Experimental Design and Statistical Methods

Workshop

SIMPLE LINEAR REGRESSION and CORRELATION

Jesús Piedrafita Arilla

jesus.piedrafita@uab.cat

Departament de Ciència Animal i dels Aliments

UAB

Items

- Correlation: degree of association
- Regression: prediction
  - The model
  - Assumptions
  - Matrix notation
  - Protocol of analysis
    - Plotting data
    - ANOVA in regression
    - Confidence intervals
    - Analysis of residuals
    - Influential observations
- Basic commands
  - cor.test
  - lm
  - anova
  - seq
  - influence.measures
- Libraries
  - car (scatterplot)

Analysis of several variables

Two main interests:

1. Estimating the **degree of association** between two variables: **CORRELATION** analysis.
2. **Predicting** the values of one variable given that we know the realised value of another variable(s): **REGRESSION** analysis. This analysis can also be used to **understand the relationship** among variables.
   a) A response variable and an independent variable: **simple (linear) regression**.
   b) A response variable and two or more independent variables: **multiple (linear) regression**.
   c) When the relationship among variables is not linear: **nonlinear regression**.
   d) If the variable is a dichotomous or binary variable: **logistic regression**.

Data example

Suppose we have recorded the age (years) and blood pressure (mm Hg) of 20 people, obtaining the data presented in the table.

Age	Blood pressure
20	120
43	128
63	141
26	126
53	134
31	128
58	136
46	132
58	140
70	144
46	128
53	136
70	146
20	124
63	143
43	130
26	124
19	121
31	126
23	123

Simple statistics

```
> ## Importing data
> BLOODP<-read.csv2("bloodpress.csv", header=T)
> attach(BLOODP)
> options(na.action=na.exclude)
> summary(BLOODP)
```

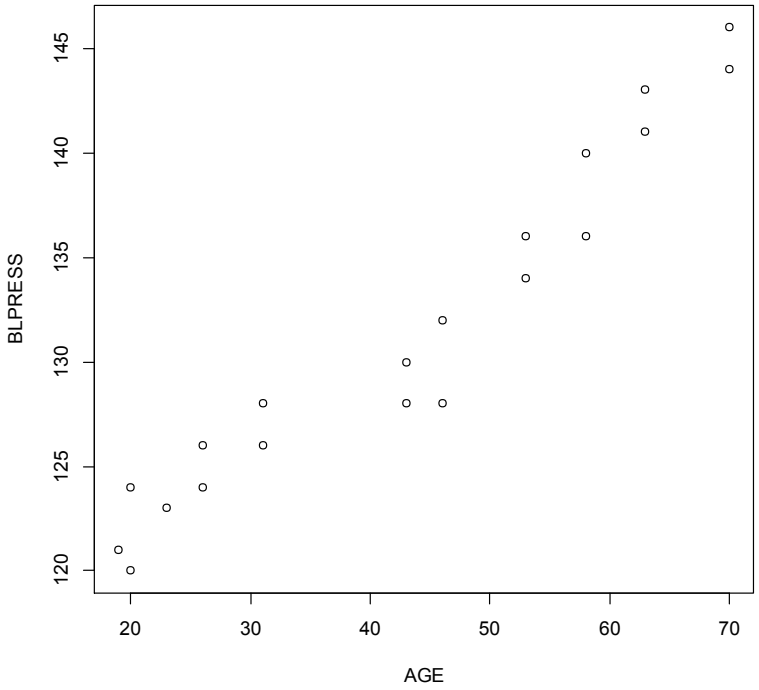
```
      AGE             BLPRESS
 Min.   :19.0   Min.   :120.0
 1st Qu.:26.0   1st Qu.:125.5
 Median :44.5   Median :129.0
 Mean   :43.1   Mean   :131.5
 3rd Qu.:58.0   3rd Qu.:137.0
 Max.   :70.0   Max.   :146.0
```

To avoid problems **in prediction** when missing values are present, we must use `options(na.action=na.exclude)`. With the current data set it would be unnecessary.

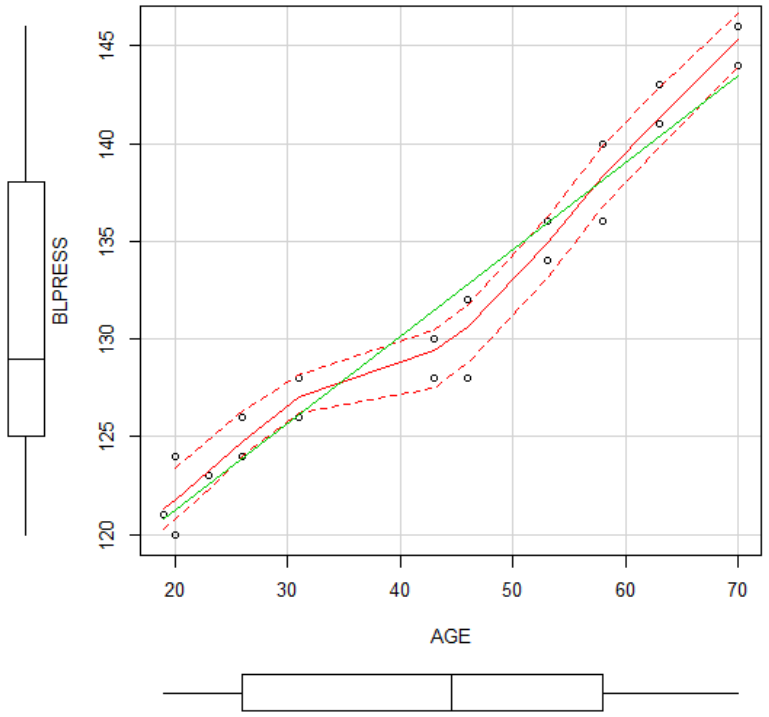
Plot of raw data

A plot for a pair of variables gives us a first impression about their relationship. It is also useful for detecting some extreme values.

```
> plot(AGE,BLPRESS)
```




```
> library(car)
> scatterplot(AGE,BLPRESS)
```



Data not obviously non linear and no evidence of non-normality (boxplots not asymmetrical). No evidence of extreme values.

Correlation (Pearson)

The correlation is a measure of the degree of association between two variables. It is calculated as

$$r = \frac{\mathrm{cov}(x,y)}{\sqrt{s_x^2 s_y^2}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

r is an estimator of ρ, the population parameter.

The denominator is the geometric mean of the sample variances estimates. This makes r to range from -1 to 1. As close is an estimate to -1 or 1, the correlation is larger.

```
> cor.test(BLPRESS, AGE)

        Pearson's product-moment correlation

data:  BLPRESS and AGE
t = 16.0262, df = 18, p-value = 4.239e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9160501 0.9869976
sample estimates:
      cor
0.966699
```

H_0: ($\rho = 0$), is rejected

r

Correlation – sample size -

The **sample size** required to have a particular correlation statistically different from 0 **depends upon the same correlation coefficient**:

$$z' = 0.5 \ln\left[\frac{1-r}{1+r}\right]$$

Fisher's classic z-transformation to normalize the distribution of Pearson correlation coefficient.

$$n = \left[\frac{z_{1-\alpha/2} + z_{1-\beta}}{\left|z'_{r_0} - z'_{r_1}\right|}\right] + 3$$

r_0 = 0 and r_1 is the magnitude of the coefficient we want to estimate.

	Sample size for a power of	
ρ	80%	90%
0.1	781	1044
0.2	194	258
0.3	85	113
0.4	47	62
0.5	29	38
0.6	20	25
0.7	14	17
0.8	10	12
0.9	7	8

Simple linear regression - the model -

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

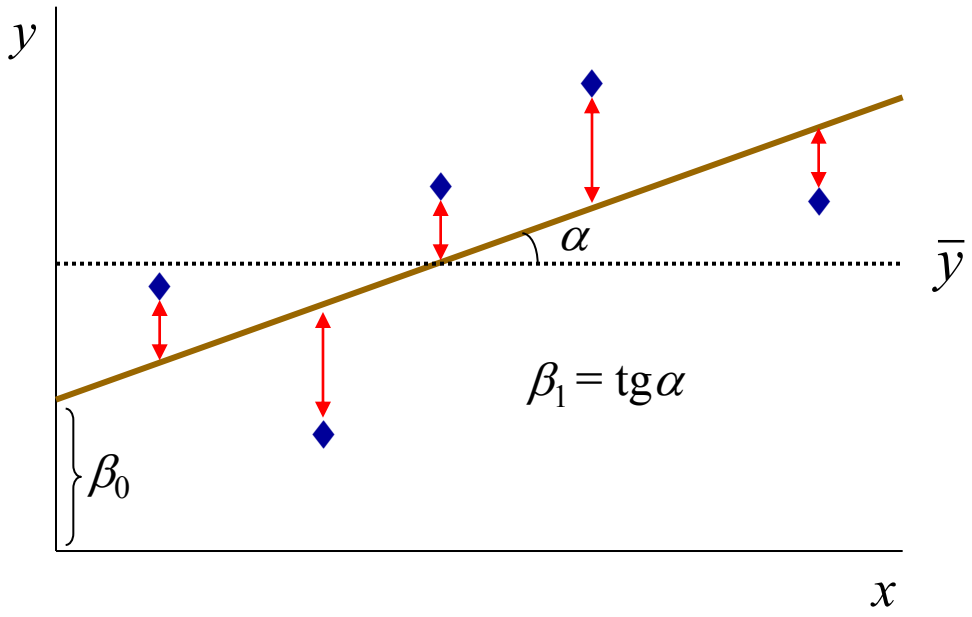
ε_i → Random error

y_i → Dependent variable

β_0 → Intercept

β_1 → Regression coefficient (slope)

x_i → Independent variable



β_1 is the increase of the dependent variable when the independent variable increases 1 unit

To estimate β_0 and β_1 we resort to the Least Squares methodology, i.e., minimize the sum of the squares of the deviations (red arrows) between actual (blue diamond) an predicted values (on the slope).

$$\hat{\beta}_1 = \frac{\text{cov}(x, y)}{s_x^2}$$

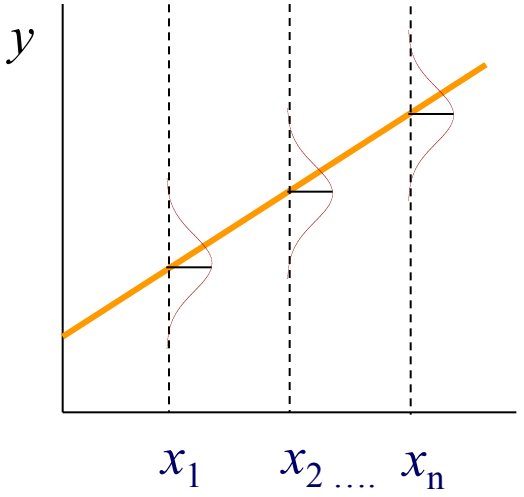
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Assumptions in regression analysis

1. The variables x and y are linearly related (definition of the model).
2. Both variables are measured for each of n observations.
3. Variable x is measured without error (fixed).
4. Variable y is a set of random observations measured with error.
5. The errors are independent and normally distributed with homogeneous variance:

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$$

Some of the above conditions can be seen in the figure. For each value (fixed) of x, there is a normal distribution of y (random), with mean on the regression line.



Matrix notation

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Note that $\mathbf{X}$ is not a matrix of 0 and 1, but contains the values of the independent variable x.

As in ANOVA, we can minimize the sum of the squared errors and then we have the normal equations:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Now $\mathbf{X}'\mathbf{X}$ is not singular and can be solved without need of a generalised inverse (or restrictions).

Simple linear regression – Protocol -

1. Decide which variable is to be y and which is to be x.
2. Plot data, y in the vertical axis.
3. Check evenness of x and y variables by a box-plot.
4. Transform x and/or y if not even.
5. Compute regression, save residuals, fitted y values and influence statistics. Calculate Durbin-Watson statistic if data are in a logical order.
6. Plot studentized or standardized residuals against fitted values (or x variable). Examine residual plots for outliers, and consider rejection of outliers with studentized or standardized residuals > 3 and go to step 5.
7. Compare influence statistics with critical values:
   - Leverage $> 2p/n$
   - Dffits (absolute value) $> 2\sqrt{(p/n)}$
   - Cook's D $> 4/n$
   - Dfbetas $> 2/\sqrt{n}$

➔ Where p = number of parameters in the model (number of β) and n = number of **data points** in the regression. If **two or more** influence statistics (among the first three) are greater than the critical values, consider rejecting points and return to step 5.

8. If outliers or leverage points are a problem, consider using a robust regression method.

(Adapted from Fry, 1993)

Simple linear regression - Results (1) -

```
> BLOODP.REG <- lm(BLPRESS ~ AGE); summary(BLOODP.REG)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7908 -1.2777  0.1688  1.8725  2.7816

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 112.31666    1.28744   87.24  < 2e-16 ***
AGE           0.44509    0.02777   16.03 4.24e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\hat{\beta}_0$ ← (Intercept)

$\hat{\beta}_1$ ← AGE

H_0: β_1=0, is rejected

For each increment of 1 year, blood pressure increases 0.4451 mm Hg

$$\hat{y}_i = 112.3167 + 0.4451\, x_i$$

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)} = \frac{0.44509}{0.02777} = 16.03$$

Simple linear regression - Results (2) -

```
Residual standard error: 2.12 on 18 degrees of freedom
Multiple R-squared: 0.9345,     Adjusted R-squared: 0.9309
F-statistic: 256.8 on 1 and 18 DF,  p-value: 4.239e-12
```

```
> anova(BLOODP.REG)

Analysis of Variance Table

Response: BLPRESS
          Df  Sum Sq Mean Sq F value    Pr(>F)
AGE        1 1154.12 1154.12  256.84 4.239e-12 ***
Residuals 18   80.88    4.49
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F = t^2$

H_0 (β_1=0) is rejected

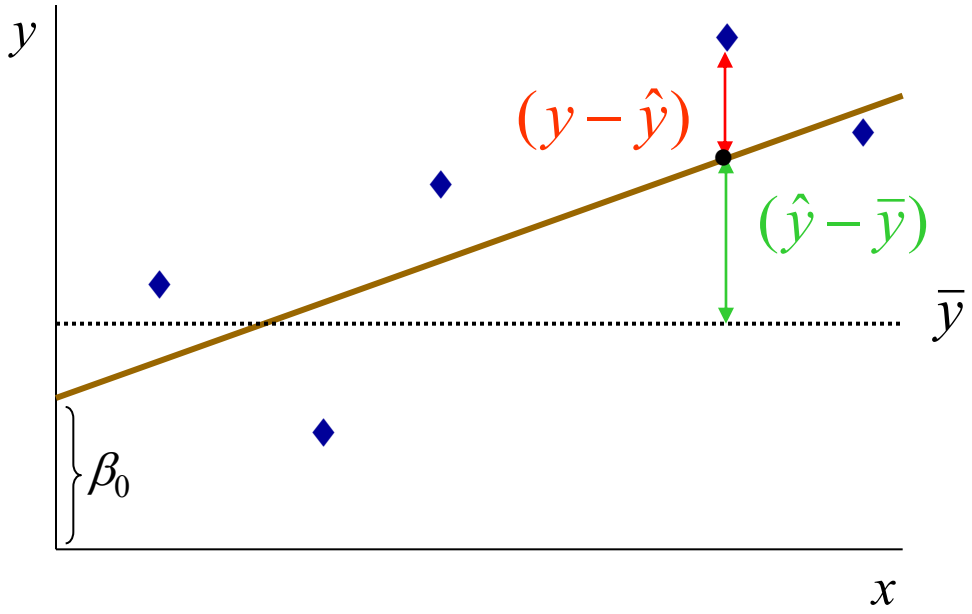
R-Squared is the square of the correlation coefficient. It represents the fraction of the total variation in blood pressure that is explained by the linear relationship with age.

$$R^2 = SS_{AGE} / SS_{TOTAL}$$

Adj R-Sq includes a correction to overcome the increment in *R-Squared* with the number of regressors (k).

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

ANOVA in regression



$$(y-\bar{y}) = (y-\hat{y}) + (\hat{y}-\bar{y})$$

Deviated to regression

Due to regression

Squaring and and summing on both sides of the equation we can arrive at the following ANOVA table:

Source	d.f.	S.S.	M.S.	E(M.S.)	F
Due to regression	1	$\hat{\beta}_1 SP_{xy}$	$\hat{\beta}_1 SP_{xy}$	$\sigma^2 + \beta_1^2 SS_x$	$MS_{\text{Reg}}/MS_{\text{Error}}$
Deviations to regression	n-2	$SS_y - \hat{\beta}_1 SP_{xy}$	$(SS_y - \hat{\beta}_1 SP_{xy})/(n-2)$	σ^2	

Simple linear regression - Results (3) -

Confidence intervals for β_1 (for β_0 is similar):

$$(\hat{\beta}_1 - t_{\alpha/2}\, s.e.(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2}\, s.e.(\hat{\beta}_1))$$

This can be done easily with R (both for b_0 and b_1):

```
> confint(BLOODP.REG,level=0.95)

                  2.5 %      97.5 %
(Intercept) 109.6118594 115.0214613
AGE           0.3867409   0.5034373
```

Simple linear regression - Results (4) -

```
> data.frame(BLOODP, Predicted=fitted(BLOODP.REG), Residuals=resid(BLOODP.REG),
+ RIstudent=rstandard(BLOODP.REG), Restudent=rstudent(BLOODP.REG))
```

	AGE	BLPRESS	Predicted	Residuals	RIstudent	Restudent
1	20	120	121.2184	-1.21844210	-0.62038822	-0.60946002
2	43	128	131.4555	-3.45549109	-1.67245066	-1.76853813
3	63	141	140.3573	0.64272718	0.32284288	0.31465921
4	26	126	123.8890	2.11102338	1.04984002	1.05300889
5	53	134	135.9064	-1.90638196	-0.93096414	-0.92733538
6	31	128	126.1144	1.88557795	0.92493113	0.92102499
7	58	136	138.1318	-2.13182739	-1.05313803	-1.05653371
8	46	132	132.7908	-0.79075835	-0.38301620	-0.37375101
9	58	140	138.1318	1.86817261	0.92289068	0.91889212
10	70	144	143.4729	0.52710357	0.27363114	0.26647648
11	46	128	132.7908	-4.79075835	-2.32047892	-2.69371696
12	53	136	135.9064	0.09361804	0.04571751	0.04443202
13	70	146	143.4729	2.52710357	1.31187546	1.34061303
14	20	124	121.2184	2.78155790	1.41627226	1.46012594
15	63	143	140.3573	2.64272718	1.32744606	1.35824022
16	43	130	131.4555	-1.45549109	-0.70445473	-0.69424391
17	26	124	123.8890	0.11102338	0.05521340	0.05366233
18	19	121	120.7734	0.22664698	0.11594922	0.11272449
19	31	126	126.1144	-0.11442205	-0.05612736	-0.05455077
20	23	123	122.5537	0.44629064	0.22434720	0.21833176

$$\hat{y}_1 = 112.317 + 0.445 \times 20 = 121.2184; \quad \hat{r}_1 = 120 - 121.2184 = -1.2184$$

RIstudent is an "Internally studentized residual", i.e., the residual divided by the own standard error (not uniform across observations).

REstudent is an "Externally studentized residual".

Only observation 11 is a weak outlier.

Some statistics useful for regression analysis

Internally studentized residual

Weak outlier, $|rs_i| > 2$ (95% confidence)

Strong outlier, $|rs_i| > 3$ (95% confidence)

$$rs_i = \frac{r_i}{\sqrt{MSE(1-h_i)}} \sim t_{N-k-1}$$

Externally studentized residuals (-*i*)

Calculated as the previous one, but removing the *i* observation to calculate the s^2. Under H_0, it follows a *t* distribution with *N*-*k*-2 df.

Leverage (h_i)

Standardized value of how much an observation deviates from the centre of the space of *x* values. Observations with high leverage can indicate an outlier in the *x* and are potentially influent. Computed as the diagonal elements of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-i}\sqrt{h_i}} = R-student_i \left[\frac{h_i}{1-h_i}\right]^{\frac{1}{2}} \qquad DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j,-i}}{s_{-i}\sqrt{c_{jj}}}$$

where c_{jj} are the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$. Analyse only *DFBETAS* corresponding to high values of *DFFITS*.

Cook's D

Essentially a *DFFITS* statistic scaled and squared to make extreme values stand out more clearly.

Simple linear regression - Results (5) -

```
> influence.measures(BLOODP.REG)
```

	dfb.1_	dfb.AGE	dffit	cov.r	cook.d	hat	inf
1	-0.23925	0.199083	-0.2475	1.251	3.17e-02	0.1416	
2	-0.15159	0.002377	-0.4057	0.842	7.36e-02	0.0500	
3	-0.05363	0.087353	0.1151	1.256	6.97e-03	0.1180	
4	0.32262	-0.248700	0.3514	1.098	6.14e-02	0.1002	
5	0.03674	-0.124512	-0.2482	1.088	3.10e-02	0.0668	
6	0.22000	-0.151823	0.2625	1.100	3.47e-02	0.0751	
7	0.10973	-0.215983	-0.3284	1.083	5.36e-02	0.0881	
8	-0.01804	-0.014580	-0.0870	1.163	3.98e-03	0.0514	
9	-0.09543	0.187846	0.2856	1.116	4.11e-02	0.0881	
10	-0.07195	0.103347	0.1224	1.346	7.90e-03	0.1742	*
11	-0.13000	-0.105085	-0.6273	0.581	1.46e-01	0.0514	*
12	-0.00176	0.005966	0.0119	1.201	7.48e-05	0.0668	
13	-0.36195	0.519927	0.6157	1.110	1.82e-01	0.1742	
14	0.57320	-0.476956	0.5930	1.031	1.65e-01	0.1416	
15	-0.23151	0.377061	0.4967	1.034	1.18e-01	0.1180	
16	-0.05951	0.000933	-0.1593	1.116	1.31e-02	0.0500	
17	0.01644	-0.012674	0.0179	1.246	1.70e-04	0.1002	
18	0.04595	-0.038599	0.0473	1.317	1.18e-03	0.1497	
19	-0.01303	0.008992	-0.0155	1.212	1.28e-04	0.0751	
20	0.07612	-0.061268	0.0804	1.266	3.41e-03	0.1193	

All values of DFFIT are below the critical value 0.63 (=2 $\sqrt{(2/20)}$), i.e., not influential observations on the predicted values. DFBETAS **(dfb.)** test influence on the parameter estimates, and do not need to be examined because DFFIT values are low.

Cook's D values are below the critical value 0.2 (=4/20).

Leverage is presented in **hat**. All values are lower than 0.2, the critical value.

Criteria to flag an observation as influential in R

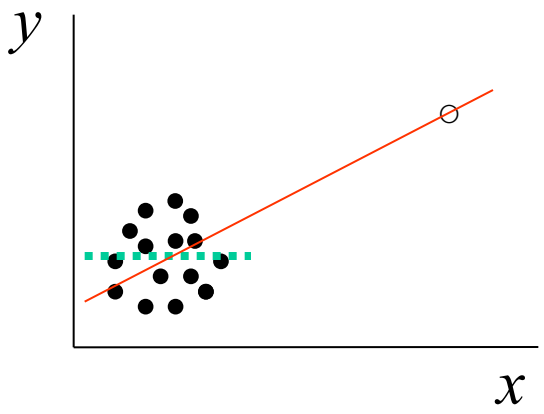
In slide 12 we presented some critical points to decide if an observation can be influential or not. These critical points are not statistical tests but rules of thumb. Furthermore, there are not agreement among statisticians on the values. In fact, R puts a **flag** (a star) on an observation, when:

- ✓ any of its absolute dfbetas value is greater than 1, or
- ✓ its absolute dffits value is greater than $3\sqrt{(p/(n-p)}$, or
- ✓ abs(1-covratio) is greater than $3p/(n-p)$, or
- ✓ its Cook's distance is greater than the 50% percentile of an F-distribution with p and $n-p$ degrees of freedom, or
- ✓ its hat value is greater than $3p/n$

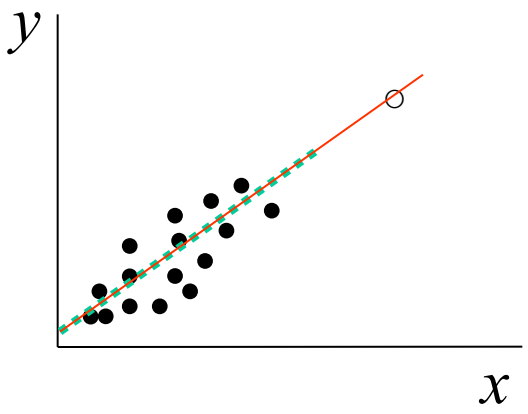
Where p denotes the number of model coefficients, including the intercept.

Some graphics about influential observations

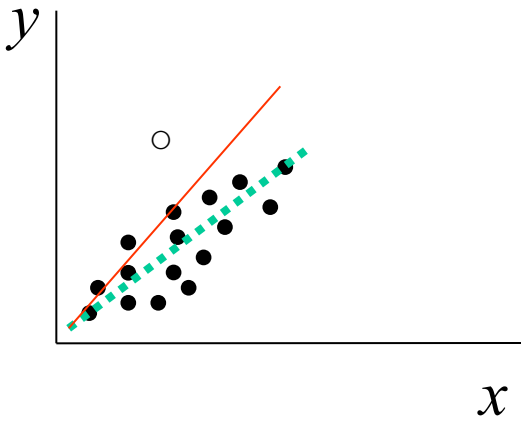
High leverage, influential



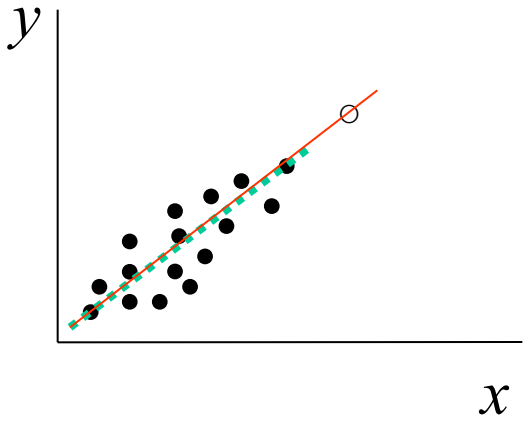
High leverage, not influential



Low leverage, influential



Low leverage, not influential



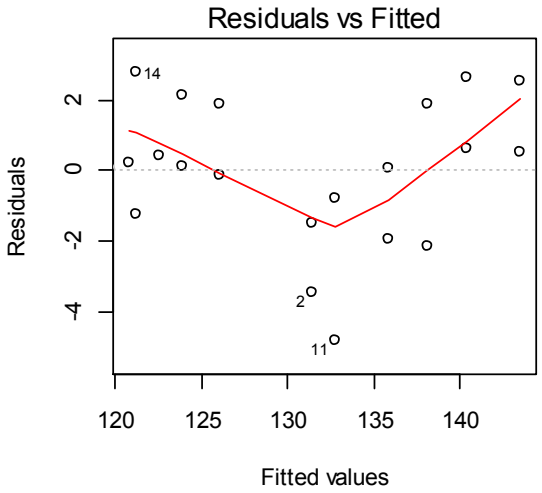
Simple linear regression - diagnostics -

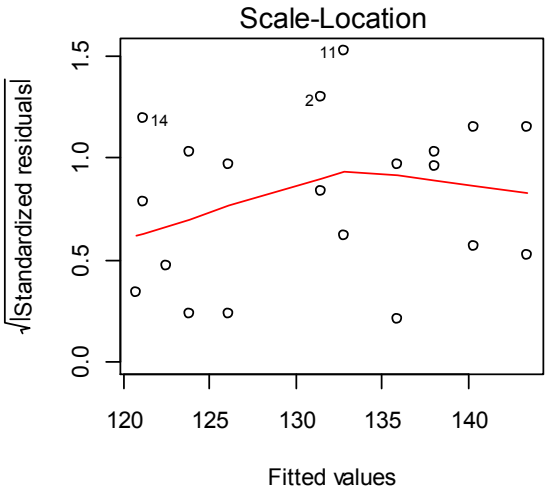
```
> layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
> plot(BLOODP.REG)
```

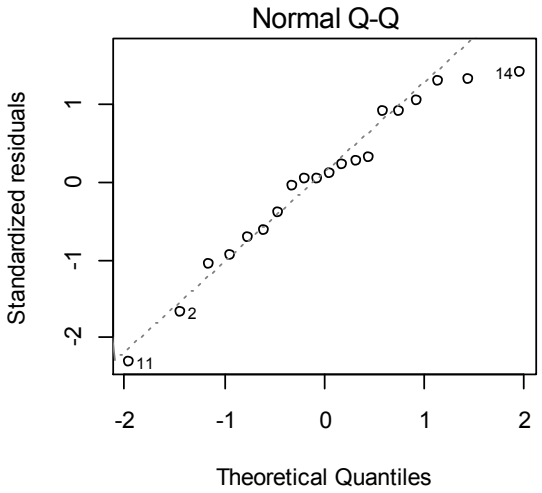
Residuals are distributed approximately at random: homogenetiy of variance met.

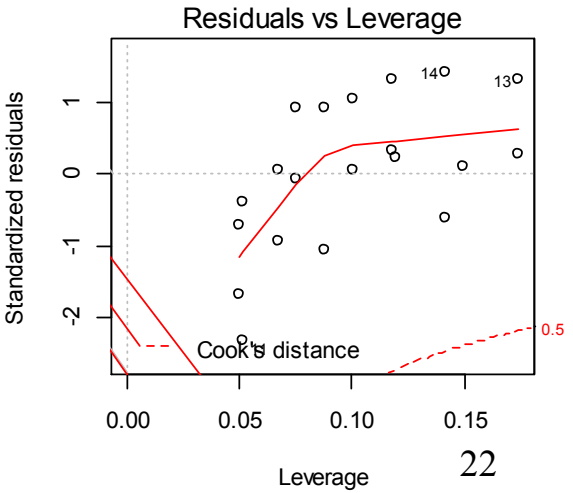
No important deviations in Q-Q plot: response variable normal.

None of the points approach the high Cook's distance contour(s): none of the observations are influential.



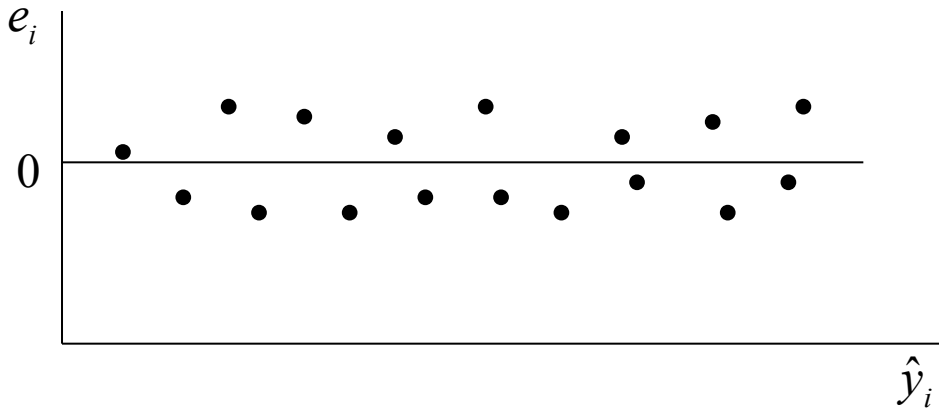




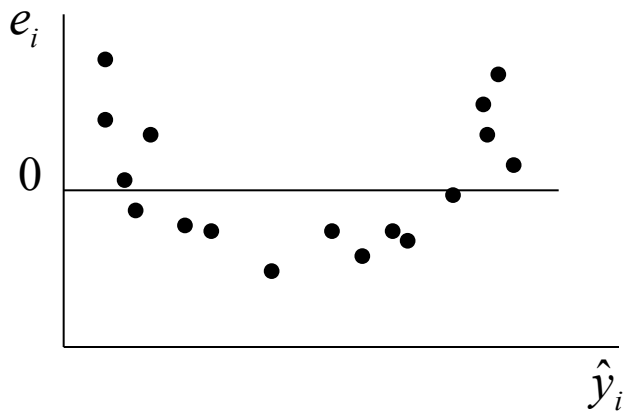


Some plots of residuals

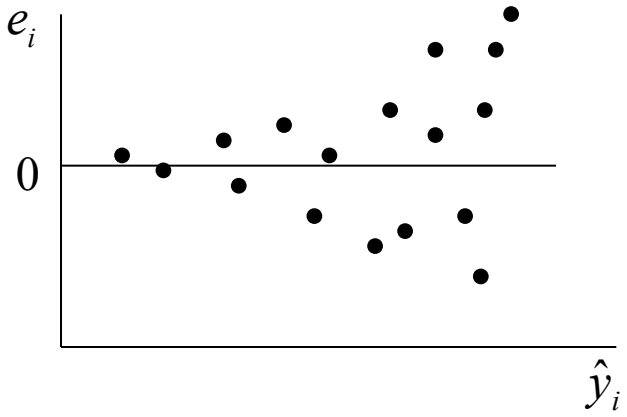
Ideal residual plot (random distribution around 0)



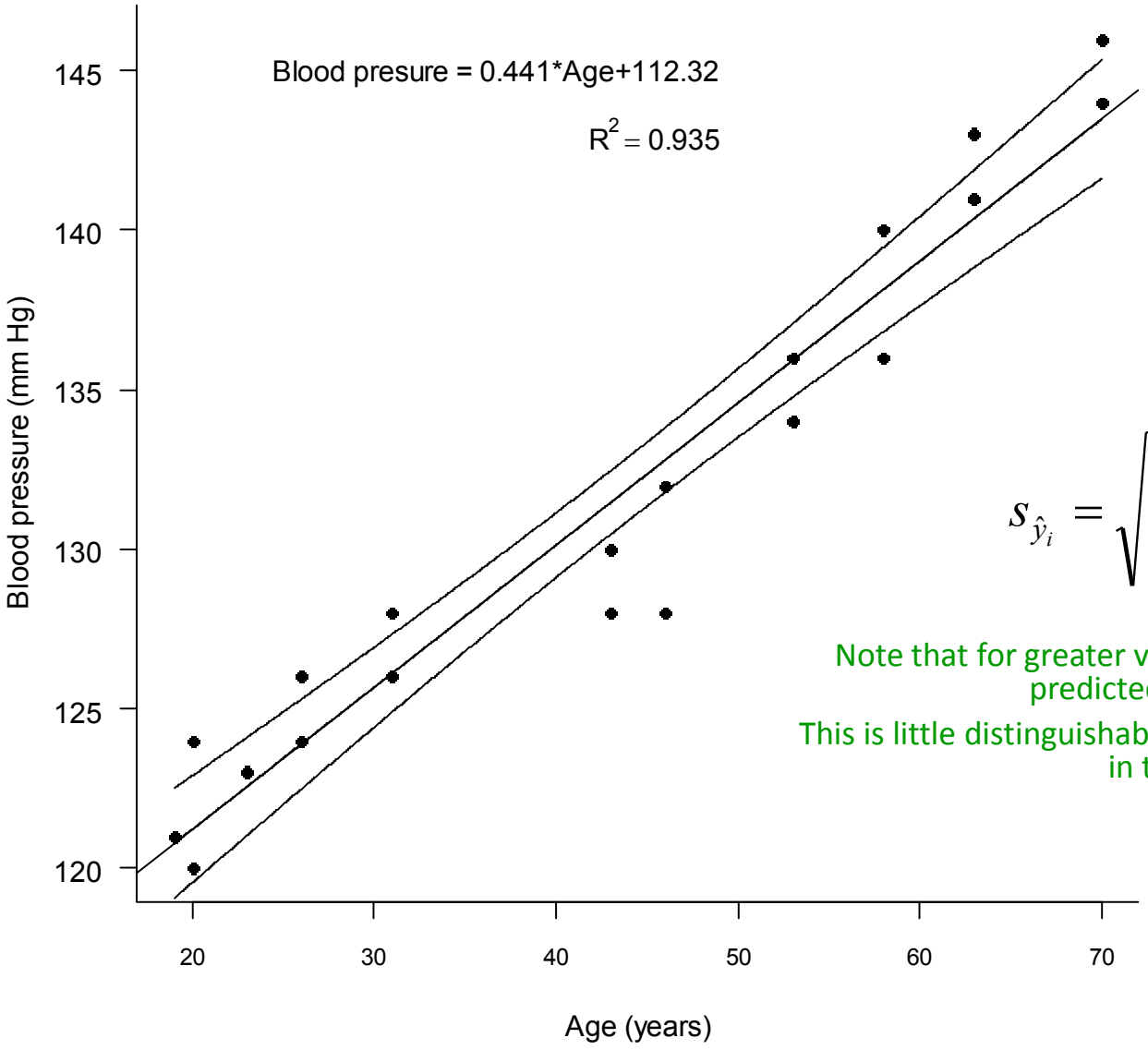
Model should involve curvature



Heterogeneous variance



Simple linear regression - Regression line and CL -



95% upper limit

Regression line

95% lower limit

Observations (•)

$$s_{\hat{y}_i} = \sqrt{MS_{RES}\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_{xx}}\right]}$$

Note that for greater values of x the standard error of predicted values is greater, and thus CL.

This is little distinguishable when the prediction is made in the interval of the observed x's.

Simple linear regression – program of the graphic -

```
> ## Summary scatterplot
> #cretate a plot with solid dots (pch=16) and no axis or labels
> plot(BLPRESS~AGE, pch=16, axes=F, xlab="", ylab="")
> #put the x-axis (axis1) with smaller label font size
> axis(1, cex.axis=.8)
> #put the x-axis label 3 lines down from the axis
> mtext(text="Age (years)", side=1, line=3)
> #put the y-axis (axis 2) with horizontal tick labels
> axis(2, las=1)
> #put the y-aix label 3 lines to the left of the axis
> mtext(text= "Blood pressure (mm Hg)", side=2, line=3)
> #add the regression line from the fitted model
> abline(BLOODP.REG)
> #add the regression formula
> text(50,145,"Blood presure = 0.441*Age+112.32", pos=2)
> #add the r squared value
> text(50,143,expression(paste(R^2==0.935)), pos=2)
> #create a sequence of 100 numbers spanning the range of ages
> x<-seq(min(AGE), max(AGE), l=1000)
> #for each value of x, calculate the upper and lower 95% confidence
> y<-predict(BLOODP.REG, data.frame(AGE=x), interval="c")
> #plot the upper and lower 95% confidence limits
> matlines(x,y, lty=1, col=1)
> #put an L-shaped box to complete the axis
> box(bty="l")
```

References

Fry J.C. 1993. *Biological Data Analysis*. IRL Press, Oxford.