

*Curso de Formación de Personal Investigador
Usuario de Animales para Experimentación*

SIMPLE LINEAR REGRESSION AND CORRELATION

Jesús Piedrafita Arilla

jesus.piedrafita@uab.cat

Departament de Ciència Animal i dels Aliments

UAB

Learning objectives

- ❑ Define the statistics related to the association between variables: correlation and regression.
- ❑ Explain how to explore the data: scatter plot (linear trend) and boxplots (evenness of the data) with R Commander.
- ❑ Define and calculate the correlation coefficient as a measure of association.
- ❑ Define linear regression as a statistic tool for prediction.
- ❑ Establish the model describing linear regression: intercept and slope, and the assumptions implied.
- ❑ Test the significance of the regression coefficients.
- ❑ Development of calculations by R Commander and interpretation of the outputs, including regression diagnostics.
- ❑ Define R^2 , the coefficient of determination.
- ❑ Distinguish outliers from influential observations and define procedures to detect them.

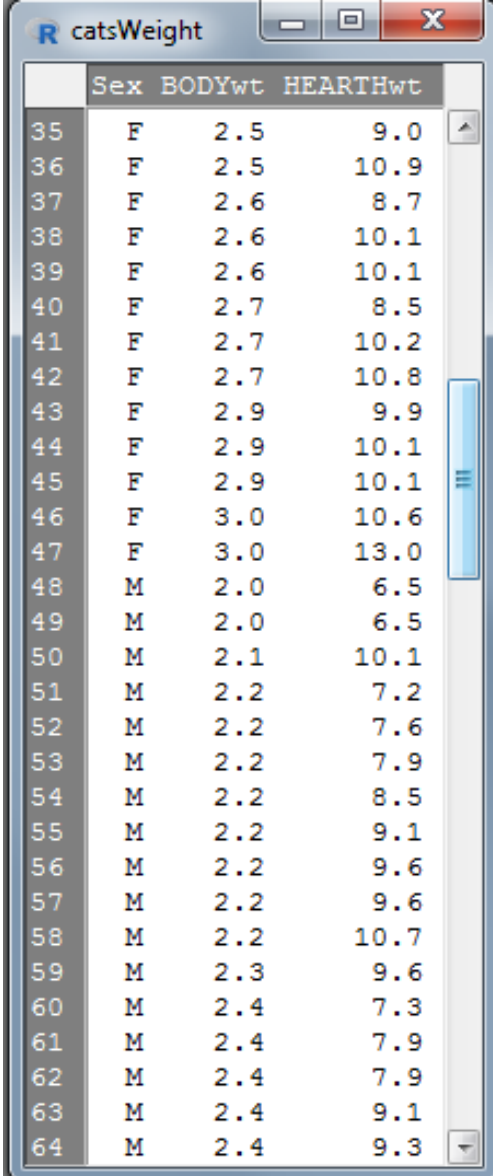
Analysis of several variables

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

To analyse the relationship of two variables we will return to the catsWeight data base. In this case we will pick up the variables body weight in kilograms (BODYwt) as the independent or explanatory variable, and hearth weight expressed in grams (HEARTwt) as the dependent or response variable.



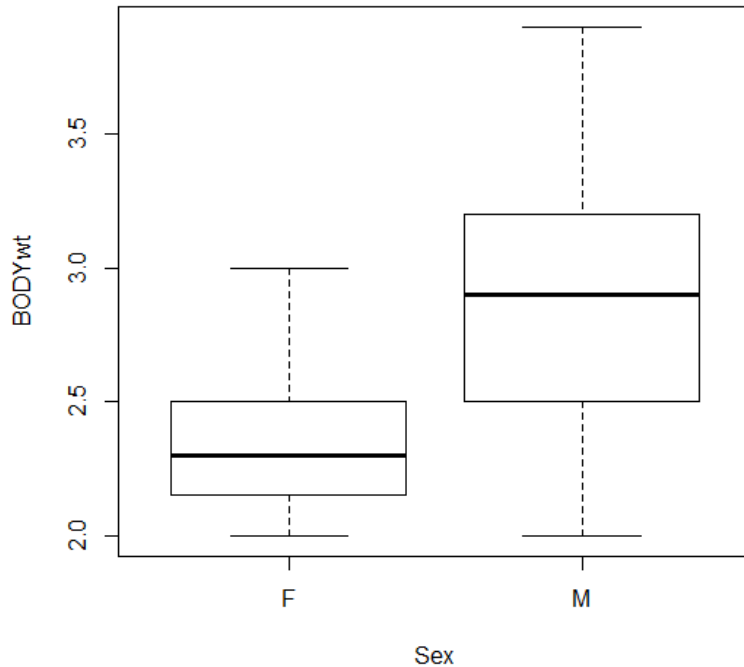
	Sex	BODYwt	HEARTHwt
35	F	2.5	9.0
36	F	2.5	10.9
37	F	2.6	8.7
38	F	2.6	10.1
39	F	2.6	10.1
40	F	2.7	8.5
41	F	2.7	10.2
42	F	2.7	10.8
43	F	2.9	9.9
44	F	2.9	10.1
45	F	2.9	10.1
46	F	3.0	10.6
47	F	3.0	13.0
48	M	2.0	6.5
49	M	2.0	6.5
50	M	2.1	10.1
51	M	2.2	7.2
52	M	2.2	7.6
53	M	2.2	7.9
54	M	2.2	8.5
55	M	2.2	9.1
56	M	2.2	9.6
57	M	2.2	9.6
58	M	2.2	10.7
59	M	2.3	9.6
60	M	2.4	7.3
61	M	2.4	7.9
62	M	2.4	7.9
63	M	2.4	9.1
64	M	2.4	9.3

Studying normality

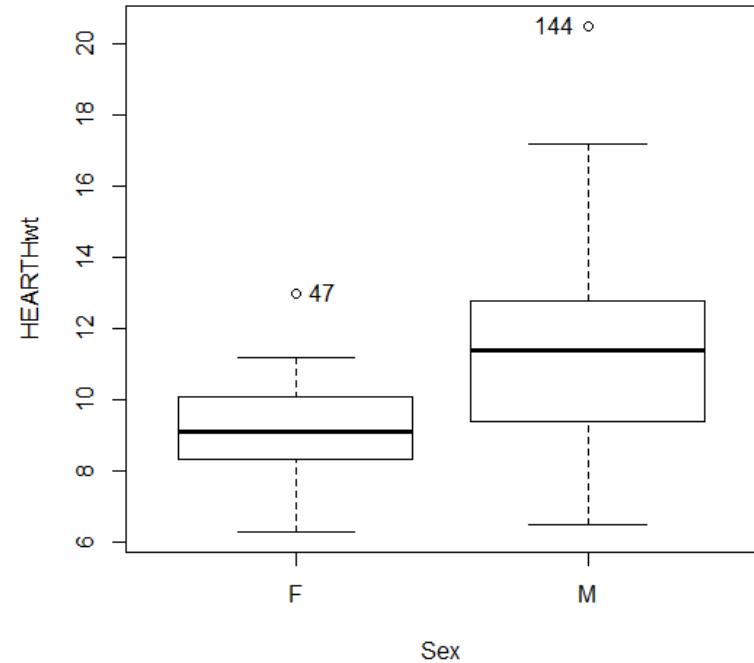
Data: catsWeight

Graphs > Boxplot; Plot by: Sex

Variable: BODYwt



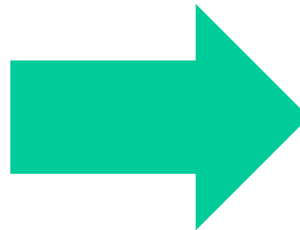
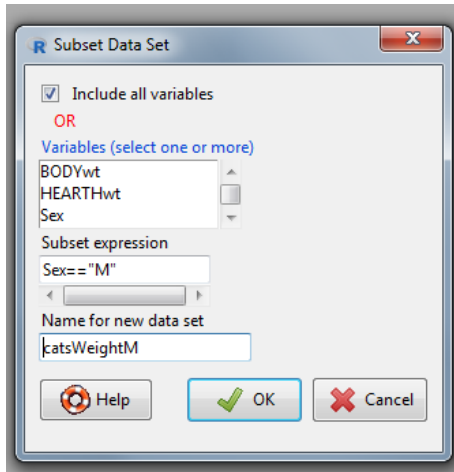
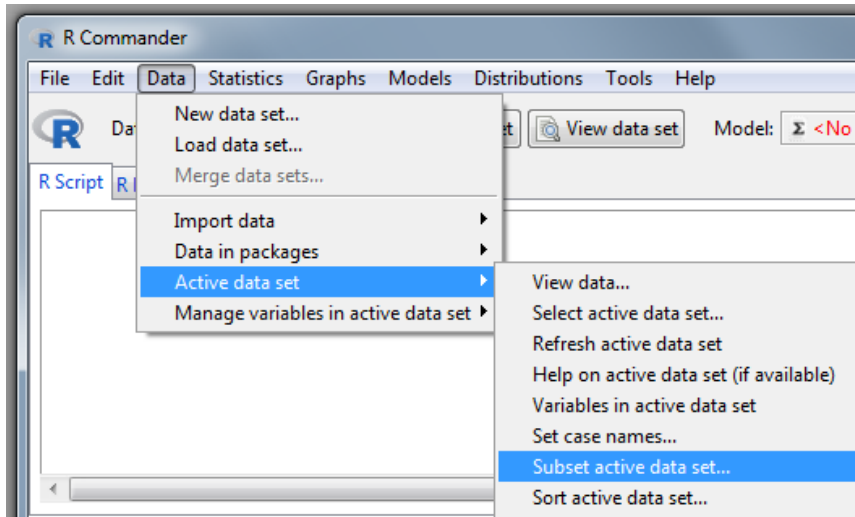
Variable: HEARTHwt



Different distributions for males and females in both variables

From now on, we will work only with the male's subset

Subsetting (for males, "M")



The screenshot shows a window titled 'catsWeightM' containing a table of data. The table has three columns: 'Sex', 'BODYwt', and 'HEARTHwt'. The rows represent individual cats, with their IDs (48-77) in the first column. All cats in this subset are male ('M').

	Sex	BODYwt	HEARTHwt
48	M	2.0	6.5
49	M	2.0	6.5
50	M	2.1	10.1
51	M	2.2	7.2
52	M	2.2	7.6
53	M	2.2	7.9
54	M	2.2	8.5
55	M	2.2	9.1
56	M	2.2	9.6
57	M	2.2	9.6
58	M	2.2	10.7
59	M	2.3	9.6
60	M	2.4	7.3
61	M	2.4	7.9
62	M	2.4	7.9
63	M	2.4	9.1
64	M	2.4	9.3
65	M	2.5	7.9
66	M	2.5	8.6
67	M	2.5	8.8
68	M	2.5	8.8
69	M	2.5	9.3
70	M	2.5	11.0
71	M	2.5	12.7
72	M	2.5	12.7
73	M	2.6	7.7
74	M	2.6	8.3
75	M	2.6	9.4
76	M	2.6	9.4
77	M	2.6	10.5

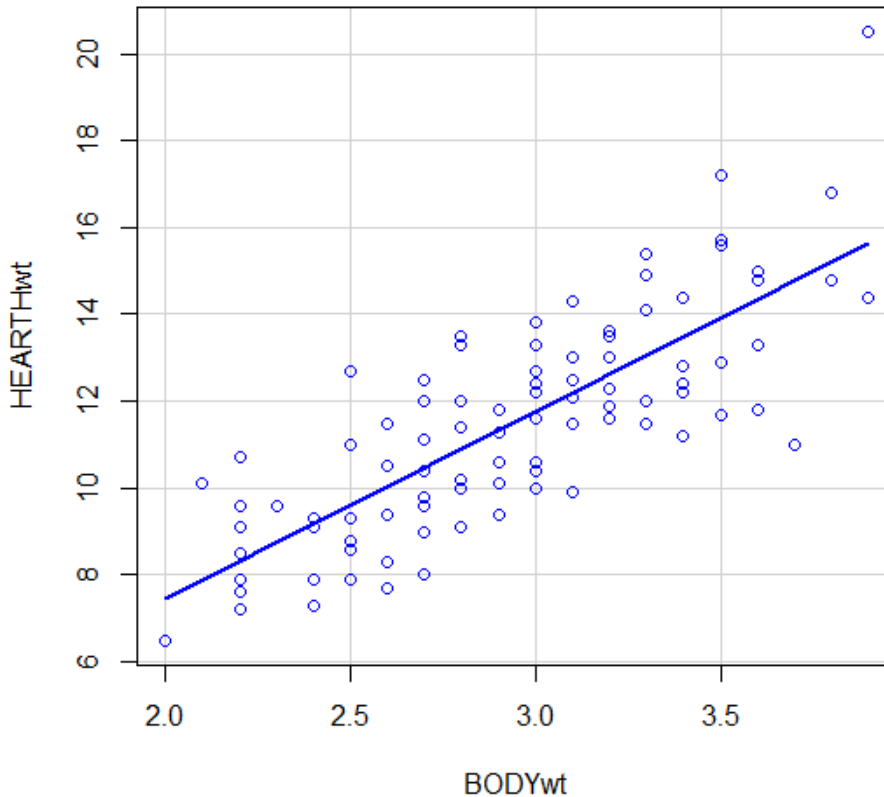
Plot of data from males

A plot for a pair of variables gives us a first impression about their relationship.

Data: catsWeightM

Graphs > Scatterplot

x-variable: BODYwt; y-variable: HEARTHwt; Options: Least-squares line



A linear relationship
is observed

Correlation (Pearson)

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

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

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.

Statistics > Summaries > Correlation test

Variables: BODYwt and HEARTHwt

Pearson's product-moment correlation

data: BODYwt and HEARTHwt

t = 12.688, df = 95, **p-value < 2.2e-16**

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.7051085 0.8569367

sample estimates:

cor

0.7930296 r

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

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.

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

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

ρ	Sample size for a power of	
	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

Correlation – sample size with R -

```
> library(pwr)  
> pwr.r.test(r = 0.5, sig.level = 0.05, power = 0.80)
```

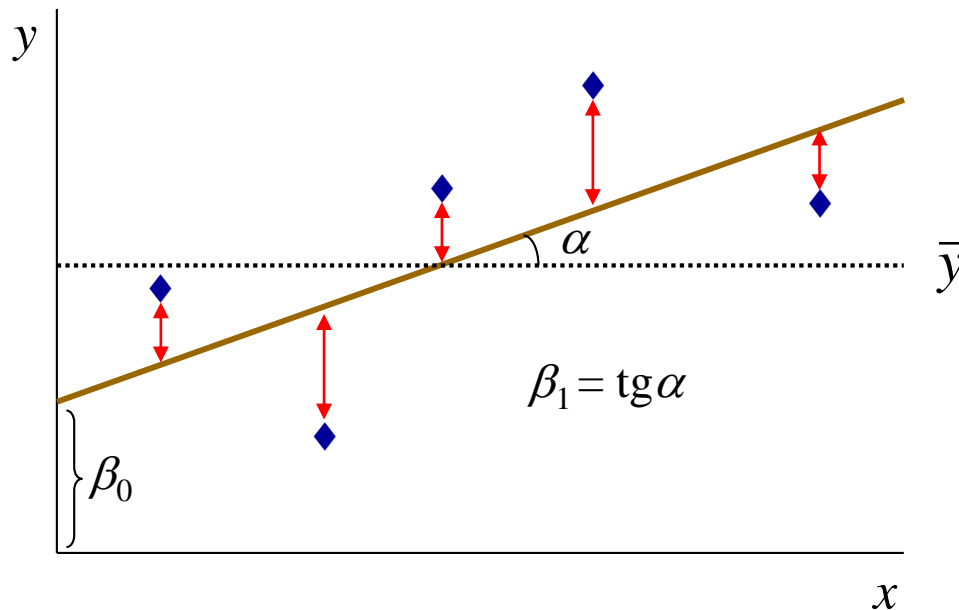
approximate correlation power calculation (arctangh
transformation)

```
      n = 28.24841  
      r = 0.5  
sig.level = 0.05  
  power = 0.8  
alternative = two.sided
```

Simple linear regression - the model -

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \rightarrow \text{Random error}$$

Dependent variable Intercept **Regression coefficient (slope)** **Independent variable**



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) and predicted values (on the slope).

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

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

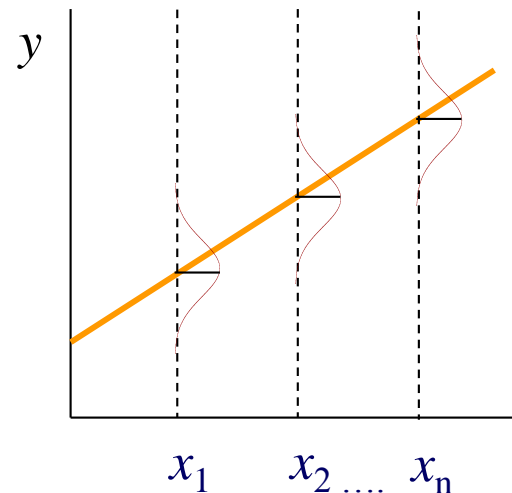
β_1 is the increase of the dependent variable when the independent variable increases 1 unit

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.



Simple linear regression - Results (1) -

Statistics > Fit models > Linear regression

Response variable: HEARTHwt; Explanatory variables: BODYwt

Coefficients:

		Estimate	Std. Error	t value	Pr(> t)
$\hat{\beta}_0$ ←	(Intercept)	-1.1841	0.9983	-1.186	0.239
$\hat{\beta}_1$ ←	BODYwt	4.3127	0.3399	12.688	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

For each increment of 1 kg of body weight,
hearth weight increases 4.3127 g

$$\hat{y}_i = -1.1841 + 4.3127x_i$$

$H_0: \beta_0 = \beta_1 = 0$, is rejected

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)} = \frac{4.3127}{0.3399} = 12.688$$

Simple linear regression - Results (1 cont.) -

Residual standard error: 1.557 on 95 degrees of freedom
Multiple R-squared: 0.6289, Adjusted R-squared: 0.625
F-statistic: 161 on 1 and 95 DF, p-value: < 2.2e-16

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. In this case 62.89% of the variation of blood pressure is explained by age.

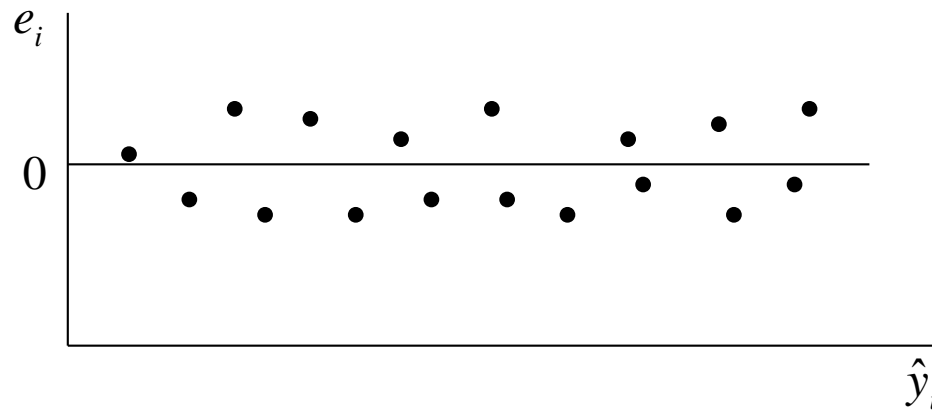
Adj R-Sq includes a correction to overcome the increment in *R-Squared* with the number of regressors (k).

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

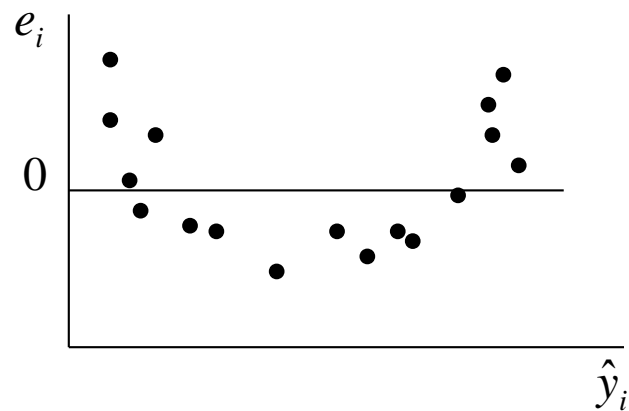
$$R_{adj}^2 = 1 - (1 - R^2) \frac{N - 1}{N - k - 1}$$

Diagnostics - Some plots of residuals

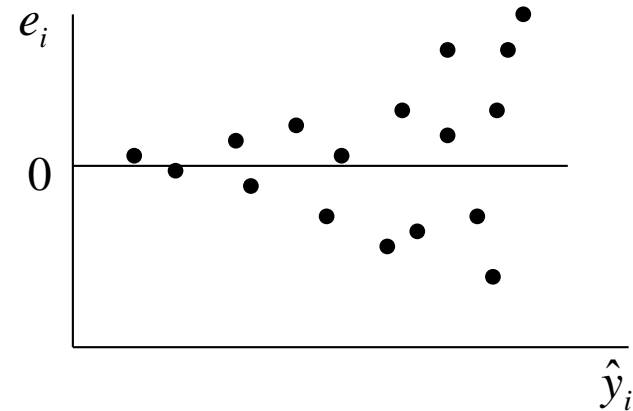
Ideal residual plot (random distribution around 0)



Model should involve curvature



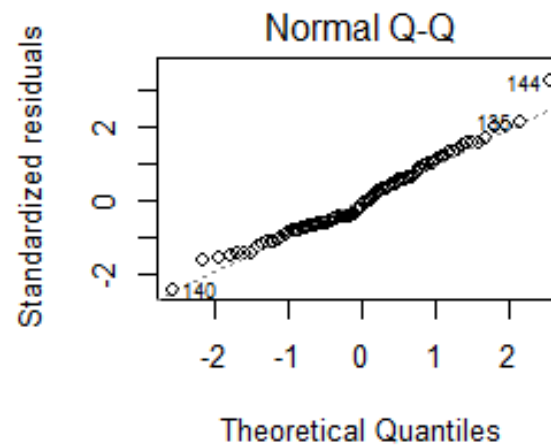
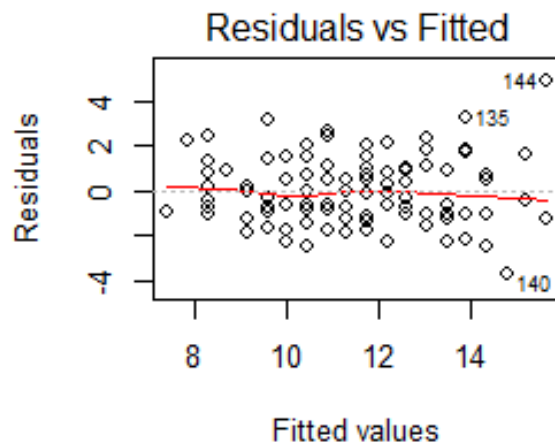
Heterogeneous variance



Simple linear regression – Results (2) -

Models > Graphs > Basic diagnostic plots

lm(HEARTHwt ~ BODYwt)

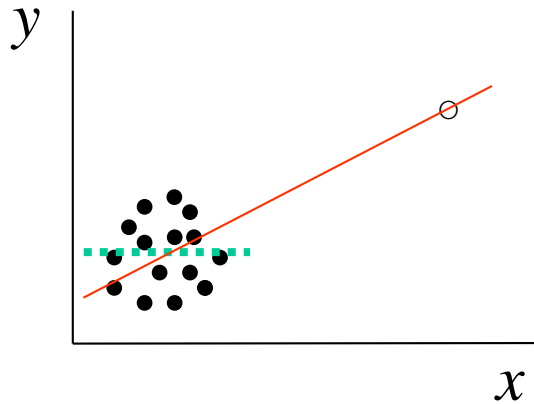


Residuals are distributed approximately at random around 0: homogeneity of variance met

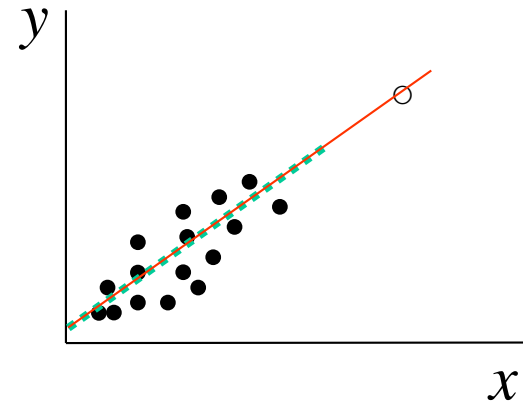
No important deviations in Q-Q plot: response variable normal

Some graphics about influential observations

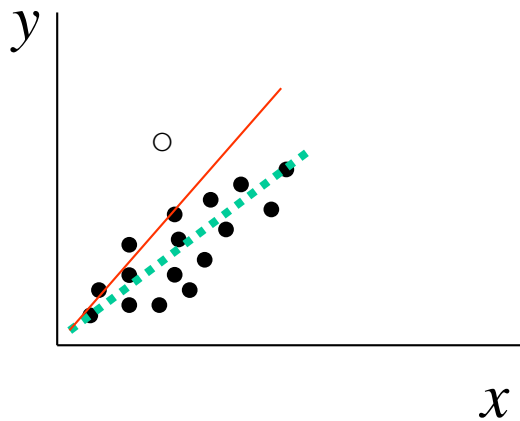
High leverage, influential



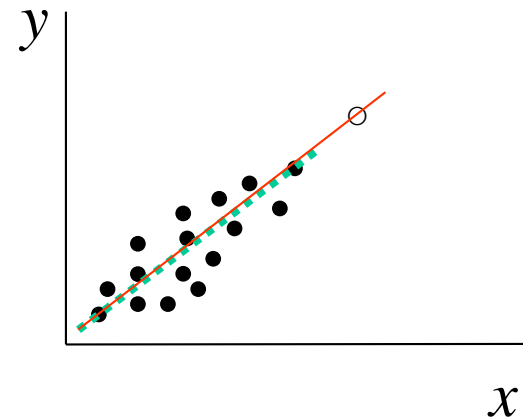
High leverage, not influential



Low leverage, influential



Low leverage, not influential



Some statistics useful for regression analysis

Standardized residual

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

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

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

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 influential.

Computed as the diagonal elements of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

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

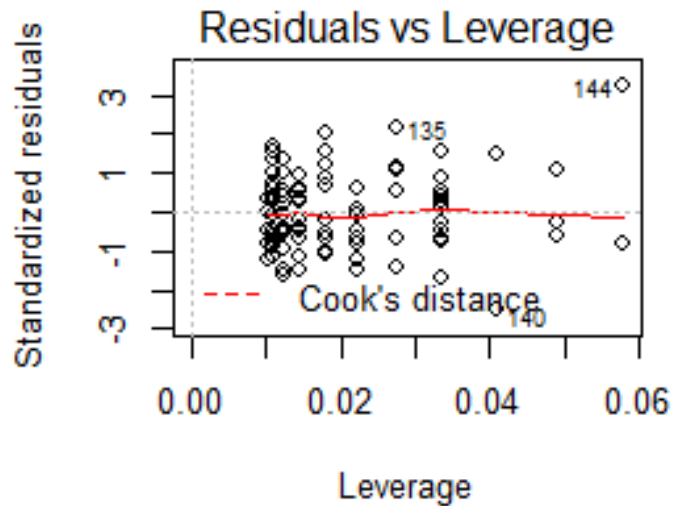
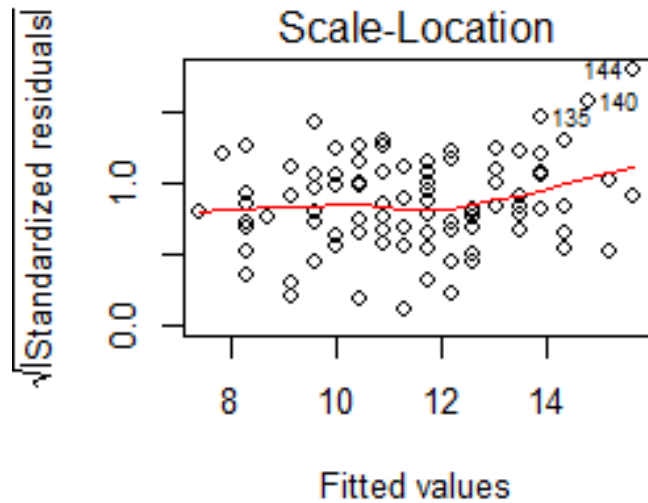
where c_{jj} are the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$. Analyse only *DFBETAS* corresponding to high values of *DFFITs*.

Cook's D

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

Simple linear regression – Results (3) -

Models > Graphs > Basic diagnostic plots



The $\sqrt{|\text{standardized residual}|}$ of obs. 140 is greater than $\sqrt{2}$ ($=1.41$): weak outlier; the one of obs. 144 is greater than $\sqrt{3}$ ($=1.73$): strong outlier

None of the points is outside the high Cook's distance contour(s), the dashed red lines: none of the observations is influential