



INCASI Working Paper Series

2018, No. 3



INCASI *International Network for
Comparative Analysis of Social Inequalities*



Promedio de los efectos marginales e interacciones en las regresiones logísticas binarias

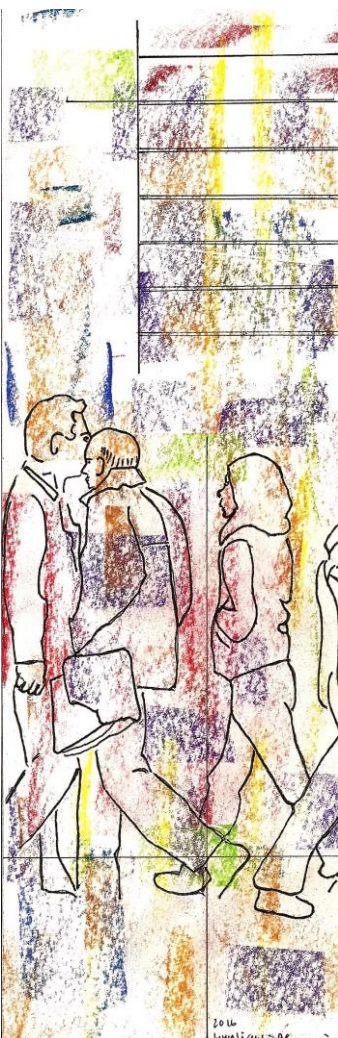
Matías Salvador Ballesteros



European
Commission

Horizon 2020
European Union funding
for Research & Innovation

Marie Skłodowska-Curie Actions (MSCA)
Research and Innovation Staff Exchange (RISE)
H2020-MSCA-RISE-2015
GA-691004



Promedio de los efectos marginales e interacciones en las regresiones logísticas binarias

Matías Salvador Ballesteros¹

¹ Instituto de Investigaciones Gino Germani,
Universidad de Buenos Aires, Argentina
Consejo Nacional de Investigaciones
Científicas y Tecnologías, Argentina
matiballesteros@yahoo.com.ar



INSTITUTO DE INVESTIGACIONES
GINO GERMANI
Facultad de Ciencias Sociales
Universidad de Buenos Aires



European
Commission

Horizon 2020
European Union funding
for Research & Innovation

Marie Skłodowska-Curie Actions (MSCA)
Research and Innovation Staff Exchange (RISE)
H2020-MSCA-RISE-2015
GA-691004

INCASI Working Paper Series is an online publication under *Creative Commons* license. Any person is free to copy, distribute or publicly communicate the work, according to the following conditions:



Attribution. All CC licenses require that others who use your work in any way must give you credit the way you request, but not in a way that suggests you endorse them or their use. If they want to use your work without giving you credit or for endorsement purposes, they must get your permission first.



NonCommercial. You let others copy, distribute, display, perform, and (unless you have chosen NoDerivatives) modify and use your work for any purpose other than commercially unless they get your permission first.



NoDerivatives. You let others copy, distribute, display and perform only original copies of your work. If they want to modify your work, they must get your permission first.

There are no additional restrictions. You cannot apply legal terms or technological measures that legally restrict doing what the license allows.

This working paper was elaborated in the context of INCASI Network, a European project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie GA, No. 691004, and coordinated by Dr. Pedro López-Roldán. This article reflects only the author's view and the Agency is not responsible for any use that may be made of the information it contains.

Digital edition: <https://ddd.uab.cat/record/189811>

Dipòsit Digital de Documents
Bellaterra, Cerdantola del Vallès (Barcelona)
Universitat Autònoma de Barcelona



Promedio de los efectos marginales e interacciones en las regresiones logísticas binarias

Matías Salvador Ballesteros

Resumen

Mood (2010; 2017) señala que entre los científicos sociales, y en particular entre los sociólogos, es frecuente la inadecuada utilización e interpretación de las regresiones logísticas. La autora cuestiona que se comparen los odds ratio (o sus logaritmos) de modelos similares entre grupos de una misma muestra, entre muestras de distintos lugares o de distintos momentos. El presente documento en primer lugar se limita a resumir lo señalado por Mood. Luego, ejemplifica dos alternativas analíticas posibles (Promedio de los Efectos Marginales e Interacción entre Variables), mostrando cómo hacerlas a partir del SPSS y del STATA. Para desarrollar los ejemplos se utilizan datos de la segunda Encuesta Nacional de Factores de Riesgos (ENFR) llevada a cabo en Argentina durante el año 2009.

Palabras Clave

Regresión logística binaria, Interacciones, Promedio de los efectos Marginales, Metodología cuantitativa

Índice

1. Introducción. 2. Odds Ratio y Promedios de los Efectos Marginales en Regresiones Logísticas Binarias. 2.1. Limitaciones de los Odds Ratio 2.2. Promedio de los efectos Marginales. 2.3. Ejemplo de lectura de regresión logística binaria con el Promedio de los Efectos Marginales. 2.4. Sintaxis con STATA. 3. Interacciones en las regresiones logísticas binarias. 3.1. Ejemplo de lectura de los coeficientes de una interacción entre dos variables cualitativas 3.2. Cambio de categoría de referencia en la interacción. 3.3. Cómo incorporar la interacción en el SPSS y STATA. 4. Referencias. 5. Anexo.

1. Introducción¹

En publicaciones recientes, Mood (2010; 2017) señala que entre los científicos sociales, y principalmente entre los sociólogos, es frecuente la utilización e interpretación inadecuada de los odds ratio en las regresiones logísticas para determinadas situaciones. En particular, la autora cuestiona que se comparen los odds ratio (o sus logaritmos) de modelos similares entre grupos de una misma muestra, entre muestras de distintos lugares o de distintos momentos. En el

presente documento en primer lugar nos limitamos a resumir lo señalado por Mood. Luego, ejemplificamos dos alternativas analíticas posibles (Promedio de los Efectos Marginales e interacción entre variables) ante el problema, mostrando cómo hacerlas a partir del SPSS (versión 17.0) y del STATA (versión 14.0). Por un lado, mostramos cómo solicitarle al STATA y cómo interpretar los coeficientes del Promedio de los Efectos Marginales (PEM, “Average Marginal Effect” o “AME”), que es la propuesta de Mood (2010 y 2017) ante las limitaciones

¹ Elaborado durante la estancia INCASI en la Universitat Autònoma de Barcelona y en la Universidad de Sevilla en el año 2017. Agradezco a Sandra Fachelli, Pedro López-Roldán e

Ildelfonso Márquez Perales por la generosidad para facilitarme bibliografía y tiempo para ayudarme a comprender la temática; así como a los evaluadores del Working Paper por todas las sugerencias realizadas.

señaladas de los odds ratio². Por otro lado, como alternativa para la comparación de coeficientes entre grupos de una misma muestra, indicamos cómo realizar a partir del SPSS y STATA una interacción entre dos variables cualitativas al interior de una regresión logística binaria, y cómo interpretar los coeficientes.

Para desarrollar los ejemplos utilizamos datos de la segunda Encuesta Nacional de Factores de Riesgos (ENFR) llevada a cabo en Argentina por el Ministerio de Salud de la Nación y el INDEC en el año 2009. La encuesta se realizó en base a un diseño probabilístico multietápico (por conglomerado y estratificado), a través de 4 etapas. Las primeras tres (departamento, área y vivienda) corresponden al “Marco de Muestreo Nacional de Viviendas”, recortándose el universo a localidades urbanas de 5.000 habitantes y más. En la cuarta etapa, se relevaron todos los hogares encontrados dentro de una vivienda y para seleccionar a la persona a entrevistar (de 18 años o más) entre los integrantes del hogar, se utilizó la tabla de Kish. En total se encuestaron a 34.732 personas (Ministerio de Salud, 2011: 17-19).

En el documento realizamos modelos de regresiones logísticas binarias para la variable dependiente “Realización de actividad física deportiva o para mejorar la condición física en los últimos 30 días”³. Para simplificar su lectura hacemos referencia a ella como de actividad física deportiva o deporte. Además, utilizamos distintas variables socioeconómicas, demográficas y geográficas que describimos en los distintos modelos. Cabe señalar que utilizamos la base ponderada sin expandir los resultados al total de la población.

Tabla 1. Distribución de frecuencia y porcentual de

variables seleccionadas y Odds Ratio de Realización de actividad física deportiva o para mejorar la condición física en los últimos treinta días entre la población residente en zonas urbanas de Argentina en el 2009.

		f	%	O.R. ¹
Sexo	Hombre	16227	46,7	-
	Mujer	18505	53,3	0,70
Grupo etario	18 a 24 años	5965	17,2	-
	25 a 34 años	7328	21,1	0,76
	35 a 49 años	8816	25,4	0,61
	50 a 65 años	7317	21,1	0,55
	66 años y más	5307	15,3	0,39
Nivel Educativo	Hasta primario incompleto	3702	10,7	-
	Hasta secundario incompleto	13881	40,0	1,99
	Hasta superior incompleto	12041	34,7	3,60
	Superior completo	5072	14,6	5,25
	Educación Especial	36	0,1	2,75
Quintil de ingreso per cápita del hogar	1er quintil	6244	18,0	-
	2do quintil	5925	17,1	1,37
	3ro quintil	5574	16,0	1,45
	4to quintil	5208	15,0	1,89
	5to quintil	5954	17,1	3,06
NBI	Sin declarar ingresos	5828	16,8	1,87
	Hogar con NBI	4699	13,5	-
Región	Hogar sin NBI	30033	86,5	1,99
	AMBA	12643	36,4	-
	Pampeana	12177	35,1	1,33
	Noroeste	3593	10,3	1,12
	Noreste	2511	7,2	1,02
	Cuyo	2234	6,4	1,28
Realización de actividad física deportiva o para mejorar la condición física	Patagónica	1575	4,5	1,36
	Sí	13347	38,4	-
	No	21385	61,6	-
Total		34732	100,0	-

¹Se consideró a la realización de actividad física como numerador y a la no realización como denominador.

Fuente: Elaboración propia en base a ENFR 2009.

² Por lo consultado en la bibliografía, el cálculo del Promedio de los Efectos Marginales (PEM) no es posible realizarlo a partir del SPSS.

Es posible realizarlo con otros Programas como el R. En el R el comando para realizar el PEM se denomina margins. Para una explicación detallada de cómo proceder puede consultarse en <https://cran.r-project.org/web/packages/margins/vignettes/Introduction.html>

³ En la base original de la ENFR 2009, en primer lugar, se le pregunta a la población por la realización de algún tipo de

actividad física en los últimos 30 días. Luego, para los que realizaron, se les consulta por distintos tipos de actividades, incluyendo si realizaron actividad física deportiva y/o para mejorar condición física. En la variable que presentamos aquí, la categoría Sí a quienes realizaron física deportiva y/o para mejorar condición física. En la categoría no contiene a quienes realizaron algún tipo de actividad física, pero no deportiva o para mejorar la condición física y a quienes no realizaron ningún tipo de actividad física.

Para tener un primer acercamiento descriptivo a la temática, en la Tabla 1 presentamos las frecuencias de las variables utilizadas y los Odds Ratio de cada una de las variables independientes con relación a la “Realización de actividad física deportiva o para mejorar la condición física en los últimos 30 días”. En la misma podemos observar que las chances relativas (Odds Ratio) de realizar actividad física deportiva son mayores entre los hombres, las personas con mayor nivel educativo, quienes no residen en el Área Metropolitana de Buenos Aires (AMBA) y en los hogares de mayores ingresos y sin Necesidades Básicas Insatisfechas (NBI).

2. Odds Ratio y Promedios de los Efectos Marginales en Regresiones Logísticas Binarias.

2.1. Limitaciones de los Odds Ratio

Mood (2010; 2017) sostiene y demuestra a partir de ejemplos empíricos, que las regresiones logísticas están afectadas por variables omitidas (“omitted variables”), aún cuando éstas no estén relacionadas con las variables independientes del modelo de la regresión. Esto es una diferencia importante con la regresión lineal, que tiene consecuencias en las interpretaciones que se pueden realizar de los modelos.

Los modelos reflejan cierto grado de heterogeneidad no observada, por lo que no es posible: comparar los odds ratios (ni los logaritmos de los odds ratio) de modelos similares entre grupos de una misma muestra (por ejemplo, realizando una regresión para hombres y otra para mujeres), entre muestras de distintos lugares (por ejemplo, de distintos países) ni de distintos momentos (por ejemplo, de un mismo país en dos años diferentes) aunque tengan las mismas variables independientes. Tampoco es posible hacerlo con modelos que contengan diferentes variables independientes en una misma muestra (Mood, 2010: 67-68; 2017: 8-9).

La heterogeneidad no observada es la variación en la variable dependiente que es causada por

variables omitidas en el modelo. Al igual que en la regresión lineal, surgen problemas cuando estas variables están correlacionadas con las independientes. Sin embargo, en la regresión logística, a diferencia de la lineal, los coeficientes también están afectados cuando las variables omitidas no están relacionadas con las variables independientes (Mood, 2010: 69).

2.2. Promedio de los efectos Marginales

Mood (2010: 73) sostiene que la solución más fácil sería trabajar la variable dependiente como cuantitativa y realizar una regresión lineal múltiple. Sin embargo, debido a las problemáticas que se estudian desde las ciencias sociales es frecuente que no se pueda transformar a la variable dependiente en una variable cuantitativa. Es por ello que la autora propone trabajar en las regresiones logísticas binarias con medidas derivadas de los Efectos Marginales (EM, “Marginal Effect” o “MFX”) de un valor determinado (generalmente la media) de todas las variables independientes de la función logística de distribución de probabilidades (“logistic probability distribution function”) correspondiente a la predicción logística de la media, y multiplicarlo por los coeficientes estimados (Mood, 2010: 75).

Una medida posible derivada de lo anterior es el Promedio de los Efectos Marginales (PEM, “Average Marginal Effect” o “AME”) que expresa el efecto promedio de la variable independiente sobre la probabilidad de que suceda la categoría de contraste de la variable dependiente ($y=1$) (Mood, 2010: 75; 2017: 4). Otra medida similar es el Promedio de los Efectos Parciales (PEP, “Average Partial Effects” o “APE”). PEM da el promedio general de las pendientes condicionales de la variable independiente, mientras que PEP el promedio condicional de las pendientes para diferentes valores de la variable independiente (Mood, 2010: 75). Estas medidas son idóneas para calcular los promedios poblacionales en términos de probabilidades. Si se está interesado en trabajar a nivel agregado son correctas. Sin embargo, si se quiere calcular las probabilidades para un determinado individuo, hay otras medidas que son más apropiadas como los

efectos marginales (EM) [no confundir con medida anterior, Promedio de los Efectos Marginal (PEM)]. Sin embargo, EM sí se ve afectada por la heterogeneidad no observada (Mood, 2010: 78).

El PEM y el PEP se ven solo muy marginalmente afectados por la heterogeneidad no observada. Ello se debe a que se basan en el promedio de los efectos marginales. Por ello son prácticamente invariantes a la exclusión de las variables independientes no relacionadas con las variables independientes ya incluidas en el modelo (Mood, 2010: 76; 2017:4-5). Sin embargo, obviamente sí se ven afectadas por la exclusión de variables independientes correlacionadas con las variables independientes que sí figuran en el modelo. En este sentido, es similar a lo que sucede con la regresión lineal múltiple, que también se ve afectada por la no inclusión de variables independientes vinculadas con las dependientes y las independientes incluidas.

Debido a que el PEM y el PEP no están afectados (al menos no más que marginalmente) por la heterogeneidad no observada que no está relacionada con las variables independientes del modelo, es posible comparar a través de modelos, grupos, muestras y años. Estas medidas son promedios poblacionales y representan el promedio condicional de los efectos de una variable independiente sobre la probabilidad que ocurra la categoría de referencia de la variable dependiente (Mood, 2010: 78). Es decir, nos indican el promedio de las diferencias en la probabilidad de pertenecer o no a la categoría de contraste de la variable dependiente, para cada una de las categorías de las variables independiente tomando como referencia una categoría determinada.

2.3. Ejemplo de lectura de regresión logística binaria con el Promedio de los Efectos Marginales

Para ejemplificar cómo se realiza la lectura del Promedio de los Efectos Marginales, realizamos

una regresión logística binaria para la variable realización de actividad física deportiva en los últimos 30 días (Tabla 2).

Tabla 2. Promedio de los Efectos Marginales (PEM) de realización de actividad física deportiva o para mejorar la condición física en los últimos treinta días¹ según variables seleccionadas entre la población residente en zonas urbanas de Argentina en el 2009

	Odds Ratio	PEM
Sexo (0=mujer)	1,46**	0,082**
Grupo etario (0= 18 a 24)		-
25 a 34 años	0,65**	-0,098**
35 a 49 años	0,53**	-0,143**
50 a 65 años	0,47**	-0,166**
66 años y más	0,35**	-0,226**
Nivel Educativo (0=primario incompleto)		-
Primario completo y secundario incompleto	1,55**	0,085**
Secundario completo y superior incompleto	2,18**	0,161**
Superior complete	3,13**	0,246**
Quintil de Ingreso per cápita del hogar (0= 1er quintil)		-
2do quintil	1,23**	0,042**
3ro quintil	1,31**	0,055**
4to quintil	1,70**	0,114**
5to quintil	2,43**	0,195**
NBI (0=con NBI)	1,35**	0,063**
Región (0=AMBA)	-	-
Pampeana	1,42**	0,073**
Noroeste	1,44**	0,078**
Noreste	1,39**	0,069**
Cuyo	1,46**	0,079**
Patagónica	1,36**	0,065**
n	30.571	

Notas: ¹0=No realizó, 1=Sí realizó. **p <0,01

Fuente: Elaboración propia en base a ENFR 2009.

Incluimos como variables independientes al Sexo (mujer = 0, categoría de referencia), grupo etario (18 a 24 años = 0, categoría de referencia), nivel educativo alcanzado⁴ (primaria incompleta referencia = 0, categoría de referencia), al quintil de ingreso per cápita del hogar⁵ (primer quintil = 0, categoría de referencia), pertenencia a un hogar con Necesidades Básicas Insatisfechas (Hogar con NBI = 0, categoría de referencia) y Región de residencia (Área Metropolitana de Buenos Aires-AMBA- = 0, categoría de

⁴ Quienes respondieron educación especial fueron excluidos del análisis (n= 36 casos, 0,1% de los casos).

⁵ Quienes no respondieron sobre los ingresos de sus hogares fueron excluidos del análisis (n= 5.828, 16,8% de los casos).

referencia). En una primera columna presentamos los Odds Ratio y en una segunda los Promedios de los Efectos Marginales (PEM). La presentación de los Odds Ratio es sólo ilustrativa, para mostrar los diferentes resultados con relación a los PEM. Sin embargo, siguiendo a Mood (2010) sólo deberíamos presentar y analizar los PEM.

A partir de la lectura de los PEM de la regresión podemos afirmar, por ejemplo, que una vez controlado el efecto del resto de las variables independientes del modelo⁶:

- Las probabilidades de realizar actividad física deportiva son 8,2 puntos porcentuales mayores para los hombres que para las mujeres
- A medida que aumentan la edad descienden las probabilidades de realizar deporte. Más concretamente se observa que la probabilidad de realizar este tipo de actividad física entre quienes tienen 66 años y más es 22,6 puntos porcentuales menor que entre quienes tienen de 18 a 24 años.
- Las probabilidades aumentan a medida que se incrementa el nivel educativo. Por ejemplo, se observa que las probabilidades de realizar prácticas deportivas aumentan 24,6 puntos porcentuales entre quienes cuentan con estudios superiores en comparación con quienes alcanzaron hasta primario incompleto.
- Las probabilidades también son más grandes a medida que aumenta el ingreso per cápita del hogar. Las probabilidades de realizar actividad física deportiva para quienes pertenecen a hogares del quinto quintil de ingreso son 19,5 puntos porcentuales mayores que las de quienes pertenecen al primer quintil.

Las probabilidades son mayores entre quienes no residen en el AMBA. Por ejemplo, las probabilidades de realizar deporte entre quienes residen en la región pampeana son 7,3 puntos

porcentuales mayores que entre quienes residen en el AMBA.

- Además, podemos afirmar que las diferencias son estadísticamente significativas trabajando con un 99% de confianza.

La realización de la regresión logística con PEM nos permitiría realizar una comparación de estos resultados con los de Argentina en otros años, con los de otros países o entre grupos al interior de esta misma encuesta.

2.4. Sintaxis con STATA

La sintaxis que se puede correr utilizando el Stata versión 14.0 y con la base de datos de la ENFR 2009⁷ para realizar la regresión logística binaria con el PEM es la siguiente:

**Primero se solicita la regresión logística con todas las variables del modelo, ponderando y solicitándole los Odds Ratio.*

```
logit ADeportiva i.SexoReg i.RANGEDAD
i.NivelEduc i.QuartilIng i.HogarNBI
i.REGION [pweight= Pondera2], or
```

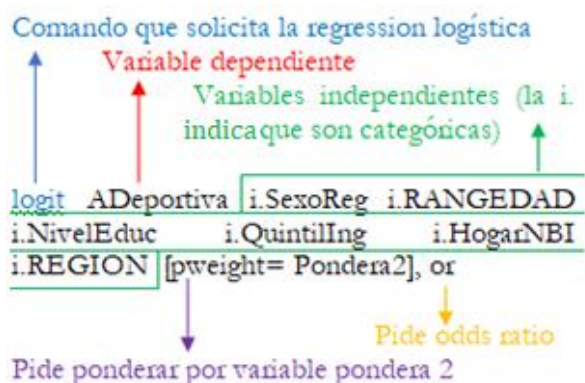
**Luego pedimos PEM para cada una de las variables independientes*

```
Margins, dydx(SexoReg RANGEDAD
NivelEduc QuintilIng Hogar NBI REGION )
```

En la primera sección de la sintaxis solicitamos al STATA que realice una regresión logística binaria. A continuación, describimos el significado de cada uno de los términos de la primera parte de la sintaxis (en la Tabla 1 del anexo se encuentra la tabla que realiza STATA cuando se ejecuta esta sintaxis).

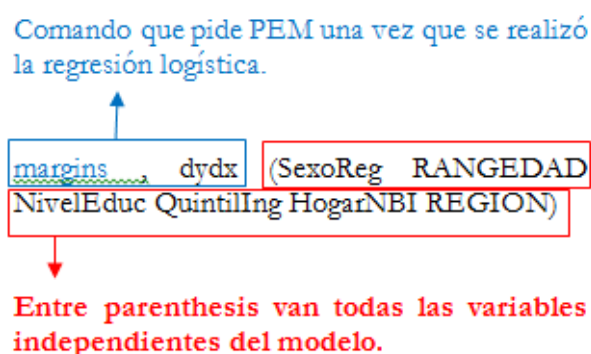
⁶ Agradezco a Santiago Poy por sus comentarios en relación a la lectura e interpretación de los resultados.

⁷ Se encuentra disponible en <https://www.indec.gob.ar/bases-de-datos.asp>



En este caso incluimos todas las variables independientes como categóricas. En caso de incluir una variable independiente como cuantitativa no debe incluirse el “i.” antes de la variable. Por otro lado, cabe señalar que el programa toma como categoría de referencia a la primera categoría de la variable. Es por ello que se debe preparar la variable según la categoría de referencia que se desee que se considere. Por ejemplo, la variable Sexo originalmente estaba codificada como 1=hombres y 2=mujeres. Debido a que en el ejemplo utilizamos a las mujeres como categoría de referencia, armamos una nueva variable (“SexoReg”) con el código 0=mujer y 1=hombre.

A continuación, describimos los términos de la segunda parte de la sintaxis (en la Tabla 2 del anexo se encuentra el resultado que arroja STATA cuando se ejecuta esa sintaxis):



3. Interacciones en las regresiones logísticas binarias

Esta sección del documento tiene como objetivo ejemplificar la interacción al interior de una regresión logística binaria entre dos variables cualitativas, interpretando los coeficientes, además de mostrar cómo se realiza en el SPSS y en STATA. Antes de avanzar con ello, nos parece interesante mencionar la reflexión de Jaccard (2001: 17) sobre las diferencias entre realizar una interacción al interior de una regresión logística y realizar regresiones logísticas diferenciales para distintos grupos. El autor plantea que es muy común si se quiere observar las diferencias en el efecto de una variable independiente sobre una variable dependiente entre dos grupos, que se realicen regresiones logísticas diferenciales. Por ejemplo, ver el efecto diferencial del ingreso en realizar o no actividad física para hombres y para mujeres. En este caso, es muy común que se realice una regresión logística para hombres y otra para mujeres. Si la variable ingreso es significativa para mujeres y no para hombres, se podría sostener que la variable solo tiene efecto y/o es más importante entre las mujeres.

Sin embargo, puede haber una lógica defectuosa detrás de ese razonamiento, si no se realiza un test estadístico formal sobre la diferencia entre los coeficientes de los dos grupos. Así podría ser que el p valor de un grupo sea 0,051 y en el otro sea 0,049. Si bien uno es significativo y el otro no, las diferencias entre los coeficientes seguramente sean mínimas. Esta lógica es incorrecta, independientemente de si se utilizan odds ratio o PEM. Es por ello por lo que Jaccard (2001: 17) sostiene que es preferible introducir la interacción en la regresión logística, porque de esa forma se genera un testeo formal de la diferencia entre los coeficientes de ambos grupos. Sin embargo, destacamos que existen test que permiten comparar la diferencia entre los coeficientes de una variable entre diferentes grupos (Mood, 2017: 18). En este sentido, si bien existen algunos test para contrastar coeficientes en la regresión logística⁸, desde aquí optamos por otro abordaje basándonos en los promedios de los efectos marginales y las interacciones. Por otro lado, es esperable que al comparar el efecto

⁸ Por ejemplo: lrtest, Hausman y Suest. Agradecemos a los evaluadores quienes sugirieron incorporar esta aclaración.

de una variable entre dos grupos no se considere exclusivamente la significancia, sino también la magnitud del efecto (por ejemplo, del PEM).

La interacción entre dos variables se genera cuando el efecto de una variable independiente sobre la dependiente varía según los valores de una tercera variable. Es decir, si una variable es “un modificador del efecto” de una variable independiente sobre la dependiente (Jaccard, 2001: 12; Silva aycaguer y Barroso Utra, 2004: 113). Las interacciones pueden ser entre dos variables o más. Sin embargo, si bien es teóricamente posible, la interacción de más de dos variables es poco frecuente, a la vez que puede comprometer “la parsimonia de los modelos y el aporte práctico que entraña no suele ser muy trascendente” (Silva Aycaguer y Barroso Utra, 2004: 113-114). Por otro lado, la interacción puede ser entre una variable cualitativa y una cualitativa, entre una cualitativa y una cuantitativa y entre dos cuantitativas.

Cabe destacar que no es habitual encontrar artículos que incorporen interacciones al interior de las regresiones logísticas binarias, aunque hay una creciente literatura sobre la interseccionalidad en las desigualdades sociales (como género y clase) que utilizan esta estrategia analítica (Veenstra, 2011; Else-Quest y Hyde, 2016; Krause y Ballesteros, 2018). A continuación veremos que en parte ello se debe a que la inclusión de interacciones puede atentar contra la parsimonia de los modelos.

3.1. Ejemplo de lectura de los coeficientes de una interacción entre dos variables cualitativas en una regresión logística

En el ejemplo que presentamos a continuación realizamos una interacción al interior de una regresión logística binaria con dos variables independientes que trabajamos como cualitativas. Una de ellas es tratada como ordinal (Quintil de ingreso per cápita del hogar⁹) y otra

nominal dicotómica (Sexo) para ver el efecto que tienen sobre la realización de actividad física deportiva. Para la variable quintil de ingreso per cápita del hogar tomamos al Primer quintil como categoría de referencia, mientras que para Sexo utilizamos a la categoría mujer como referencia.

En el Modelo 1 de la Tabla 3 se presentan los odds ratio de la regresión logística binaria con las dos variables independientes sin realizar ninguna interacción. La misma nos permite afirmar que los hombres tienen 1,4 veces más chances que las mujeres de realizar actividad física deportiva. También que quienes pertenecen a hogares del quinto quintil tienen 3 veces más chances de practicar deporte que los que pertenecen al primer quintil de ingreso.

El Modelo 1 muestra que las mujeres pertenecientes a los hogares de menores recursos son las que menos chances tienen de realizar actividad física deportiva. Sin embargo, lo que no muestra es si existe una interacción entre las variables. En la literatura se sostiene que las diferencias de género en las posibilidades de realizar actividad física deportiva son más importantes entre los sectores de menores recursos debido al mayor peso que para las mujeres de estos sectores tiene la carga del trabajo doméstico y la crianza de los hijos y por el poco dinero para destinar al ocio (Hargreaves, 1993; García Ferrando, 2006). Esta idea también puede plantearse al revés, las diferencias debido a los ingresos son más intensas entre las mujeres que entre los hombres. Para poner a prueba esta hipótesis es que nos sirve incorporar la interacción.

Como puede observarse en el Modelo 2 de la Tabla 3, para analizar la interacción entre las variables cualitativas, se crean nuevas dummies. Para entender cuántas *dummies* se crean, puede ser útil pensar en los grados de libertad que tiene un cuadro que cruza a las dos variables independientes¹⁰. Recordemos que la fórmula para ver cuántos grados de libertad tiene un

⁹ En el ejemplo nos interesa trabajarla como una variable ordinal contrastando lo que sucede en cada una de las categorías de los quintiles de ingreso al compararlo con el primer quintil (categoría de referencia), y no incluirla como una variable cuantitativa.

¹⁰ De esta forma lo mencionó Pedro López-Roldán la primera vez que hablamos sobre el tema, y creo que es más pedagógica que otras que aparecen en la literatura.

cuadro es $(\text{Filas}-1) \times (\text{Columnas}-1)$. En este caso sería $(\text{las 5 categorías del quintil}-1) \times (\text{las 2 categorías de sexo}-1)$: $(5-1) \times (2-1) = 4 \times 1 = 4$.

Como puede observarse en la Tabla 4 se crean cuatro variables dummies. Las dummies creadas reflejan las distintas intersecciones posibles entre las categorías de las dos variables que no incluyan a ninguna de las dos categorías de referencia. Dado que tomamos como referencia a las mujeres y a quienes pertenecen al primer quintil, las variables dummies que grafican la interacción son las siguientes: 1. Hombres del 2do quintil, 2. Hombres del 3er quintil, 3. Hombres del 4to quintil y 4. Hombres del 5to quintil. En la Tabla 4 se puede apreciar con más claridad la construcción de las dummies de la interacción.

Para analizar el efecto que genera la interacción en la regresión describimos, por un lado, lo que sucede con las variables originales y, por el otro, lo que acontece con los términos de la interacción. Como puede apreciarse en el Modelo 2 de la Tabla 3, con la incorporación de la interacción se cambia la magnitud del efecto de las variables originales. Debido a la presencia de las variables dummies de la interacción, los coeficientes de las dos variables independientes originales reflejan el efecto de esa variable, pero solo en la categoría de referencia de la otra variable independiente (Jaccard, 2001: 20).

Veámoslo con el ejemplo para que se entienda más claramente. En el Modelo 2 de la Tabla 3, el odds ratio señala que los hombres tienen 2 veces más chances que las mujeres de realizar actividad física deportiva (recordemos que en la regresión original era 1,4 veces). Lo que sucede es que ese coeficiente ahora solo señala el efecto de ser mujer, para la categoría de referencia de la otra variable (primer quintil de ingreso per cápita del hogar). Es decir, lo que nos señala ahora ese coeficiente es que, en el primer quintil de ingreso, las mujeres tienen 2 veces más chances que los hombres de no realizar actividad física deportiva. Algo similar sucede con el quintil de ingreso per cápita del hogar. Ahora lo que nos indica es cuál es el efecto del ingreso entre las mujeres. Así, puede interpretarse que las mujeres del quinto quintil de ingreso per cápita tienen 4,1

veces más chances de realizar actividad física que los del primer quintil (en el Modelo 1 era 3,04).

Es por lo anterior que Jaccard (2001: 21) señala que los coeficientes de las variables independientes que intervienen en una interacción dependen de cuál es la categoría de referencia que se defina de la otra variable. Como veremos con un ejemplo más adelante (sección 3.2., Tabla5), los coeficientes del quintil de ingreso per cápita del hogar serían diferentes si nosotros hubiéramos tomado como categoría de referencia a los hombres y no a las mujeres. También hubiera sido diferente el coeficiente de la variable sexo si tomáramos como referencia al quinto quintil y no al primer quintil de ingreso per cápita del hogar.

Tabla 3. Odds Ratio de realización de actividad física deportiva o para mejorar la condición física en los últimos treinta días¹ según variables seleccionadas entre la población residente en zonas urbanas de Argentina en el 2009.

	Modelo 1	Modelo 2	Modelo 3
Sexo (0=mujer)	1,44**	2,04**	2,08**
Quintil de ingreso per cápita del hogar (0= 1er quintil)	-	-	-
Quintil 2	1,36**	1,56**	1,42**
Quintil 3	1,45**	1,71**	1,55**
Quintil 4	1,89**	2,40**	2,19**
Quintil 5	3,04**	4,10**	3,29**
Sexo y Quintil de ingreso per cápita del hogar (0=mujer y 1er quintil)	-	-	-
Hombre y Quintil 2	-	0,77**	0,76**
Hombre y Quintil 3	-	0,72**	0,71**
Hombre y Quintil 4	-	0,62**	0,61**
Hombre y Quintil 5	-	0,55**	0,55**
Nivel Educativo (0=primario incompleto)	-	-	-
Primario completo y secundario incompleto	-	-	1,54**
Secundario completo y superior incompleto	-	-	2,19**
Superior completo	-	-	3,09**
NBI (0=con NBI)	-	-	1,35**
Grupo etario (0= 18 a 24)	-	-	-
25 a 34 años	-	-	0,65**
35 a 49 años	-	-	0,53**
50 a 65 años	-	-	0,47**
66 años y más	-	-	0,35**
Región (0=AMBA)	-	-	-
Pampeana	-	-	1,41**
Noroeste	-	-	1,44**
Noreste	-	-	1,39**

	Modelo 1	Modelo 2	Modelo 3
Cuyo	-	-	1,45**
Patagónica	-	-	1,36**
Constante	0,31**	0,26**	0,17**
Pseudo R2 de Nagelkerke	0,056	0,059	0,130
BIC	-278532	-278497	-279942
AIC	1,215	1,213	1,161
n	30.571		

Notas: 0=No realizó, 1=Sí realizó. **p <0,01

Fuente: elaboración propia en base a ENFR 2009

Tabla 4. Forma en que construyen las variables dummies en una interacción entre dos variables cualitativas.

Quintil de ingreso per capital del hogar	Sexo	
	Hombre	Mujer
1er quintil	0	0
2do quintil	1 (Dummy 1)	0
3er quintil	1 (Dummy 2)	0
4to quintil	1 (Dummy 3)	0
5to quintil	1 (Dummy 4)	0

Fuente: Elaboración propia en base a ENFR 2009

Ahora bien, antes de ejemplificar lo que sucede cuando cambiamos la categoría de referencia, veamos qué nos indican los términos de la interacción en la regresión. Estos indican las diferencias del efecto de una variable, según la categoría de la otra variable. Es una medida que está basada en la ratio de un odds ratio. Por ejemplo, la variable Dummy 1 (hombres del segundo quintil de ingreso) nos indica la ratio entre el odds ratio de realizar deportes entre los hombres y las mujeres del segundo quintil de ingreso con el odds ratio de realizar deportes entre los hombres y las mujeres del primer quintil de ingreso. La Dummy 1 se puede ver reflejada de la siguiente forma:

$$\frac{\text{odds ratio de } \frac{\text{hombres}}{\text{mujeres}} \text{ del segundo quintil de ingreso en la realización de deporte}}{\text{odds ratio de } \frac{\text{hombres}}{\text{mujeres}} \text{ del primer quintil de ingreso en la realización de deporte}}$$

Dicho de otra forma, nos indica las diferencias género en las chances relativas (odds ratio) de realizar deporte en el segundo quintil de ingreso con relación a las diferencias de género en el primer quintil. Si el indicador es igual a 1, o si pasa por el 1 dentro intervalo de confianza del indicador, las diferencias de género entre el segundo y el primer quintil no serían significativas. Por el contrario, si no pasa por el 1 son significativas. Si es mayor a 1 indica que las diferencias de género son mayores en el segundo quintil de ingreso que en el primero. Si son menores, que son mayores en el segundo quintil en comparación con el primero.

Para entender lo anterior, es decir cómo interpretar los términos de la interacción en una regresión logística binaria entre dos variables cualitativas x y z, Jaccard (2001:22-23) plantea una regla general. En primer lugar, propone seleccionar a una de las variables para que sea la variable independiente focal (*"focal independent variable"*, x por ejemplo) y que la otra sea la variable moderadora (*"moderator variable"*, z por ejemplo). Luego plantea que el coeficiente de los términos de la interacción es una razón de dos odds ratio: se centra en las chances pronosticadas (*predicted odds*) para el grupo clasificado como 1 en la variable dummy de x dividido por las probabilidades pronosticadas para el grupo de referencia en x. Y se divide el resultado de esa odds ratio dentro del grupo clasificado como 1 para la variable z (numerador) con el resultado de esa odds ratio para el grupo de referencia de la variable z (denominador).

En este caso (Tabla 3), lo que nos permite decir es que se corrobora nuestra hipótesis de trabajo y las diferencias de género en las chances relativas de realizar deporte son significativamente más importantes entre los sectores de menores recursos (primer quintil de ingreso per cápita del hogar) que en el resto. Contrastando los datos con la dummy 4, podemos decir que las diferencias de género en la práctica del deporte en favor de los hombres son 1,8¹¹ veces menos grandes entre quienes

¹¹ El odds ratio es 0,55. Por lo tanto, para hacer la lectura más sencilla realizamos el siguiente cálculo: $\frac{1}{0,55} = 1,82$

pertenecen a hogares del quinto quintil que entre los del primer quintil. Nótese que a medida que aumenta el quintil de ingreso, descienden las diferencias de género en comparación con las existentes en el primer quintil de ingreso. Entonces lo que nos permite corroborar esta regresión es que las diferencias de género en las chances de realizar deporte aumentan a medida que disminuyen los ingresos de los hogares per cápita en los que residen las personas.

Cabe destacar que la interacción de sólo dos variables ha implicado la incorporación de cuatro variables dummies. En este sentido, es que la incorporación de interacciones puede atentar contra la parsimonia de los modelos. De hecho, tanto el BIC como el AIC consideran que ajusta levemente mejor el Modelo 1 que el Modelo 2. En este sentido, a la regresión se le podrían incorporar la interacción de otras variables de interés teórico como sexo con grupo etario y con nivel educativo (agregarían cuatro y tres variables dummies más respectivamente) lo que también atentaría contra la parsimonia del modelo¹².

Para facilitar el desarrollo del ejemplo incluimos en los Modelos 1 y 2 solo las variables que son parte de la interacción. Sin embargo, como puede observarse en el Modelo 3, también es posible incluir una interacción entre dos variables en una regresión en la que haya más variables independientes. Allí puede observarse que tienen más chances de realizar actividad física deportiva quienes tienen mayores niveles educativos, los más jóvenes, las mujeres, lo que no residen en el Gran Buenos Aires, los que no pertenecen a un hogar con NBI y los pertenecientes a los hogares de mayores ingresos. Además, que, incluso controlando por todas las variables anteriores, las diferencias de género son más importantes entre los sectores del 1er quintil de ingreso per cápita del hogar que en el resto de los quintiles.

3.2. Cambio de categoría de referencia en la interacción

Retomando la idea que señalamos anteriormente, fijémonos lo que sucede cuando cambiamos la categoría de referencia de una de las variables independientes. A continuación (Tabla 5) presentamos los mismos modelos que en la Tabla 3, pero en la interacción tomamos como referencia a las mujeres (igual que antes) y a quienes pertenecen a hogares del quinto quintil de ingreso per cápita (antes habíamos tomado al primero).

Tabla 5. Odds Ratio de realización de actividad física deportiva o para mejorar la condición física en los últimos treinta días¹ según variables seleccionadas entre la población residente en zonas urbanas de Argentina en el 2009.

	Modelo 1	Modelo 2	Modelo 3
Sexo (0=mujer)	1,44**	1,12*	1,14*
Quintil de ingreso per cápita del hogar (0= 5to quintil)	-	-	-
Quintil 1	0,33**	0,24**	0,30**
Quintil 2	0,45**	0,38**	0,43**
Quintil 3	0,48**	0,42**	0,47**
Quintil 4	0,62**	0,59**	0,66**
Sexo y Quintil de ingreso per cápita del hogar (0=mujer y 5to quintil)	-	-	-
Hombre y Quintil 1	-	1,83**	1,86
Hombre y Quintil 2	-	1,39**	1,41
Hombre y Quintil 3	-	1,31**	1,32
Hombre y Quintil 4	-	1,12**	1,12
Nivel Educativo (0=primario incompleto)	-	-	-
Primario completo y secundario incompleto	-	-	1,54**
Secundario completo y superior incompleto	-	-	2,19**
Superior complete	-	-	3,09**
NBI (0=con NBI)	-	-	1,35**
Grupo etario (0= 18 a 24)	-	-	-
25 a 34 años	-	-	0,65**
35 a 49 años	-	-	0,53**
50 a 65 años	-	-	0,47**
66 años y más	-	-	0,35**
Región (0=AMBA)	-	-	-

¹² Es por este motivo que es más frecuente encontrar artículos que por ejemplo realicen regresiones logísticas diferenciales por sexo y que no incluyan las interacciones al interior de la regresión. Por ejemplo, desde la Epidemiología Social Europea se sostiene la primera modalidad de análisis debido a que las “diferencias en los determinantes de la salud según sexo pueden evidenciarse,

bien incluyendo en los modelos términos de interacción o bien realizando el análisis por separado. [Sin embargo] Dada la marcada estructuración por género de la sociedad, es posible que tuvieran que introducirse varios términos de interacción, lo que dificultaría la interpretación de los resultados” (Rohlf et al., 2000: 152).

	Modelo 1	Modelo 2	Modelo 3
Pampeana	-	-	1,41**
Noroeste	-	-	1,44**
Noreste	-	-	1,39**
Cuyo	-	-	1,45**
Patagónica	-	-	1,36**
Constante			0,17**
Pseudo R2 de Nagelkerke	0,056	0,059	0,130
BIC	-278532	-278497	-279942
AIC	1,215	1,213	1,161
n		30.571	

Notas: '0=No realizó, 1=Si realizó.

** p< 0,01; *p<0,05; + p > 0,05

Fuente: Elaboración propia en base a ENFR 2009.

El Modelo 1 es igual al de la Tabla 3, con la diferencia que para el quintil de ingresos los contrastes se realizan con relación al quinto quintil. Sin embargo, nótese que las chances relativas entre el quinto y el primer quintil continúan iguales. El odds ratio del primer quintil es 0,33. Si dividimos 1 por 0,33 da 3,03 (el mismo valor que tenía el odds ratio del quinto quintil de ingreso en comparación con el primer quintil, en el Modelo 1 de la regresión de la Tabla 3)¹³. Es decir, ahora el coeficiente se puede interpretar como que las mujeres pertenecientes al primer quintil de ingreso tienen 3,03 veces menos chances de realizar deporte que las del quinto quintil. A partir del Modelo 1 de la Tabla 3 podíamos decir lo mismo, las mujeres del quinto quintil tienen 3,04 veces más chances de realizar actividad física deportiva que las del 1er quintil.

Con relación al Modelo 2, veamos lo que sucede con cada uno de los términos. En cuanto al quintil de ingreso per cápita, sucede lo mismo que lo descrito para el Modelo 1. Ahora el coeficiente se puede interpretar como que las mujeres pertenecientes al primer quintil de ingreso tienen 4,10¹⁴ veces menos chances de realizar deporte que las del quinto quintil (en el Modelo 2 de la Tabla 3 podíamos decir lo mismo, las mujeres del quinto quintil tienen 4,10

veces más chances de realizar actividad física deportiva que las del 1er quintil). El efecto no cambió debido a que para la otra variable (sexo) continuamos tomando la misma categoría de referencia (mujeres).

En cambio, nótese que el valor de los coeficientes del sexo descendió considerablemente. Ahora lo que estamos viendo es el efecto del género en el quinto quintil de ingreso per cápita del hogar. Los hombres del quinto quintil de ingreso per cápita del hogar tienen veces 1,12 veces más chances de realizar deportes que las mujeres del quinto quintil (en el Modelo 2 de la regresión anterior –Tabla 3– el valor del odds ratio fue 2,04). Además, ahora la relación es significativa trabajando con un 95% de confianza y ya no con el 99%.

Algo similar sucede si cambiamos la categoría de referencia del sexo. En caso de considerar a los hombres como categoría de referencia, los odds ratio de la variable quintil de ingreso per cápita del hogar disminuyen: el odds ratio entre el primero y el quinto quintil pasa de 4,10 a 2,25 (consideramos innecesario incluir esta nueva regresión en el documento debido a que ya fue ejemplificado cambiando la referencia del ingreso).

En el Modelo 2 de la Tabla 5 también puede observarse que cambian los coeficientes que dan cuenta de la interacción, ya que ahora las categorías de referencia son mujer y quinto quintil de ingreso per cápita. Nótese, por ejemplo, que la variable dummy hombre y cuarto quintil no es significativa. Eso da cuenta que no hay diferencias significativas al comparar las desigualdades de género entre el cuarto y el quinto quintil. Por otro lado, el coeficiente del primer quintil es 1,83. En la Tabla 3, el valor del quinto quintil es 0,545¹⁵. Si dividimos 1 por 0,545 nos da 1,83. Es decir, ahora el coeficiente se interpreta como que las desigualdades de género (en el sentido de que los hombres tienen más

¹³ En realidad el valor de la Tabla 3 es 3,04 y no 3,03. La diferencia surge por el redondeo de los valores.

¹⁴ El valor que se muestra en la Tabla 5 es 0,24 por el redondeo, pero si agregamos un decimal más es 0,244. Por eso es que 1 dividido 0,244 da 4,10.

¹⁵ En la Tabla 3 se muestra el valor 0,55 por el redondeo, pero para hacer más exacto el cálculo tomaremos un decimal más y por eso consideramos el valor 0,545.

chances de practicar deporte que las mujeres), son 1,83 veces más grandes en el primer quintil que en el quinto. Las diferencias en las desigualdades de género entre el primer y quinto quintil de ingreso son las mismas, independientemente de cuál quintil de ingreso se tome de referencia.

Por último, vemos en el Modelo 3 de la Tabla 5 que el resto de las variables independientes (Nivel educativo, Grupo etario, NBI y Región) tienen el mismo efecto que en el Modelo 3 de la Tabla 4. Es decir, el cambio de la categoría de referencia de una interacción no afecta al efecto del resto de las variables que se incluyen en la regresión y no son parte de la interacción.

3.3. Cómo incorporar la interacción en el SPSS y STATA

En el caso de SPSS

Incorporar la interacción de dos variables en una regresión logística en el SPSS es muy sencillo. En el caso de realizar la regresión logística con sintaxis, simplemente hay que incluir a las dos variables con las que se realiza la interacción por separado, y luego incluirlas nuevamente con el signo * entre medio de las dos. A continuación, presentamos la sintaxis del ejemplo que utilizamos en la Tabla 3 de este documento, en el que introdujimos las variables de cada uno de los modelos por pasos. En el paso dos (Modelo 2 de la Tabla 3) puede observarse que se introduce la interacción (SEXO*QuintilIng).

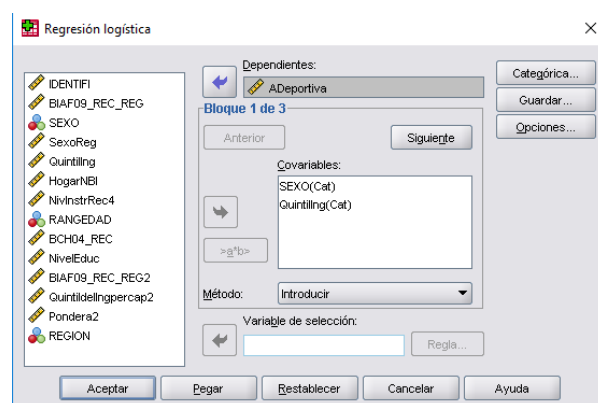
```
LOGISTIC REGRESSION VARIABLES ADeportiva
/METHOD=ENTER SEXO QuintilIng
/METHOD=ENTER QuintilIng*SEXO
/METHOD=ENTER NivInstrRec4 HogarNBI RANGEDAD REGION
/CONTRAST (SEXO)=Indicator
/CONTRAST (QuintilIng)=Indicator(1)
/CONTRAST (NivInstrRec4)=Indicator(1)
/CONTRAST (HogarNBI)=Indicator(1)
/CONTRAST (RANGEDAD)=Indicator(1)
/CONTRAST (REGION)=Indicator(1)
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

En la sintaxis se indica que las categorías de referencia son primer quintil de ingreso, primaria incompleta, Hogar con NBI, grupo etario de 18 a 24 años y el AMBA (por eso el 1 entre paréntesis ya que en la variable original estas son las primeras categorías). A su vez también se

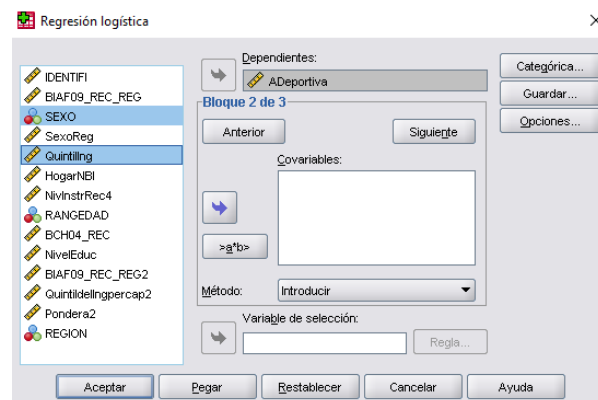
considera a las mujeres como categoría de referencia (en la variable original es la última categoría, por eso no se agrega nada. Si se quisiera que la referencia fueran los hombres habría que poner 1 entre paréntesis).

Para hacer el mismo procedimiento con los comandos interactivos que permite el SPSS, hay que comenzar igual que con cualquier regresión logística binaria: ir a Analizar / Regresión / Regresiones logísticas binarias.

Luego incluir la variable dependiente donde dice variable dependiente y las independientes del primer modelo donde dice variables independientes:



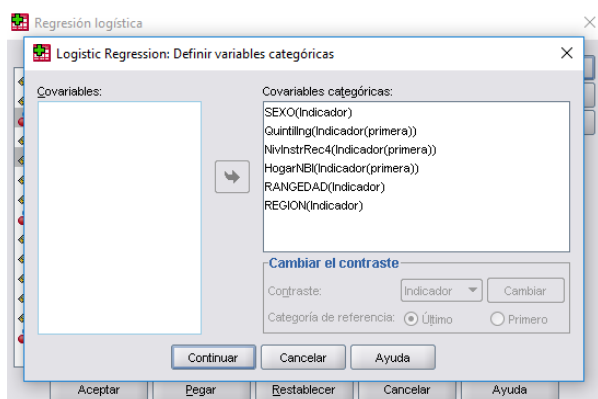
A continuación, se presiona el botón “siguiente” y se introducen las variables que se solicitan para el segundo modelo. En este caso nosotros introdujimos la interacción. Para ello hay que posicionarse con el mouse sobre una de las dos variables independientes que se quieren incorporar en la interacción. Luego se deja presionada la tecla “Control” (o “ctrl”) del teclado y se selecciona la otra variable independiente:



Como se puede observar en esta imagen, y a diferencia de la primera imagen, la opción `>a*b>` aparece disponible (las letras aparecen negras, antes estaban grises). Se debe seleccionar esa opción para que la interacción se incorpore a la regresión.

A continuación, presionamos nuevamente la opción siguiente e incluimos las variables a incorporar en el tercer Modelo (NBI, Nivel Educativo, Grupo etario y Región).

Por último, seleccionamos la opción categórica e indicamos cuáles de las variables son categóricas y si solicitamos que la categoría de referencia sea la primera o la última. En este caso, todas las variables fueron incorporadas al modelo como categóricas:



Por último, se debe presionar Continuar y Aceptar. Como resultado el SPSS nos muestra distintos coeficientes de la regresión logística para cada uno de los modelos. En la Tabla 3 del anexo se encuentra la salida para el tercer Modelo de la regresión, en la que se encuentran los odds ratio, el error estándar y la significancia de cada una de las variables independientes con relación a la dependiente.

En el caso de STATA

La incorporación de la interacción en el STATA también es muy sencilla. Simplemente hay que incorporar el símbolo `##` entre las dos variables independientes que se incorporan a la interacción. A continuación, presentamos la sintaxis utilizada para el Modelo 3 de la Tabla 3:

`Logit ADeportiva i.SexoReg i.RANGEDAD
i.NivelEduc i.QuintilIng i.HogarNBI`

`i.REGION i.SexoReg##i.QuintilIng [pweight=Pondera2] , or`

La sintaxis es muy similar a la descrita en la primera parte de la sintaxis presentada en la sección 3.2, donde se le solicita al STATA realizar la regresión logística binaria previo a que se calcule los PEM. Sin embargo, la diferencia es que en este caso se incorpora la interacción al interior de la regresión logística con el siguiente término:

`i.SexoReg##i.QuintilIng`

En el mismo se encuentran ambas variables (SexoReg es el sexo y QuintilIng es el Quintil de ingreso per cápita del hogar), precedidas de “i.” (que indica que las variables son cualitativas) y el símbolo `##` (indica la solicitud de interacción entre ambas variables).

En la Tabla 4 del anexo se encuentra la salida para el tercer Modelo de la regresión, en la que se encuentran los odds ratio, el error estándar y la significancia para cada una de las variables independientes con relación a la dependiente. Nótese que los coeficientes de los odds ratio son idénticos a los que arroja el SPSS. En cambio, el error estándar que informa es levemente mayor. Ello se debe a que en este caso el STATA informa el error estándar robusto “Robust standard error”. Para este ejemplo, cuando se solicitó la regresión sin ponderar, tanto el SPSS como el STATA informaron el error estándar. Cuando se pondera la base, el SPSS continúa reportando el error estándar, mientras que el STATA informa el error estándar robusto. Este es un punto importante para tener en cuenta ya que las diferencias si se considera el error estándar convencional (que informa el SPSS) o el error estándar robusto (que en este caso informó el STATA cuando la base está ponderada) puede incidir en la significancia de las distintas variables. En nuestro ejemplo, la única diferencia se encuentra en que podemos aceptar la dummy hombre quintil 2 con un 95% de confianza (con el error estándar robusto que informa el STATA) en lugar de con un 99% (con el error estándar que informa el SPSS). Para una discusión sobre este tema puede verse Freedman (2006) y Giles (2013). Destacamos de todas

formas que esta discusión excede a la interacción y a lo discutido en el presente documento, y que el STATA informa el error estándar robusto en regresiones en las que no se trabaja con interacciones y está vinculado con la heterocedasticidad de los modelos.

4. Referencias

- Else-Quest, N. M.; Hyde, J.S. (2016). Intersectionality in quantitative psychological research: II. Methods and techniques. *Psychology of Women Quarterly*, 40, 3, 319-336.
- Freedman, D. A. (2006). On the so-called “Huber Sandwich Estimator” and “Robust Standard Errors”. *The American Statistician*, 60, 299– 302
- García Ferrando, M. (2006). Veinticinco años de análisis del comportamiento deportivo de la sociedad española (1980-2005). *Revista Internacional de Sociología*, LXIV, 44, 15-38.
- Giles, D. (2013). Robust Standard Errors for Nonlinear Models. [internet]. [citado 2017 Dic 20]. Disponible en: <http://davegiles.blogspot.com.ar/2013/05/robust-standard-errors-for-nonlinear.html>
- Hargreaves, J. (1993). “Problemas y promesa en el ocio y los deportes femeninos”. En J.M. Brohm et al. (Comp.) *Materiales de sociología del deporte*. Madrid: Ediciones La Piqueta.
- Jaccard, J. (2001). *Interaction Effects in Logistic Regression*. Sage University Papers Series on Quantitative Applications in the Social Sciences, vol. 135. Thousand Oaks, CA: Sage Publications.
- Krause, M. y Ballesteros, M.S. (2018). Interseccionalidad en desigualdades en salud en Argentina: discusiones teórico-metodológicas a partir de una encuesta poblacional. *Hacia la promoción de la salud*, 23, 2. (En prensa).
- Ministerio de Salud (2011). *Segunda Encuesta Nacional de Factores de Riesgo*. Bs As: MSAL.
- Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, 26,1, 67-82.
- Mood, C. (2017). Logistic regression: Uncovering unobserved heterogeneity. [internet]. [citado 2017 Dic 8]. Disponible en: http://www.su.se/polopoly_fs/1.341160.1501927239!/menu/standard/file/Logit2%20%286%29.pdf
- Rohlfs, I.; Borrell, C.; Anitua, C.; Artazcoz, L.; Colomer, C.; Escribá, V.; et al. (2000). La importancia de la perspectiva de género en las encuestas de salud. *Gac Sanit.*, 14, 2, 146-155
- Silva Aycaguer, L. C.; Barroso Utra, I. M. (2004). *Regresión logística*. Madrid: La Muralla.
- Veenstra, G. (2011). Race, gender, class, and sexual orientation: intersecting axes of inequality and self-rated health in Canada. *International journal for equity in health* 10, 1, 3.

5. Anexo

Tabla 1. Salida de STATA para los Odds Ratio de realización de actividad física deportiva o para mejorar la condición física en los últimos treinta días según variables seleccionadas entre la población residente en zonas urbanas de Argentina en el 2009.

ADeportiva	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
SexoReg						
VarOn	1.463258	.059055	9.43	0.000	1.351972	1.583704
RANGEDAD						
25 a 34 aOs	.6490319	.0445442	-6.30	0.000	.5673442	.7424812
35 a 49 aOs	.5282684	.0351327	-9.60	0.000	.4637086	.6018165
50 a 65 aOs	.4749525	.0345708	-10.23	0.000	.4118064	.5477814
66 aOs y mOs	.349987	.0275742	-13.33	0.000	.2999086	.4084275
NivelEduc						
Primario completo y secundario incompleto	1.545065	.1190369	5.65	0.000	1.328517	1.796909
Secundario completo y superior incompleto	2.179981	.1804819	9.41	0.000	1.853451	2.564036
Superior completo	3.131461	.2910706	12.28	0.000	2.609922	3.75722
QuintilIng						
2	1.230287	.0823206	3.10	0.002	1.079074	1.402691
3	1.306845	.0908666	3.85	0.000	1.140352	1.497647
4	1.704569	.1219777	7.45	0.000	1.481505	1.961218
5	2.425041	.1759643	12.21	0.000	2.103559	2.795654
HogarNBI						
Hogar sin NBI	1.34987	.0917607	4.41	0.000	1.181488	1.542248
REGION						
Pampeana	1.416065	.0802355	6.14	0.000	1.267224	1.582388
Noroeste	1.444637	.0870744	6.10	0.000	1.28367	1.62579
Noreste	1.390321	.0889427	5.15	0.000	1.226483	1.576046
Cuyo	1.455734	.1006573	5.43	0.000	1.271234	1.667012
Patagónica	1.361623	.0828311	5.07	0.000	1.208582	1.534043
_cons	.1989485	.0225685	-14.23	0.000	.1592873	.248485

Fuente: Elaboración propia en base a ENFR 2009

Tabla 2. Salida de STATA para Promedio de los Efectos Marginales (PEM) de realización de actividad física deportiva o para mejorar la condición física en los últimos treinta días según variables seleccionadas entre la población residente en zonas urbanas de Argentina en el 2009.

	Delta-method		z	P> z	[95% Conf. Interval]	
	dy/dx	Std. Err.				
SexoReg						
VarOn	.0816795	.0086659	9.43	0.000	.0646946	.0986643
RANGEDAD						
25 a 34 años	-.0982444	.0155455	-6.32	0.000	-.128713	-.0677758
35 a 49 años	-.1431233	.0148946	-9.61	0.000	-.1723163	-.1139304
50 a 65 años	-.1655119	.0160276	-10.33	0.000	-.1969254	-.1340983
66 años y más	-.2259225	.0165771	-13.63	0.000	-.2584131	-.1934319
NivelEduc						
Primario completo y secundario incompleto	.0854272	.0142649	5.99	0.000	.0574685	.113386
Secundario completo y superior incompleto	.1614989	.015972	10.11	0.000	.1301942	.1928035
Superior completo	.245983	.0190359	12.92	0.000	.2086733	.2832926
QuintilIng						
2	.0422318	.0136042	3.10	0.002	.0155681	.0688955
3	.0550331	.014265	3.86	0.000	.0270741	.0829921
4	.1135362	.0150959	7.52	0.000	.0839488	.1431236
5	.1945951	.0155781	12.49	0.000	.1640626	.2251276
HogarNBI						
Hogar sin NBI	.0627275	.0138663	4.52	0.000	.0355501	.0899049
REGION						
Pampeana	.0732491	.0117118	6.25	0.000	.0502944	.0962037
Noroeste	.0776102	.0124924	6.21	0.000	.0531256	.1020948
Noreste	.0692557	.0133056	5.20	0.000	.0431772	.0953343
Cuyo	.0792842	.0145646	5.44	0.000	.0507381	.1078303
Patagónica	.0647307	.0126627	5.11	0.000	.0399122	.0895492

Fuente: Elaboración propia en base a ENFR 2009

Tabla 3. Salida de SPSS para los Odds Ratio de realización de actividad física deportiva o para mejorar la condición física en los últimos treinta días según variables seleccionadas entre la población residente en zonas urbanas de Argentina en el 2009.

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 ^a	SEXO(1)	,734	,059	152,844	1	,000	2,084
	QuintilIng			422,580	4	,000	
	QuintilIng(1)	,348	,060	33,617	1	,000	1,417
	QuintilIng(2)	,439	,061	51,112	1	,000	1,551
	QuintilIng(3)	,782	,063	156,450	1	,000	2,186
	QuintilIng(4)	1,190	,063	361,499	1	,000	3,288
	QuintilIng * SEXO			65,445	4	,000	
	QuintilIng(1) by SEXO(1)	-,279	,082	11,526	1	,001	,756
	QuintilIng(2) by SEXO(1)	-,338	,083	16,532	1	,000	,713
	QuintilIng(3) by SEXO(1)	-,496	,083	35,517	1	,000	,609
	QuintilIng(4) by SEXO(1)	-,606	,080	57,557	1	,000	,545
	NivInstrRec4			434,209	3	,000	
	NivInstrRec4(1)	,434	,051	71,821	1	,000	1,543
	NivInstrRec4(2)	,783	,054	208,803	1	,000	2,188
	NivInstrRec4(3)	1,128	,061	338,680	1	,000	3,088
	HogarNBI(1)	,301	,043	48,970	1	,000	1,351
	RANGEDAD			546,864	4	,000	
	RANGEDAD(1)	-,429	,041	111,904	1	,000	,651
	RANGEDAD(2)	-,638	,040	256,805	1	,000	,528
	RANGEDAD(3)	-,746	,043	303,639	1	,000	,474
	RANGEDAD(4)	-1,056	,049	468,792	1	,000	,348
	REGION			158,799	5	,000	
	REGION(1)	,344	,031	124,043	1	,000	1,411
	REGION(2)	,363	,045	63,949	1	,000	1,437
	REGION(3)	,328	,052	39,057	1	,000	1,388
	REGION(4)	,375	,053	49,709	1	,000	1,454
	REGION(5)	,309	,062	25,097	1	,000	1,362
	Constante	-1,791	,073	594,737	1	,000	,167

Fuente: Elaboración propia en base a ENFR 2009

Tabla 4. Salida de STATA para los Odds Ratio de realización de actividad física deportiva o para mejorar la condición física en los últimos treinta días según variables seleccionadas entre la población residente en zonas urbanas de Argentina en el 2009.

ADeportiva	Odds Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
SexoReg						
VarOn	2.083917	.1839131	8.32	0.000	1.752907	2.477433
RANGEDAD						
25 a 34 aOs	.6509073	.0446773	-6.26	0.000	.568976	.7446365
35 a 49 aOs	.5284749	.035156	-9.59	0.000	.4638734	.602073
50 a 65 aOs	.4742469	.0345198	-10.25	0.000	.411194	.5469685
66 aOs y mOs	.3479477	.027408	-13.40	0.000	.2981704	.4060349
NivelEduc						
primario completo y secundario incompleto	1.543076	.1192421	5.61	0.000	1.326204	1.795414
secundario completo y superior incompleto	2.187744	.1818268	9.42	0.000	1.858882	2.574786
Superior completo	3.08838	.2875354	12.11	0.000	2.573249	3.706633
QuintilIng						
2	1.416766	.1272948	3.88	0.000	1.188006	1.689575
3	1.551203	.1431946	4.76	0.000	1.294472	1.858852
4	2.185559	.2019214	8.46	0.000	1.823563	2.619413
5	3.288213	.3042678	12.86	0.000	2.742811	3.942068
HogarNBI						
Hogar sin NBI	1.351143	.092084	4.42	0.000	1.182197	1.544234
REGION						
Pampeana	1.41108	.0799121	6.08	0.000	1.262835	1.576728
Noroeste	1.436953	.0867261	6.01	0.000	1.276641	1.617395
Noreste	1.387723	.0888517	5.12	0.000	1.224061	1.573268
Cuyo	1.454374	.1006831	5.41	0.000	1.269841	1.665724
Patagónica	1.361657	.0827126	5.08	0.000	1.208822	1.533816
SexoReg#QuintilIng						
VarOn#2	.75629	.0974877	-2.17	0.030	.5874438	.9736668
VarOn#3	.7131643	.0923354	-2.61	0.009	.5533273	.9191725
VarOn#4	.6087876	.0779075	-3.88	0.000	.4737359	.7823397
VarOn#5	.5454347	.0652396	-5.07	0.000	.4314498	.6895332
cons	.1667167	.0199324	-14.98	0.000	.1318897	.2107402

Fuente: Elaboración propia en base a ENFR 2009