## Simple Linear Regression: Introduction to Diagnostics

### Anscombe Data Frame (1.973)

A data frame proposed by Anscombe in 1.973 and composed by 4 sets of 11 (x,y) points. All of them provide the same estimates for the intercept and the slope in the least squared line. Goodness of fit in terms of the coefficient of determination (R2) provided for the 4 sets is the same common value 0.667. Data set are clearly different between them, a simple regression line is well-suited for data set A, but it is not suitable for sets B, C and D. Observations that are identified by outliers in its residual (ordinary residual or easier studentized residuals rstudent(model)), high leverage observations (hatvalues(model)) and influent data identified by atypical values in Cook’s distance (cooks.distance(model)) are diagnostic indicators that are easily understood in simple regression and are very useful in the diagnosis of general multiple regression models estimated by least squares.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **XA** | **YA** | **XB** | **YB** | **XC** | **YC** | **XD** | **YD** |
| *10* | *8,04* | *10* | *9,14* | *10* | *7,46* | *8* | *6,58* |
| *8* | *6,95* | *8* | *8,14* | *8* | *6,77* | *8* | *5,76* |
| *13* | *7,58* | *13* | *8,74* | *13* | *12,74* | *8* | *7,71* |
| *9* | *8,81* | *9* | *8,77* | *9* | *7,11* | *8* | *8,84* |
| *11* | *8,33* | *11* | *9,26* | *11* | *7,81* | *8* | *8,47* |
| *14* | *9,96* | *14* | *8,10* | *14* | *8,84* | *8* | *7,04* |
| *6* | *7,24* | *6* | *6,13* | *6* | *6,08* | *8* | *5,25* |
| *4* | *4,26* | *4* | *3,10* | *4* | *5,39* | *19* | *12,50* |
| *12* | *10,84* | *12* | *9,13* | *12* | *8,15* | *8* | *5,56* |
| *7* | *4,82* | *7* | *7,26* | *7* | *6,42* | *8* | *7,91* |
| *5* | *5,68* | *5* | *4,74* | *5* | *5,73* | *8* | *6,89* |

1. An studentized residuals is an atypical value if its absolute value is greater tan 2 or 3, depending on the sample size.
2. A Cook’s distance for an observation is atypical if it is greater than 4/(n-2) according to Chatterjee-Hadi criteria. A boxplot can be used to identify outliers in Cook’s distance, since not general rules are valid for any problem.
3. A *leverage* (hii) for an observation is high if it is greater than *2p/n. Theoretical máximum is 1.*

For every data set A to D plot and discuss results in the follow items:

1. Simple linear regression Y vs X. Plot residuals (studentized), leverages, Cook’s distance (cooks.distance(model)).
2. Identify pattern in residuals.
3. Check the presence of outliers in residuals (boxplot or statistical distribution).
4. Check the presence of observations with high leverages.
5. Check the presence of observations that are influent data.
6. Copy scatterplots of (x,y) data and residuals vs fitted values. Draw the estimated regression line.

|  |  |
| --- | --- |
|  |  |

***Y vs. X Residuals vs. Fitted values***

r = R2 = .

(recta ) = -----------------------------------------------------------------------------------------

|  |  |
| --- | --- |
|  |  |

***Y vs. X Residuals vs. Fitted values***

r = R2 = .

(recta ) = -----------------------------------------------------------------------------------------

|  |  |
| --- | --- |
|  |  |

***Y vs. X Residuals vs. Fitted values***

r = R2 = .

(recta ) = -----------------------------------------------------------------------------------------

|  |  |
| --- | --- |
|  |  |

***Y vs. X Residuals vs. Fitted values***

r = R2 = .

(recta ) = -----------------------------------------------------------------------------------------

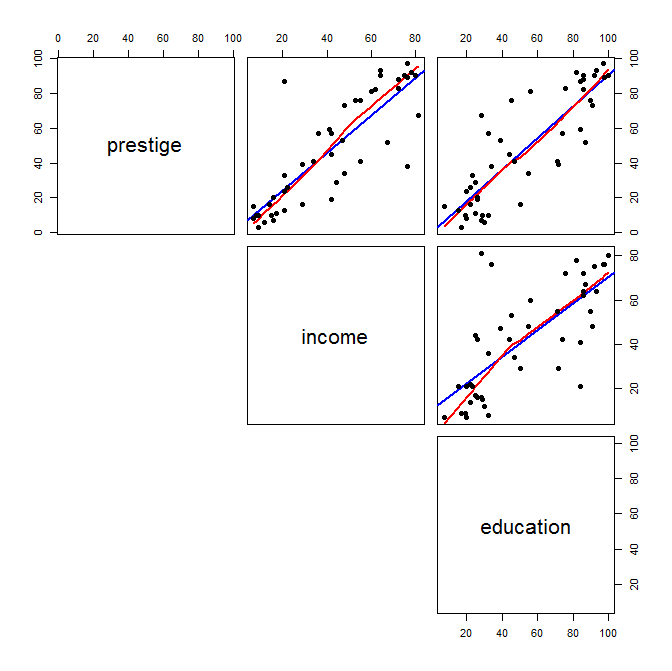
# Multiple regression:

****

## Example: Duncan data on prestige of professions or weight vs height in Davis

Use data(Duncan) in library(car). Open also data(Prestige).

Study correlations between numeric variables appearing in the work space. Explicative variables are income and education. Response variables is prestige and we have to propose a multiple regression model to explain the prestige of jobs.

****

### Suggested steps

* Correlation matrix in R, cor(duncan1, use="pairwise.complete.obs" )

1. Matrix of 2 by 2 scatterplots.
2. Forward regression from the nul model with a direction forward option in method step().

> duncan1.lm0 <- lm( prestige ~1, data=duncan1)

> summary(duncan1.lm0)

> step(duncan1.lm0, ~income+education, direction=”forward”, data=duncan1)

1. Backward regression from the model with INCOME+EDUCATION in backward direction option in method step().

> duncan1.lm2 <- lm( prestige ~ income+education, data=duncan1)

> summary(duncan1.lm2)

> step(duncan1.lm2, direction=”backward”,data=duncan1)

1. Use method step(.) in R from the nul model to the maximal model with direction specification “both” (it is the default)

> duncan1.lml <- lm( prestige ~income+education, data=duncan1)

> summary(duncan1.lm1)

> duncan1.lm<- step(duncan1.lm1, ~income+education, data=duncan1)

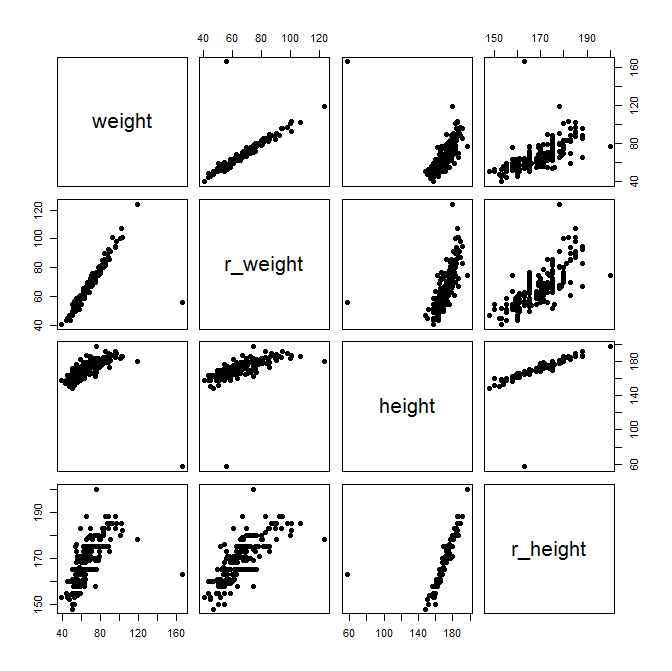
1. Linear correlation between a response variable and explicative variables might not be significative once some of the explicative variables are already included in the model.
2. ***A touch on diagnostics:*** Check outliers in residuals and influent data in the selected model. Compute histogram of studentized residuals (rstudent(model)), leverage (hatvalues(model)) and Cook’s distance (cooks.distance(model)).
3. R2 and global regression test .
4. Residual analysis:

* Detection of o*utliers*.
* Scatterplot of studentitzed residual *vs.* .
* Scatterplot of studentitzed residual *vs.* .
* Detection of *a priori* (leverage, hatvalues()) and *a posterior influent data (cooks.distance())*.
* Scatterplot of studentitzed residual *vs.* *leverage*.
* Scatterplot of studentitzed residual *vs.* Cook’s distance.

## Example: weight vs height in Davis

The Davis data frame has 200 rows and 5 columns. The subjects were men and women engaged in regular exercise. There are some missing data. This data frame contains the following columns:

* sex: A factor with levels: F, female; M, male.
* weight:Measured weight in kg.
* height: Measured height in cm.
* r\_weight : Reported weight in kg.
* r\_height : Reported height in cm.



Firstly, we examine the relationship between the reported weight and the actual weight in order to assess how data behaves. Pay attention to outliers.

Secondly, we focus on the classical relationship between weight (Y) and height (X): does a quadratic fit hold? Why?

### Suggested steps

* Correlation matrix in R, cor(Davis, use="pairwise.complete.obs" )

1. Matrix of 2 by 2 scatterplots.
2. Multiple regression weight (Y) vs r\_weight (Y). Interpret the regression equation and quality of the fit
3. Multiple regression weight (Y) vs height (X). Interpret the regression equation and quality of the fit
4. Multiple regression weight (Y) vs poly(height,2) (X). Can you Interpret the regression equation and quality of the fit?