## 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 ) = -----------------------------------------------------------------------------------------