

1. Se maximice la varianza contenida en las variables originales, es decir, que se capture la mayor cantidad de información posible.

2. Las componentes sean mutuamente incorrelacionadas, evitando redundancias.

La interpretación de los resultados del ACP es sencilla y se basa en los siguientes puntos clave:

1. **Distancia entre observaciones**: Las observaciones cercanas en el gráfico tienen características similares, mientras que las observaciones alejadas son más disímiles.
2. **Proximidad a los ejes**: Cuanto más cerca esté una observación de un eje, mayor será la influencia de ese componente principal en la observación. Si una observación está lejos del eje, captará menos información de ese componente.
3. **Distancia del origen**: Observaciones cercanas al origen de los ejes representan características promedio, mientras que aquellas más alejadas del origen reflejan características más extremas o diferenciadas.

Este enfoque proporciona una forma visual y cuantitativa de entender las relaciones y diferencias entre las observaciones en un espacio de menor dimensión.

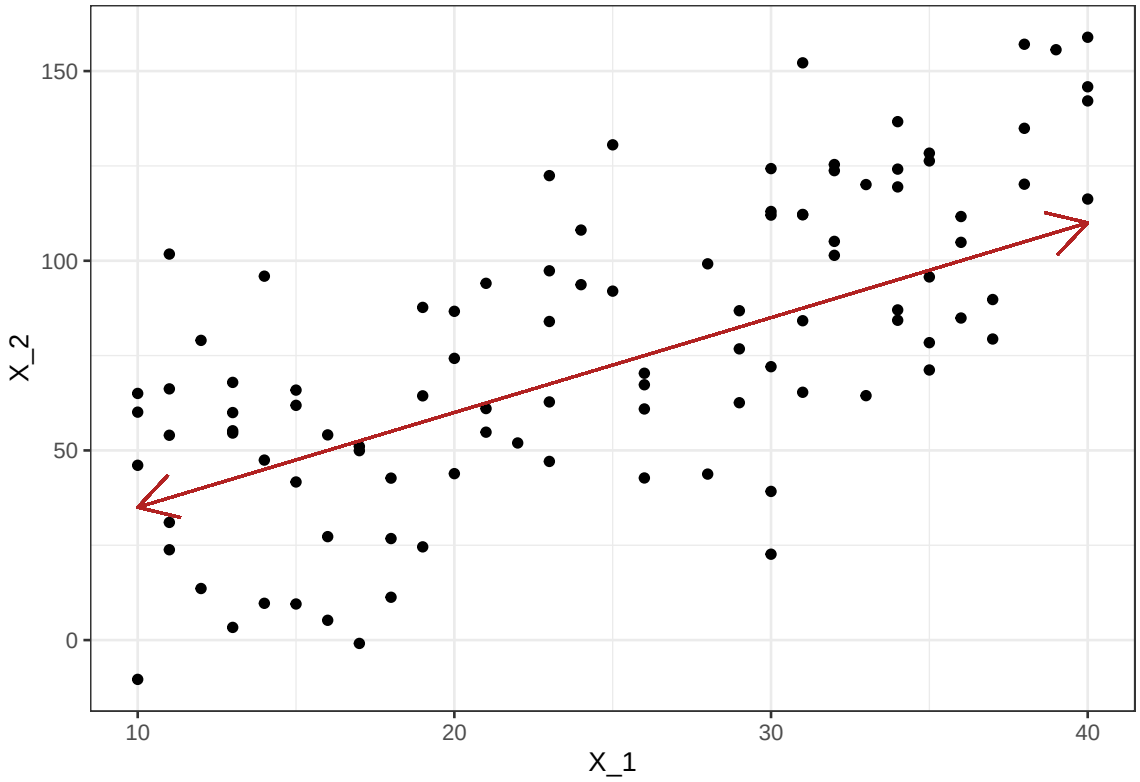
4.1. Calculo de las componentes: interpretación geométrica

Una forma intuitiva de comprender el **Análisis en Componentes Principales (ACP)** es interpretar las componentes principales desde un punto de vista geométrico. Supongamos que tenemos un conjunto de observaciones para las cuales disponemos de dos variables (X_1 y X_2).

La **primera componente principal** (Z_1) se define como el vector que sigue la dirección en la cual las observaciones presentan la mayor variabilidad (representado por una línea roja en un gráfico de dispersión). Esta dirección maximiza la varianza de los datos proyectados, capturando así la mayor cantidad de información posible en una sola dimensión.

La **proyección** de cada observación sobre esta dirección corresponde al valor de la primera componente principal para esa observación, también conocido como **principal component score**.

Esta interpretación geométrica permite visualizar cómo el ACP reduce la dimensionalidad del conjunto de datos, enfocándose en las direcciones que contienen más información.



La **segunda componente principal** (Z_2) sigue la dirección en la cual los datos muestran la mayor varianza posible, bajo la restricción de que esta dirección sea **ortogonal** (perpendicular) a la primera componente (Z_1).

La condición de ortogonalidad garantiza que las componentes principales no estén correlacionadas entre sí. Es decir, las componentes capturan diferentes aspectos de la variabilidad en los datos sin redundancia, ya que sus direcciones forman un ángulo recto entre sí en el espacio geométrico.

Esta propiedad es fundamental para el **Análisis en Componentes Principales (ACP)**, ya que permite descomponer la variabilidad de los datos en direcciones independientes, facilitando la interpretación y reducción de la dimensionalidad.

4.2. Cálculo de los componentes: método matemático

Cada **componente principal** (Z_i) se obtiene mediante una **combinación lineal** de las variables originales. Estas nuevas componentes pueden interpretarse como variables derivadas, obtenidas al combinar las variables originales de una manera específica.

La **primera componente principal** de un grupo de variables $(X_1, X_2, ..., X_p)$ es la combinación lineal normalizada de dichas variables que captura la mayor varianza, es decir, la que contiene la mayor cantidad de información posible. Esta se expresa como:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

El hecho de que la combinación lineal esté **normalizada** implica que:

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1$$

Los coeficientes $\phi_{11}, \ldots, \phi_{p1}$ se conocen como **cargas** o *loadings* en inglés. Estas cargas nos ayudan a interpretar cada componente principal, ya que indican la importancia o el peso que tiene cada variable original en la componente. Por ejemplo, ϕ_{11} es el *loading* de la variable X_1 en la primera componente principal. Los *loadings* permiten entender qué tipo de información recoge cada componente, mostrando cuáles variables originales tienen mayor influencia.

Dado un conjunto de datos $\mathbf{X}$ con n observaciones y p variables, el proceso para calcular la primera componente principal incluye los siguientes pasos:

- **Centralización de las variables**: A cada valor se le resta la media de su respectiva variable, de modo que todas las variables tengan media cero.
- **Maximización de la varianza**: Se resuelve un problema de optimización para encontrar los valores de las cargas (*loadings*) que maximicen la varianza en los datos proyectados sobre la primera componente.

Una vez calculada la primera componente principal (Z_1), se procede a calcular la segunda (Z_2), repitiendo el mismo proceso, pero añadiendo la restricción de que la nueva componente no puede estar correlacionada con la primera. Esto implica que Z_1 y Z_2 deben ser **ortogonales** (independientes). Este proceso se repite de forma iterativa hasta calcular todas las componentes posibles (hasta un máximo de $\min(n-1, p)$) o hasta el punto donde se decida detener el análisis.

Las componentes principales se obtienen a través de un método matemático llamado **descomposición en valores singulares (DVS)**. Cada componente está asociada a un valor llamado **valor propio**, que indica la cantidad de información (o varianza) explicada por esa componente. El **orden de importancia** de las componentes viene determinado por la magnitud de sus valores propios.

4.3. Reproducibilidad de los componentes

El proceso de estimación de las **componentes principales** genera siempre los mismos resultados, independientemente del software utilizado. Esto significa que los valores de los *loadings* obtenidos serán los mismos en cualquier programa estadístico.

La única diferencia que puede ocurrir es que el **signo** de todos los *loadings* esté invertido. Esto sucede porque el vector de *loadings* determina la **dirección** de la componente principal, y dicha dirección es la misma independientemente del signo. En otras palabras, la componente principal define una línea en el espacio de las variables que se extiende en ambas direcciones, por lo que cambiar el signo no afecta su interpretación.

De manera similar, los **valores de las componentes principales** para cada observación (*principal component scores*) también serán consistentes entre diferentes programas, salvo por la posible inversión del signo.

4.4. Proporción de varianza explicada

Una de las preguntas más frecuentes que surge tras estimar las componentes principales es: **¿cuánta información es capaz de capturar cada una de las componentes principales obtenidas?**

Para responder a esta pregunta, se recurre a la **proporción de varianza explicada** por cada componente principal. Esta métrica indica qué porcentaje de la variabilidad total de los datos originales es capturado por una componente específica.

Tanto la **proporción de varianza explicada** como la **proporción de varianza explicada acumulada** son herramientas clave para decidir cuántas componentes principales deben ser utilizadas en los análisis posteriores.

Claramente, si se calculan todas las componentes principales de una base de datos, se conserva, aunque transformada, toda la información presente en los datos originales. En este caso, la **proporción de varianza explicada acumulada** de todas las componentes sumará siempre 1 (o el 100%), ya que se ha capturado la totalidad de la varianza original.

Estas métricas permiten determinar de manera eficiente el número de componentes a elegir sin perder una cantidad significativa de información.

4.5. Número óptimo de componentes principales

Dado que el objetivo del **Análisis en Componentes Principales (PCA)** es reducir la dimensionalidad, es de interés utilizar el número mínimo de componentes que sea suficiente para explicar los datos de manera adecuada. Sin embargo, no existe un único método o criterio que determine el número óptimo de componentes principales a utilizar. Algunos de los enfoques más utilizados son los siguientes:

- **Criterio del eigenvalue mayor que 1**: Seleccionar aquellas componentes cuyos valores propios (*eigenvalues*) sean mayores que 1, ya que estas componentes explican más varianza que una variable original individual.
- **Proporción de varianza explicada**: Elegir suficientes componentes para alcanzar un umbral deseado de varianza explicada, típicamente el **75%** de la variabilidad total.
- **Scree plot**: Utilizar un gráfico de lineas (*scree plot*) que representa los valores propios de cada componente. Se selecciona el número de componentes correspondientes al punto en el que se observa un "codo" en la gráfica, es decir, donde la pendiente se suaviza y los valores propios disminuyen significativamente.

Estos métodos son los más comunes para determinar cuántas componentes principales son necesarias para mantener la mayor cantidad de información relevante en los datos.

4.6. Interpretación de las Componentes

Una de las características más importantes del **Análisis en Componentes Principales (ACP)** es la posibilidad de interpretar las componentes que sintetizan la información contenida en las variables originales. Sin embargo, esta interpretación no es automática y puede resultar compleja. Afortunadamente, existen algunas herramientas que ayudan al analista en esta tarea.

4.6.1. Círculo de Correlaciones

El **círculo de correlaciones** es un gráfico que representa tanto las variables originales como las componentes principales. Los ejes del gráfico corresponden a las componentes, mientras que las variables originales se representan mediante flechas.

- La **longitud de las flechas** indica qué tan bien está representada una variable por las componentes principales. Cuanto más larga sea la flecha, mejor representa la componente a esa variable.
- La **distancia entre las flechas y los ejes de las componentes** se interpreta como la relevancia de la variable en la definición de la componente. Cuanto más cerca esté una flecha de un componente, mayor será su contribución a la construcción de dicha componente.

En resumen, para interpretar las componentes, se debe prestar atención a las flechas más largas y cercanas a los ejes de las componentes.

4.6.2. Correlación entre Componentes y Variables Originales

Las **correlaciones elevadas** entre las variables originales y una componente específica indican que estas variables son importantes en la definición de dicha componente. Las variables altamente correlacionadas con una componente principal contribuyen significativamente a su interpretación y nos permiten comprender mejor la naturaleza de la componente.

4.7. Representación gráfica de los individuos e interpretación

Una vez interpretadas las componentes principales, podemos proceder a interpretar el **gráfico de las observaciones**. Esta interpretación se basa en los siguientes aspectos:

- **Distancia entre puntos**: Las observaciones representadas por puntos cercanos indican que comparten características similares, mientras que los puntos distantes representan observaciones con características distintas.
- **Distancia desde el centro de los ejes**: Las observaciones que se encuentran cerca del origen de los ejes representan características promedio en relación con el significado de las componentes principales. Cuanto más alejadas estén del origen, más extremas son sus características.
- **Proximidad a las componentes**: Las observaciones cercanas a los ejes de las componentes principales están mejor caracterizadas por dichas componentes, mientras que las observaciones más alejadas están menos representadas por esas componentes.

Esta interpretación gráfica nos permite entender cómo se relacionan las observaciones entre sí y con las componentes principales, proporcionando una visión más clara de la estructura subyacente en los datos.

4.8 Índices de calidad de la representación

La interpretación de las variables y observaciones en el Análisis en Componentes Principales (ACP) puede complementarse con dos medidas de calidad: la **contribución de las variables a los componentes** (*Contribución*) y la **representación de las variables por los componentes** (cos^2). Estas medidas proporcionan información adicional sobre la relevancia y la precisión de las representaciones obtenidas.

- **Contribución**: Esta métrica indica qué tan importantes son las variables o las observaciones para la construcción de las componentes principales. Una mayor contribución sugiere que la variable o la observación tiene un peso significativo en la definición de una componente, lo que facilita su interpretación.
- cos^2: Esta medida refleja qué tan bien están representadas las variables o las observaciones por las componentes principales. Un valor alto de cos^2 indica que la variable o la observación está bien representada por el componente, mientras que un valor bajo sugiere que su representación es menos precisa.

Estas medidas son fundamentales para evaluar la calidad de las proyecciones y asegurarse de que las variables y observaciones clave sean bien representadas en el análisis.

4.9. La importancia de la matriz de correlaciones.

El análisis de la **matriz de correlaciones** es un paso crucial antes de realizar un **Análisis en Componentes Principales (ACP)**, ya que permite verificar la condición necesaria y suficiente para aplicar el método. Para que el ACP sea válido, las variables deben mostrar patrones de correlación entre grupos, es decir, deben estar correlacionadas. Si no existe correlación entre las variables, el ACP no es aplicable.

Dos casos extremos ilustran este concepto:

1. **Correlaciones iguales a 1**: Si todas las correlaciones entre las variables son iguales a 1, no tiene sentido aplicar el ACP. Esto indica que cada variable mide exactamente lo mismo y proporciona la misma información. En este caso, bastaría con seleccionar una sola variable como representante del conjunto, ya que no se gana nada con la reducción dimensional.
2. **Correlaciones iguales a 0**: Si todas las correlaciones son iguales a 0, significa que las variables son completamente independientes y no comparten información. En este escenario, tampoco tiene sentido aplicar el ACP, ya que cada variable mide un concepto diferente. Para representar adecuadamente los datos, sería necesario conservar todas las variables originales.

Por lo tanto, la matriz de correlaciones nos ayuda a identificar si el ACP es un método adecuado para reducir la dimensionalidad y obtener representaciones útiles de los datos.

5. Como realizar un ACP: pasos a seguir

El punto de partida de un **Análisis en Componentes Principales (ACP)** es la matriz de datos que estamos analizando. Es importante recordar que las variables deben ser cuantitativas para aplicar esta metodología.

Pasos a seguir:

1. **Analizar la estructura de correlaciones**: Observar la matriz de correlaciones entre las variables originales para verificar si el ACP es adecuado.
2. **Aplicar el método**: Ejecutar el ACP para transformar los datos y generar las componentes principales.
3. **Seleccionar el número de componentes**: Determinar cuántas componentes principales utilizar, basándose en criterios como la varianza explicada o el scree plot.
4. **Interpretar las componentes**: Utilizar el **círculo de correlaciones** para interpretar las componentes principales e intentar identificar un nombre para cada una que resuma sus características.
5. **Interpretar el gráfico de las observaciones**: Analizar el gráfico de los individuos para entender cómo se relacionan las observaciones en el espacio de las componentes.
6. **Extraer conclusiones**: Sacar conclusiones basadas en la representación gráfica y en la interpretación de las componentes y las observaciones.

Consideraciones:

- Los pasos 1, 2, 3 y 4 son esenciales si el objetivo es utilizar el ACP para **reducir el número de variables originales** y emplear las componentes obtenidas en otros análisis, como **regresión lineal** o **clustering**.
- Los pasos 5 y 6 son más relevantes cuando se busca crear un **mapa de posicionamiento**, por ejemplo, para entender la relación entre individuos o objetos en un contexto de marketing.

Este proceso guía al analista desde la verificación inicial de las correlaciones hasta la extracción de conclusiones significativas basadas en las representaciones gráficas del ACP.

6. El mapa perceptual (*biplot*)

El **mapa perceptual de posicionamiento** es una técnica de investigación útil para comprender la posición que ocupa la marca en la mente del consumidor, identificar los ventajaso desventajas que ofrece cada una de las marcas y cómo se diferencian entre sí, con el objetivo de diseñar estrategias de marketing más efectivas,

Ventajas del Mapa Perceptual:

1. **Conocer a la competencia**: Permite identificar la posición de las marcas competidoras con respecto a los ideales de los consumidores, lo que facilita entender qué tan cerca o lejos está la competencia del propio negocio.

2. **Conocer el ideal de los consumidores**: Facilita comprender cuál es el producto o servicio ideal según los consumidores, y ayuda a evaluar si la oferta actual está alineada con ese ideal. Esto permite desarrollar estrategias de marketing ajustadas para reducir cualquier distancia percibida.

3. **Descubrir nuevos segmentos de mercado**: Ayuda a detectar si existen segmentos de mercado potencialmente atractivos que no están siendo aprovechados por la empresa.

4. **Conocer la posición de la empresa en el mercado**: Ofrece una visión clara de la posición actual de la empresa en el mercado, lo que permite decidir si continuar con la misma estrategia o modificarla para mejorar los resultados.

5. **Identificar los valores asociados al producto**: Proporciona información sobre los atributos y valores que los consumidores asocian con el producto o servicio.

Relación con el Análisis en Componentes Principales (ACP):

El **mapa perceptual** es una herramienta que combina los dos gráficos principales del **Análisis en Componentes Principales (ACP)**: el **círculo de correlaciones** y el **gráfico de marcas**. La interpretación de este mapa es similar a la del gráfico de individuos en el ACP, pero con un enfoque específico en el posicionamiento de las marcas en relación con los atributos clave.

En el mapa perceptual:

- Las **flechas** representan las **características o atributos** de las marcas, es decir, las variables originales que definen los componentes.
- Los **puntos** representan las **marcas** u observaciones.

La **proximidad** entre una marca y una flecha indica qué tan fuertemente está asociada esa marca con un atributo específico. Marcas cercanas entre sí tienen características similares desde el punto de vista de los consumidores, mientras que las marcas distantes reflejan percepciones significativamente diferentes.

Esta visualización proporciona una comprensión clara del **posicionamiento** de las marcas en relación con los atributos percibidos por los consumidores, ayudando a identificar diferencias clave, fortalezas y oportunidades en el mercado.

7. Aplicaciones

7.1 Mapa perceptual de posicionamiento (*brand rating Survey*)

Análisis en Componentes Principales aplicado a la Percepción de Marcas

En este estudio, se analiza cómo el **Análisis en Componentes Principales (ACP)** puede ser utilizado para interpretar los resultados de una encuesta sobre la percepción de los consumidores respecto a un conjunto de marcas, con el fin de construir un **mapa de posicionamiento**. Los datos provienen de un panel de consumidores de la empresa **TNS**, quienes mensualmente expresan su valoración de diversas marcas de productos mediante una escala Likert de 1 a 10.

TNS es una de las consultorías más grandes del mundo en investigación de mercados y, en España, realiza análisis para pequeñas, medianas empresas y multinacionales.

Contexto del Estudio

En diciembre de 2017, la "empresa B" experimentó una fuerte caída en sus ventas, a pesar de haber lanzado varias campañas navideñas impulsadas por el departamento comercial. Buscando entender las razones de esta crisis, el CEO de la "empresa B" sospecha que uno de los problemas podría ser que los clientes prueban los productos pero no repiten la compra. Además, han surgido nuevas empresas que han promocionado agresivamente sus marcas, lo que podría estar afectando el mercado. Otro punto a considerar es que las campañas de marketing de "empresa B" se centraron en promocionar el producto como algo muy novedoso.

Deseando investigar más a fondo las causas del descenso de las ventas y verificar si la estrategia de marketing es adecuada para el segmento de mercado, el CEO solicitó a **TNS** realizar el siguiente análisis:

1. Determinar si los consumidores volverían a comprar los productos de la marca B después de probarlos.
2. Evaluar si las nuevas empresas están atacando el mismo segmento de mercado.
3. Analizar si la estrategia de marketing de la empresa B es adecuada.

Aspectos a Evaluar

Para el estudio, **TNS** identificó **9 aspectos** clave relacionados con la percepción de las marcas, así como **10 marcas competidoras** que operan en el mercado español (A, B, C, D, E, F, G, H, I, J).

Los **9 aspectos** evaluados fueron:

1. **"perform"**: La marca tiene un buen rendimiento.
2. **"leader"**: La marca es percibida como líder.
3. **"latest"**: La marca es percibida como novedosa.
4. **"fun"**: La marca es divertida.
5. **"serious"**: La marca es seria.
6. **"bargain"**: La marca es percibida como una ganga.
7. **"value"**: Los productos de la marca tienen buen valor.
8. **"trendy"**: La marca está de moda.
9. **"rebuy"**: Los consumidores comprarían de nuevo productos de la misma marca.

Posteriormente, **TNS** realizó una encuesta donde se pidió a los consumidores que expresaran sus preferencias por cada marca en relación a los 9 aspectos mencionados, utilizando una escala de valoración de 1 a 10.

Objetivos del ACP

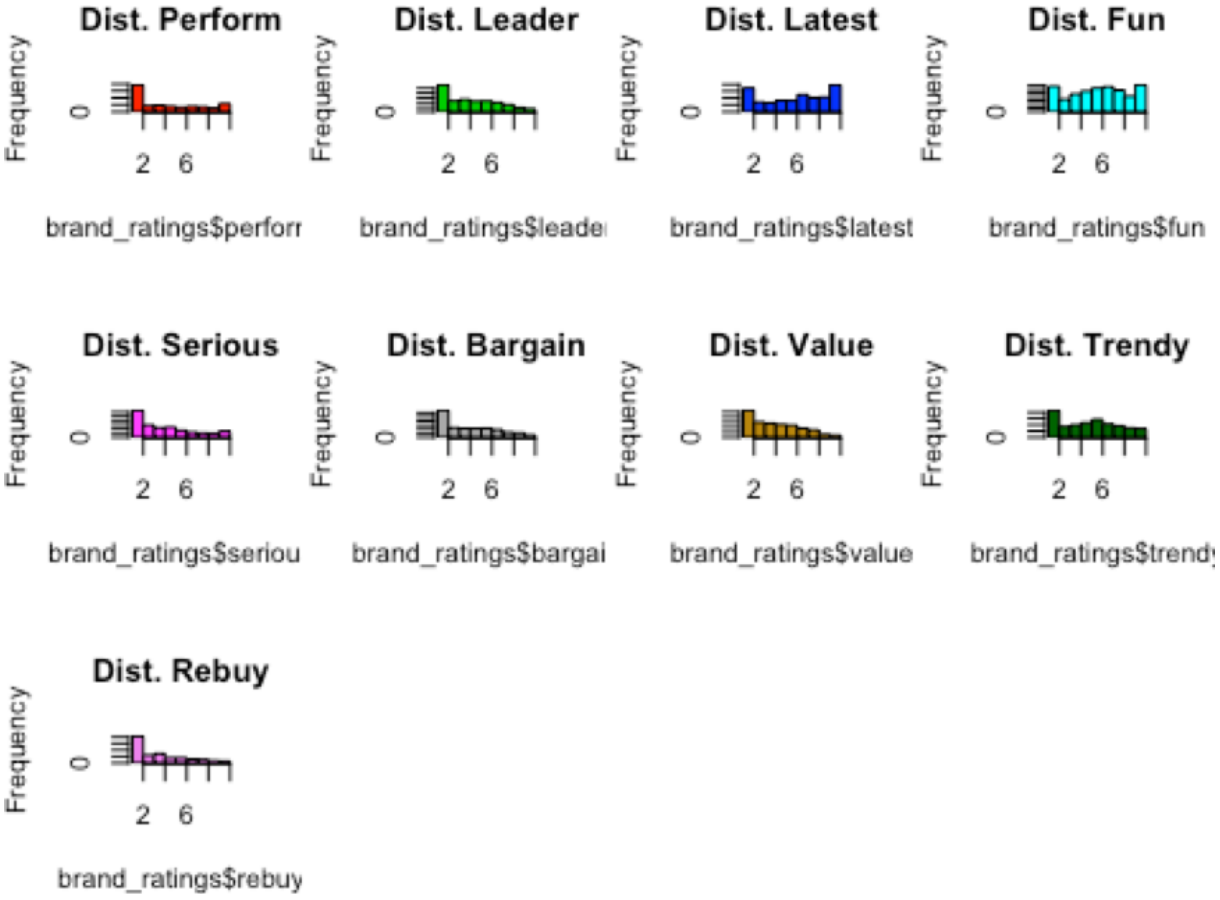
Mediante el ACP, se buscará sintetizar la información de los 9 aspectos clave y las 10 marcas para construir un mapa perceptual que permita visualizar el posicionamiento de las marcas en relación con estos atributos. Este análisis permitirá a la "empresa B" obtener insights sobre su percepción en el mercado y tomar decisiones estratégicas basadas en los resultados.

7.1.2 Análisis descriptiva y manipulación

Una vez realizada la encuesta, **TNS** llevó a cabo un **análisis descriptivo** para obtener una visión preliminar de la naturaleza de los datos. Dado que todas las variables son numéricas, se calcularon los histogramas correspondientes para cada una de ellas.

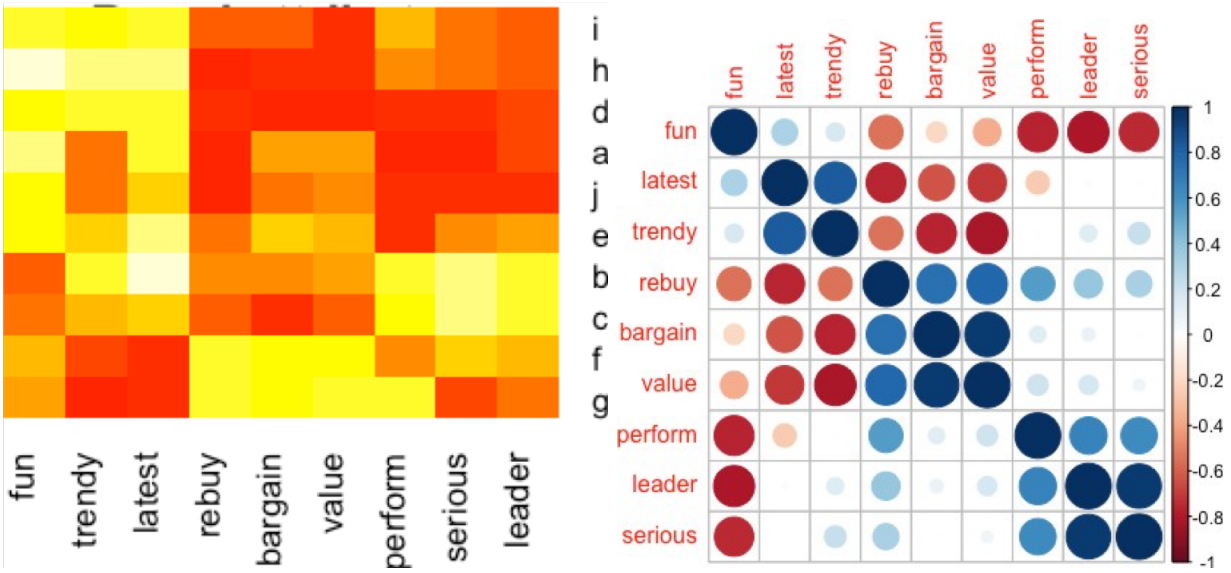
Al observar las distribuciones, **TNS** detectó que la mayoría de las variables presentaban una **distribución asimétrica a la izquierda**, lo que indica una mayor concentración de respuestas en los valores más bajos de la escala, un patrón común en las encuestas de preferencias.

Además, **TNS** verificó que todos los encuestados respondieron a todas las preguntas, por lo que no se encontraron **datos faltantes** en el conjunto de datos.



Posteriormente, **TNS** identificó otro problema en los datos: todas las preferencias estaban **desagregadas**. Es decir, para cada atributo, la encuesta recogía las preferencias de los consumidores de manera individual para cada marca. Sin embargo, para extraer conclusiones relevantes sobre la posición de la marca "B" y compararla con las demás, era necesario **agregar los datos a nivel de las marcas**.

Finalmente, **TNS** realizó un análisis de las **relaciones entre marcas y atributos**, así como de las **correlaciones entre los distintos atributos**, con el fin de identificar patrones que pudieran ser útiles en el análisis del posicionamiento de las marcas.



Los gráficos confirmaron dos aspectos clave:

1. Existía una **relación significativa** entre las marcas y los atributos evaluados.
2. Los atributos mostraban **patrones claros de correlación** entre sí.

7.1.2 ACP

A continuación, **TNS** aplicó el **Análisis en Componentes Principales (ACP)**, y como primer paso calculó el número de componentes necesarias para resumir adecuadamente los distintos atributos.

Componente	Valor propio (Eigenvalue)	Porcentaje de varianza	Porcentaje acumulado de varianza
Comp 1	4.55	50.62%	50.62%
Comp 2	3.01	33.44%	84.06%
Comp 3	0.59	6.57%	90.63%
Comp 4	0.37	4.20%	94.84%
Comp 5	0.25	2.88%	97.72%
Comp 6	0.13	1.49%	99.22%
Comp 7	0.04	0.51%	99.73%
Comp 8	0.02	0.23%	99.97%
Comp 9	0.00	0.02%	100.00%

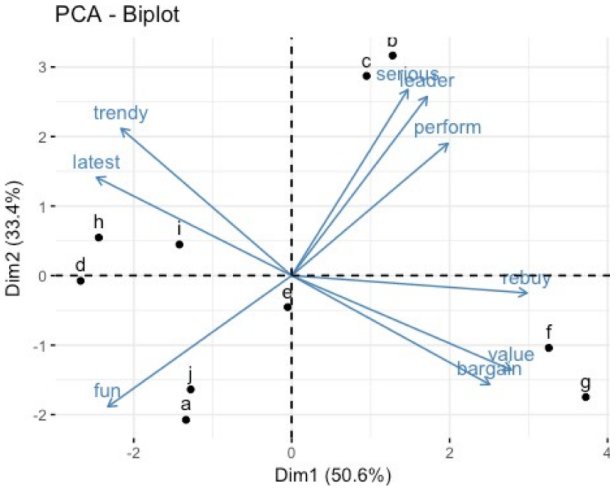
Table 13: Número de componentes y su importancia

Al observar los **valores propios**, **TNS** concluyó que con solo **dos componentes** era posible obtener una representación adecuada de la información contenida en la encuesta, ya que estas dos componentes explicaban el **84.06%** de la varianza total de los datos.

7.1.3 Mapa perceptual

Una vez identificado el número óptimo de componentes, **TNS** procedió a crear el **mapa perceptual**. Este análisis permitió a la empresa responder a las preguntas planteadas por el CEO de manera

efectiva.

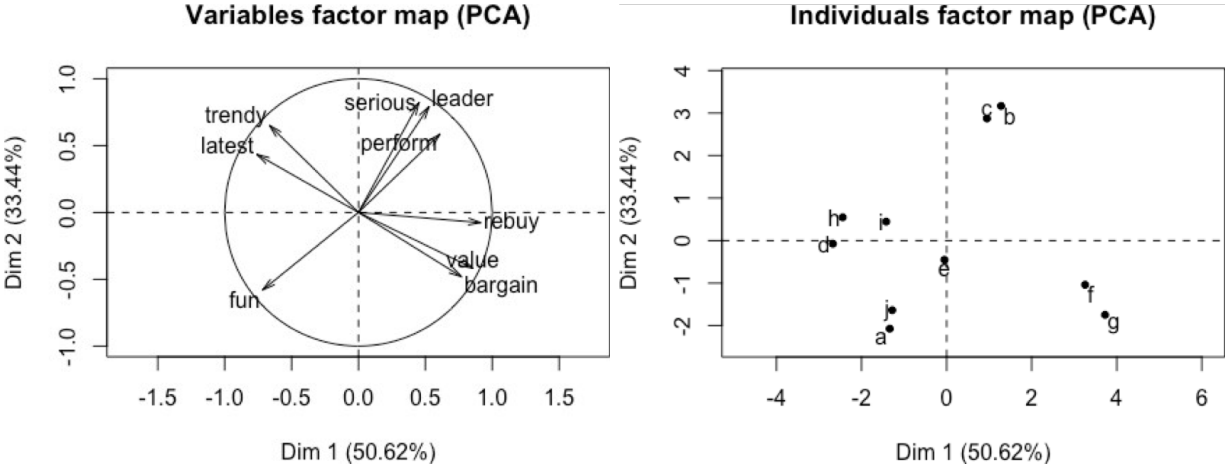


La empresa B tenia nueva un competidor que estaba afectando a su segmento de mercado. La marca venia percibida como una marca leader y de buena calidad, pero los consumidores indicaban que no lo la hubieran comprado otra vez. Las campañas de marketing estaban mal construidas porque, si bien la marca venia promocionada como novedosa no venia percibida de esta forma.

7.1.4 Mapa perceptual como construirlo

El análisis reveló que la **empresa B** tenía un nuevo competidor que estaba afectando significativamente su segmento de mercado. Si bien la marca era percibida como una **líder** y de **buena calidad**, los consumidores indicaron que **no la volverían a comprar**.

Además, las campañas de marketing resultaron ineficaces, ya que, aunque la marca se promocionaba como **novedosa**, no era percibida de esa manera por los consumidores.



7.2 Securitas: posicionamiento

Securitas Direct es una empresa especializada en seguridad (vigilancia y patrullaje móvil), monitoreo, consultoría e investigación, con sede en Estocolmo, Suecia. El grupo cuenta con más de 300,000 empleados distribuidos en 53 países alrededor del mundo.

Con más de 25 años de experiencia, el grupo **Securitas Direct** nació en Suecia en 1988, como parte del grupo Securitas. Diez años después, la empresa de alarmas **Securitas Direct** comenzó a operar de manera independiente.

Desde sus inicios, **Securitas Direct** ha experimentado un crecimiento constante y una expansión significativa en Europa. Actualmente, la empresa está presente en países como Bélgica, Dinamarca, Finlandia, Italia, Países Bajos, Noruega, Portugal, España, Suecia y el Reino Unido. Además, ha logrado una expansión notable en América del Sur, con oficinas en Chile, Brasil y Perú.

Siendo ya líder en Europa, en los últimos años la empresa ha centrado su atención en el mercado estadounidense, que representa un segmento de gran interés debido a las políticas de seguridad vigentes en el país.

Actualmente, **Securitas Direct** está llevando a cabo varios estudios de investigación de mercado para identificar la mejor estrategia para entrar en este nuevo segmento. En particular, el equipo de marketing está trabajando para determinar en qué estados sería más conveniente abrir nuevas filiales.

Para alcanzar este objetivo, la empresa dispone de diversas fuentes de información, que incluyen el número de arrestos por cada 100,000 residentes en los 50 estados de EE. UU., segmentado por los delitos de **Asalto** (*Assault*), **Asesinato** (*Murder*) y **Violación** (*Rape*). También se ha recopilado el porcentaje de la población que vive en áreas urbanas (*UrbanPop*), lo que proporcionará una visión adicional de los estados más adecuados para la expansión.

7.2.1 Data-set

El conjunto de datos está compuesto por **cuatro variables cuantitativas** e informa sobre el número de arrestos por cada 100,000 residentes en los cincuenta estados de EE. UU. Las variables incluidas son:

- **Asalto** (*Assault*): número de arrestos por asalto.
- **Asesinato** (*Murder*): número de arrestos por asesinato.
- **Violación** (*Rape*): número de arrestos por violación.
- **UrbanPop**: porcentaje de la población que vive en áreas urbanas.

Este conjunto de datos proporciona información clave para analizar las tasas de criminalidad en relación con la urbanización en los diferentes estados.

```
# leo mis datos
securitas_USA = read.csv(file="securitasUSA.csv", header = TRUE)
x=securitas_USA[,-1]
# uso los nombre de los estados como etiquetas de las lineas de mi data
rownames(x) = t(securitas_USA[,1])
# Realizo un resumen de las variables
summary(x)
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
```

```
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

Dado que todas las variables son cuantitativas, se ha decidido realizar un **Análisis en Componentes Principales (ACP)** con el objetivo de identificar visualmente los estados más adecuados para la apertura de nuevas filiales.

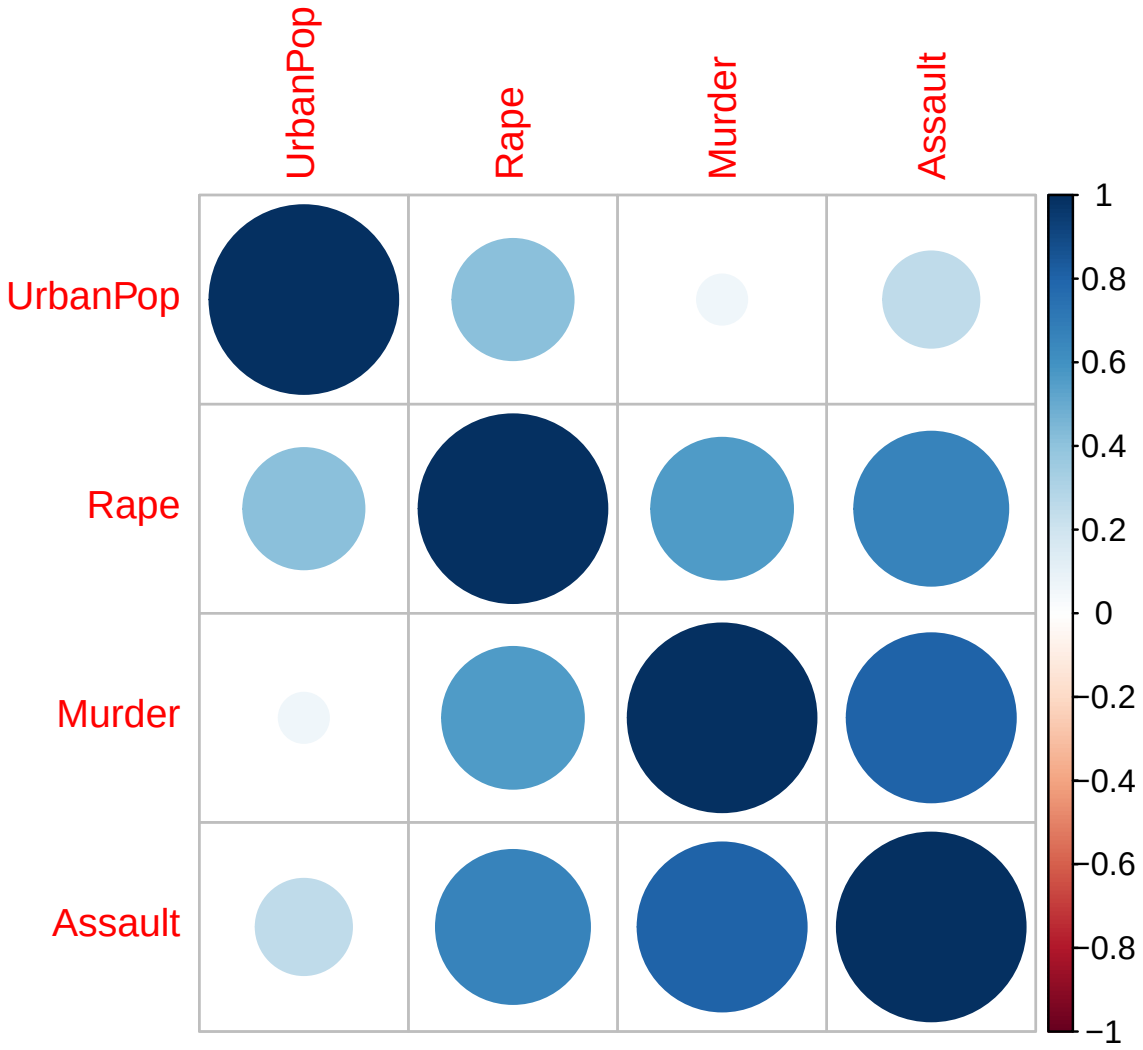
7.2.2 ACP Step 1: análisis de las correlaciones

Una forma visual de analizar las correlaciones entre las variables cuantitativas es utilizando la función `corrplot()` de la librería **corrplot**. La interpretación es bastante sencilla: cuanto más grandes son las circunferencias, mayor es la correlación entre las variables.

- El **color azul** indica correlaciones **positivas**.
- El **color rojo** indica correlaciones **negativas**.

A mayor tamaño de la circunferencia, más fuerte es la correlación entre las variables.

```
# Cargo la librería
suppressMessages(library(corrplot))

# La función "corrplot" permite visualizar las correlaciones
corrplot(cor(x), order = "hclust")
```



Se puede observar que la variable **Assault** está positivamente correlacionada con **Murder** y **Rape**, mientras que presenta correlaciones más bajas con **UrbanPop**. Por su parte, **UrbanPop** muestra correlaciones muy bajas con **Murder** y **Assault**. Además, las variables **Murder** y **Rape** están fuertemente correlacionadas entre sí.

Estas correlaciones sugieren la existencia de patrones claros entre las variables. El hecho de que **UrbanPop** tenga correlaciones muy bajas con **Murder** y **Assault** indica que, al aplicar el ACP, es probable que estas variables estén **contrapuestas** en los componentes principales.

7.2.3 ACP Step 2: Aplico la metodologia

Para realizar un **Análisis en Componentes Principales (PCA)** en R, se utiliza la función `PCA()` de la librería **FactoMineR**. Esta librería incluye varias funciones que permiten aplicar diferentes metodologías de análisis multivariante.

La función `PCA()` requiere como argumento principal los datos, que en el ejemplo están representados por el objeto x. Además, el parámetro `graph=FALSE` se utiliza para evitar la generación de gráficos en la primera ejecución, si no son necesarios de inmediato.

```r
# Cargar la librería
suppressMessages(library(FactoMineR))

# Aplicar PCA a los datos (x), sin generar gráficos
pca <- PCA(x, graph = FALSE)
```

7.2.4 ACP step 3: identifico el numero de Componentes

Se observa que hay dos valores propios mayores que 1, siguiendo el criterio de **eigenvalue > 1**. Además, las dos primeras componentes explican el **74%** de la variabilidad total contenida en los datos, de acuerdo con el criterio de **porcentaje de variabilidad explicada**.

Con esta información, decidimos seleccionar dos componentes para realizar el análisis. En R, podemos visualizar los valores propios y elegir las componentes utilizando el comando `pca$eig`. La función `round()` se usa para redondear los valores a un número específico de decimales.

```r
# Visualizar los valores propios redondeados
round(pca$eig, 2)
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1       2.48                 62.01                             62.01
## comp 2       0.99                 24.74                             86.75
## comp 3       0.36                  8.91                             95.66
## comp 4       0.17                  4.34                            100.00
```

```r
round(pca$eig,3)
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1      2.480                62.006                            62.006
## comp 2      0.990                24.744                            86.750
## comp 3      0.357                 8.914                            95.664
## comp 4      0.173                 4.336                           100.000
```

7.2.5 ACP Step 4: realizo el calculo de la ACP con dos componentes las calculo y las interpreto

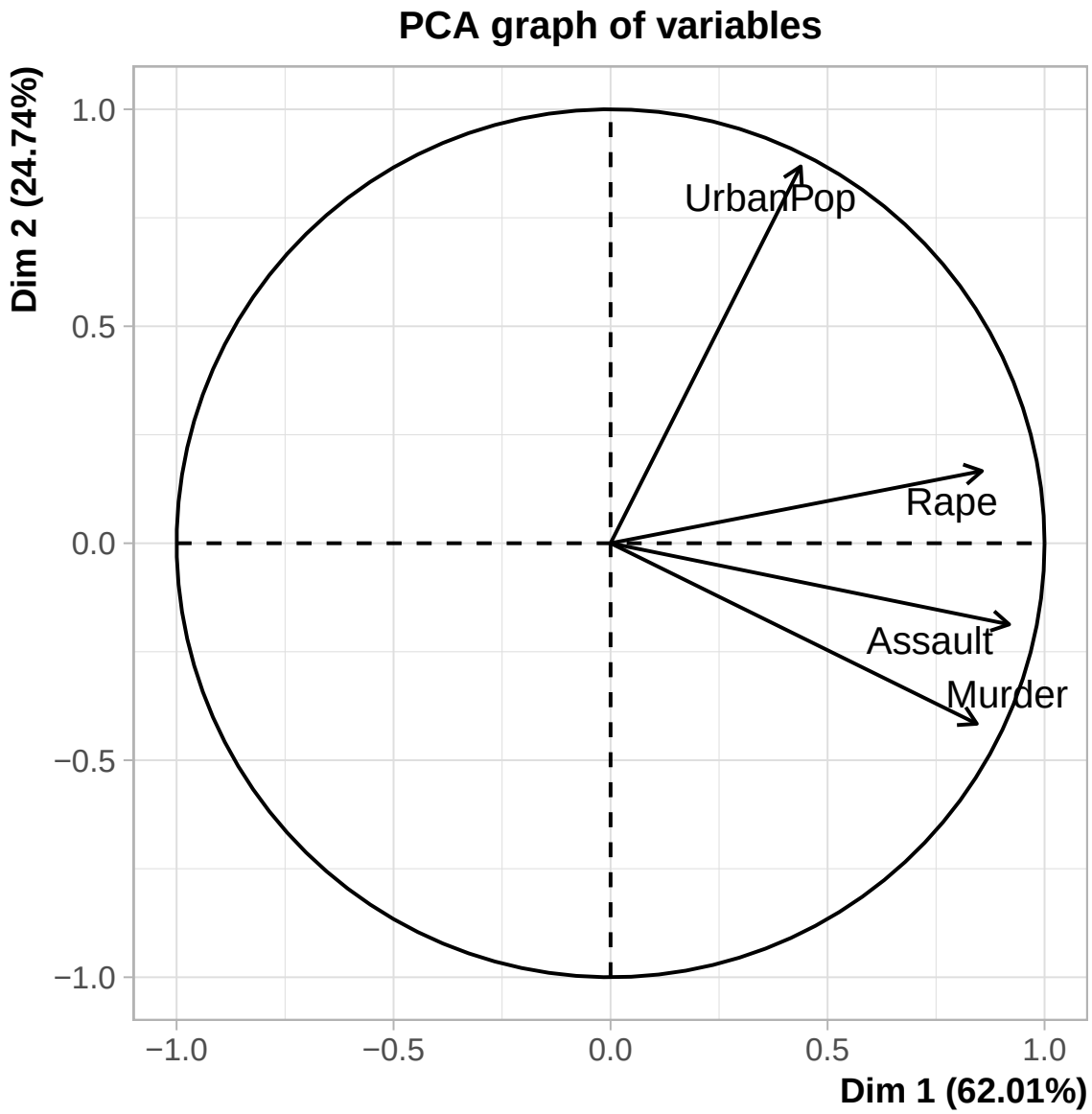
El gráfico representa las correlaciones entre las variables originales y las dos primeras componentes principales (Componente 1 y Componente 2). Una flecha larga que forma un ángulo pequeño con una componente indica una alta correlación entre la variable y dicha componente.

Se observa que la **Componente 1** está altamente correlacionada con las variables **Assault**, **Murder**, y **Rape**, ya que todas las variables apuntan en la misma dirección, hacia la derecha del gráfico. A partir de esta observación, podríamos interpretar la Componente 1 como un **indicador del nivel de criminalidad** en los estados de EE.UU. Los estados ubicados en la parte derecha del gráfico estarán caracterizados por un **alto nivel de criminalidad**, mientras que los estados en la parte izquierda estarán asociados a un **bajo nivel de criminalidad**.

Por otro lado, la **Componente 2** está correlacionada únicamente con la variable **UrbanPop**. Esta componente puede interpretarse como un **indicador de la tasa de población urbana**. Los estados que se encuentran en la parte superior del gráfico tendrán una **alta tasa de población urbana**, mientras que aquellos en la parte inferior presentarán una **baja tasa de población urbana**.

En R para calcular un ACP con solo dos componentes podemos utilizar el parámetro *ncp=2 (dos es el numero de las componentes)*

```
# x son mis datos y ncp son el numero de las componentes
pca2 = PCA(x, ncp=2, graph=FALSE)
# Para realizar el analisis y la interpretación de las componentes:
## plot
plot.PCA(pca2, axes=c(1, 2), choix="var")
```



El circulo de las correlaciones se obtiene mediante la función *plot.PCA(pca2, axes=c(1, 2),*

choix="var"). El parametro *axes=c(1, 2)* indica que queremos visualizar las primeras dos componentes. Las correlaciones, los contributos y los cos^2 los obtenemos respectivamente con los comandos: *pca2varcor*, *pca2varcontrib*, *pca2varcos2*.

```
var_measures = round(cbind(pca2$var$cor, pca2$var$contrib, pca2$var$cos2),3)
colnames(var_measures) = c("corCP1","corCP2","contribCP1","contribCP2",
                           "cos2CP1","cos2CP2")
var_measures
```

```
##          corCP1 corCP2 contribCP1 contribCP2 cos2CP1 cos2CP2
## Murder    0.844 -0.416     28.719     17.488   0.712   0.173
## Assault   0.918 -0.187     34.010      3.534   0.844   0.035
## UrbanPop  0.438  0.868      7.739     76.179   0.192   0.754
## Rape      0.856  0.166     29.532      2.800   0.732   0.028
```

Se observa que las variables **Assault**, **Murder** y **Rape** presentan correlaciones elevadas con la **primera componente**. Tanto las **contribuciones** como los valores de cos^2 confirman que estas tres variables son significativas para la interpretación de la primera componente y que están bien representadas por ella.

Por otro lado, la variable **UrbanPop** muestra un patrón similar con la **segunda componente**, lo que indica que es la variable más relevante en la interpretación de dicha componente.

7.2.6 ACP Step 5: realizo el gráfico de los individuos y lo interpreto

Observando el gráfico y recordando que la **primera componente** indica el **nivel de criminalidad** (los estados ubicados a la derecha tienen una tasa de criminalidad más alta) y que la **segunda componente** está relacionada con la **tasa de población** (los estados en la parte superior son los más poblados), podemos identificar los siguientes estados como posibles candidatos:

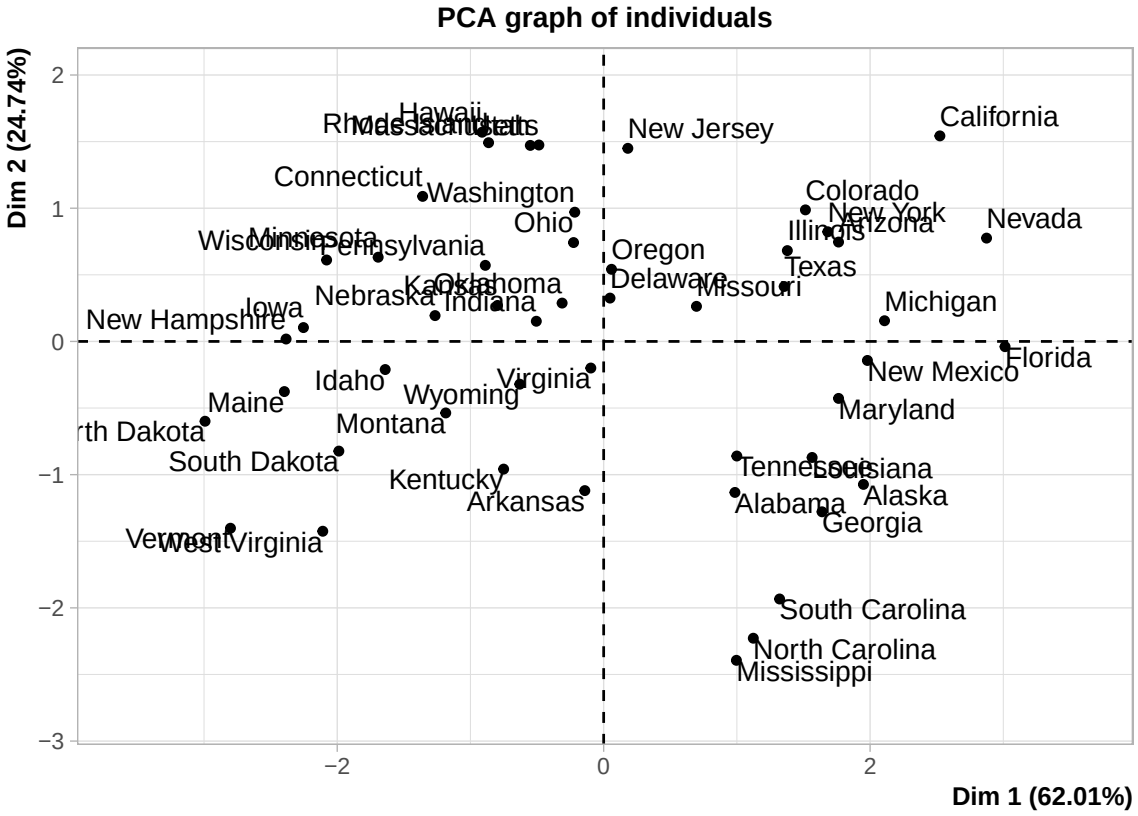
- **California** y **Nevada** destacan como los principales candidatos, ya que presentan altos niveles de criminalidad y una alta densidad de población.
- También se consideran **Florida**, **Michigan**, y **New México** como potenciales opciones.

Por otro lado:

- **South Carolina** y **North Carolina**, aunque tienen una alta tasa de criminalidad, presentan una baja densidad de población, lo que podría hacerlos menos atractivos.
- **Virginia**, **North Dakota** y **South Dakota** son los estados con las tasas de criminalidad más bajas, por lo que no serían interesantes para este análisis como posibles candidatos.

El gráfico se obtiene mediante la función *plot.PCA(pca2, axes=c(1, 2), choix="ind")*

```
# Para realizar el analisis y la interpretación de los individuos:
plot.PCA(pca2, axes=c(1, 2), choix="ind")
```



7.3 My Global Company (Promociones)

Promociones y Estrategias de Marketing

Las promociones (descuentos, sorteos, regalos, acumulación de puntos, etc.) son uno de los recursos más importantes del marketing. A través de ellas, las empresas pueden dar a conocer sus productos, generar una necesidad en el mercado, mejorar el posicionamiento de su marca, incrementar las ventas, atraer la atención de los clientes y mejorar su imagen de marca.

Por ello, si se desea vender un producto o servicio, es fundamental conocer las diferentes **estrategias de promoción** disponibles y elegir la más adecuada para el negocio.

Existen varias estrategias de marketing relacionadas con las promociones:

1. Estrategias de impulso
Consisten en incentivar a las personas encargadas de vender el producto para que realicen su labor de la mejor manera posible.

2. Estrategias híbridas o combinadas
En esta estrategia se combinan elementos de las estrategias de impulso y atracción, es decir, se ofrecen incentivos tanto a los vendedores como a los consumidores finales.

3. Estrategias de atracción
A diferencia de las estrategias de impulso, donde el objetivo es el vendedor, aquí el foco está en el consumidor final. En este grupo se incluyen promociones como descuentos, regalos, obsequios, etc.

Myglobal, una empresa de consultoría, ha sido contratada por una multinacional de gran distribución para estudiar esta última categoría y analizar cómo los consumidores perciben y evalúan las promociones.

7.3.1 Data-set

Después de una fase cualitativa, se seleccionaron siete tipos de promociones y se diseñó un cuestionario que fue completado por diez personas. Cada encuestado indicó su grado de interés por cada promoción en una escala del 1 al 10. Además, se recogieron algunas características socioeconómicas de los encuestados, como el **sexo**. Las promociones evaluadas fueron:

- **P1: "cantidad"**: mayor cantidad de producto.
- **P2: "valedisc"**: vale descuento para la próxima compra.
- **P3: "descuento"**: descuento directo en el precio señalado en la etiqueta.
- **P4: "sorteo"**: posibilidad de participar en un sorteo de un regalo.
- **P5: "puntos"**: acumulación de puntos para canjear por regalos.
- **P6: "obsequio"**: obsequio incluido con el producto.
- **P7: "tvcon"**: fichas para participar en un concurso de televisión.

El principal objetivo de este análisis es verificar cómo se agrupan los consumidores en función de las preferencias expresadas por cada una de las promociones seleccionadas. El **Análisis en Componentes Principales (ACP)** puede ser una herramienta útil para este propósito.

No todas las variables en el estudio son cuantitativas. Además de las preferencias por las promociones, se recogió el **sexo** de los encuestados. Esta variable no será incluida en el análisis principal, pero será utilizada como una **variable descriptiva** para interpretar los resultados una vez completado el análisis.

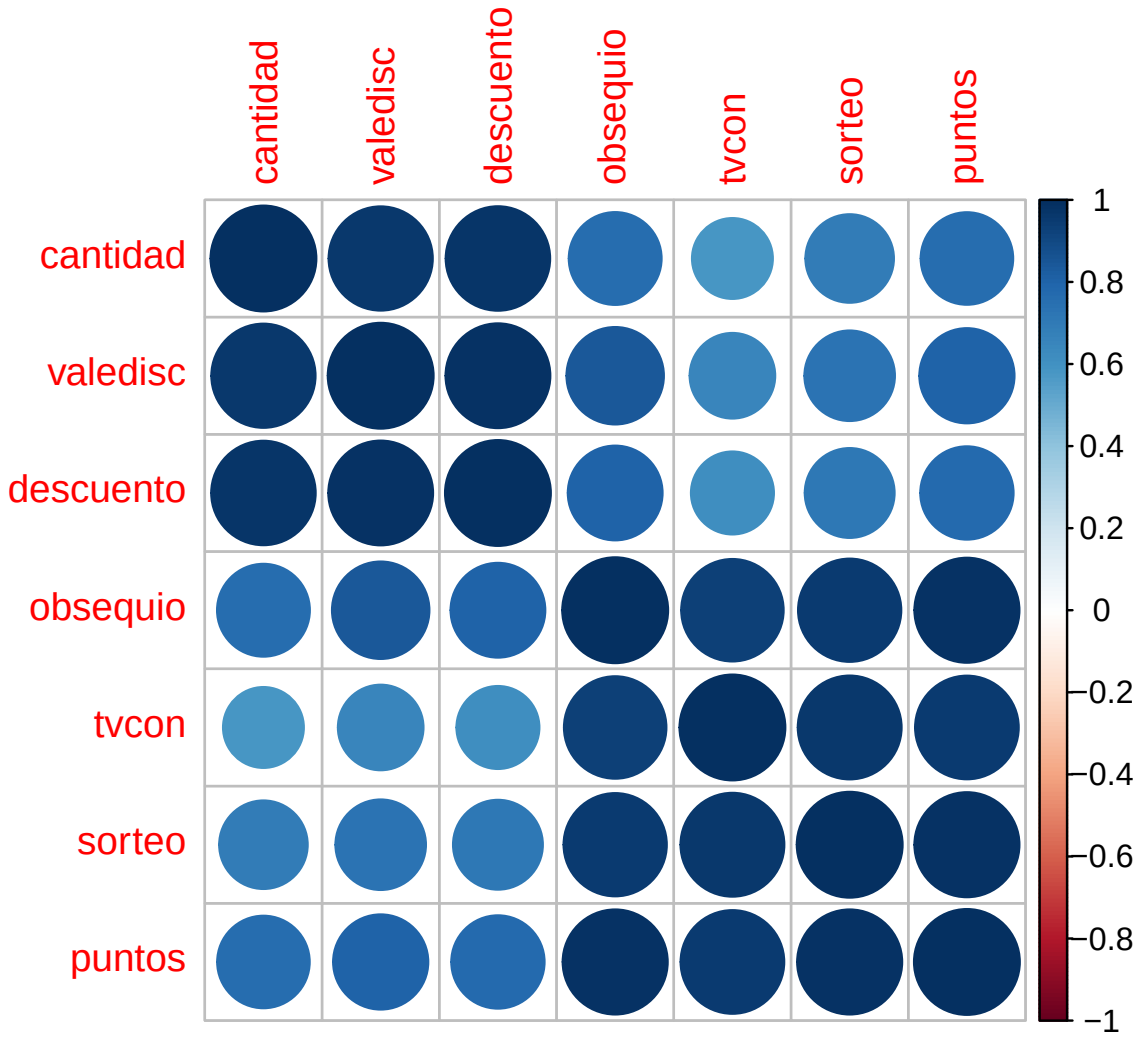
```
# leo mis datos
x = read.csv(file="MyGlobalCompany.csv", header = TRUE,sep=";")
summary(x)
```

```
##     cantidad        valedisc      descuento        sorteo        puntos
##  Min.   : 4.00   Min.   :4.00   Min.   :4.00   Min.   :1.0   Min.   :1.00
##  1st Qu.: 5.00   1st Qu.:4.25   1st Qu.:5.00   1st Qu.:3.0   1st Qu.:2.25
##  Median : 6.00   Median :5.50   Median :5.50   Median :5.0   Median :4.50
##  Mean   : 6.70   Mean   :6.20   Mean   :6.40   Mean   :5.1   Mean   :4.60
##  3rd Qu.: 8.75   3rd Qu.:8.00   3rd Qu.:8.75   3rd Qu.:7.0   3rd Qu.:6.75
##  Max.   :10.00   Max.   :9.00   Max.   :9.00   Max.   :9.0   Max.   :8.00
##     obsequio       tvcon         genero
##  Min.   :2.0   Min.   :1.00   Length:10
##  1st Qu.:3.0   1st Qu.:1.25   Class :character
##  Median :4.5   Median :3.00   Mode  :character
##  Mean   :4.9   Mean   :4.00
##  3rd Qu.:7.0   3rd Qu.:5.75
##  Max.   :8.0   Max.   :9.00
```

7.3.2 ACP step 1: análisis de las correlaciones

El primer paso para realizar el análisis en componentes principales consiste en el analizar la estructura de correlaciones entre las variables originales.

```
#Cargo la libreria
suppressMessages(library(corrplot))
# La función "corrplot" permite dibujar las correlaciones
corrplot(cor(x[,-8]), order="hclust")
```



Al observar las correlaciones, aunque todas son muy elevadas, es posible identificar claramente dos grupos de variables:

- El **primer grupo** está compuesto por las variables **cantidad**, **valedisc**, y **descuento**.
- El **segundo grupo** incluye las variables **sorteo**, **puntos**, **obsequio**, y **tvcon**.

Dado que las correlaciones entre todas las variables son tan elevadas, es probable que con una o, como máximo, dos componentes principales se pueda explicar la mayor parte de la variabilidad presente en los datos.

7.3.3 ACP step 2: aplico la metodología

Para aplicar un **Análisis en Componentes Principales (PCA)** en R, utilizamos la función `PCA()` de la librería **FactoMineR**.

En este caso, dado que hay una variable categórica (el método PCA no admite variables categóricas), debemos especificar que esta información solo se usará de forma descriptiva en el análisis. En R, podemos indicar que la variable **sexo** (que es la variable número 8 en el conjunto de datos) se debe tratar como una variable descriptiva mediante el parámetro `quali.sup = 8`.

El código para aplicar el PCA sería el siguiente:

```
library(FactoMineR)
#
pca = PCA(x, quali.sup=8, graph=FALSE)
```

7.3.4 ACP step 3: identifico el numero de Componentes

```
round(pca$eig,3)
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1      5.997                 85.678                             85.678
## comp 2      0.891                 12.724                             98.403
## comp 3      0.063                  0.896                             99.299
## comp 4      0.025                  0.364                             99.663
## comp 5      0.021                  0.305                             99.967
## comp 6      0.002                  0.028                             99.996
## comp 7      0.000                  0.004                            100.000
```

Se observa que hay un **valor propio** mayor que 1, lo que cumple con el criterio de **eigenvalue > 1**. Además, la **primera componente** explica ya el **85%** de la variabilidad total contenida en los datos, de acuerdo con el criterio de **porcentaje de variabilidad explicada**.

Teniendo en cuenta esta información, decidimos seleccionar una única componente para realizar el análisis.

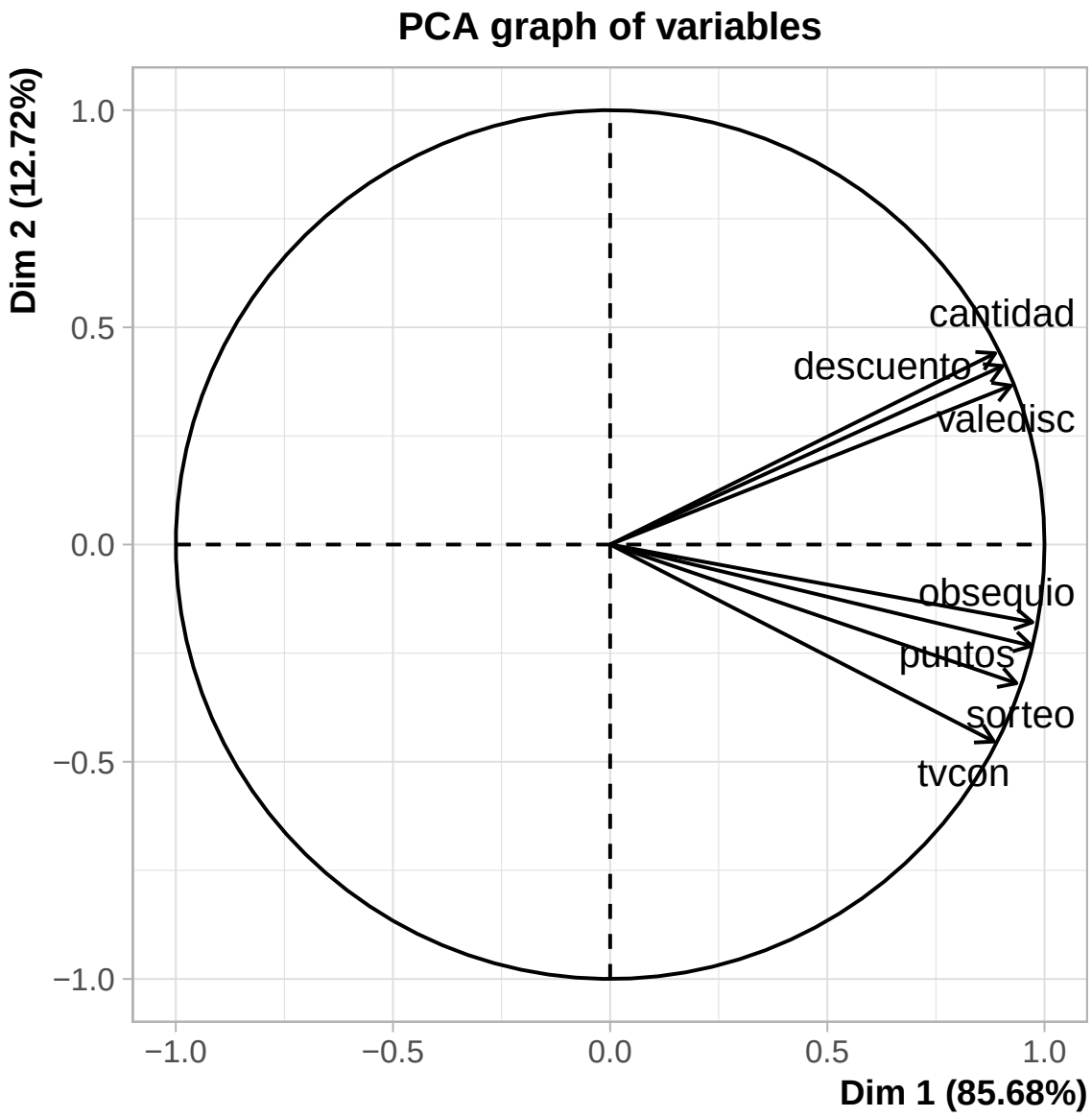
7.3.5 ACP step 4: realizo el calculo de la ACP con dos componentes las calculo y las interpreto

```
pca2 = PCA(x, quali.sup=8, ncp=2, graph=FALSE)
pca2$var$cor
```

```
##                Dim.1      Dim.2
## cantidad  0.8865789  0.4404405
## valedisc  0.9230087  0.3658077
```

```
## descuento 0.9036179  0.4110273
## sorteo    0.9355476 -0.3192779
## puntos    0.9704841 -0.2331002
## obsequio  0.9723214 -0.1789017
## tvcon     0.8834573 -0.4535260
```

```
plot.PCA(pca2, axes=c(1, 2), choix="var")
```



El gráfico muestra las correlaciones entre las variables originales y las dos primeras componentes (Componente 1 y Componente 2). Una flecha larga que forma un ángulo pequeño con una componente indica una alta correlación entre la variable y la componente.

Se puede observar que la **Componente 1** está altamente correlacionada con todas las promociones, y todas las variables apuntan en la misma dirección, hacia la derecha del gráfico. A partir de estas

observaciones, podríamos interpretar esta componente como la **actitud de los consumidores hacia las promociones**.

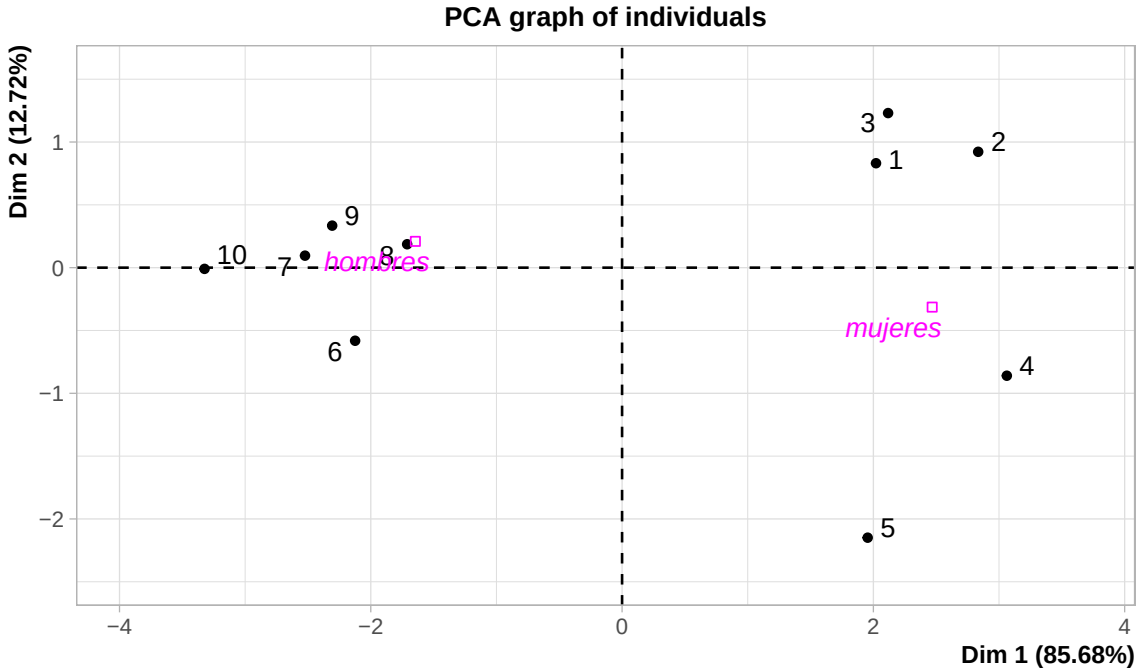
```
var_measures = round(cbind(pca2$var$cor, pca2$var$contrib, pca2$var$cos2),3)
colnames(var_measures) = c("corCP1","corCP2","contribCP1","contribCP2",
                           "cos2CP1","cos2CP2")
var_measures
```

```
##           corCP1 corCP2 contribCP1 contribCP2 cos2CP1 cos2CP2
## cantidad   0.887  0.440     13.106     21.779   0.786   0.194
## valedisc   0.923  0.366     14.205     15.023   0.852   0.134
## descuento  0.904  0.411     13.614     18.967   0.817   0.169
## sorteo     0.936 -0.319     14.594     11.445   0.875   0.102
## puntos     0.970 -0.233     15.704      6.100   0.942   0.054
## obsequio   0.972 -0.179     15.763      3.593   0.945   0.032
## tvcon      0.883 -0.454     13.014     23.092   0.780   0.206
```

Se observa que todas las variables presentan correlaciones elevadas con la **primera componente**. Tanto las **contribuciones** como los valores de cos^2 confirman que todas las variables son significativas para la definición de la primera componente y que están bien representadas por ella.

7.4.6 ACP step 5: realizo el gráficos de los individuos y lo interpreto

```
plot.PCA(pca2, axes=c(1, 2), choix="ind")
```



Al observar el gráfico y recordando que la **primera componente** representa la **actitud hacia las promociones** (con los consumidores más a la derecha mostrando una mejor actitud), se puede ver cómo los individuos se distribuyen a lo largo de esta componente. En la parte derecha del gráfico se encuentran predominantemente las mujeres, mientras que en la parte izquierda están los hombres.

Por lo tanto, podemos concluir que la multinacional debería centrar sus promociones en las mujeres, ya que este grupo muestra un mayor grado de interés en esta estrategia de marketing.

7.5 ACP para reducir las dimensiones y estimar una regresión linear (Principal Components Regression).

El método **Principal Components Regression (PCR)** consiste en ajustar un modelo de regresión lineal por mínimos cuadrados utilizando como predictores las componentes generadas a partir de un **Análisis en Componentes Principales (PCA)**. De esta manera, con un número reducido de componentes se puede explicar la mayor parte de la información contenida en los datos.

En estudios observacionales, es común contar con un gran número de variables que pueden ser utilizadas como predictores. Sin embargo, un alto número de predictores no necesariamente implica una mayor cantidad de información útil. Si las variables están correlacionadas entre sí, la información que aportan es redundante y puede violar la condición de no colinealidad requerida en la regresión lineal.

El PCA permite eliminar la información redundante y reducir el número de variables. Al emplear las componentes principales como predictores, se puede mejorar el modelo de regresión. Es importante destacar que, aunque la **regresión por componentes principales** reduce el número de predictores en el modelo, no debe considerarse como un método de selección de variables, ya que todas las variables originales son necesarias para el cálculo de las componentes.

7.5.1 Caso de Satisfacción Estudiantil

Un gran problema en las universidades es la **retención de estudiantes de primer año**. Por diversas razones, muchos de estos estudiantes no regresan para su segundo año. Si se pudieran identificar las causas principales que afectan la retención, se podrían implementar estrategias de mejora que ayuden a estos estudiantes a completar su educación universitaria.

El conjunto de datos contiene **147 observaciones** sobre las siguientes **32 variables**. Las diez primeras variables son variables de segmentación, mientras que el resto se refiere a cinco conceptos latentes:

1. **Imagen**
2. **Calidad específica**
3. **Calidad genérica**
4. **Valor**
5. **Satisfacción**

Descripción de Variables

- **Imagen**: Percepción general de los estudiantes sobre las escuelas de TIC (reconocida internacionalmente, variedad de cursos, liderazgo en investigación).

- **Calidad específica**: Percepción sobre la calidad de las habilidades específicas adquiridas durante el primer año en la universidad (competencias técnicas relacionadas con la materia de estudio).
- **Calidad genérica**: Percepción sobre la calidad de las habilidades generales adquiridas durante el primer año (habilidades para resolver problemas, habilidades de comunicación).
- **Valor**: Ventajas o beneficios que los exalumnos pueden obtener del título universitario (trabajo bien remunerado, motivación laboral, perspectivas de mejora y promoción).
- **Satisfacción**: Grado de satisfacción de los estudiantes respecto a la formación recibida en el primer año.

Descripción de los Ítems de la Encuesta

- **Imagen**:
  - **ima1 MV**: Es la mejor universidad para estudiar empresariales.
  - **ima2 MV**: Es internacionalmente reconocida.
  - **ima3 MV**: Cuenta con una amplia gama de cursos.
  - **ima4 MV**: Los profesores son de alta calidad.
  - **ima5 MV**: Las instalaciones y el equipamiento son excelentes.
  - **ima6 MV**: Es líder en investigación.
  - **ima7 MV**: Es bien considerada por las empresas.
  - **ima8 MV**: Está orientada a nuevas necesidades y tecnologías.
- **Calidad**:
  - **Calidad específica**:
    - **quaf1 MV**: Habilidades básicas.
    - **quaf2 MV**: Habilidades técnicas específicas.
    - **quaf3 MV**: Habilidades aplicadas.
  - **Calidad genérica**:
    - **qutr1 MV**: Habilidades en resolución de problemas.
    - **qutr2 MV**: Formación en gestión empresarial.
    - **qutr3 MV**: Habilidades de comunicación oral y escrita.
    - **qutr4 MV**: Planificación y gestión del tiempo adquiridas.
    - **qutr5 MV**: Habilidades de trabajo en equipo.
- **Valor**:
  - **val1 MV**: Me ha permitido encontrar un trabajo bien remunerado.
  - **val2 MV**: Tengo buenas perspectivas de mejora y promoción.
  - **val3 MV**: Me ha permitido encontrar un trabajo que me motiva.
  - **val4 MV**: La capacitación recibida es la base sobre la cual desarrollaré mi carrera.
- **Satisfacción**:
  - **sat1 MV**: Estoy satisfecho con mi carrera.

Análisis

Se procederá a estimar un modelo de **regresión lineal** utilizando la satisfacción como variable de respuesta y las demás variables como predictores. También se verificará la presencia de **multicolinealidad**. En caso de encontrar problemas, se considerará el **Análisis en Componentes Principales (ACP)** como alternativa.

Además, se estimarán distintas ACP para cada grupo de preguntas, se elegirá el número de componentes para cada uno y se interpretarán. Finalmente, se volverá a estimar un modelo de regresión utilizando las componentes en lugar de las variables originales y se proporcionará una interpretación de los resultados.

```
# leo mis datos
x = read.csv(file="satisfaction.csv", header = TRUE,sep=",")
x = x[,-c(1:2)]
```

7.5.2 Método clásico: Estimación por método de regresión lineal

Para conocer el nivel de satisfacción, como primer paso podemos estimar una regresión clásica de la variable satisfacción en función de las otras variables de la encuesta.

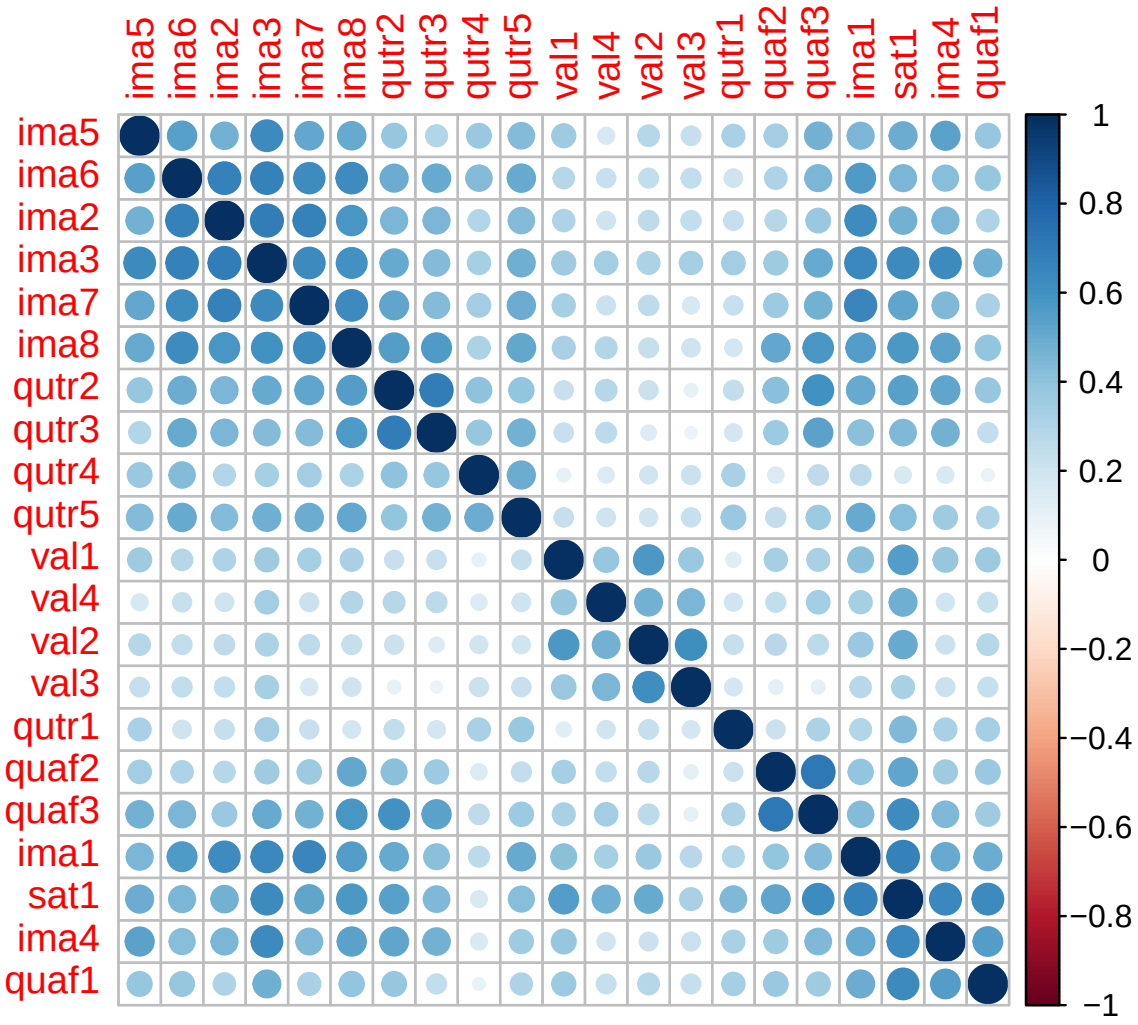
```
modelo <- lm(sat1 ~ ., data = x[-c(1:11)])
summary(modelo)
```

```
##
## Call:
## lm(formula = sat1 ~ ., data = x[-c(1:11)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4859 -0.5000 -0.0006  0.5535  3.0024
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.437697   0.541058  -2.657 0.008891 **
## ima2        -0.004750   0.062599  -0.076 0.939636
## ima3         0.113517   0.077199   1.470 0.143914
## ima4         0.175036   0.069849   2.506 0.013477 *
## ima5        -0.035459   0.061736  -0.574 0.566736
## ima6        -0.043026   0.059699  -0.721 0.472414
## ima7         0.075619   0.065938   1.147 0.253611
## ima8         0.076768   0.070176   1.094 0.276053
## quaf1        0.218973   0.059070   3.707 0.000312 ***
## quaf2        0.010281   0.060521   0.170 0.865374
## quaf3        0.148905   0.063398   2.349 0.020380 *
## qutr1        0.171217   0.060332   2.838 0.005288 **
## qutr2        0.056485   0.062048   0.910 0.364367
```

```
## qutr3       -0.007941   0.063034  -0.126 0.899948
## qutr4       -0.113611   0.049706  -2.286 0.023930 *
## qutr5        0.039731   0.050920   0.780 0.436688
## val1         0.112730   0.049393   2.282 0.024134 *
## val2         0.157254   0.058781   2.675 0.008451 **
## val3        -0.054826   0.047073  -1.165 0.246325
## val4         0.126913   0.046886   2.707 0.007728 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9223 on 127 degrees of freedom
## Multiple R-squared:  0.7768, Adjusted R-squared:  0.7434
## F-statistic: 23.26 on 19 and 127 DF,  p-value: < 2.2e-16
```

Para evaluar el nivel de satisfacción, como primer paso, podemos estimar un modelo de **regresión lineal** clásico de la variable **satisfacción** en función de las demás variables de la encuesta.

```r
#Cargo la libreria
suppressMessages(library(corrplot))
# La función "corrplot" permite dibujar las correlaciones
corrplot(cor(x[,-c(1:10)]), order="hclust")
```



7.5.2 Metodo PCA

Para reducir el número de variables, una opción es utilizar el método **Análisis en Componentes Principales (PCA)**. En una primera etapa, seleccionaremos el número adecuado de componentes, las interpretaremos y, en una segunda etapa, utilizaremos estas componentes como variables predictoras en el nuevo modelo.

A continuación, estimaremos las respectivas componentes para cada grupo de variables utilizando la función `PCA()` en R. Cabe destacar que, en el caso de las variables de **calidad genérica** y **calidad específica**, realizamos un único análisis debido a la similitud entre las variables.

```
data = x[,-c(1:10,32)]
pca_image = PCA(data[,c(1:8)], ncp=2, graph=FALSE)
pca_qual = PCA(data[,c(9:16)],ncp=2, graph=FALSE)
pca_val = PCA(data[,c(17:20)],ncp=2, graph=FALSE)
```

Numero de las componentes Al observar los valores propios, es evidente que con una sola componente podemos explicar la mayor parte de la información relacionada con **Valor** e **Imagen**.

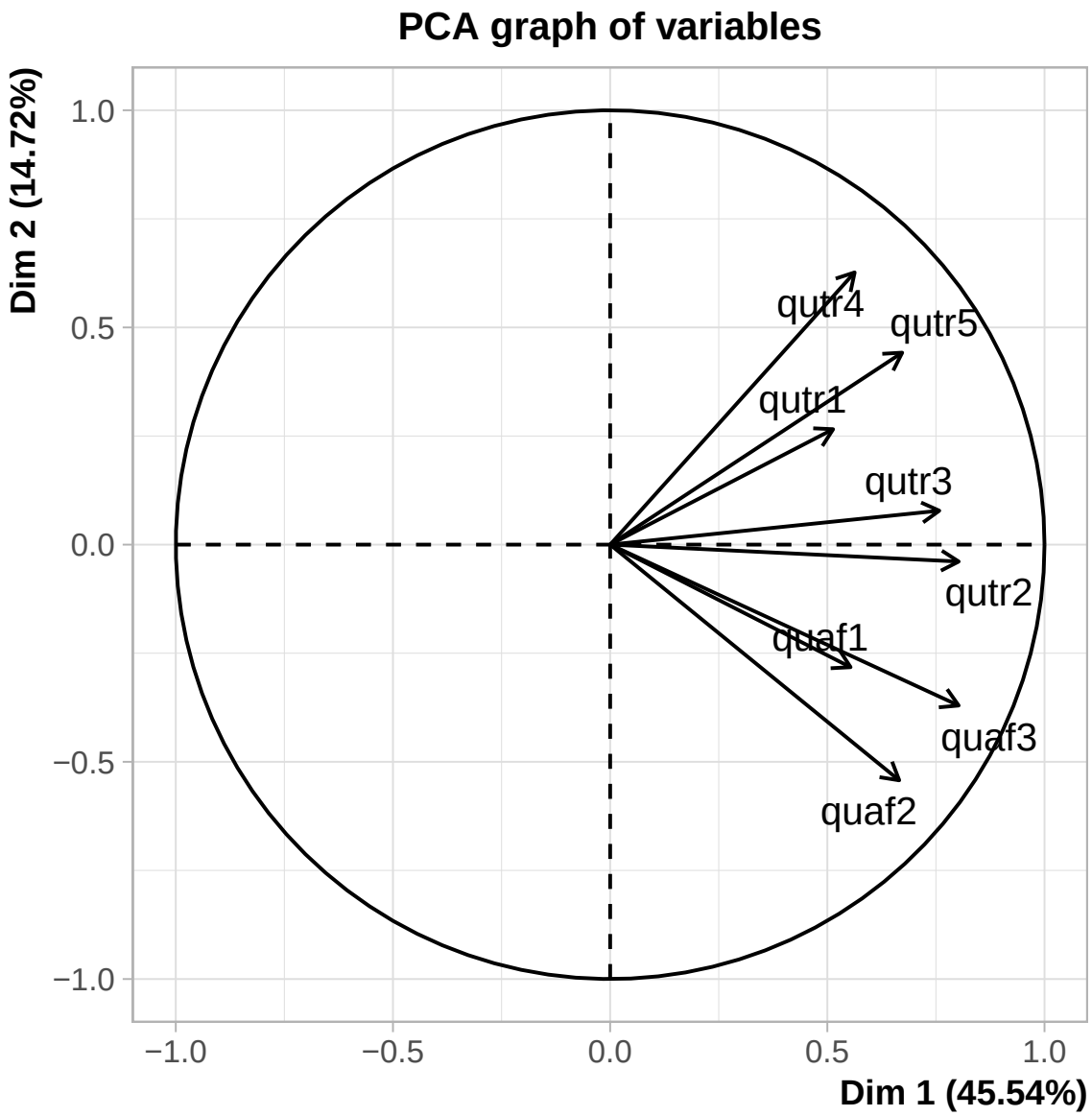
En el caso de **Calidad**, se consideran dos componentes de acuerdo con el criterio de **eigenvalue > 1**.

```
eigen=round(rbind(pca_image$eig[,1],pca_qual$eig[,1],pca_qual$eig[,1]),3)
rownames(eigen)= c("Image","Qual","Val")
eigen
```

```
##       comp 1 comp 2 comp 3 comp 4 comp 5 comp 6 comp 7 comp 8
## Image  5.046  0.726  0.535  0.462  0.411  0.324  0.264  0.234
## Qual   3.644  1.177  0.992  0.684  0.530  0.483  0.274  0.217
## Val    3.644  1.177  0.992  0.684  0.530  0.483  0.274  0.217
```

Interpretación de las componentes En cuanto a la interpretación, las variables que representan **Valor** e **Imagen** se pueden considerar como indicadores de estos conceptos. Para la interpretación de la **Calidad**, recurrimos al **círculo de correlaciones**.

```
plot.PCA(pca_qual, axes=c(1, 2), choix="var")
```



El **círculo de correlaciones** indica que la **primera componente** se puede interpretar como un índice de **calidad**. En cambio, la **segunda componente** diferencia entre la calidad **genérica** (en la parte alta) y la calidad **específica** (en la parte baja). Por lo tanto, podemos denominar esta componente como **tipología de calidad: Específica vs Genérica**.

Estimación regresión lineal: componentes utilizadas como variables predictoras Ahora podemos volver a estimar el modelo de regresión utilizando las componentes en lugar de las variables originales.

```
data.pca = data.frame(pca_image$ind$coord[,1],pca_qual$ind$coord[,c(1,2)],
                      pca_val$ind$coord[,1],x$sat1)

colnames(data.pca)= c("Image","Qual","Spec.vs.Gen","Val","Sat")
```

```r
modelo2 <- lm(Sat ~ ., data = data.pca)
summary(modelo2)
```

```
##
## Call:
## lm(formula = Sat ~ ., data = data.pca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1973 -0.7044  0.0550  0.7814  2.6383
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.01361    0.08433  83.171  < 2e-16 ***
## Image        0.24184    0.05985   4.041 8.70e-05 ***
## Qual         0.33634    0.06904   4.871 2.92e-06 ***
## Spec.vs.Gen -0.36403    0.07785  -4.676 6.74e-06 ***
## Val          0.35494    0.06053   5.864 3.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.022 on 142 degrees of freedom
## Multiple R-squared:  0.6933, Adjusted R-squared:  0.6847
## F-statistic: 80.24 on 4 and 142 DF,  p-value: < 2.2e-16
```

El modelo muestra que todos los coeficientes son altamente significativos (p-valor < 0.001), a excepción de la intersección, que no lo es. Esto es normal, dado que todas las componentes están estandarizadas (promedio 0, desviación estándar 1).

El coeficiente más importante es la **tipología de calidad** (-0.36), seguido por **valor** (0.35) y **calidad** (0.33). **Imagen** resulta ser la variable menos relevante, con un coeficiente de (0.24), aunque sigue siendo significativo. Es interesante notar que la tipología de calidad tiene un valor negativo, lo que indica que la calidad específica tiene un impacto negativo sobre la satisfacción, en contraste con las calidades genéricas.

El valor de R^2 es elevado y el **F-test** es significativo.

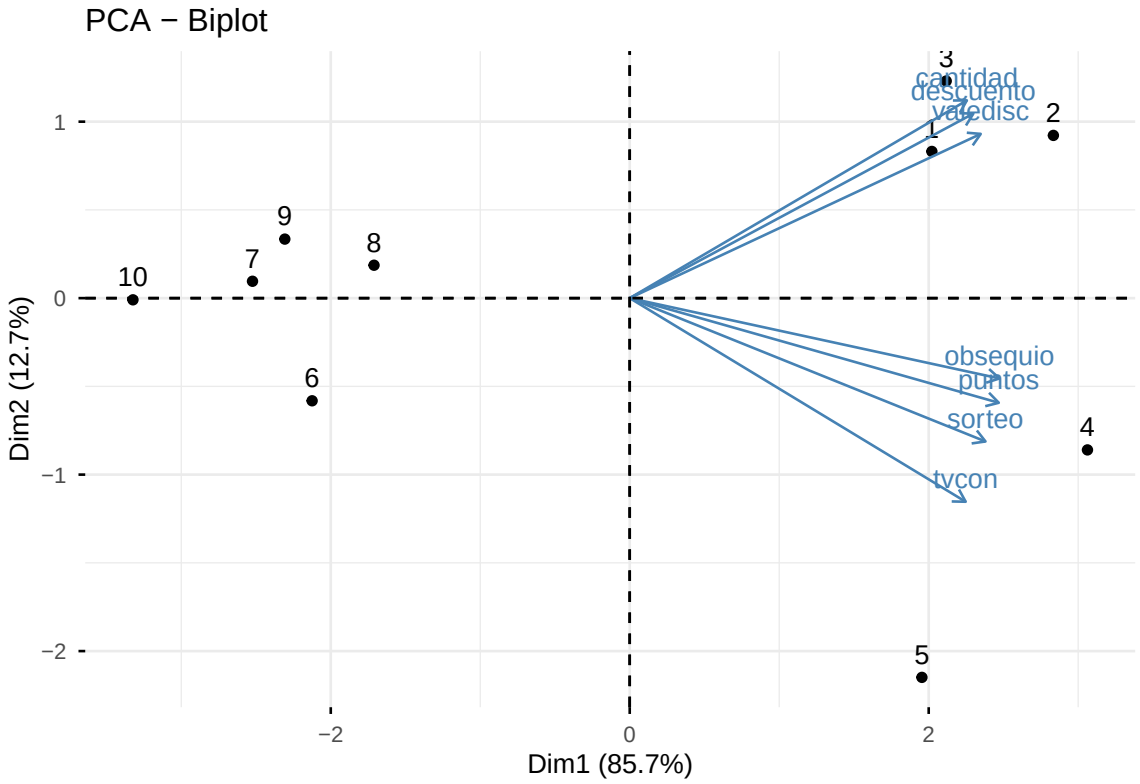
Considerando los aspectos más relevantes a potenciar para aumentar la satisfacción y mejorar la retención, podríamos concluir que se debe dar mayor énfasis a las calidades genéricas, seguidas del valor.

Anexo 1: Como Construir un Mapa Perceptual en R.

En R el mapa perceptual se construye mediante la función textit{fviz_pca_biplot(pca2)} de la liberia *library(factoextra)*,

Considerando el ejemplo de las promociones obtenemos:

```
suppressMessages(library(factoextra))
fviz_pca_biplot(pca2)
```



Anexo 2: Demostración Teórica del Cálculo de Autovalores y Autovectores en PCA

Esta demostración muestra cómo se calculan los autovalores y autovectores de una matriz, así como las componentes principales asociadas.

1. Definición de Autovalores y Autovectores

Dada una matriz cuadrada A de dimensión $n \times n$, un número λ se denomina **autovalor** de A si existe un vector no nulo $\mathbf{v}$ (llamado **autovector**) tal que se cumple la siguiente relación:

$$A\mathbf{v} = \lambda\mathbf{v}$$

2. Reformulación del Problema

La relación se puede reescribir como:

$$A\mathbf{v} - \lambda\mathbf{v} = 0$$

De forma compacta, esto se expresa como:

$$(A - \lambda I)\mathbf{v} = 0$$

donde I es la matriz identidad de dimensión n.

3. Condición de No-Trivialidad

Para que esta ecuación tenga soluciones no triviales (es decir, $\mathbf{v} \neq 0$), el determinante de $(A - \lambda I)$ debe ser igual a cero:

$$\det(A - \lambda I) = 0$$

4. Cálculo de la Ecuación Característica

Consideremos la siguiente matriz A:

$$A = \begin{pmatrix} 4 & 2 \\ 1 & 3 \end{pmatrix}$$

Primero, formamos la matriz $(A - \lambda I)$:

$$A - \lambda I = \begin{pmatrix} 4 - \lambda & 2 \\ 1 & 3 - \lambda \end{pmatrix}$$

Calculamos el determinante:

$$\det(A - \lambda I) = (4 - \lambda)(3 - \lambda) - (2)(1)$$

Expandiendo el determinante:

$$\begin{aligned} &= (4 - \lambda)(3 - \lambda) - 2 \\ &= 12 - 4\lambda - 3\lambda + \lambda^2 - 2 \\ &= \lambda^2 - 7\lambda + 10 \end{aligned}$$

5. Resolución de la Ecuación Característica

Ahora resolvemos la ecuación cuadrática:

$$\lambda^2 - 7\lambda + 10 = 0$$

Utilizamos la fórmula cuadrática:

$$\lambda = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

donde $a = 1$, $b = -7$, y $c = 10$:

$$\begin{aligned}\lambda &= \frac{7 \pm \sqrt{(-7)^2 - 4 \cdot 1 \cdot 10}}{2 \cdot 1} \\ &= \frac{7 \pm \sqrt{49 - 40}}{2} \\ &= \frac{7 \pm \sqrt{9}}{2} \\ &= \frac{7 \pm 3}{2}\end{aligned}$$

Obteniendo los autovalores:

$$\lambda_1 = 5, \quad \lambda_2 = 2$$

6. Cálculo de Autovectores

Para encontrar los autovectores asociados a cada autovalor, sustituimos λ en la ecuación $(A - \lambda I)\mathbf{v} = 0$.

Autovector para $\lambda_1 = 5$:

$$A - 5I = \begin{pmatrix} 4-5 & 2 \\ 1 & 3-5 \end{pmatrix} = \begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix}$$

Resolviendo la ecuación:

$$\begin{pmatrix} -1 & 2 \\ 1 & -2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0$$

Esto da lugar a las ecuaciones:

1. $-v_1 + 2v_2 = 0$
2. $v_1 - 2v_2 = 0$

De aquí, deducimos que $v_1 = 2v_2$. Entonces, un autovector correspondiente a $\lambda_1 = 5$ es:

$$\mathbf{v_1} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

Autovector para $\lambda_2 = 2$:

$$A - 2I = \begin{pmatrix} 4-2 & 2 \\ 1 & 3-2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix}$$

Resolviendo la ecuación:

$$\begin{pmatrix} 2 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0$$

Esto da lugar a las ecuaciones:

1. $2v_1 + 2v_2 = 0$
2. $v_1 + v_2 = 0$

De aquí, deducimos que $v_1 = -v_2$. Entonces, un autovector correspondiente a $\lambda_2 = 2$ es:

$$\mathbf{v_2} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

7. Cálculo de Componentes Principales

Las componentes principales se obtienen proyectando los datos originales sobre los autovectores. Si X es la matriz de datos centrados, las componentes se calculan como:

$$Z = XV$$

donde V es la matriz de autovectores.

Análisis de correspondencias

1. Introducción

El análisis de correspondencias es una técnica descriptiva y exploratoria cuyo objetivo es resumir una gran cantidad de datos en un número reducido de dimensiones, minimizando la pérdida de información. En este sentido, su objetivo es similar al de los métodos factoriales, como el análisis de componentes principales (ACP), pero con la diferencia de que el análisis de correspondencias se aplica a **variables categóricas** y **ordinales**.

El **análisis de correspondencias simples** es comúnmente utilizado para la representación gráfica de datos que se presentan en **tablas de contingencia** con dos variables nominales o ordinales.

Cuando los datos están organizados en una tabla de contingencia con dos variables cualitativas, donde las categorías de una variable se sitúan en las filas y las de la otra en las columnas, el análisis de correspondencias permite resumir la información presente en estas filas y columnas. Esto se logra proyectando los datos en un subespacio reducido, en el que se pueden representar simultáneamente los puntos fila y los puntos columna. Este enfoque facilita obtener conclusiones sobre las relaciones entre las dos variables nominales u ordinales de origen.

El análisis de correspondencias simples se puede extender al caso de **múltiples variables nominales**, conocido como **análisis de correspondencias múltiples**. Este método sigue los mismos principios generales del análisis de correspondencias simples, pero se aplica a tablas de contingencia multidimensionales. Generalmente, una de las variables representa ítems o individuos, mientras que el resto son variables cualitativas o ordinales que describen características de estos.

Aplicaciones en Marketing

El análisis de correspondencias, tanto simple como múltiple, es ampliamente utilizado en **Marketing** para distintos fines:

- **Análisis del comportamiento del consumidor (preferencias)**:
  - ¿Están determinados atributos de los coches relacionados con ciertas marcas?
  - ¿Existe alguna relación entre la disposición de los consumidores a contratar servicios Premium y su nivel económico?
- **Posicionamiento de empresas basado en preferencias de los consumidores**:
  - ¿Existe alguna correlación entre estrategias comerciales y variables como la provincia, la edad o el género de los consumidores?
- **Identificación de tipologías de individuos en relación con variables cualitativas**:
  - Por ejemplo, patrones de consumo o perfiles de clientes.

En resumen, el análisis de correspondencias es una técnica descriptiva que permite crear un **mapa perceptual** de las categorías de las variables analizadas en un espacio reducido de pocas dimensiones (habitualmente 2). La **distancia** entre los puntos representados en este espacio refleja la fuerza de las relaciones de dependencia y similitud entre las categorías, proporcionando una visión clara y compacta de los datos.

2. Presentación intuitiva

El **Análisis de Correspondencias (AC)** es una técnica estadística utilizada para analizar, desde una perspectiva gráfica, las **relaciones de dependencia e independencia** entre un conjunto de variables categóricas.

2.1. Punto de Partida: Tabla de Contingencia

El punto de partida para el análisis de correspondencias es una **tabla de contingencia**, que representa el cruce entre dos variables categóricas. Esta tabla contiene las frecuencias observadas para cada combinación de categorías de las dos variables.

A partir de estas frecuencias observadas, el método realiza diversas transformaciones de los datos con el fin de obtener una **matriz numérica**. Esta matriz puede ser utilizada en métodos matemáticos que permiten, de manera similar al ACP, **calcular componentes principales** y representar las modalidades de las variables en un sistema de **ejes cartesianos**.

2.2. Representación Gráfica

El objetivo del Análisis de Correspondencias es asociar a cada modalidad de las variables categóricas un punto en el sistema de ejes cartesianos. La **proximidad o distancia entre los puntos** refleja las **relaciones de dependencia y semejanza** entre las categorías analizadas.

Este enfoque gráfico facilita la interpretación visual de los datos, ya que permite identificar patrones y asociaciones entre las categorías de las variables estudiadas.

En resumen, el Análisis de Correspondencias proporciona una forma eficiente de representar gráficamente las relaciones entre variables categóricas, aprovechando transformaciones matemáticas que permiten resumir la información en componentes principales. Esta técnica es especialmente útil en estudios donde se desea explorar la **dependencia** o **independencia** entre categorías y visualizar estas relaciones de manera clara y comprensible.

3. Transformaciones de las frecuencias observadas: frecuencias relativas, perfiles columnas y filas

El **Análisis de Correspondencias** comienza con la transformación de las modalidades de las variables categóricas. Este proceso se realiza en tres pasos, tomando como punto de partida dos variables categóricas X e Y, con sus respectivas modalidades $\{x_1, \dots, x_i\}$ y $\{y_1, \dots, y_j\}$.

3.1. Pasos de la Transformación

3.1.1. Cálculo de la Tabla de Contingencia

El primer paso consiste en construir una **tabla de contingencia** que cruza las distintas modalidades de las dos variables categóricas. El resultado es n_{ij}, que representa la **frecuencia observada** para la combinación de la modalidad i de la variable X y la modalidad j de la variable Y.

Tabla de Contingencia: n_{ij}

	1	2	3	4	Marginal fila
1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
Marginal columna	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	N

Matriz de frecuencias absolutas

3.1.2. Cálculo de las Frecuencias Relativas

A continuación, se calculan las **frecuencias relativas**. Esto se realiza dividiendo cada valor n_{ij} por el número total de observaciones N, obteniendo así f_{ij}, donde:

$f_{ij} = \frac{n_{ij}}{N}$

	1	2	3	4	Marginal fila
1	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1.}$
2	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2.}$
3	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3.}$
Marginal columna	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	1

Matriz de frecuencias relativas

Dónde $f_{ij} = \frac{n_{ij}}{N}$

3.1.3. Cálculo de los Perfiles de Filas y Columnas

El tercer paso consiste en calcular los **perfiles de filas** y **perfiles de columnas**. Esto se logra dividiendo cada frecuencia observada n_{ij} por el total marginal de la fila $n_{i.}$ o el total marginal de la columna $n_{.j}$.

- **Perfil de Filas**: $\frac{n_{ij}}{n_{i.}}$
- **Perfil de Columnas**: $\frac{n_{ij}}{n_{.j}}$

	1	2	3	4
1	f11/**f1.**	f12/f1.	f13/f1.	f14/f1.
2	f21/f2.	f22/f2.	f23/f2.	f24/f2.
3	f13/f3.	f23/f3.	f33/f3.	f34/f3.
4	f14/f4.	f23/f4.	f33/f4.	f44/f4.

Marginales rigas

	1	2	3	4
1	f11/**f.1**	f12/f.2	f13/f.3	f14/f.4
2	f21/f.1	f22/f.2	f23/f.3	f24/f.4
3	f13/f.1	f23/f.2	f33/f.3	f34/f.4
4	f14/f.1	f23/f.2	f33/f.3	f44/f.4

Marginales columnas

3.2. Aplicación de la Descomposición en Valores Singulares

Como resultado de estos pasos, obtenemos dos matrices numéricas que pueden ser analizadas mediante la **descomposición en valores singulares (SVD)**. Este procedimiento nos permite calcular las **componentes principales** y generar las **representaciones gráficas** que muestran las relaciones entre las modalidades de las variables categóricas en un espacio reducido.

4. Test de dependencia entre las variables categóricas.

Un requisito fundamental para aplicar el **Análisis de Correspondencias** es que las variables categóricas estén **relacionadas**. Para determinar si existe dependencia entre las variables, se utiliza el **test de Chi-cuadrado (X^2)**, que evalúa la asociación entre dos variables categóricas a través de una tabla de contingencia.

4.1. Hipótesis del Test de X^2

El test de X^2 se basa en las siguientes hipótesis:

- **Hipótesis nula (H_0)**: Las variables categóricas son independientes, es decir, no existe relación entre las modalidades de las variables.
- **Hipótesis alternativa (H_1)**: Las variables categóricas son dependientes, es decir, existe una relación significativa entre las modalidades.

4.2. Cálculo del Estadístico X^2

El **estadístico** X^2 se calcula a partir de las frecuencias observadas en la tabla de contingencia, que cruza las categorías de las dos variables. Este valor compara las **frecuencias observadas (n_{ij})** con las **frecuencias esperadas (n^*_{ij})** bajo la hipótesis de independencia. La fórmula es la siguiente:

$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$

Donde: - n_{ij} es la frecuencia observada para la combinación de la modalidad i de la variable X y la modalidad j de la variable Y. - n^*_{ij} es la frecuencia esperada bajo la hipótesis de independencia, que se calcula como:

$n^*_{ij} = \frac{n_{i.} \times n_{.j}}{N}$

Donde: - $n_{i.}$ es el total de frecuencias de la fila i (frecuencia marginal de fila), - $n_{.j}$ es el total de frecuencias de la columna j (frecuencia marginal de columna), - N es el total de observaciones.

4.3. Criterio de Decisión

Una vez calculado el estadístico X^2, se compara con un valor crítico de la distribución de Chi-cuadrado con un nivel de significancia $\alpha = 0.05$ y los grados de libertad correspondientes: Grados de libertad $= (k-1)(m-1)$

Donde k y m son el número de categorías de las variables X e Y, respectivamente.

- Si el **p-value** asociado al estadístico X^2 es **menor** que el nivel de significancia $\alpha = 0.05$, se rechaza la hipótesis nula (H_0) y se concluye que existe **dependencia** entre las variables.
- Si el p-value es mayor que α, no hay suficiente evidencia para rechazar H_0, y se asume que las variables son **independientes**.

Si el test de Chi-cuadrado muestra que las variables son dependientes, se puede proceder con el **Análisis de Correspondencias**. Este método proyecta las relaciones de dependencia entre las modalidades de las variables en un espacio de menor dimensión, facilitando la interpretación visual y la comprensión de las asociaciones.

El test de X^2 es esencial para garantizar que el análisis de correspondencias se aplique de manera adecuada, ya que la técnica está diseñada para representar gráficamente las relaciones de dependencia entre las variables.

En resumen, el **test de Chi-cuadrado (X^2)** es una herramienta clave para evaluar la dependencia entre variables categóricas antes de aplicar el Análisis de Correspondencias. Si el test indica que las variables están relacionadas, el análisis puede ayudar a visualizar estas relaciones en un espacio reducido, permitiendo una interpretación clara de los datos.

5. Elegimos el numero de las componentes

El número de **componentes principales** en el **Análisis de Correspondencias** se determina de manera similar al **Análisis de Componentes Principales (ACP)**. El objetivo es encontrar el menor número de componentes que permita explicar la mayor proporción de la variabilidad de los datos, simplificando su estructura sin perder información relevante.

5.1. Variabilidad Explicada

En el contexto del Análisis de Correspondencias, la **variabilidad explicada** por cada componente suele ser considerablemente menor que en el ACP. Mientras que en el ACP, valores del **60% al 80%** de variabilidad explicada pueden considerarse aceptables para retener un número determinado de componentes, en el caso del Análisis de Correspondencias los valores típicos de variabilidad explicada son mucho más bajos.

- En el Análisis de Correspondencias, **valores de** 30% **o** 40% **de variabilidad explicada** se consideran bastante **elevados** y suficientes para interpretar las relaciones entre las variables categóricas.

  Esto se debe a la naturaleza de las **variables categóricas**, que suelen presentar una mayor dispersión en la información y, por lo tanto, es más difícil capturar gran parte de la variabilidad con un número reducido de componentes.

5.2. Interpretación de Componentes

Al igual que en el ACP, se debe analizar el porcentaje de **varianza acumulada** para cada componente. Generalmente, se seleccionan las primeras componentes que explican una proporción significativa de la variabilidad, aunque el criterio de selección puede ser más laxo debido a los bajos porcentajes de variabilidad explicada en este tipo de análisis.

- **Componentes seleccionados**: El número de componentes que se retienen dependerá del balance entre la **cantidad de variabilidad explicada** y la **simplicidad** del modelo.

6. Interpretación del grafico de los individuos

En el **Análisis de Correspondencias**, los gráficos resultantes permiten visualizar las relaciones entre las categorías de las variables. La interpretación de estos gráficos se basa en varios criterios importantes que nos ayudan a comprender mejor las asociaciones y dependencias entre las modalidades representadas. A continuación, describimos los criterios clave para la interpretación:

6.1. Proximidad entre categorías

La **proximidad entre las categorías** en el gráfico se interpreta en términos de **asociación o dependencia**. Cuanto más cercanas estén dos categorías en el gráfico, mayor es su relación o dependencia.

- Categorías cercanas comparten características similares o están más asociadas en los datos.

6.2. Posición respecto al origen

Una categoría que coincide con el **perfil promedio** se ubicará en el **centro del espacio**, cerca del **origen** del gráfico. Si una categoría está **alejada del origen**, esto indica que difiere significativamente del perfil promedio.

- Cuanto más alejada esté una categoría del origen, más singular es su perfil con respecto al conjunto de datos.

6.3. Puntos extremos y contrapuestos

Los **puntos más extremos** en el gráfico son aquellos que representan categorías más alejadas del promedio y, por tanto, son **más importantes para la interpretación**. Las categorías contrapuestas pueden sugerir perfiles muy diferenciados.

6.4. Índices de $\cos^2$ y de contribución absoluta

Para evaluar la calidad de la representación de cada categoría en los ejes principales, utilizamos dos indicadores: **el coseno al cuadrado** ($\cos^2$) y la **contribución absoluta**.

$\cos^2$ **(Cosenos al Cuadrado)** El **coseno al cuadrado** ($\cos^2$) nos indica la **calidad de la representación** de un punto sobre un eje en particular. Cuanto más cercano a 1 sea el valor de $\cos^2$, mejor estará representado el punto en ese eje.

La fórmula del coseno al cuadrado es:

$$\cos^2_{ik} = \frac{\text{Coordenada del punto en el eje } k^2}{\sum_{l=1}^{p} \text{Coordenada del punto en el eje } l^2}$$

Donde: - $\cos^2_{ik}$ es el coseno al cuadrado de la categoría i en el eje k. - La coordenada del punto en el eje k indica la posición de la categoría i en el factor k. - El denominador es la suma de los cuadrados de las coordenadas del punto en todos los ejes.

Cuanto más próximo a 1 sea el $\cos^2$, mejor estará representada la categoría en ese eje. Si el valor es bajo, significa que el punto no está bien representado en ese factor y su interpretación en ese eje debe tomarse con precaución.

Contribución Absoluta La **contribución absoluta** indica **cuánto una modalidad contribuye** a explicar la variabilidad de un determinado factor (componente principal). Las modalidades que tienen una mayor contribución son las más relevantes para la interpretación del factor correspondiente.

La fórmula para calcular la contribución de una modalidad i al factor k es:

$\text{Contribución}_{ik} = \frac{f_i \times d^2_{ik}}{\lambda_k}$

Donde: - f_i es la **frecuencia relativa** de la categoría i (proporción de ocurrencias en la muestra). - d_{ik} es la **coordenada** de la categoría i en el eje k. - λ_k es el **autovalor** asociado al eje k, que representa la cantidad de inercia explicada por ese factor.

Cuanto mayor sea la contribución de una categoría al factor, mayor será su importancia en la explicación de ese eje, y mejor estará representada en el gráfico. Si una categoría tiene una baja contribución, su interpretación en ese factor es menos significativa.

7. Análisis de las correspondencias múltiples

El **Análisis de Correspondencias Múltiples** (ACM) se basa en aplicar un análisis de correspondencias a la **matriz de Burt**, que se define como $B = Z'Z$. La **matriz de Burt** se construye mediante la superposición de "cajas" de datos. En los **bloques diagonales** de la matriz de Burt aparecen las **matrices diagonales** que contienen las frecuencias marginales de cada variable analizada. Por su parte, los elementos **fuera de la diagonal** contienen las **tablas de frecuencias cruzadas** correspondientes a todas las combinaciones posibles (de dos en dos) entre las variables.

La matriz de Burt se genera a partir de la **matriz disyuntiva completa** (denominada matriz Z), que se obtiene transformando todas las variables categóricas en variables dummy. A continuación, se muestra un ejemplo para ilustrar este proceso.

7.1. Ejemplo

Consideremos una tabla formada por 10 individuos de una empresa, clasificados según su **género**, los **años en la empresa** y los **ingresos obtenidos**:

Individuos	Género	Años	Ingreso
1	Mujer	5	Medio
2	Mujer	3	Alto
3	Hombre	4	Bajo
4	Mujer	1	Bajo
5	Mujer	2	Medio
6	Hombre	5	Alto
7	Mujer	2	Medio
8	Hombre	3	Bajo
9	Hombre	1	Alto
10	Mujer	4	Medio

7.2. Matriz Disyuntiva Completa (Matriz Z)

A partir de esta tabla original, construimos la **tabla disyuntiva** o **matriz $Z**", con tantas columnas como categorías en las variables analizadas.

Género		Años					Ingresos		
Mujer	Hombre	1	2	3	4	5	Bajo	Medio	Alto
1	0	0	0	0	0	1	0	1	0
1	0	0	0	1	0	0	0	0	1
0	1	0	0	0	1	0	1	0	0
1	0	1	0	0	0	0	1	0	0
1	0	0	1	0	0	0	0	1	0
0	1	0	0	0	0	1	0	0	1
1	0	0	1	0	0	0	0	1	0
0	1	0	0	1	0	0	1	0	0
0	1	1	0	0	0	0	0	0	1
1	0	0	0	0	1	0	0	1	0

En la **matriz disyuntiva completa (Z)**, si alguna de las variables es continua, esta debe transformarse en nominal, agrupándola en intervalos a los que se asignan rangos de valores discretos.

- Las **frecuencias marginales** de las filas de la matriz disyuntiva corresponden al número total de preguntas o categorías (s).
- Las **frecuencias marginales** de las columnas representan el número de sujetos que han seleccionado una modalidad específica (j) de la pregunta (q).

7.3. Relación entre Variables y la Matriz de Burt

La relación entre cada variable con las demás en la matriz disyuntiva completa permite construir la **matriz de Burt**. La matriz de Burt contiene todas las **tablas de contingencia simples** para cada par de variables categóricas (combinaciones de dos a dos).

- Para **n individuos** que han respondido preguntas sobre dos variables nominales con *p1* y *p2* modalidades, respectivamente, realizar un **análisis de correspondencias simples** sobre la tabla de contingencia (*p1*, *p2*) es equivalente a analizar una **tabla binaria** con n filas y (*p1* + *p2*) columnas, que describen las respuestas.

El resultado sería:

		Género		Años					Ingresos		
		M	H	1	2	3	4	5	B	M	A
→ Género →	M	1	0	0	0	0	0	1	0	1	0
	H	1	0	0	0	1	0	0	0	0	1
	1	0	1	0	0	0	1	0	1	0	0
	2	1	0	1	0	0	0	0	1	0	0
Z = Años	3	1	0	0	1	0	0	0	0	1	0
	4	0	1	0	0	0	0	1	0	0	1
↦	5	1	0	0	1	0	0	0	0	1	0
	B	0	1	0	0	1	0	0	1	0	0
Ingresos ↦	M	0	1	1	0	0	0	0	0	0	1
	A	1	0	0	0	0	1	0	0	1	0

		Género		Años					Ingresos		
		M	H	1	2	3	4	5	B	M	A
→ Género →	M	1	1	0	1	1	0	1	0	0	1
	H	0	0	1	0	0	1	0	1	1	0
	1	0	0	0	1	0	0	0	0	1	0
	2	0	0	0	0	1	0	1	0	0	0
Z' = Años	3	0	1	0	0	0	0	0	1	0	0
	4	0	0	1	0	0	0	0	0	0	1
↦	5	1	0	0	0	0	1	0	0	0	0
	B	0	0	1	1	0	0	0	1	0	0
Ingresos ↦	M	1	0	0	0	1	0	1	0	0	1
	A	0	1	0	0	0	1	0	0	1	0

Al cruzar las variables dos a dos, la **matriz disyuntiva** se convierte en una **matriz de Burt** que contiene todas las tablas de contingencia simples entre las variables.

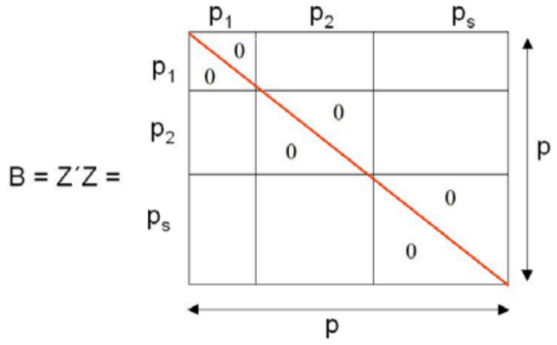
7.4. Matriz de Burt ($B = Z'Z$)

A partir de la **matriz disyuntiva completa** (Z), podemos construir la **matriz de contingencia de Burt** (B). Esta es una **matriz simétrica** de orden (p, p), donde cada bloque contiene una **submatriz** con las tablas de contingencia de las variables, cruzadas dos a dos.

- Los **bloques en la diagonal** contienen las **tablas de contingencia** de cada variable consigo

misma.

- Los bloques fuera de la diagonal contienen las **tablas de contingencia cruzadas** entre las variables.



Ejemplo de **matriz de Burt** para las variables consideradas:

MATRIZ DE BURT

Z'.Z =

		Género		Años					Ingresos		
		M	H	1	2	3	4	5	B	M	A
Género	M	6	0	1	2	1	1	1	1	4	1
	H	0	4	1	0	1	1	1	2	0	2
Años	1	1	1	2	0	0	0	0	1	0	1
	2	2	0	0	2	0	0	0	0	2	0
	3	1	1	0	0	2	0	0	1	0	1
	4	1	1	0	0	0	2	0	1	1	0
	5	1	1	0	0	0	0	2	0	1	1
Ingresos	B	1	2	1	0	1	1	0	3	0	0
	M	4	0	0	2	0	1	1	0	4	0
	A	1	2	1	0	1	0	1	0	0	3

7.5 Análisis e Interpretación

Una vez construida la **matriz de Burt**, el **análisis e interpretación** se realizan de manera similar a como se hace en el **análisis de correspondencias simples**. No obstante, el análisis en el contexto del **Análisis de Correspondencias Múltiples** tiende a ser más complejo debido a la mayor cantidad de variables involucradas.

7.5.1. Pasos de Interpretación:

1. **Proximidad entre categorías**: Se observa la cercanía o lejanía entre las categorías de las variables en los gráficos resultantes. Esto refleja la **asociación o dependencia** entre las categorías.
2. **Contribución de cada categoría**: Se evalúan los **índices de contribución** para determinar qué categorías son más importantes en la explicación de los ejes o factores principales. Las categorías con mayores contribuciones serán clave en la interpretación de los gráficos.
3. **Cálculo del coseno al cuadrado** ($\cos^2$): Este indicador permite evaluar la **calidad de la representación** de cada categoría en los ejes principales. Cuanto más cercano a 1 sea el valor de $\cos^2$, mejor representada estará la categoría en el gráfico.

En resumen, el **Análisis de Correspondencias Múltiples** permite representar gráficamente las relaciones entre múltiples variables categóricas, facilitando la interpretación visual de la dependencia entre las mismas.

8. Ejemplos

8.1 Tareas Domesticas

En las últimas décadas, la estructura y las dinámicas familiares han experimentado transformaciones significativas, especialmente entre las parejas jóvenes. Factores como la igualdad de género, la creciente participación de la mujer en el mercado laboral y la mayor valoración del tiempo personal han impulsado cambios en la manera en que las responsabilidades del hogar y las tareas familiares se distribuyen entre los miembros de la pareja.

Este estudio tiene como objetivo analizar las **tendencias actuales en la repartición de las tareas familiares** entre las parejas jóvenes, con el fin de identificar patrones y comportamientos emergentes que puedan ofrecer **oportunidades valiosas para el marketing**. La repartición de responsabilidades en el hogar no solo influye en la dinámica familiar, sino también en los hábitos de consumo, las decisiones de compra y las prioridades en el estilo de vida.

Entender cómo las parejas jóvenes distribuyen sus tareas y responsabilidades es fundamental para las empresas que buscan **diseñar productos, servicios y campañas de marketing** alineadas con las nuevas realidades sociales. Las marcas que ofrecen soluciones prácticas y adaptadas a las necesidades de los hogares modernos tienen una ventaja competitiva en un mercado donde las expectativas de eficiencia, igualdad y conveniencia están en constante crecimiento.

A través de este estudio, se explorarán aspectos clave como: - **¿Cómo se reparten las tareas del hogar?** ¿Existen diferencias según el género, el nivel de ingresos o la situación laboral de cada miembro de la pareja? - **¿Qué productos o servicios resultan más atractivos para las parejas que buscan optimizar su tiempo y gestionar sus responsabilidades familiares de manera equitativa? - ¿Cómo influyen las decisiones sobre la distribución de tareas en el comportamiento de compra?**

El presente estudio investiga la **distribución de 13 tareas domésticas** entre las parejas y cómo estas se reparten. Aunque la **tabla de contingencia** que utilizamos no es muy grande, su análisis visual ofrece una interpretación clara de los **perfiles de fila y columna**. Sin embargo, lo más importante es que este caso proporciona un excelente ejemplo para mostrar cómo el **Análisis de Correspondencias** puede identificar **asociaciones** entre los niveles de las variables categóricas involucradas.

Objetivos del Análisis El análisis se desarrollará en **tres pasos** clave:

1. **Exploración de la tabla de contingencia**: En primer lugar, observaremos la tabla de contingencia que muestra la distribución de las tareas domésticas entre las parejas.
2. **Verificación mediante el test de Chi-cuadrado (X^2)**: En este segundo paso, evaluaremos si existe una **relación significativa** entre las dos variables principales: el **tipo de tarea** y su **distribución** dentro de la pareja. El test de X^2 permitirá confirmar si las variables están relacionadas o si son independientes.

3. **Análisis de Correspondencias**: Finalmente, realizaremos un Análisis de Correspondencias para identificar las **relaciones más importantes** entre las categorías de las variables. Esto nos permitirá obtener una visión clara de las **asociaciones** más relevantes entre los tipos de tareas y cómo se reparten dentro de las parejas.

```
library(FactoMineR)
library(gplots)
library(factoextra)

# leo mis datos
data(housetasks)
x=housetasks
```

En el análisis de datos categóricos, las **tablas de contingencia** son una herramienta esencial para examinar la relación entre dos o más variables. Sin embargo, cuando estas tablas contienen mucha información, puede ser difícil interpretarlas visualmente. Una forma de simplificar la interpretación es mediante el uso del comando `balloonplot` en **R**, que permite generar una representación visual de las tablas de contingencia utilizando **circunferencias**.

Uso de `balloonplot` en R

Mediante el comando `balloonplot(t(x))`, donde `x` es la tabla de contingencia transpuesta, podemos generar una **visualización intuitiva** de los datos. En esta representación:

- **Circunferencias más grandes** indican una **mayor relación** o frecuencia observada entre las variables categóricas.
- **Circunferencias más pequeñas** indican una relación más débil o menos frecuente.

Este tipo de gráfico es particularmente útil cuando queremos resaltar las asociaciones más importantes entre las categorías de las variables, haciendo que la interpretación sea rápida y efectiva.

```
dt = as.table(as.matrix(x))
balloonplot(t(dt), main ="", xlab ="", ylab="", show.margins = TRUE)
```

	Wife	Alternating	Husband	Jointly	
Laundry	156	14	2	4	176
Main_meal	124	20	5	4	153
Dinner	77	11	7	13	108
Breakfeast	82	36	15	7	140
Tidying	53	11	1	57	122
Dishes	32	24	4	53	113
Shopping	33	23	9	55	120
Official	12	46	23	15	96
Driving	10	51	75	3	139
Finances	13	13	21	66	113
Insurance	8	1	53	77	139
Repairs		3	160	2	165
Holidays		1	6	153	160
	600	254	381	509	1744

a que la tabla de contingencia no es muy grande, es posible observar algunas relaciones claras entre las variables.

Es evidente que las tareas del hogar como **Lavandería**, **Comida principal** y **Cena** son realizadas con mayor frecuencia por la **Esposa**. Por otro lado, las tareas relacionadas con **reparaciones** y **conducción** suelen ser realizadas principalmente por el **Marido**. En cuanto a la planificación de **vacaciones**, estas están mayormente asociadas con la opción **conjuntamente**, lo que indica una repartición más equitativa en este tipo de actividad.

El test de X^2 confirma la existencia de una relación significativa entre las dos variables (tipos de tareas y distribución de tareas). Recordemos las hipótesis planteadas para el test:

- H_0: Las variables son independientes.
- H_1: Las variables son dependientes.

Dado que el p-value obtenido es **0**, con un nivel de significancia $\alpha = 0.05$, podemos **rechazar la hipótesis nula** (H_0) y concluir que **existe dependencia** entre el tipo de tareas y su distribución entre los miembros de la pareja.

En **R**, podemos realizar el test de X^2 utilizando el siguiente comando: *chisq.test(x)*

```
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 1944.5, df = 36, p-value < 2.2e-16
```

En R, podemos realizar el **Análisis de Correspondencias** utilizando la función `CA()` de la librería **FactMineR**. Esta función proporciona diferentes salidas útiles para interpretar los resultados del análisis:

- **ca$eig**: Para elegir el número de componentes.
- **ca***row* ∗ ∗*y* ∗ ∗*ca***col**: Para analizar los perfiles de filas (líneas) y columnas, respectivamente.

Paso 1: Selección del Número de Componentes

El primer paso en el Análisis de Correspondencias es determinar el **número de componentes** que se deben utilizar. Esto se realiza analizando la variabilidad explicada por cada componente a través de los valores propios (autovalores).

```
library(FactoMineR)
ca = CA(x, graph = FALSE)
ca$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.5428893               48.69222                           48.69222
## dim 2  0.4450028               39.91269                           88.60491
## dim 3  0.1270484               11.39509                          100.00000
```

En este ejemplo, con **dos componentes** se alcanza un **88.60%** de la variabilidad explicada de los datos, lo cual es un valor suficientemente elevado para realizar el análisis.

Paso 2: Análisis de los Cosenos Cuadrados ($\cos^2$) y Contribuciones

Una vez seleccionados los componentes, es importante analizar los $\cos^2$ y las **contribuciones** de las filas y columnas. Estos valores nos permiten identificar qué modalidades y tareas son las más significativas en cada componente.

```
ca$row
```

Análisis de Perfiles de Filas

```
## $coord
##                 Dim 1      Dim 2       Dim 3
## Laundry    -0.9918368  0.4953220 -0.31672897
## Main_meal  -0.8755855  0.4901092 -0.16406487
## Dinner     -0.6925740  0.3081043 -0.20741377
## Breakfeast -0.5086002  0.4528038  0.22040453
## Tidying    -0.3938084 -0.4343444 -0.09421375
## Dishes     -0.1889641 -0.4419662  0.26694926
```

```
## Shopping   -0.1176813 -0.4033171  0.20261512
## Official    0.2266324  0.2536132  0.92336416
## Driving     0.7417696  0.6534143  0.54445849
## Finances    0.2707669 -0.6178684  0.03479681
## Insurance   0.6470759 -0.4737832 -0.28936051
## Repairs     1.5287787  0.8642647 -0.47208778
## Holidays    0.2524863 -1.4350066 -0.12958665
##
## $contrib
##                 Dim 1      Dim 2       Dim 3
## Laundry    18.2867003  5.5638913  7.96842443
## Main_meal  12.3888433  4.7355230  1.85868941
## Dinner      5.4713982  1.3210221  2.09692603
## Breakfeast  3.8249284  3.6986131  3.06939857
## Tidying     1.9983518  2.9656441  0.48873403
## Dishes      0.4261663  2.8441170  3.63429434
## Shopping    0.1755248  2.5151584  2.22335679
## Official    0.5207837  0.7956201 36.94038942
## Driving     8.0778371  7.6468564 18.59638635
## Finances    0.8750075  5.5585460  0.06175066
## Insurance   6.1470616  4.0203590  5.25263863
## Repairs    40.7300940 15.8806509 16.59639139
## Holidays    1.0773030 42.4539986  1.21261994
##
## $cos2
##                 Dim 1      Dim 2       Dim 3
## Laundry    0.73998741 0.18455213 0.075460467
## Main_meal  0.74160285 0.23235928 0.026037873
## Dinner     0.77664011 0.15370323 0.069656660
## Breakfeast 0.50494329 0.40023001 0.094826699
## Tidying    0.43981243 0.53501508 0.025172490
## Dishes     0.11811778 0.64615253 0.235729693
## Shopping   0.06365362 0.74765514 0.188691242
## Official   0.05304464 0.06642648 0.880528877
## Driving    0.43201860 0.33522911 0.232752289
## Finances   0.16067678 0.83666958 0.002653634
## Insurance  0.57601197 0.30880208 0.115185951
## Repairs    0.70673575 0.22587147 0.067392778
## Holidays   0.02979239 0.96235977 0.007847841
##
## $inertia
##  [1] 0.13415976 0.09069235 0.03824633 0.04112368 0.02466697 0.01958732
##  [7] 0.01497017 0.05330000 0.10150885 0.02956446 0.05793584 0.31287411
## [13] 0.19631064
```

Al observar los **contributos** y los **cosenos cuadrados** ($\cos^2$), notamos que la **primera componente** está asociada principalmente con las tareas de **Lavandería**, **Comida principal**, seguidas de **Cena** y **Desayuno**. La **segunda componente**, por otro lado, está asociada principalmente

con la **Conducción**.

```
ca$col
```

Análisis de Perfiles de Columnas

```
## $coord
##                   Dim 1      Dim 2       Dim 3
## Wife        -0.83762154  0.3652207 -0.19991139
## Alternating -0.06218462  0.2915938  0.84858939
## Husband      1.16091847  0.6019199 -0.18885924
## Jointly      0.14942609 -1.0265791 -0.04644302
##
## $contrib
##                  Dim 1     Dim 2      Dim 3
## Wife        44.462018 10.312237 10.8220753
## Alternating  0.103739  2.782794 82.5492464
## Husband     54.233879 17.786612  6.1331792
## Jointly      1.200364 69.118357  0.4954991
##
## $cos2
##                   Dim 1     Dim 2       Dim 3
## Wife        0.801875947 0.1524482 0.045675847
## Alternating 0.004779897 0.1051016 0.890118521
## Husband     0.772026244 0.2075420 0.020431728
## Jointly     0.020705858 0.9772939 0.002000236
##
## $inertia
## [1] 0.3010185 0.1178242 0.3813729 0.3147248
```

En cuanto a la **distribución de las tareas**, la **primera componente** se asocia tanto a las tareas realizadas por la **Mujer** como por el **Hombre**, mientras que la **segunda componente** está más relacionada con las tareas que se realizan **conjuntamente**.

Paso 3: Interpretación de los Resultados

Al analizar los componentes, se puede observar lo siguiente:

- En la **primera componente**, hay una clara **contraposición entre el Hombre y la Mujer**, con sus respectivas tareas. Las tareas del hogar como **Lavandería**, **Comida principal** y **Cena** son realizadas principalmente por la **Esposa**. En cambio, tareas como **Reparaciones** y **Conducción** son realizadas principalmente por el **Marido**.

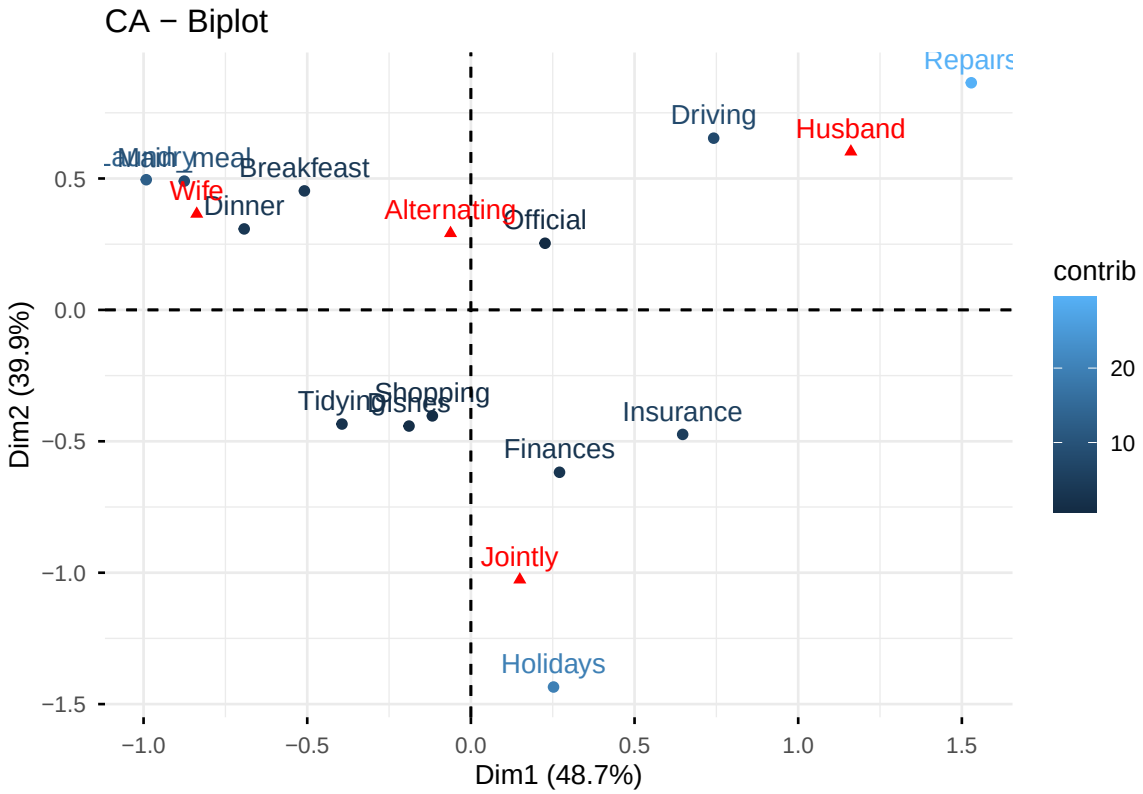
- En la **segunda componente**, hay una contraposición entre las tareas que se realizan **conjuntamente** y las que se realizan de forma **separada**. Actividades como las **vacaciones** se asocian frecuentemente a la columna **conjuntamente**, mientras que otras tareas como las **tareas oficiales** se realizan alternándose entre los miembros de la pareja.

Esta interpretación confirma las observaciones iniciales de la **tabla de contingencia**.

Paso 4: Visualización del Análisis de Correspondencias

Para visualizar los perfiles conjuntamente, podemos generar un gráfico de **biplot** que represente las filas y columnas en el mismo gráfico. En **R**, podemos utilizar la función `fviz_ca_biplot()` de la librería **factoextra** para este propósito:

```
library(factoextra)
fviz_ca_biplot(ca, col.row = "contrib")
```



Este gráfico permite visualizar de forma conjunta las relaciones entre las tareas y su distribución entre los miembros de la pareja, destacando las categorías más significativas en cada componente.

8.2 Tipologías de productos VS tipos de clientes. Análisis de las preferencias de los consumidores

En este ejemplo exploraremos las posibles relaciones entre distintas **categorías de consumidores** y algunos **tipos de productos**. El objetivo principal es investigar si existe una **relación de depen-**

dencia entre los productos y los consumidores, y comprender las **preferencias** de los consumidores hacia los distintos productos.

Los **consumidores** están identificados por letras (A, B, C, D, E, F, G) y representan siete diferentes categorías. Por otro lado, los **productos** se clasifican en cuatro tipos, identificados como **p1**, **p2**, **p3**, y **p4**.

Objetivo del Caso de Estudio

La empresa interesada en este análisis busca dos objetivos principales: 1. Verificar si existe una **relación de dependencia** entre los tipos de productos y los consumidores. 2. Investigar las **preferencias de los consumidores** hacia los productos para comprender qué tipo de cliente prefiere o compra los distintos productos.

Para ello, se dispone de las preferencias de una muestra de **1268 consumidores**.

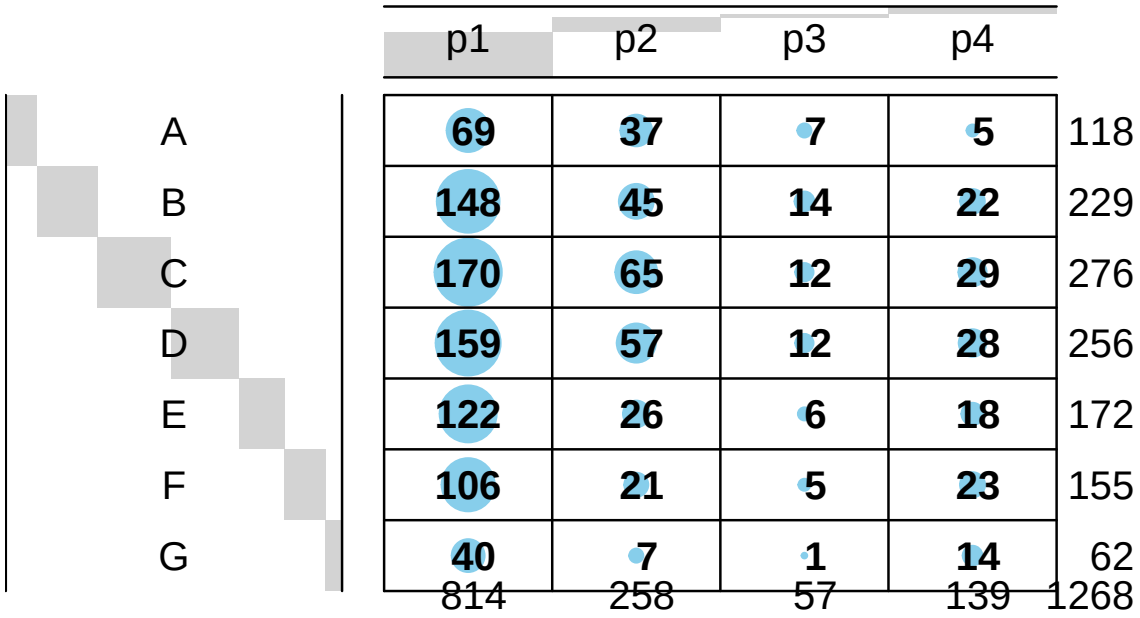
1. Estadística Descriptiva

El primer paso consiste en analizar las posibles relaciones entre las categorías de consumidores y los tipos de productos utilizando **estadística descriptiva**. Este análisis preliminar nos permitirá observar cualquier patrón evidente en las preferencias de los consumidores.

```
# cargo las liberias necesarias
library(FactoMineR)
library(gplots)
library(factoextra)

# leo mis datos
consumer_Product = read.csv(file="consumer_Product.csv", header = TRUE,sep=",")
x=consumer_Product[,-1]
# uso los nombre de los tipos de consumidores como etiquetas de las
# lineas de mi datos
rownames(x) = t(consumer_Product[,1])
```

Podemos observar que el producto 1 es el mas apreciado por los consumidores en particular por los tipos de consumidores B, C, D. Los productos p2 y p3 también son mas apreciados por los consumidores A, B, C, D mientras que en el caso del producto p4 las preferencias son mas distribuidas.

```
dt = as.table(as.matrix(x))
balloonplot(t(dt), main ="", xlab ="", ylab="", show.margins = TRUE)
```



En el segundo paso, se utiliza el **test de Chi-cuadrado** (X^2) para verificar si existe una **relación significativa** entre las dos variables: los **tipos de consumidores** y los **tipos de productos**. Las hipótesis para el test son:

- H_0: Las variables son independientes (no hay relación entre consumidores y productos).
- H_1: Las variables son dependientes (existe una relación significativa entre consumidores y productos).

El test se lleva a cabo con un nivel de significancia $\alpha = 0.05$.

Ya que el p-valor es igual a 0.0027 podemos rechazar la hipótesis nula y concluir que las variables son dependientes (existe una relación significativa entre consumidores y productos).

```
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 39.087, df = 18, p-value = 0.002774
```

Finalmente, se realiza un **Análisis de Correspondencias** para explorar las **relaciones más importantes** entre los tipos de productos y las categorías de consumidores. Este análisis permitirá visualizar las asociaciones más relevantes y extraer conclusiones sobre las **preferencias de los consumidores**.

```
ca = CA(x, graph=FALSE)
ca$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## dim 1 0.025964511              84.229284                           84.22928
## dim 2 0.003809127              12.356870                           96.58615
## dim 3 0.001052352               3.413846                          100.00000
```

En el caso del **Análisis de Correspondencias**, el primer paso es determinar el **número de componentes** a retener. En este análisis, hemos observado que con **dos componentes** se logra explicar un **96.58%** de la variabilidad total de los datos, lo cual es un valor lo suficientemente elevado para proceder con la interpretación.

Una vez seleccionado el número de componentes, el siguiente paso es analizar los **cosenos al cuadrado** (cos^2) y las **contribuciones**. Estos indicadores nos permiten identificar cuáles son los **productos** y **tipos de clientes** más significativos en la construcción de los componentes.

- El cos^2 nos informa sobre la **calidad de la representación** de un punto (producto o cliente) en los ejes seleccionados. Cuanto más cercano a 1 esté el valor de cos^2, mejor estará representado el punto en el plano.
- Las **contribuciones** muestran la **importancia** de cada modalidad (producto o cliente) en la definición de los componentes, permitiéndonos identificar las categorías más relevantes en el análisis.

Este enfoque nos ayudará a entender mejor las asociaciones clave entre los productos y los diferentes tipos de clientes en el espacio reducido generado por el análisis.

```
ca$row
```

```
## $coord
##         Dim 1       Dim 2        Dim 3
## A  0.33104845  0.01797362 -0.024388630
## B  0.03552894 -0.05461529  0.059357867
## C  0.06269292  0.04647640 -0.015659408
## D  0.03796448  0.03298613  0.006364536
## E -0.10196655 -0.10235180 -0.046040964
## F -0.20601242 -0.01654219 -0.005672455
## G -0.39922296  0.14971975  0.012513846
##
## $contrib
##        Dim 1      Dim 2      Dim 3
## A 39.2794804  0.7892390  5.2598888
## B  0.8780121 14.1422708 60.4660767
## C  3.2949340 12.3432458  5.0720028
## D  1.1207155  5.7671078  0.7771291
## E  5.4318096 37.3056556 27.3235413
## F 19.9810875  0.8781588  0.3737607
## G 30.0139609 28.7743222  0.7276006
##
## $cos2
```

```
##       Dim 1      Dim 2        Dim 3
## A 0.9916944 0.00292325 0.0053823208
## B 0.1624904 0.38396509 0.4535444911
## C 0.6203604 0.34093541 0.0387041676
## D 0.5608408 0.42339691 0.0157622535
## E 0.4521920 0.45561538 0.0921926281
## F 0.9928458 0.00640149 0.0007527264
## G 0.8759417 0.12319763 0.0008606485
##
## $inertia
## [1] 0.0102841405 0.0014029846 0.0013790588 0.0005188429 0.0031189026
## [6] 0.0052253750 0.0088966856
```

En el caso de los **consumidores**, podemos observar que **A**, **F**, y **G** están bien representados por la **primera componente** y son los que más contribuyen a su explicación. Por otro lado, los consumidores **E**, **B**, **C**, y **D** están mejor representados por la **segunda componente**.

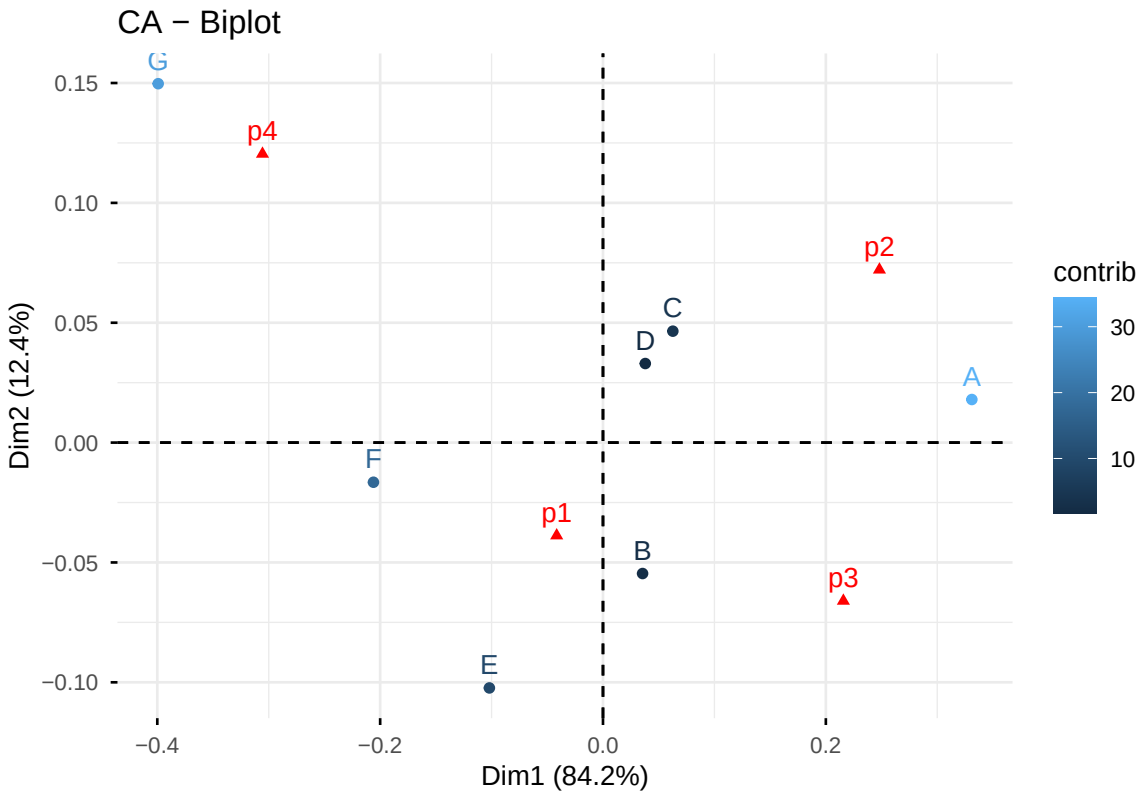
```
ca$col
```

```
## $coord
##          Dim 1       Dim 2       Dim 3
## p1 -0.04156158 -0.03878894 -0.01006256
## p2  0.24811045  0.07209681 -0.01373413
## p3  0.21574454 -0.06604735  0.13880268
## p4 -0.30560296  0.12041667  0.02750052
##
## $contrib
##        Dim 1     Dim 2     Dim 3
## p1  4.270798 25.356859  6.176760
## p2 48.240347 27.765602  3.647048
## p3  8.058516  5.148028 82.298187
## p4 39.430339 41.729511  7.878005
##
## $cos2
##        Dim 1      Dim 2       Dim 3
## p1 0.5182299 0.45139239 0.030377707
## p2 0.9195377 0.07764466 0.002817614
## p3 0.6632886 0.06216325 0.274548178
## p4 0.8595812 0.13345813 0.006960710
##
## $inertia
## [1] 0.002139768 0.013621377 0.003154516 0.011910329
```

En cuanto a los **productos**, los productos **p2** y **p4** están bien representados por la **primera componente**, mientras que **p1** está mejor representado por la **segunda componente**. En el caso del producto **p3**, su mejor representación se encuentra en la **tercera componente**.

Para visualizar estas relaciones, generamos un gráfico biplot utilizando la siguiente función:

```r
fviz_ca_biplot(ca, col.row = "contrib")
```



Interpretación del Gráfico

Al observar el gráfico biplot, se puede destacar lo siguiente: - El producto **p2** es claramente preferido por la categoría de consumidores **A**, mientras que es elegido muy poco por los consumidores **F**. - Existe una relación fuerte entre los consumidores **G** y el producto **p4**. - Los consumidores **C** y **D** muestran características más promedio, ya que se encuentran bastante cerca del **centro de los ejes**, lo que indica que no están fuertemente asociados a ningún producto en particular. - Los consumidores **B** tienen una mayor preferencia por los productos **p1** y **p3**.

Esta visualización nos permite identificar claramente las relaciones más importantes entre los diferentes consumidores y los productos.

8.3 Encuesta Consumo Tea

Este caso de estudio se basa en una encuesta realizada a **300 consumidores** para analizar sus hábitos y opiniones sobre el **consumo de té**. La encuesta se dividió en dos bloques principales:

En este primer bloque, los participantes respondieron preguntas relacionadas directamente con su **consumo de té**. Estas preguntas incluyen aspectos como: - **Frecuencia de consumo**: ¿Con qué frecuencia consumen té? - **Preferencia por tipos de té**: ¿Prefieren té negro, verde, de hierbas, entre otros? - **Opiniones sobre el consumo de té**: ¿Qué piensan del té como una bebida saludable o como una alternativa a otras bebidas?

El segundo bloque incluye preguntas **más descriptivas**, donde se recogió información adicional sobre los participantes. Estas variables permiten segmentar a los consumidores y entender sus hábitos desde un contexto socio-demográfico. Las preguntas en este bloque abarcan: - **Sexo**: Masculino o femenino. - **Edad**: Variable continua que representa la edad de los consumidores. - **Categoría socio-profesional**: Nivel de ocupación o sector laboral de los participantes. - **Práctica deportiva**: Si los participantes practican deportes regularmente o no.

Excepto por la variable **edad**, todas las demás son **variables categóricas**, lo que facilita su análisis mediante técnicas de análisis de correspondencias o segmentación por grupos.

Este estudio permite explorar tanto las **preferencias de consumo** de té como identificar posibles **patrones de comportamiento** asociados a las características socio-demográficas de los consumidores, proporcionando información valiosa para segmentar mejor el mercado del té. Mejoras: Estructura organizada: El texto está clar

MCA

```
# Loading FactoMineR
library(FactoMineR)
data(tea)
# MCA with the graphs given by default
mca <- MCA(tea, quanti.sup=19, quali.sup=c(20:36), graph=FALSE)
```

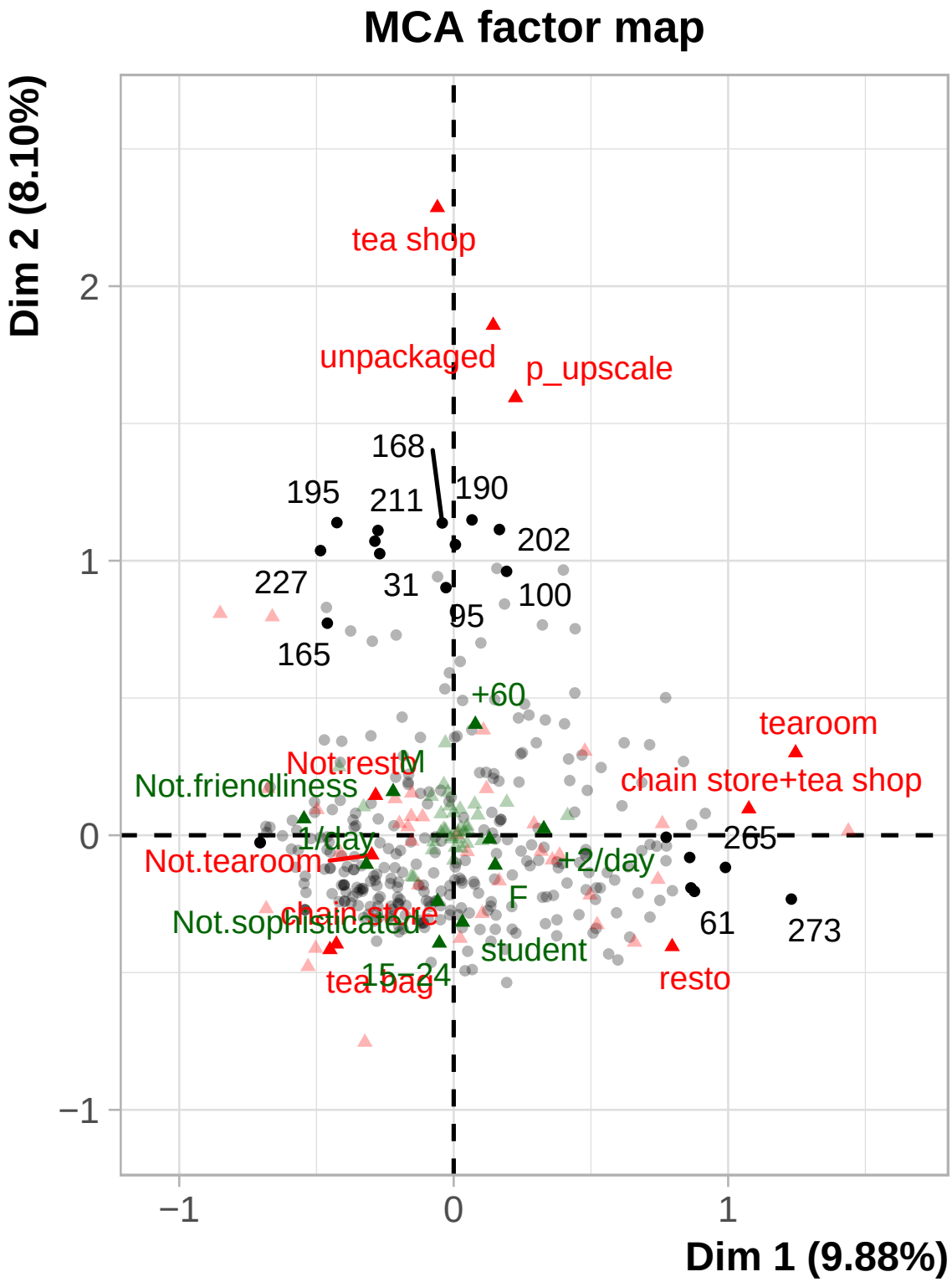
Dado que la mayoría de las variables de la encuesta son **cualitativas**, se decidió realizar un **Análisis de Correspondencias Múltiples (ACM)** para explorar las relaciones entre las variables y los individuos. En este análisis, se consideraron todas las **variables descriptivas** (sexo, edad, categoría socio-profesional y práctica deportiva) como **variables suplementarias**.

Observaciones Iniciales Sin entrar en los detalles técnicos del análisis, se puede observar que la **interpretación de los gráficos** resultantes es **bastante compleja**. Esto se debe a que tanto los **individuos** como las **variables** tienden a distribuirse de manera relativamente **homogénea** alrededor del centro de los ejes, lo que dificulta la identificación de patrones claros o asociaciones directas entre los grupos.

Consideraciones a partir de Observaciones Extremas A pesar de la distribución homogénea en general, aún podemos extraer **algunas conclusiones útiles** al observar las posiciones **extremas** en los gráficos. Las observaciones situadas en los extremos de los ejes suelen ser las más relevantes, ya que tienden a representar individuos o variables que se **diferencian significativamente** del promedio.

Estas observaciones extremas pueden proporcionar información clave para: - **Identificar grupos de consumidores específicos** con comportamientos o características particulares. - **Resaltar las categorías de productos o perfiles demográficos** que muestran preferencias o comportamientos marcadamente diferentes.

```r
# Graph with some labels
plot(mca, autoLab="y",cex=0.7, select="cos2 20", selectMod="cos2 10")
```



En este análisis, destacamos algunos **patrones de consumo** específicos entre los individuos:

- Los individuos **265** y **273** son **bebedores frecuentes** de té y lo consumen en diversas ocasiones, es decir, en cualquier situación.
- Por otro lado, los individuos **200** y **262** consumen té de manera más restringida, limitándose a beberlo **en casa**, ya sea **durante el desayuno** o **por la noche**.

Relaciones entre las Variables y Dimensiones En cuanto a las variables del estudio, observamos que **"precio"**, **"dónde"**, y **"cómo"** están fuertemente relacionadas con las **primeras dos dimensiones** del análisis. Sin embargo, para obtener una interpretación más profunda y detallada de estas relaciones, es necesario una **representación gráfica** de las categorías, que nos permita visualizar con mayor claridad cómo se distribuyen y contrastan estas variables.

Interpretación de las Dimensiones La **primera dimensión** parece reflejar una contraposición entre varias categorías. Por un lado, tenemos: - **"Salón de té"** - **"Tienda de cadena + tienda de té"** - **"Bolsa de té + sin embalar"** - **"Pub"** - **"Restaurante"** - **"Trabajo"**

Estas categorías se asocian con contextos más sociales o comerciales, y se oponen a las siguientes categorías, que implican la **ausencia de estas situaciones**: - **"No amigos"** - **"No restaurante"** - **"No funciona"** - **"No hogar"**

Esta dimensión también parece diferenciar entre los **bebedores regulares de té** y los **bebedores ocasionales**.

La **segunda dimensión** establece una contraposición entre: - **"Tienda especializada"** - **"Sin embalaje"** - **"Precio exclusivo"**

Estas categorías, que implican un consumo más selectivo y premium, se oponen a las **categorías más comunes** relacionadas con el consumo de té.

Aunque la **interpretación** de los resultados del **Análisis de Correspondencias Múltiples (ACM)** puede ser complicada debido a la distribución homogénea de las categorías y los individuos en los ejes, toda la **información contenida en las variables** sigue siendo útil. Esta información puede aprovecharse de manera efectiva para realizar un **Análisis Clúster**, utilizando las **coordenadas de los individuos** en los ejes del ACM.

Análisis Clúster basado en el ACM

El **Análisis Clúster** nos permite identificar **grupos específicos de consumidores** que comparten características o patrones similares en sus respuestas y comportamientos. Al utilizar las **coordenadas sobre los ejes** obtenidas en el ACM, podemos agrupar a los consumidores según su **proximidad en el espacio multidimensional** generado por el análisis.

Este enfoque combina los beneficios del ACM y el Análisis Clúster: - **ACM**: Reduce la dimensionalidad del espacio de datos y representa las relaciones entre las variables de manera gráfica. - **Clúster**: Identifica **subgrupos homogéneos** dentro de los consumidores, lo que facilita la segmentación.

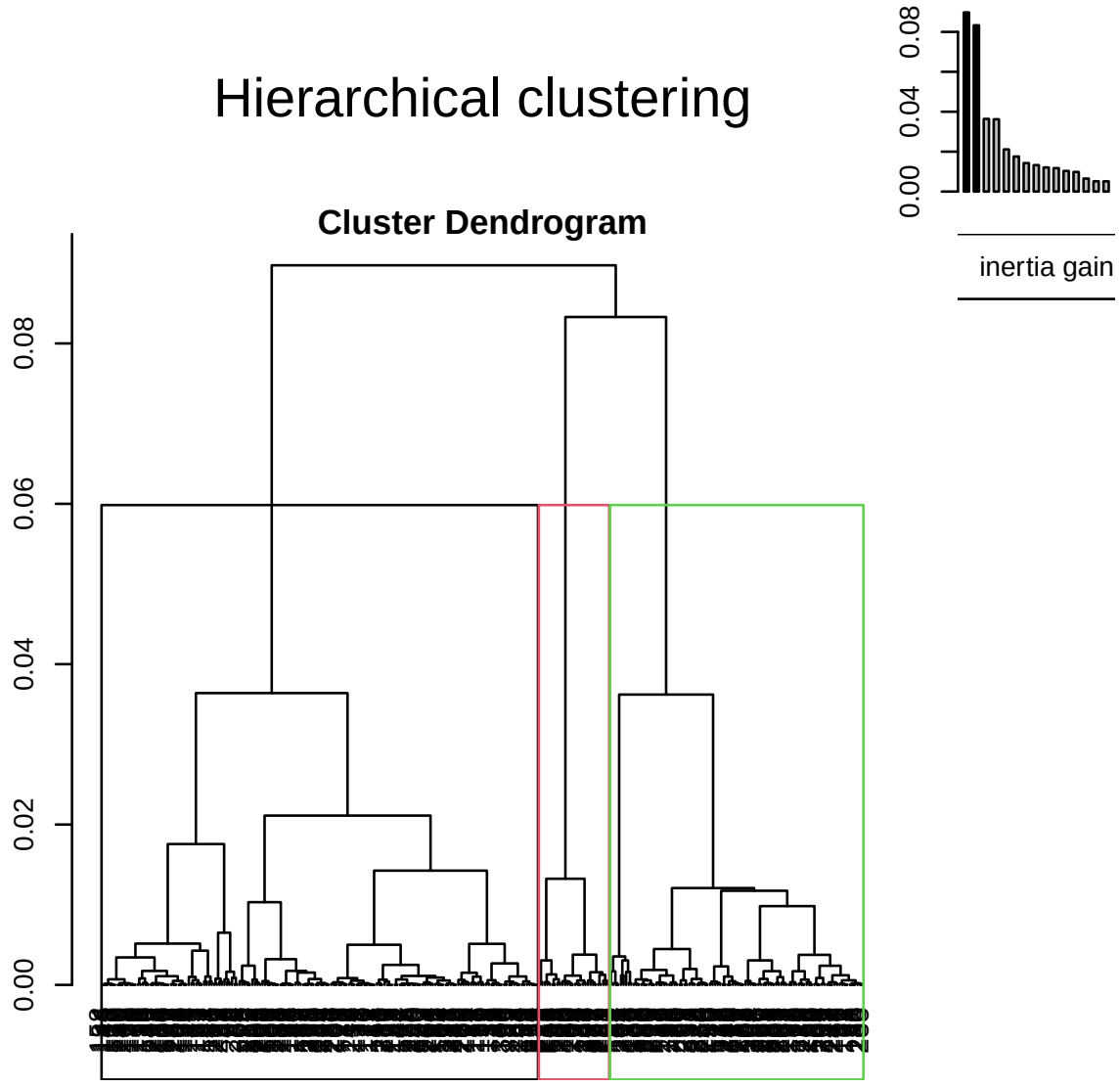
Este método tiene varias ventajas: 1. **Identificación de patrones**: A pesar de la dificultad en la interpretación de los gráficos del ACM, el Análisis Clúster nos permite identificar patrones claros en el comportamiento de los consumidores. 2. **Segmentación basada en coordenadas**: Las **coordenadas sobre los ejes** proporcionan una base sólida para segmentar a los individuos en grupos significativos. 3. **Aplicación práctica**: Estos grupos pueden ser utilizados para adaptar

estrategias de marketing, crear productos específicos para cada segmento, o mejorar la comunicación con los consumidores según sus características.

```
# Análisis cluster
hc <- HCPC(mca, nb.clust=-1, graph=FALSE)
#Numero de grupos identificados por el algoritmo:
hc$call$t$nb.clust
```

```
## [1] 3
```

```
plot(hc, choice= "tree")
```




```
as.data.frame(round(hc$desc.var$category$`1`,3))
```

```
##                                Cla/Mod Mod/Cla Global p.value v.test
```

```
## where=chain store              82.292  92.398 64.000   0.000 12.207
## how=tea bag                    81.765  81.287 56.667   0.000 10.099
## tearoom=Not.tearoom            69.008  97.661 80.667   0.000  8.896
## price=p_branded                83.158  46.199 31.667   0.000  6.405
## friends=Not.friends            77.885  47.368 34.667   0.000  5.401
## pub=Not.pub                    64.979  90.058 79.000   0.000  5.383
## resto=Not.resto                64.253  83.041 73.667   0.000  4.203
## tea.time=Not.tea time          68.702  52.632 43.667   0.000  3.602
## price=p_private label          90.476  11.111  7.000   0.001  3.352
## sugar=sugar                    66.897  56.725 48.333   0.001  3.339
## frequency=1/day                69.474  38.596 31.667   0.003  2.976
## work=Not.work                  61.972  77.193 71.000   0.007  2.693
## age_Q=15-24                    68.478  36.842 30.667   0.008  2.672
## always=Not.always              62.437  71.930 65.667   0.009  2.607
## price=p_unknown                91.667   6.433  4.000   0.012  2.525
## price=p_cheap                 100.000   4.094  2.333   0.019  2.355
## lunch=Not.lunch                59.766  89.474 85.333   0.022  2.294
## frequency=1 to 2/week          72.727  18.713 14.667   0.022  2.287
## How=alone                      61.538  70.175 65.000   0.032  2.145
## slimming=slimming              71.111  18.713 15.000   0.038  2.073
## friendliness=Not.friendliness  68.966  23.392 19.333   0.041  2.047
## friendliness=friendliness      54.132  76.608 80.667   0.041 -2.047
## slimming=No.slimming           54.510  81.287 85.000   0.038 -2.073
## lunch=lunch                    40.909  10.526 14.667   0.022 -2.294
## always=always                  46.602  28.070 34.333   0.009 -2.607
## work=work                      44.828  22.807 29.000   0.007 -2.693
## How=lemon                      30.303   5.848 11.000   0.001 -3.227
## sugar=No.sugar                 47.742  43.275 51.667   0.001 -3.339
## How=other                       0.000   0.000  3.000   0.000 -3.523
## tea.time=tea time              47.929  47.368 56.333   0.000 -3.602
## price=p_variable               41.964  27.485 37.333   0.000 -4.033
## resto=resto                    36.709  16.959 26.333   0.000 -4.203
## frequency=+2/day               42.520  31.579 42.333   0.000 -4.321
## pub=pub                        26.984   9.942 21.000   0.000 -5.383
## friends=friends                45.918  52.632 65.333   0.000 -5.401
## where=tea shop                 10.000   1.754 10.000   0.000 -5.570
## how=unpackaged                 13.889   2.924 12.000   0.000 -5.616
## how=tea bag+unpackaged         28.723  15.789 31.333   0.000 -6.674
## price=p_upscale                15.094   4.678 17.667   0.000 -6.863
## tearoom=tearoom                 6.897   2.339 19.333   0.000 -8.896
## where=chain store+tea shop     12.821   5.848 26.000   0.000 -9.357
```

El **64%** de la muestra total compra su té en una **tienda de cadena (chain store)**. Dentro de este grupo, el **82.3%** pertenece al **Clúster 1**, lo que indica que esta modalidad de compra es predominante en este segmento.

- Dentro del **Clúster 1**, **el 92%** de los consumidores compra té en una tienda de cadena.
- Prefieren comprar el té en **bolsas**, en lugar de otras formas de presentación.

- No frecuentan **salones de té** y no toman té en **pubs**.
- No suelen beber té en compañía de **amigos**.
- Tienen una fuerte preferencia por comprar **té de marca**.

```
as.data.frame(round(hc$desc.var$category$`2`,3))
```

```
##                                      Cla/Mod Mod/Cla Global p.value v.test
## where=tea shop                        86.667  81.250 10.000   0.000 10.904
## how=unpackaged                        66.667  75.000 12.000   0.000  9.157
## price=p_upscale                       50.943  84.375 17.667   0.000  8.895
## resto=Not.resto                       13.575  93.750 73.667   0.003  2.934
## Tea=green                             27.273  28.125 11.000   0.004  2.845
## sophisticated=sophisticated           13.488  90.625 71.667   0.008  2.649
## sex=M                                 16.393  62.500 40.667   0.010  2.593
## escape.exoticism=Not.escape-exoticism 14.557  71.875 52.667   0.022  2.294
## escape.exoticism=escape-exoticism      6.338  28.125 47.333   0.022 -2.294
## how=tea bag+unpackaged                 4.255  12.500 31.333   0.012 -2.520
## sex=F                                  6.742  37.500 59.333   0.010 -2.593
## sophisticated=Not.sophisticated        3.529   9.375 28.333   0.008 -2.649
## Tea=Earl Grey                          6.736  40.625 64.333   0.004 -2.851
## where=chain store+tea shop             2.564   6.250 26.000   0.004 -2.895
## resto=resto                            2.532   6.250 26.333   0.003 -2.934
## age_Q=15-24                            2.174   6.250 30.667   0.001 -3.427
## price=p_branded                        2.105   6.250 31.667   0.000 -3.538
## price=p_variable                       2.679   9.375 37.333   0.000 -3.668
## how=tea bag                            2.353  12.500 56.667   0.000 -5.394
## where=chain store                      2.083  12.500 64.000   0.000 -6.302
```

El **Clúster 2** agrupa a los consumidores que compran té en **tiendas especializadas**. Este grupo se diferencia claramente del primero por las siguientes características:

- Prefieren comprar **té sin confeccionar** (a granel).
- Optan por **té de alto costo**, lo que refleja un interés por productos premium o exclusivos.
- Son consumidores más exigentes, lo que sugiere una preferencia por productos de calidad superior, a menudo asociados con marcas o tiendas especializadas en té.

```
as.data.frame(round(hc$desc.var$category$`3`,3))
```

```
##                             Cla/Mod Mod/Cla Global p.value v.test
## where=chain store+tea shop   84.615  68.041 26.000   0.000 11.356
## how=tea bag+unpackaged       67.021  64.948 31.333   0.000  8.530
## tearoom=tearoom              77.586  46.392 19.333   0.000  7.925
## friends=friends              44.898  90.722 65.333   0.000  6.753
## price=p_variable             55.357  63.918 37.333   0.000  6.503
## resto=resto                  60.759  49.485 26.333   0.000  6.110
## pub=pub                      63.492  41.237 21.000   0.000  5.731
```

```
## How=other                  100.000   9.278  3.000   0.000  4.175
## frequency=+2/day            44.882  58.763 42.333   0.000  3.945
## tea.time=tea time           41.420  72.165 56.333   0.000  3.843
## work=work                   47.126  42.268 29.000   0.001  3.422
## lunch=lunch                 52.273  23.711 14.667   0.003  2.943
## sugar=No.sugar              39.355  62.887 51.667   0.007  2.678
## sex=F                       38.202  70.103 59.333   0.009  2.628
## How=lemon                   51.515  17.526 11.000   0.017  2.393
## home=home                   33.333 100.000 97.000   0.028  2.197
## breakfast=breakfast         38.194  56.701 48.000   0.038  2.071
## breakfast=Not.breakfast     26.923  43.299 52.000   0.038 -2.071
## home=Not.home                0.000   0.000  3.000   0.028 -2.197
## How=alone                   27.692  55.670 65.000   0.021 -2.309
## frequency=1/day             23.158  22.680 31.667   0.020 -2.324
## price=p_private label        9.524   2.062  7.000   0.016 -2.413
## sex=M                       23.770  29.897 40.667   0.009 -2.628
## sugar=sugar                 24.828  37.113 48.333   0.007 -2.678
## lunch=Not.lunch             28.906  76.289 85.333   0.003 -2.943
## Tea=green                    9.091   3.093 11.000   0.001 -3.211
## frequency=1 to 2/week       11.364   5.155 14.667   0.001 -3.386
## work=Not.work               26.291  57.732 71.000   0.001 -3.422
## tea.time=Not.tea time       20.611  27.835 43.667   0.000 -3.843
## where=tea shop               3.333   1.031 10.000   0.000 -3.970
## price=p_branded             14.737  14.433 31.667   0.000 -4.575
## pub=Not.pub                 24.051  58.763 79.000   0.000 -5.731
## resto=Not.resto             22.172  50.515 73.667   0.000 -6.110
## friends=Not.friends          8.654   9.278 34.667   0.000 -6.753
## how=tea bag                 15.882  27.835 56.667   0.000 -6.971
## tearoom=Not.tearoom         21.488  53.608 80.667   0.000 -7.925
## where=chain store           15.625  30.928 64.000   0.000 -8.173
```

El **Clúster 3** se caracteriza por ser un **mix** de las preferencias observadas en los dos primeros grupos. Sin embargo, lo que distingue a este grupo es su **preferencia por consumir té en compañía de amigos**. Este clúster muestra un patrón de consumo más social, en comparación con los otros grupos que consumen té de manera más individual.

Resumiendo:

- El **Clúster 1** agrupa a los consumidores más tradicionales en cuanto a la compra de té, quienes prefieren adquirirlo en tiendas de cadena, generalmente en formato de bolsa, y no participan en contextos sociales como salones de té o pubs para su consumo. Este grupo muestra una **alta lealtad a las marcas** comerciales, lo que representa una oportunidad para las empresas de té en términos de segmentación y personalización de las ofertas.
- El **Clúster 2** identifica a los consumidores que buscan una experiencia más **premium**, comprando té sin confeccionar y a precios elevados en **tiendas especializadas**.
- El **Clúster 3** representa una mezcla de los dos grupos anteriores, pero con una clara preferencia por consumir té en **contextos sociales**, en compañía de amigos.

Anexo 1: Como se calculan los perfiles lineas y columnas

En esta demostración teórica, se explicará cómo obtener los perfiles de filas y columnas a partir de una tabla de contingencia en el análisis de correspondencias.

1. Tabla de Contingencia

Supongamos que tenemos una **tabla de contingencia** N de tamaño $I \times J$, donde: - I es el número de categorías de la primera variable (filas). - J es el número de categorías de la segunda variable (columnas). - n_{ij} representa la frecuencia observada en la intersección de la categoría i de la primera variable con la categoría j de la segunda variable.

El total de la tabla se denota como n, es decir:

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$

2. Cálculo de las Frecuencias Relativas

El primer paso para obtener los perfiles es convertir las frecuencias absolutas en **frecuencias relativas**. Esto se hace dividiendo cada n_{ij} entre el total n:

$$f_{ij} = \frac{n_{ij}}{n}$$

Donde f_{ij} es la frecuencia relativa de la celda (i, j).

3. Perfiles de Filas

El **perfil de fila** está compuesto por las frecuencias relativas condicionales de cada categoría de fila. Para la fila i, su perfil se obtiene dividiendo cada celda n_{ij} por el total de la fila $n_{i.}$, que es la suma de todas las frecuencias de esa fila:

$$p_{ij} = \frac{n_{ij}}{n_{i.}} \quad \text{donde} \quad n_{i.} = \sum_{j=1}^{J} n_{ij}$$

De esta manera, el **perfil de la fila** i es el vector:

$$(p_{i1}, p_{i2}, \dots, p_{iJ})$$

Este vector describe cómo la categoría de la fila i se distribuye en las categorías de las columnas.

4. Perfiles de Columnas

De manera similar, el **perfil de columna** está compuesto por las frecuencias relativas condicionales de cada categoría de columna. Para la columna j, su perfil se obtiene dividiendo cada celda n_{ij} por el total de la columna $n_{.j}$, que es la suma de todas las frecuencias de esa columna:

$$q_{ij} = \frac{n_{ij}}{n_{.j}} \quad \text{donde} \quad n_{.j} = \sum_{i=1}^{I} n_{ij}$$

De esta forma, el **perfil de la columna** j es el vector:

$$(q_{1j}, q_{2j}, \cdots, q_{Ij})$$

Este vector describe cómo la categoría de la columna j se distribuye en las categorías de las filas.

5. Ejemplo Numérico

Consideremos una tabla de contingencia N de 3 filas (categorías de una variable) y 3 columnas (categorías de la segunda variable):

$$N = \begin{pmatrix} 10 & 15 & 25 \\ 30 & 20 & 50 \\ 40 & 35 & 75 \end{pmatrix}$$

El total de la tabla es:

$$n = 10 + 15 + 25 + 30 + 20 + 50 + 40 + 35 + 75 = 300$$

Perfiles de Filas

Para la primera fila:

$$n_{1.} = 10 + 15 + 25 = 50$$

Los perfiles de fila se calculan dividiendo cada valor de la fila 1 entre $n_{1.}$:

$$p_{11} = \frac{10}{50} = 0.2, \quad p_{12} = \frac{15}{50} = 0.3, \quad p_{13} = \frac{25}{50} = 0.5$$

El perfil de la fila 1 es:

$$(0.2, 0.3, 0.5)$$

Perfiles de Columnas

Para la primera columna:

$$n_{\cdot 1} = 10 + 30 + 40 = 80$$

Los perfiles de columna se calculan dividiendo cada valor de la columna 1 entre $n_{\cdot 1}$:

$$q_{11} = \frac{10}{80} = 0.125, \quad q_{21} = \frac{30}{80} = 0.375, \quad q_{31} = \frac{40}{80} = 0.5$$

El perfil de la columna 1 es:

$$(0.125, 0.375, 0.5)$$

Anexo 2: Como se calculan los perfiles lineas y columnas

En este documento, explicaremos cómo obtener la **matriz de Burt** y proporcionaremos un ejemplo de su cálculo.

1. Matriz Disyuntiva Completa

El primer paso en el ACM es construir la **matriz disyuntiva completa**, que convierte las variables categóricas en **variables dummy**. Si tenemos I individuos y p variables con distintas modalidades, la matriz disyuntiva completa Z tendrá:

- **Filas**: Corresponden a los **individuos** (total I).
- **Columnas**: Corresponden a las **modalidades de las variables** (total $\sum k$, donde k es el número de categorías por variable).

Cada celda de la matriz disyuntiva es 1 si el individuo pertenece a esa categoría, y 0 en caso contrario.

Ejemplo

Supongamos que tenemos 3 individuos y 2 variables categóricas:

- Variable 1: con categorías A_1, A_2
- Variable 2: con categorías B_1, B_2, B_3

La matriz disyuntiva completa Z para estos datos sería:

$$Z = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

2. Matriz de Burt

La **matriz de Burt** se obtiene multiplicando la **matriz disyuntiva completa** Z por su transpuesta Z':

$$B = Z'Z$$

La matriz de Burt es una **matriz simétrica** que organiza la información de las variables de manera que: - Los **bloques diagonales** contienen las **tablas de contingencia** de cada variable consigo misma, es decir, las frecuencias marginales. - Los **bloques fuera de la diagonal** contienen las **tablas de contingencia cruzadas** entre pares de variables.

Estructura de la Matriz de Burt

La estructura de la matriz de Burt es la siguiente:

$$B = \begin{pmatrix} A_1 \times A_1 & A_1 \times B_1 & A_1 \times B_2 & \dots & A_1 \times B_p \\ A_2 \times A_2 & A_2 \times B_1 & A_2 \times B_2 & \dots & A_2 \times B_p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ B_1 \times B_1 & B_1 \times B_2 & B_1 \times B_3 & \dots & B_1 \times B_p \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Cada bloque de la matriz contiene una subtabla de contingencia. Por ejemplo: - El bloque $A_1 \times A_1$ contiene la tabla de contingencia de la primera variable categórica consigo misma. - El bloque $A_1 \times B_1$ contiene la tabla de contingencia cruzada entre las modalidades A_1 y B_1.

Ejemplo de Cálculo

Para el ejemplo anterior, donde Z es:

$$Z = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

La matriz de Burt $B = Z'Z$ sería:

$$B = \begin{pmatrix} 2 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Interpretación

- Los **bloques diagonales** en la matriz de Burt contienen las **frecuencias marginales** de cada categoría. Por ejemplo, el primer bloque de la diagonal muestra que la primera categoría de la primera variable aparece dos veces en la muestra.
- Los **bloques fuera de la diagonal** representan las **relaciones entre las categorías** de las diferentes variables, es decir, las **frecuencias conjuntas** de las categorías de dos variables.

Anexo 3 Calculo de las componentes en el analisis de las correspondencias:

En este documento, se presenta una demostración teórica sobre cómo se calculan las **componentes principales** en un Análisis de Correspondencias, a partir de la tabla de contingencia.

1. Tabla de Contingencia

Dado un conjunto de variables categóricas, el Análisis de Correspondencias comienza con una **tabla de contingencia** N, donde: - I representa el número de categorías de la primera variable (filas). - J representa el número de categorías de la segunda variable (columnas). - n_{ij} es la frecuencia observada en la intersección de la categoría i de la primera variable y la categoría j de la segunda variable.

El total de la tabla se denota como n, es decir:

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$

2. Frecuencias Relativas

El siguiente paso es calcular las **frecuencias relativas** de cada celda en la tabla de contingencia:

$$f_{ij} = \frac{n_{ij}}{n}$$

Donde f_{ij} representa la proporción de la frecuencia observada n_{ij} respecto al total de la tabla.

Las **frecuencias marginales** de las filas y columnas se calculan como:

$$f_{i.} = \sum_{j=1}^{J} f_{ij} \quad \text{(frecuencia marginal de la fila } i\text{)}$$

$$f_{.j} = \sum_{i=1}^{I} f_{ij} \quad \text{(frecuencia marginal de la columna } j\text{)}$$

3. Matriz de Inercia

La **inercia** en el Análisis de Correspondencias es una medida de la variabilidad o dispersión en los datos. La inercia total es la suma de los cuadrados de las diferencias entre las frecuencias observadas y las frecuencias esperadas bajo la hipótesis de independencia entre las filas y las columnas.

La **frecuencia esperada** bajo la hipótesis de independencia entre las filas y las columnas es:

$$f_{ij}^* = f_{i.} \times f_{.j}$$

La inercia total se define como:

$$\text{Inercia total} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}$$

4. Matriz de Residuos (Matriz de Desviación)

La **matriz de desviación** o **matriz de residuos** se utiliza para medir las desviaciones entre las frecuencias observadas y las frecuencias esperadas. Para cada celda de la tabla de contingencia, calculamos el valor de:

$$d_{ij} = \frac{f_{ij} - f_{ij}^*}{\sqrt{f_{ij}^*}}$$

La matriz de residuos contiene los valores de d_{ij} y refleja qué tan lejos están las frecuencias observadas de las frecuencias esperadas.

5. Descomposición en Valores Singulares

El siguiente paso es aplicar la **descomposición en valores singulares (SVD)** a la **matriz de residuos**. Este es el paso clave para calcular las **componentes principales** en el Análisis de Correspondencias.

Si representamos la matriz de residuos como D, la **SVD** de D nos permite descomponer la matriz en tres matrices: una matriz de vectores singulares de filas, una matriz diagonal con los valores singulares, y una matriz de vectores singulares de columnas:

$$D = U \Sigma V'$$

Donde: - U contiene los **vectores singulares** de las filas (las coordenadas de las filas en los ejes principales). - V contiene los **vectores singulares** de las columnas (las coordenadas de las columnas en los ejes principales). - Σ es una matriz diagonal que contiene los **valores singulares** asociados a cada componente.

Los **valores singulares** en Σ permiten calcular la **inercia explicada** por cada componente. Cuanto mayor sea un valor singular, más variabilidad explica el componente correspondiente.

6. Cálculo de las Coordenadas

Las **coordenadas** de las filas y las columnas en el espacio de componentes principales se calculan a partir de los vectores singulares. Las coordenadas de las filas en el componente k se calculan como:

$$\text{Coordenada de fila } i \text{ en el eje } k = u_{ik} \times \sigma_k$$

Donde: - u_{ik} es el elemento i-ésimo del vector singular correspondiente a la fila i en el componente k. - σ_k es el valor singular correspondiente al componente k.

De manera similar, las coordenadas de las columnas en el componente k se calculan como:

$$\text{Coordenada de columna } j \text{ en el eje } k = v_{jk} \times \sigma_k$$

Donde: - v_{jk} es el elemento j-ésimo del vector singular correspondiente a la columna j en el componente k.

7. Interpretación de las Componentes

Las **componentes principales** obtenidas a través de la descomposición en valores singulares nos permiten proyectar las categorías de las filas y las columnas en un **espacio reducido**. Los **valores singulares** indican la **importancia** de cada componente, y las **coordenadas** permiten visualizar cómo se relacionan las categorías entre sí.

Inercia Explicada

La **inercia explicada** por cada componente es una medida de cuánta variabilidad en los datos es capturada por ese componente. Se calcula como:

$$\text{Inercia explicada por el componente } k = \frac{\sigma_k^2}{\text{Inercia total}}$$

PLS-SEM

1. Introducción

La primera fase de una investigación científica consiste en definir el fenómeno que se desea analizar. Esta definición formal del objeto de estudio es una tarea compleja y, a la vez, fundamental, ya que condiciona estrictamente las medidas que se podrán emplear durante la investigación.

En general, es posible distinguir dos grandes tipos de fenómenos. El primer grupo lo constituyen los conceptos que son observables directamente en la realidad.

En este caso, las medidas aplicables se concretan en la recopilación de información sobre el objeto, considerando tanto propiedades físicas (como el volumen, la temperatura y la longitud), como características cualitativas (por ejemplo, el color, el olor y la forma).

La segunda tipología incluye aquellos fenómenos que no son observables de manera directa y que hacen referencia a conceptos abstractos y teóricos. Ejemplos de este tipo de fenómenos se encuentran en diversas disciplinas como la psicología, la sociología y la economía. Términos como motivación, satisfacción o estrato social son comunes en estas áreas, pero resulta difícil especificarlos de manera concreta.

En estos casos, el trabajo del investigador se torna más complejo, pues requiere la habilidad de identificar el conjunto de relaciones entre diferentes variables que permitan obtener, de manera indirecta, una "medida" del fenómeno.

En este capítulo se ofrece una introducción al concepto de variable latente, con el objetivo de responder a algunas preguntas clave sobre su naturaleza, las distintas formas de medirlas, la manera en que se relacionan entre sí, y los dos enfoques principales desarrollados para estudiarlas y estimarlas. Estos enfoques constituyen un paso esencial para comprender los temas tratados en los capítulos siguientes. Finalmente, se presenta una sección sobre la simbología y las notaciones que se emplearán a lo largo del trabajo.

2. Variables latentes

Con frecuencia, un investigador se enfrenta al problema de que la variable que desea analizar no puede medirse de manera directa. Ejemplos de este tipo de variables incluyen la satisfacción, la motivación, la fidelidad, y en general, todas aquellas actitudes relacionadas con los comportamientos humanos. Estas variables, conocidas como variables latentes, son muy comunes en disciplinas sociales como la sociología, psicología, economía y política. La preponderancia de ejemplos provenientes de la psicología se debe a que este concepto fue utilizado por primera vez en dicha área.

En el ámbito estadístico, las variables latentes son empleadas en diversos análisis y técnicas de modelización con aplicaciones en múltiples campos del conocimiento. Este uso multidisciplinar ha conducido a la aparición de varias definiciones de lo que constituye una variable latente.

Una variable latente se puede definir a partir de las siguientes características:

- No es posible observarla directamente en la realidad.
- Puede considerarse una herramienta de reducción de datos, ya que permite expresar de manera resumida el conjunto de factores que la representan.

- Es útil para comprender las relaciones entre variables que no pueden explicarse de forma directa.
- Puede ser considerada como una variable hipotética.

Sin embargo, la característica más distintiva de una variable latente es que no puede observarse directamente en la realidad. Esta consideración implica la necesidad de identificar un conjunto de indicadores (variables manifiestas) que permitan medirla de forma indirecta.

3. ¿Cómo se construye una variable latente? Relaciones entre variables latentes y manifiestas.

Para construir una variable latente, es fundamental tener en cuenta dos aspectos clave de su definición: primero, una variable latente no es observable directamente en la realidad; segundo, es necesario identificar un conjunto de indicadores (variables manifiestas) que permitan recabar una medida indirecta de esta. Con estos dos elementos en mente, el siguiente paso es comprender cómo se relacionan las variables latentes y las manifiestas. Existen dos formas principales de relación:

1. **Mediante los efectos que la variable latente ejerce sobre los indicadores**.
2. **Mediante los diferentes indicadores que se asumen como "causa" de la variable latente**.

El primer caso se conoce como **modalidad reflexiva**, donde se considera que las variables manifiestas son causadas por la variable latente. El segundo caso se denomina **modalidad formativa**, ya que en este escenario, las variables manifiestas generan la variable latente.

La diferencia fundamental entre las modalidades reflexiva y formativa radica en la relación de causa y efecto entre los indicadores y el constructo. Para ilustrar mejor este concepto, consideremos un ejemplo.

Imaginemos que el objetivo de una investigación es medir el estado de embriaguez en un grupo de jóvenes. Dado que este es un concepto abstracto y no observable directamente, el investigador deberá identificar un conjunto de indicadores que le permitan medir la variable latente. Podría optar por la **modalidad formativa** midiendo, por ejemplo, el nivel de vino, cerveza o whisky en la sangre de los jóvenes. Alternativamente, podría emplear la **modalidad reflexiva**, midiendo el tiempo que un joven en estado de embriaguez puede caminar en línea recta, cuánto tiempo puede mantener la vista en un mismo punto, o cuánto tiempo puede mantenerse en equilibrio sobre una pierna, entre otros indicadores.

Las modalidades reflexiva y formativa para medir la variable latente en este ejemplo se ilustran en la Figura 1.1.

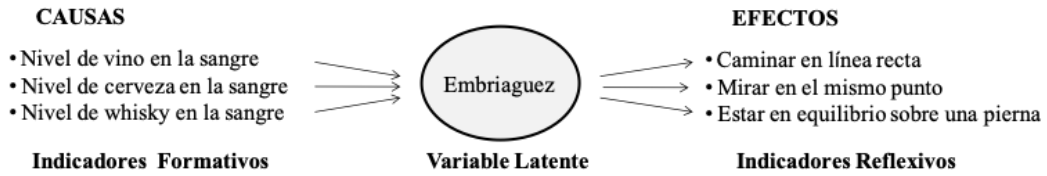


Figura 1.1. Ejemplos de variable latente medida mediante indicadores reflexivos y formativos.

4. ¿Cómo se relacionan las variables latentes? Modelos de ecuaciones estructurales

Un aspecto particularmente interesante en la investigación de variables latentes es la forma en que estas se relacionan entre sí. En la mayoría de los estudios, el objetivo del investigador no es simplemente estimar una variable, sino analizar las relaciones dentro de un sistema complejo de variables latentes. Un ejemplo de esto es el modelo del **European Customer Satisfaction Index (ECSI)**, que mide la satisfacción del cliente teniendo en cuenta un conjunto de variables: Imagen, Expectativas del Cliente, Calidad Percibida del Producto, Calidad del Servicio y Valor del Servicio.

Sin entrar en detalle sobre el modelo ECSI, lo que resulta relevante es observar cómo el objetivo no es solo estimar la satisfacción del cliente, sino también comprender cómo esta variable se relaciona con otras en el sistema.

Para alcanzar este objetivo, es necesario expresar las relaciones entre las variables de manera algebraica. La herramienta empleada para esto es el **modelo de ecuaciones estructurales (SEM)**.

El SEM se divide en dos componentes principales: el **modelo estructural** y el **modelo de medida**.

4.1. Modelo estructural

El modelo estructural especifica las relaciones lineales entre las variables latentes del sistema. Estas relaciones pueden expresarse de la siguiente forma:

$$\eta = B\eta + \Gamma\xi + \zeta$$

Donde: - η representa las variables latentes endógenas. - B es una matriz de coeficientes que define las relaciones entre las variables latentes endógenas. - Γ es una matriz de coeficientes que define las relaciones entre las variables latentes exógenas (ξ) y las endógenas (η). - ζ representa el término de error del modelo estructural.

4.2. Modelo de medida

El modelo de medida especifica las relaciones entre las variables latentes y las variables observadas (indicadores). Estas relaciones se representan de la siguiente forma:

Para las variables latentes exógenas (ξ):

$$x = \Lambda_x\xi + \delta$$

Y para las variables latentes endógenas (η):

$$y = \Lambda_y\eta + \varepsilon$$

Donde: - x representa las variables observadas que corresponden a las variables latentes exógenas (ξ). - y representa las variables observadas que corresponden a las variables latentes endógenas (η). - Λ_x y Λ_y son matrices de carga factorial que definen las relaciones entre las variables latentes y las observadas. - δ y ε son términos de error del modelo de medida.

4.3. Ejemplo aplicado: Modelo ECSI

En el modelo **ECSI**, la satisfacción del cliente (η_1) depende de varias variables latentes exógenas, como la Imagen (ξ_1), las Expectativas del Cliente (ξ_2) y la Calidad Percibida (ξ_3). Las ecuaciones del modelo estructural podrían expresarse así:

$$\eta_1 = \Gamma_1\xi_1 + \Gamma_2\xi_2 + \Gamma_3\xi_3 + \zeta_1$$

El modelo de medida, que vincula las variables observadas con las latentes, sería:

$$x_1 = \Lambda_x\xi_1 + \delta_1$$
$$y_1 = \Lambda_y\eta_1 + \varepsilon_1$$

Aquí, x_1 podría representar indicadores observados como las calificaciones de Imagen de la empresa, mientras que y_1 serían indicadores observados de la satisfacción del cliente.

5. Dos enfoques posibles

En el ámbito de las técnicas desarrolladas para investigar las relaciones entre variables latentes, es posible identificar dos grandes enfoques: **Hard-Modeling** y **Soft-Modeling**. Cada uno de estos enfoques ofrece herramientas distintas para estimar modelos de ecuaciones estructurales (SEM).

5.1. SEM Clásico (Hard-Modeling)

El **SEM-ML** (Maximum Likelihood Approach to Structural Equation Modeling), también conocido como **LISREL** (Linear Structural Relations), representa el enfoque clásico dentro del Hard-Modeling. Este método se basa en la estructura de la matriz de varianza-covarianza y es ampliamente utilizado para estimar parámetros que reflejan relaciones causales entre las variables latentes.

No obstante, el SEM clásico presenta varias limitaciones importantes:

1. **Soluciones impropias**: Pueden surgir problemas como varianzas negativas o coeficientes de correlación mayores que uno, frecuentemente debido a una mala especificación del modelo.
2. **Ambigüedad en los factores**: Dos modelos pueden presentar índices de ajuste similares, pero con correlaciones opuestas entre factores, lo que dificulta la discriminación entre ellos.
3. **Convergencia**: El algoritmo puede no llegar a la convergencia, especialmente en modelos complejos.

Adicionalmente, el SEM clásico requiere que se cumplan ciertas suposiciones, como la normalidad multivariante de los datos y el uso de muestras grandes. A pesar de estos requisitos, su principal ventaja es que proporciona estimaciones consistentes y óptimas, siempre que los datos sigan las condiciones adecuadas.

5.2. PLS-SEM (Soft-Modeling)

Como alternativa al SEM clásico, el **PLS-PM** (Partial Least Squares Path Modeling) surge dentro del enfoque Soft-Modeling. Este método fue desarrollado para superar algunas de las limitaciones del SEM clásico. A diferencia de SEM-ML, **PLS-SEM** no requiere suposiciones tan estrictas sobre la distribución de los datos y puede trabajar eficazmente con muestras más pequeñas. Además, permite el uso de variables categóricas mediante la creación de variables dummy.

Entre las ventajas de PLS-SEM se incluyen:

- No requiere que los datos sigan una distribución normal.
- Funciona bien con muestras de tamaño reducido, siempre que haya más observaciones que variables en cada bloque de medición.
- Permite el uso de variables categóricas.

Sin embargo, PLS-SEM presenta inconvenientes relacionados con la calidad de las estimaciones, ya que su consistencia solo está garantizada a medida que el tamaño del conjunto de datos aumenta.

5.3. Comparación entre SEM Clásico y PLS-SEM

Existe un debate en la comunidad científica sobre si es adecuado o no utilizar PLS-SEM para la estimación de modelos de ecuaciones estructurales. Este debate divide a los investigadores en dos grupos. Por un lado, están aquellos que consideran que el PLS no es un enfoque suficientemente riguroso para estimar modelos SEM y reconocen el **CB-SEM** (Covariance-Based SEM, como el SEM clásico) como la única opción válida para estimar modelos. Por otro lado, hay quienes reconocen las ventajas del PLS-SEM, especialmente cuando el CB-SEM no puede aplicarse debido a la naturaleza de los datos o al tipo de teoría que se desea explorar o probar.

En lugar de posicionarse en este debate, es importante destacar que ambos enfoques tienen sus ventajas y desventajas. El **SEM clásico** es más adecuado cuando se cumplen las suposiciones sobre la distribución de los datos y el tamaño de la muestra, y cuando el objetivo es la estimación precisa de parámetros en modelos bien especificados. Sin embargo, **PLS-SEM** es una herramienta más flexible, útil cuando no se cumplen las condiciones necesarias para el SEM clásico, o cuando el principal objetivo es la predicción en lugar de la inferencia causal.

5.4. Ejemplo Aplicado

Consideremos el modelo **ECSI (European Customer Satisfaction Index)**, que mide la satisfacción del cliente mediante variables latentes como la Imagen de la empresa, Expectativas del cliente, Calidad percibida del producto y Valor percibido. Si el objetivo del investigador es obtener estimaciones precisas de las relaciones causales entre estas variables latentes y la satisfacción del cliente, el SEM clásico sería la mejor opción, siempre y cuando se disponga de una muestra grande y los datos sigan una distribución normal.

Por el contrario, si los datos no cumplen con estas condiciones estrictas, o si el objetivo principal es predecir la satisfacción del cliente a partir de estas variables latentes, el **PLS-SEM** sería más adecuado debido a su capacidad para manejar muestras más pequeñas y datos con distribuciones no normales.

6. PLS-PM el modelo

El **PLS-PM** (Partial Least Squares Path Modeling) es una metodología estadística desarrollada para el análisis de modelos estructurales de variables latentes. A diferencia de LISREL, el objetivo principal del PLS es obtener la mejor predicción posible de las variables latentes sin preocuparse por explicar todas las covariaciones entre los indicadores del modelo. De acuerdo con este enfoque, el PLS estima los parámetros de manera que la varianza residual de todas las variables dependientes sea mínima.

6.1. Especificación del Modelo

El modelo PLS se divide en dos componentes principales:

1. **Modelo estructural (modelo interno)**: Relaciona las variables latentes endógenas con otras variables latentes en el sistema.
2. **Modelo de medidas (modelo externo)**: Analiza las relaciones entre las variables manifiestas y las variables latentes.

6.1.1. Modelo Estructural

El **modelo estructural** describe las relaciones de causa-efecto entre las variables latentes. Estas asociaciones pueden ser representadas por un sistema de ecuaciones lineales recursivo. En este contexto, las variables latentes (VL) pueden actuar como variables de respuesta (variables latentes endógenas) o como variables explicativas (variables latentes exógenas).

La estructura general del modelo estructural se expresa de la siguiente manera:

$$\eta = B\eta + \beta\xi + \zeta$$

Donde: - η representa las variables latentes endógenas. - B es la matriz de coeficientes que define las relaciones entre las variables latentes endógenas. - β es la matriz de coeficientes que define las relaciones entre las variables latentes exógenas (ξ) y las endógenas (η). - ζ es el término de error del modelo estructural.

6.1.2. Modelo de Medidas

Dado que una **variable latente** no puede ser observada directamente, es necesario identificar un conjunto de **variables manifiestas** que permitan obtener una estimación indirecta de la variable latente. Existen dos enfoques principales para relacionar las VL con las variables manifiestas (VM): el **método reflexivo** y el **método formativo**.

Método Reflexivo En el método reflexivo, las **variables latentes** son consideradas como las causas de las variables manifiestas. Esto significa que las variables manifiestas son “manifestaciones” o expresiones de las variables latentes.

La relación entre las variables latentes (ξ_j) y las variables manifiestas (x_{jk}) se puede expresar de la siguiente manera:

$$x_{jk} = \lambda_{jk}\xi_j + \delta_{jk}$$

Donde: - x_{jk} es la variable manifiesta asociada a la variable latente ξ_j. - λ_{jk} es la carga factorial que indica la influencia de la variable latente sobre la variable manifiesta. - δ_{jk} es el término de error del modelo reflexivo.

Este modelo es útil cuando se considera que las variables manifiestas son efectos o consecuencias de las variables latentes.

Método Formativo En el método formativo, la **variable latente** es vista como una combinación lineal de las variables manifiestas. Es decir, las variables manifiestas "generan" la variable latente.

La relación se puede expresar de la siguiente manera:

$$\xi_j = \sum_k \beta_{jk} x_{jk} + \zeta_j$$

Donde: - ξ_j es la variable latente que se forma a partir de las variables manifiestas x_{jk}. - β_{jk} es el peso de cada variable manifiesta en la formación de la variable latente. - ζ_j es el término de error del modelo formativo.

Este modelo es adecuado cuando las variables manifiestas son consideradas causas que generan la variable latente.

6.2. Ejemplo Teórico

Imaginemos que queremos modelar la **satisfacción del cliente** en una empresa. En este caso, la satisfacción del cliente (η_1) es la variable latente endógena y está relacionada con varias variables latentes exógenas como la **calidad percibida del producto** (ξ_1) y la **imagen de la empresa** (ξ_2).

6.2.1. Modelo estructural (relación entre variables latentes)

El modelo estructural que relaciona la satisfacción del cliente (η_1) con la calidad percibida (ξ_1) y la imagen de la empresa (ξ_2) podría representarse como:

$$\eta_1 = \beta_1\xi_1 + \beta_2\xi_2 + \zeta_1$$

Donde: - η_1 es la satisfacción del cliente. - ξ_1 es la calidad percibida. - ξ_2 es la imagen de la empresa. - β_1 y β_2 son los coeficientes que representan las relaciones causales. - ζ_1 es el término de error.

6.2.2. Modelo de medidas (método reflexivo)

Supongamos que la satisfacción del cliente (η_1) se mide a través de tres variables manifiestas: **recompra** (y_1), **recomendación** (y_2) y **lealtad** (y_3). Estas variables manifiestas son consideradas como manifestaciones de la satisfacción latente, lo que nos lleva a un modelo reflexivo:

$$y_1 = \lambda_1 \eta_1 + \varepsilon_1$$
$$y_2 = \lambda_2 \eta_1 + \varepsilon_2$$
$$y_3 = \lambda_3 \eta_1 + \varepsilon_3$$

Donde: - y_1, y_2 y y_3 son las variables manifiestas (recompra, recomendación y lealtad). - λ_1, λ_2 y λ_3 son las cargas factoriales. - ε_1, ε_2 y ε_3 son los términos de error.

6.2.3. Modelo de medidas (método formativo)

Por otro lado, si modelamos la calidad percibida del producto (ξ_1) como una combinación de varias características del producto, el modelo formativo sería:

$$\xi_1 = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \zeta_2$$

Donde: - x_1, x_2 y x_3 son variables manifiestas que describen diferentes aspectos del producto (como durabilidad, diseño y funcionalidad). - β_1, β_2 y β_3 son los coeficientes que miden el peso de cada variable manifiesta en la formación de la calidad percibida. - ζ_2 es el término de error.

7. Algoritmo PLS-PM

7.1. Descripción Intuitiva

El **PLS-PM (Partial Least Squares Path Modeling)** es un algoritmo iterativo diseñado para maximizar la varianza explicada de las variables dependientes. Este método no requiere supuestos sobre la normalidad de los datos ni tamaños de muestra grandes, lo que lo hace adecuado para una amplia gama de aplicaciones.

A nivel intuitivo, el algoritmo del PLS-PM sigue estos pasos:

1. **Inicialización**: Se asignan valores iniciales a las variables latentes, generalmente utilizando una combinación lineal de sus indicadores.
2. **Estimación de los pesos**: Se calculan los pesos de los indicadores para cada variable latente, con el objetivo de maximizar la covarianza entre las variables latentes y sus manifestaciones.
3. **Estimación de las variables latentes**: Con los pesos calculados, se actualizan las estimaciones de las variables latentes como una combinación ponderada de los indicadores.
4. **Repetición del proceso**: Los pasos anteriores se repiten iterativamente hasta que las estimaciones convergen.
5. **Estimación de los coeficientes estructurales**: Finalmente, se estiman los coeficientes del modelo estructural, que describen las relaciones entre las variables latentes.

7.2. Descripción Técnica

A continuación, describimos los pasos del algoritmo PLS-PM de forma técnica:

7.2.1. Paso 1: Inicialización

Dado un conjunto de variables manifiestas $X = [x_1, x_2, ..., x_k]$, para cada bloque de variables manifiestas asociado a una variable latente ξ_j, se asignan valores iniciales a ξ_j. Por lo general, esto se hace tomando una combinación lineal de los indicadores:

$$\xi_j^{(0)} = \frac{1}{k} \sum_{i=1}^{k} x_{ij}$$

7.2.2. Paso 2: Estimación de los pesos

Los pesos de los indicadores se estiman de forma iterativa para maximizar la covarianza entre la variable latente ξ_j y sus manifestaciones x_{ij}. Los pesos se estiman de la siguiente manera:

$$w_{ij} = \frac{\sum_{i=1}^{k} \text{cov}(\xi_j, x_{ij})}{\sum_{i=1}^{k} x_{ij}^2}$$

7.2.3. Paso 3: Estimación de las variables latentes

Una vez calculados los pesos, las variables latentes se actualizan como una combinación ponderada de sus variables manifiestas:

$$\xi_j^{(t+1)} = \sum_{i=1}^{k} w_{ij} x_{ij}$$

Este proceso se repite iterativamente hasta que las estimaciones de ξ_j convergen, es decir, hasta que $\|\xi_j^{(t+1)} - \xi_j^{(t)}\|$ es suficientemente pequeño.

7.2.4. Paso 4: Estimación de los coeficientes estructurales

Una vez que las variables latentes han sido estimadas, se procede a estimar los coeficientes del modelo estructural. Estos coeficientes describen las relaciones lineales entre las variables latentes y se calculan utilizando regresión ordinaria entre las variables latentes:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Donde X es la matriz de las variables latentes exógenas y y son las variables latentes endógenas.

7.3. Ejemplo Calculado a Mano

Vamos a calcular un pequeño ejemplo paso a paso. Supongamos que tenemos un conjunto de datos con una variable latente ξ_1 que está relacionada con dos variables manifiestas x_1 y x_2.

7.3.1. Paso 1: Inicialización

Supongamos que las variables manifiestas x_1 y x_2 tienen los siguientes valores:

$$x_1 = [2, 3, 4, 5]$$
$$x_2 = [1, 2, 3, 4]$$

Calculamos la media de las variables manifiestas para obtener una estimación inicial de la variable latente ξ_1:

$$\xi_1^{(0)} = \frac{x_1 + x_2}{2} = \frac{[2, 3, 4, 5] + [1, 2, 3, 4]}{2} = [1.5, 2.5, 3.5, 4.5]$$

7.3.2. Paso 2: Estimación de los pesos

Ahora calculamos los pesos w_1 y w_2 utilizando la covarianza entre ξ_1 y x_1, y entre ξ_1 y x_2. Supongamos que las covarianzas son:

$$\text{cov}(\xi_1, x_1) = 1, \quad \text{cov}(\xi_1, x_2) = 0.8$$

Los pesos se calculan como:

$$w_1 = \frac{\text{cov}(\xi_1, x_1)}{\sum x_1^2} = \frac{1}{54} = 0.0185$$
$$w_2 = \frac{\text{cov}(\xi_1, x_2)}{\sum x_2^2} = \frac{0.8}{30} = 0.0267$$

7.3.3. Paso 3: Actualización de las variables latentes

Con los pesos calculados, actualizamos la variable latente ξ_1:

$$\xi_1^{(1)} = w_1 \cdot x_1 + w_2 \cdot x_2$$
$$\xi_1^{(1)} = 0.0185 \cdot [2, 3, 4, 5] + 0.0267 \cdot [1, 2, 3, 4] = [0.059, 0.112, 0.165, 0.218]$$

7.3.4. Paso 4: Estimación de los coeficientes estructurales

Finalmente, si quisiéramos estimar las relaciones entre las variables latentes, podríamos ajustar un modelo de regresión utilizando los valores de ξ_1 y otras variables latentes (si existieran), siguiendo la fórmula:

$$\hat{\beta} = (X'X)^{-1}X'y$$

8. Validación del Modelo PLS

La validación del modelo PLS se divide en dos partes: la validación del **modelo de medidas** (para las relaciones entre variables latentes y sus indicadores) y la validación del **modelo estructural** (para las relaciones entre las variables latentes). A continuación, explicamos los criterios de validación para ambos componentes.

8.1. Validación del Modelo de Medidas

La validación del modelo de medidas depende de si la relación entre las variables latentes (LV) y las variables manifiestas (MV) sigue un enfoque reflexivo o formativo. Cada modalidad tiene sus propios criterios de validación.

8.1.1. Validación del Modelo de Medidas: Modalidad Reflexiva

Cuando las LV causan las MV, es decir, cuando se aplica el modelo reflexivo, los indicadores reflejan el constructo latente. Por tanto, es crucial verificar los siguientes aspectos:

1. **Unidimensionalidad de los Indicadores**
2. **Explicación de los Indicadores por las Variables Latentes**

1. Unidimensionalidad de los Indicadores El objetivo es asegurar que los indicadores dentro de cada bloque midan una única variable latente. Para evaluar esto, se pueden aplicar tres criterios:

- **Análisis de Componentes Principales (PCA)**: Se realiza un análisis de componentes principales de los indicadores. Para que un bloque sea unidimensional, el primer autovalor debe ser mayor que 1, y el segundo autovalor debe ser mucho menor que 1. Esto sugiere que una sola componente principal explica la mayor parte de la varianza de los indicadores.

  Ejemplo: Si aplicamos PCA y obtenemos autovalores de 2.5 y 0.4 para un bloque de indicadores, podemos concluir que el bloque es unidimensional.

- **Alfa de Cronbach**: Evalúa la consistencia interna de los indicadores dentro de un bloque. Si el alfa es mayor que 0.7, el bloque se considera unidimensional.

  Fórmula del Alfa de Cronbach:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_{y_i}^2}{\sigma_{y_{total}}^2}\right)$$

Donde:

- k es el número de indicadores.
- $\sigma_{y_i}^2$ es la varianza del i-ésimo indicador.
- $\sigma_{y_{total}}^2$ es la varianza total.

Ejemplo Teórico: Supongamos que tenemos tres indicadores con varianzas de 1.2, 1.3 y 1.4. La varianza total es 4.5. El alfa de Cronbach sería:

$$\alpha = \frac{3}{2}\left(1 - \frac{1.2 + 1.3 + 1.4}{4.5}\right) = 0.73$$

2. Explicación de los Indicadores por las Variables Latentes El **AVE (Average Variance Extracted)** es otro criterio importante, que mide el porcentaje de varianza explicada de los indicadores por su variable latente. Se recomienda que el AVE sea mayor a 0.50, lo que implica que más del 50% de la varianza de los indicadores es explicada por la variable latente.

Fórmula del AVE:

$$AVE = \frac{\sum_{i=1}^{k}\lambda_i^2}{k}$$

Donde: - λ_i son las cargas factoriales de los indicadores. - k es el número de indicadores.

Ejemplo Teórico: Supongamos que tenemos tres indicadores con cargas factoriales $\lambda_1 = 0.8$, $\lambda_2 = 0.75$ y $\lambda_3 = 0.7$. El AVE sería:

$$AVE = \frac{0.8^2 + 0.75^2 + 0.7^2}{3} = \frac{0.64 + 0.5625 + 0.49}{3} = 0.564$$

Dado que el AVE es mayor que 0.50, podemos concluir que más del 50% de la varianza de los indicadores es explicada por la variable latente.

8.1.2.Validación del Modelo de Medidas: Modalidad Formativa

En el modelo formativo, las variables manifiestas **forman** la variable latente. Aquí, es importante analizar los pesos de los indicadores para determinar la importancia de cada uno en la formación de la variable latente. Los pesos indican qué tanto contribuye cada indicador a la variable latente.

Interpretación: Se espera que los indicadores más importantes tengan mayores pesos absolutos. Sin embargo, los indicadores con pesos bajos pueden seguir siendo relevantes si están teóricamente justificados.

Ejemplo Teórico: Supongamos que tenemos tres indicadores con los siguientes pesos:

- $w_1 = 0.6$

- $w_2 = 0.4$
- $w_3 = 0.2$

Aquí, el indicador w_1 es el que más contribuye a la formación de la variable latente, mientras que w_3 tiene la menor contribución.

8.2. Validación del Modelo Estructural

La validación del modelo estructural se enfoca en analizar las relaciones entre las variables latentes, similares a un análisis de regresión lineal múltiple. Los principales criterios son:

1. **Coeficientes Estructurales (β)**
2. **R-cuadrado (R^2)**
3. **Significancia de los Coeficientes (Bootstrap)**

1. Coeficientes Estructurales (β) Los coeficientes estructurales (β) representan las relaciones causales entre las variables latentes. Estos coeficientes se interpretan de forma similar a los coeficientes de regresión lineal. Valores altos de β sugieren una relación fuerte entre las variables latentes.

Fórmula para los Coeficientes Estructurales:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Donde: - X es la matriz de las variables latentes exógenas. - y es la variable latente endógena.

Ejemplo Teórico: Supongamos que tenemos una relación entre dos variables latentes, donde el coeficiente estimado es $\beta = 0.75$. Esto indica una relación fuerte y positiva entre las dos variables latentes.

2. R-cuadrado (R^2) El R^2 mide el poder predictivo del modelo, es decir, el porcentaje de la variabilidad de la variable latente endógena explicada por las variables latentes exógenas. Un R^2 cercano a 1 indica un buen ajuste del modelo.

Fórmula del R^2:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Donde: - y_i son los valores observados. - $\hat{y}_i$ son los valores predichos por el modelo. - $\bar{y}$ es la media de y_i.

Ejemplo Teórico: Si el R^2 es 0.80, esto significa que el 80% de la variabilidad en la variable latente endógena es explicada por las variables latentes exógenas.

3. Significancia de los Coeficientes (Bootstrap) Para evaluar la significancia estadística de los coeficientes estructurales (β), se usa un procedimiento de bootstrap. Este método consiste en tomar múltiples muestras de los datos originales, estimar los coeficientes para cada muestra y calcular los intervalos de confianza.

Si los intervalos de confianza no contienen el valor 0, se puede concluir que el coeficiente es estadísticamente significativo.

Ejemplo Teórico: Supongamos que, después de realizar el bootstrap, obtenemos un intervalo de confianza de 0.60 a 0.90 para el coeficiente β. Como el intervalo no contiene 0, podemos concluir que el coeficiente es significativo.

8.2.1. Criterios Recientes para Evaluar el Modelo Estructural en PLS-PM

En los últimos años, se han introducido criterios más avanzados para evaluar el **modelo estructural** en el PLS-PM, complementando los coeficientes estructurales (β) y el R^2. A continuación, describimos los criterios más relevantes junto con ejemplos teóricos.

f^2 - Tamaño del Efecto El f^2 mide el tamaño del efecto de una variable exógena sobre una variable endógena. Se interpreta de la siguiente manera:

- $f^2 > 0.02$: efecto pequeño.
- $f^2 > 0.15$: efecto medio.
- $f^2 > 0.35$: efecto grande.

Fórmula del f^2:

$$f^2 = \frac{R^2_{incluido} - R^2_{excluido}}{1 - R^2_{incluido}}$$

Donde $R^2_{incluido}$ es el R^2 con el predictor incluido en el modelo y $R^2_{excluido}$ es el R^2 cuando se excluye dicho predictor.

Ejemplo Teórico: Supongamos que tenemos una variable exógena X_1 que influye sobre una variable endógena Y. Al calcular el R^2 con X_1 incluido en el modelo obtenemos $R^2_{incluido} = 0.50$. Luego, calculamos R^2 sin X_1, obteniendo $R^2_{excluido} = 0.40$. El f^2 se calcula de la siguiente forma:

$$f^2 = \frac{0.50 - 0.40}{1 - 0.50} = \frac{0.10}{0.50} = 0.20$$

Dado que el $f^2 = 0.20$, interpretamos que el tamaño del efecto de X_1 sobre Y es **medio**.

Q^2 - Relevancia Predictiva El Q^2 de Stone-Geisser evalúa el poder predictivo del modelo mediante el **procedimiento de blindfolding**. Si el Q^2 es positivo, indica que el modelo tiene relevancia predictiva para la variable endógena.

Fórmula del Q^2:

$$Q^2 = 1 - \frac{\sum(y_{observado} - y_{predicho})^2}{\sum(y_{observado} - \bar{y})^2}$$

Ejemplo Teórico: Supongamos que, para la variable endógena Y, tenemos los siguientes valores observados: $y_{observado} = [10, 12, 15, 13]$, y los valores predichos por el modelo son: $y_{predicho} = [9, 11, 14, 13]$. La media de los valores observados es $\bar{y} = 12.5$.

Primero, calculamos los errores:

$$\sum(y_{observado} - y_{predicho})^2 = (10-9)^2 + (12-11)^2 + (15-14)^2 + (13-13)^2 = 1+1+1+0 = 3$$

Luego, calculamos la varianza total de los valores observados:

$$\sum(y_{observado} - \bar{y})^2 = (10-12.5)^2 + (12-12.5)^2 + (15-12.5)^2 + (13-12.5)^2 = 6.25+0.25+6.25+0.25 = 13$$

Finalmente, calculamos el Q^2:

$$Q^2 = 1 - \frac{3}{13} = 1 - 0.23 = 0.77$$

Dado que el $Q^2 = 0.77$, podemos concluir que el modelo tiene **alta relevancia predictiva**.

SRMR - Ajuste Global del Modelo El **Standardized Root Mean Square Residual (SRMR)** mide el ajuste global del modelo. Un valor de SRMR inferior a 0.08 indica un buen ajuste del modelo estructural.

Fórmula del SRMR:

$$SRMR = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(r_i - \hat{r_i})^2}$$

Donde: - r_i son los residuos observados. - $\hat{r_i}$ son los residuos predichos. - p es el número de indicadores.

Ejemplo Teórico: Supongamos que tenemos tres indicadores con los siguientes residuos observados y predichos:

- $r_1 = 0.10, \hat{r_1} = 0.05$
- $r_2 = 0.20, \hat{r_2} = 0.15$
- $r_3 = 0.15, \hat{r_3} = 0.10$

Calculamos el SRMR:

$$SRMR = \sqrt{\frac{1}{3}\left((0.10-0.05)^2 + (0.20-0.15)^2 + (0.15-0.10)^2\right)}$$

$$SRMR = \sqrt{\frac{1}{3}(0.0025 + 0.0025 + 0.0025)} = \sqrt{\frac{0.0075}{3}} = \sqrt{0.0025} = 0.05$$

Dado que el SRMR es inferior a 0.08, podemos concluir que el modelo tiene **un buen ajuste global**.

Estos criterios avanzados, como el f^2 para el tamaño del efecto, el Q^2 para la relevancia predictiva y el SRMR para el ajuste global del modelo, proporcionan una evaluación más completa y robusta del modelo estructural en PLS-PM, especialmente en escenarios de investigación más complejos. Cada uno de estos criterios ofrece información clave para interpretar la calidad del modelo y su capacidad para predecir y explicar las relaciones entre las variables.

8.3. Tabla de Criterios para Evaluar el Modelo Estructural en PLS-PM

Criterio	¿Para qué se utiliza?	¿Cómo se interpreta?
β **Coeficientes Estructurales**	Evalúa las relaciones entre variables latentes exógenas y endógenas.	Valores altos (cercanos a 1 o -1) indican una relación fuerte; valores cercanos a 0 indican una relación débil.
R^2	Mide el poder explicativo del modelo, es decir, cuánto de la varianza de la variable endógena es explicada por las exógenas.	Un R^2 cercano a 1 indica un poder explicativo alto; R^2 cercano a 0 indica bajo poder explicativo.
f^2	Evalúa el tamaño del efecto de una variable exógena sobre una endógena.	$f^2 > 0.02$ indica un efecto pequeño, $f^2 > 0.15$ efecto medio, y $f^2 > 0.35$ efecto grande.
Q^2	Mide la capacidad predictiva del modelo (usando el procedimiento de blindfolding).	Un Q^2 positivo indica que el modelo tiene relevancia predictiva; valores cercanos a 0 o negativos sugieren que el modelo carece de capacidad predictiva.
SRMR	Evalúa el ajuste global del modelo estructural.	Valores de SRMR menores a 0.08 indican un buen ajuste del modelo.
Alfa de Cronbach	Evalúa la consistencia interna de los indicadores dentro de un bloque (unidimensionalidad).	Un valor de alfa mayor a 0.70 indica una buena consistencia interna; menor a 0.70 sugiere inconsistencia.
AVE (Average Variance Extracted)	Mide el porcentaje de varianza de los indicadores explicado por la variable latente.	Un AVE mayor a 0.50 indica que la variable latente explica más del 50% de la varianza de sus indicadores.

9. PLS-PM in R

El Modelado de Ecuaciones Estructurales de Mínimos Cuadrados Parciales (PLS-PM, por sus siglas en inglés) es una técnica de modelado multivariante usada para analizar relaciones entre variables

latentes. En este documento, exploraremos cómo implementar un modelo PLS-PM utilizando la librería `cSEM` en R.

9.1. Instalación y carga de librerías

Para utilizar `cSEM`, primero debes instalarla si no lo has hecho ya. Puedes instalarla directamente desde CRAN con el siguiente código:

```
# Instala la librería cSEM
install.packages("cSEM")
```

Después de la instalación, cargamos la librería:

```
# Cargar la librería cSEM
library(cSEM)
```

9.2. Descripción del modelo

Supongamos que queremos modelar la relación entre tres variables latentes: **Calidad del Producto**, **Satisfacción del Cliente**, y **Lealtad del Cliente**. Las relaciones entre estas variables se pueden representar en un diagrama estructural, donde la calidad del producto afecta la satisfacción del cliente, y ambas afectan la lealtad del cliente.

9.2.1. Definición del modelo

El modelo puede definirse en R usando las funciones proporcionadas por `cSEM`. Primero, debemos definir las variables latentes y las manifestaciones de cada una.

```
# Definición del modelo estructural
model <- "
  # Variables latentes
  Calidad_Producto =~ x1 + x2 + x3
  Satisfaccion_Cliente =~ y1 + y2 + y3
  Lealtad_Cliente =~ z1 + z2 + z3

  # Relaciones estructurales
  Satisfaccion_Cliente ~ Calidad_Producto
  Lealtad_Cliente ~ Calidad_Producto + Satisfaccion_Cliente
"
```

- `Calidad_Producto`, `Satisfaccion_Cliente` y `Lealtad_Cliente` son variables latentes.
- `x1`, `x2`, `x3`, `y1`, `y2`, `y3`, `z1`, `z2`, `z3` son las manifestaciones observables (indicadores) de estas variables.

9.3. Estimación del modelo

Para estimar el modelo definido, usamos la función `csem()`. Esta función requiere al menos dos parámetros: los datos y el modelo.

```
# Datos de ejemplo (se deben proporcionar o simular)
data <- your_data_frame_here  # Reemplazar con tu conjunto de datos

# Estimación del modelo PLS-PM
result <- csem(data = data, model = model)
```

Los principales parámetros son:

- `data`: El conjunto de datos que contiene los indicadores observables.
- `model`: El modelo estructural definido anteriormente.

9.4. Resultados del modelo

Una vez que se ha estimado el modelo, se pueden extraer diferentes tipos de resultados, como las cargas de los indicadores, los coeficientes de las relaciones estructurales, y los índices de ajuste. Aquí mostramos algunas de las funciones más importantes para analizar los resultados:

9.4.1. Cargas de los indicadores

```
# Obtener las cargas de los indicadores
loadings <- result$Estimates$Loadings
print(loadings)
```

9.4.2. Coeficientes estructurales

```
# Obtener los coeficientes estructurales
path_coefficients <- result$Estimates$Path_coefficients
print(path_coefficients)
```

9.4.3. Índices de ajuste

```
# Obtener los índices de ajuste
fit_indices <- assess(result)
print(fit_indices)
```

9.5. Ejemplo completo

A continuación se muestra un ejemplo completo de la estimación y validación de un modelo PLS-PM con datos simulados:

```
# Simulación de datos (solo para propósitos de ejemplo)
set.seed(123)
data <- data.frame(
  x1 = rnorm(100), x2 = rnorm(100), x3 = rnorm(100),
  y1 = rnorm(100), y2 = rnorm(100), y3 = rnorm(100),
  z1 = rnorm(100), z2 = rnorm(100), z3 = rnorm(100)
)

# Definición del modelo estructural
model <- "
  Calidad_Producto =~ x1 + x2 + x3
  Satisfaccion_Cliente =~ y1 + y2 + y3
  Lealtad_Cliente =~ z1 + z2 + z3

  Satisfaccion_Cliente ~ Calidad_Producto
  Lealtad_Cliente ~ Calidad_Producto + Satisfaccion_Cliente
"

# Estimación del modelo PLS-PM
result <- csem(data = data, model = model)

# Mostrar resultados
print(result$Estimates$Loadings)            # Cargas de los indicadores
print(result$Estimates$Path_coefficients)   # Coeficientes estructurales
print(assess(result))
```