Chapter 5: Model diagnostics

Dr. Abbas Rammal

Bachelor's degree in Mathematics
Option: DATA SCIENCE
October 2023

Plan

1. Introduction
2. The assumptions of the regression model
3. Graphical analysis of residues
4. Remedies
5. Outlier detection

Introduction

- The statistical properties and inference rely largely on assumptions about errors.
- It is necessary to ensure conformity with the assumptions.
- **Objective:** The objective of this chapter is to verify whether these hypotheses are valid.

The assumptions of the regression model

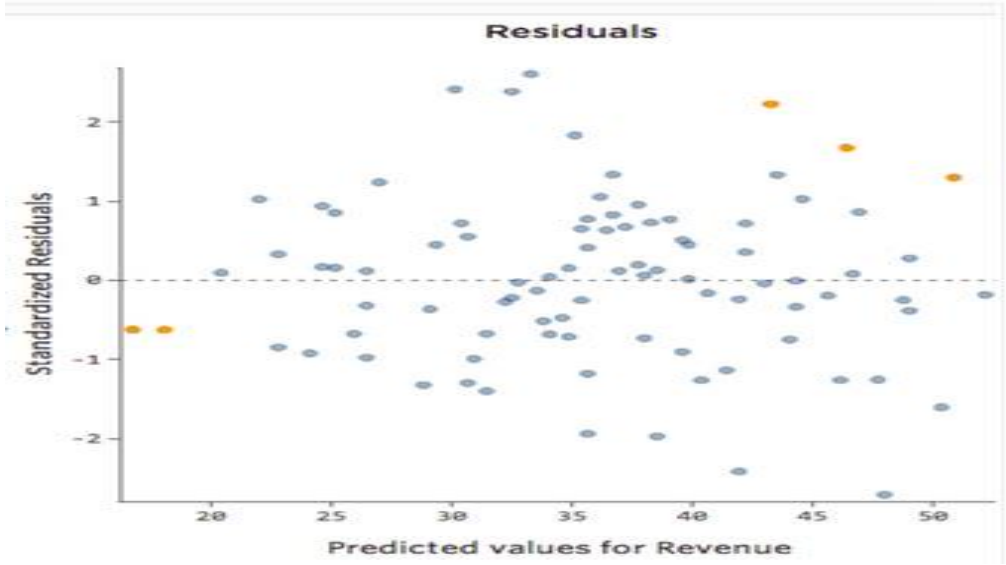
- Linearity of the model.
- Independence of errors.
- Homoscedasticity of errors (constant variance)
- Normality of errors.

Graphical analysis of residues

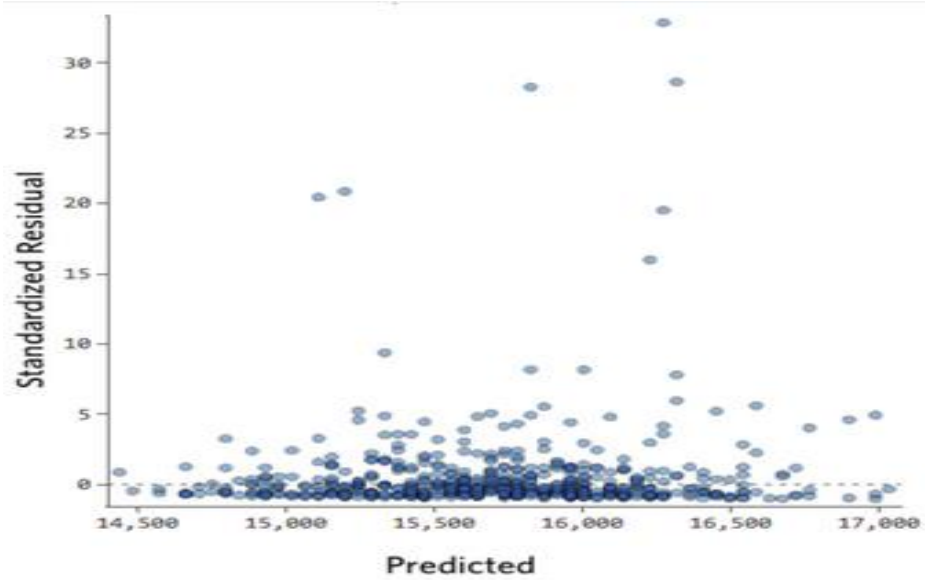
- Among the most used graphic methods, we have:
- The graph of residuals versus estimated values $\widehat{y_i}$.
- The graph of residuals versus data relating to explanatory variables x_{ij}.
- The graph of residuals versus time (for time series).
- QQ-plot of residuals.

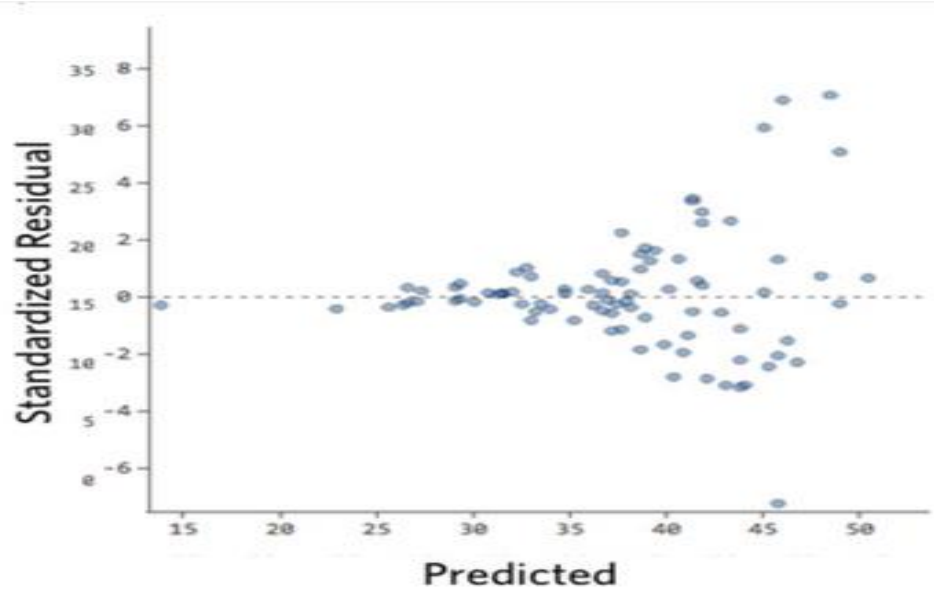
Plot of residuals versus estimated values (Residual Plot)

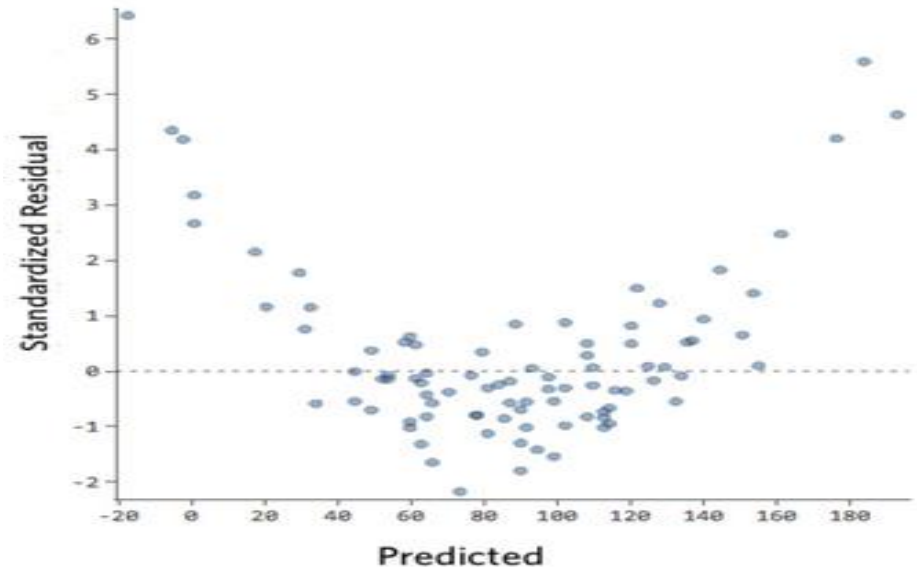
- The most useful way to plot the residuals is to use the predicted $\widehat{y_i}$ values on the x-axis and the residuals (e_i) on the y-axis.

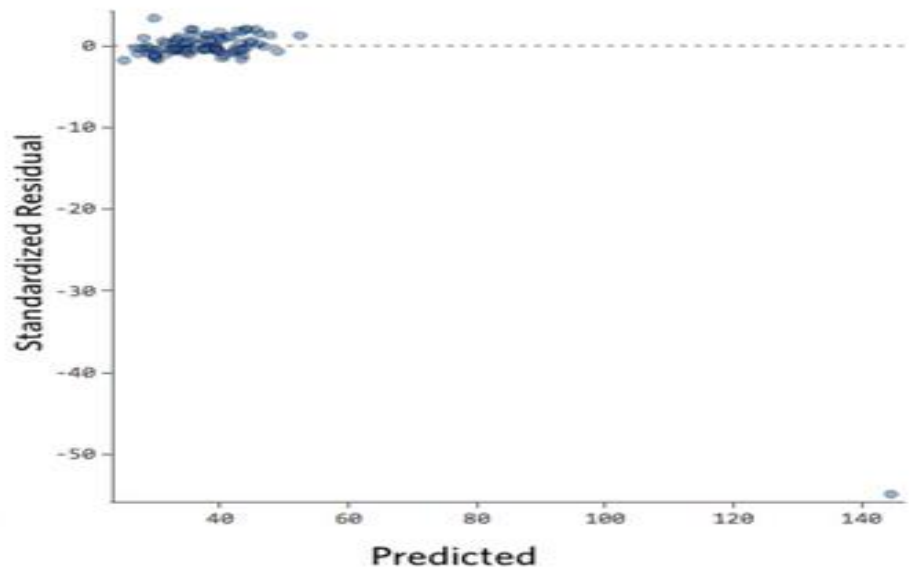


- The distance of the line to 0 corresponds to the incorrect prediction of this value.
- A horizontal behavior of the residuals is expected to consider the model as satisfactory. The residual graph shows a fairly random pattern. There are no clear models.


Standardized Residual
Predicted
14,500
15,000
15,500
16,000
16,500
17,000


Standardized Residual
Predicted


Standardized Residual
Predicted


Standardized Residual
Predicted

- These graphs are not evenly distributed vertically, or they have an outlier, or they have a clear shape.
- If you can detect a clear pattern or trend in your residuals, your model can be improved.

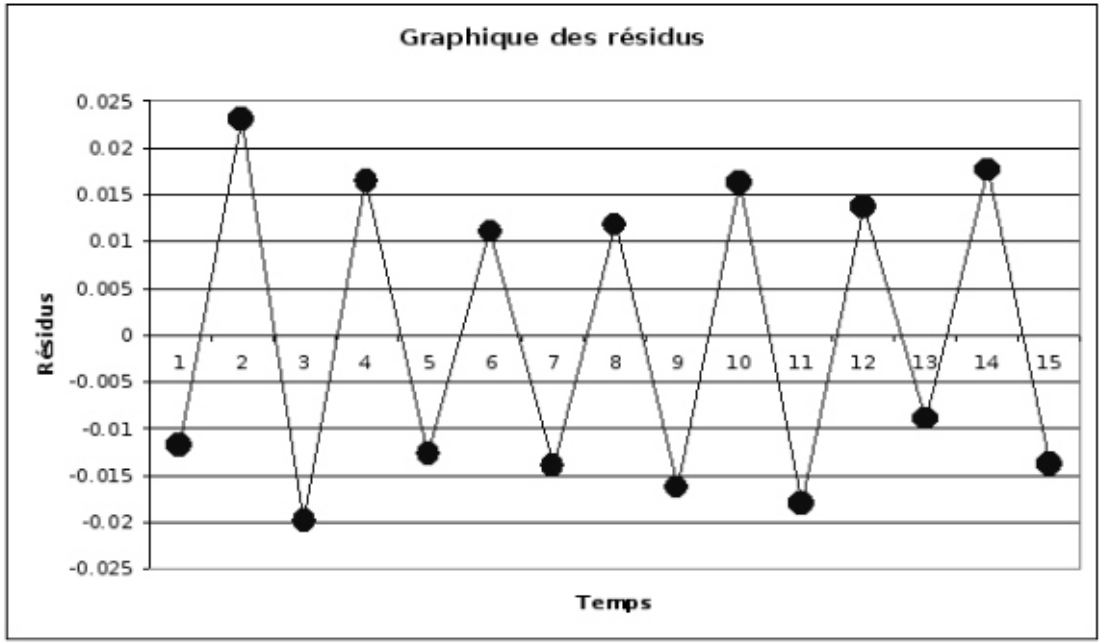
1. **Problem 1: Heteroscedasticity of errors: Figures 1 and 2**
2. **Problem 2: Non linearity: Figure 3**
3. **Problem 3: Outliers Figure 4**

Plot of residuals versus explanatory variables

- This type of chart looks similar to the previous one.
- It consists of using the values of the explanatory variables x_{ij} on the x-axis and the residuals (e_i) on the y-axis.
- It allows you to choose the correct form to use in the model for each of the variables considered.
- A horizontal behavior of the residues is expected to consider the model as satisfactory.

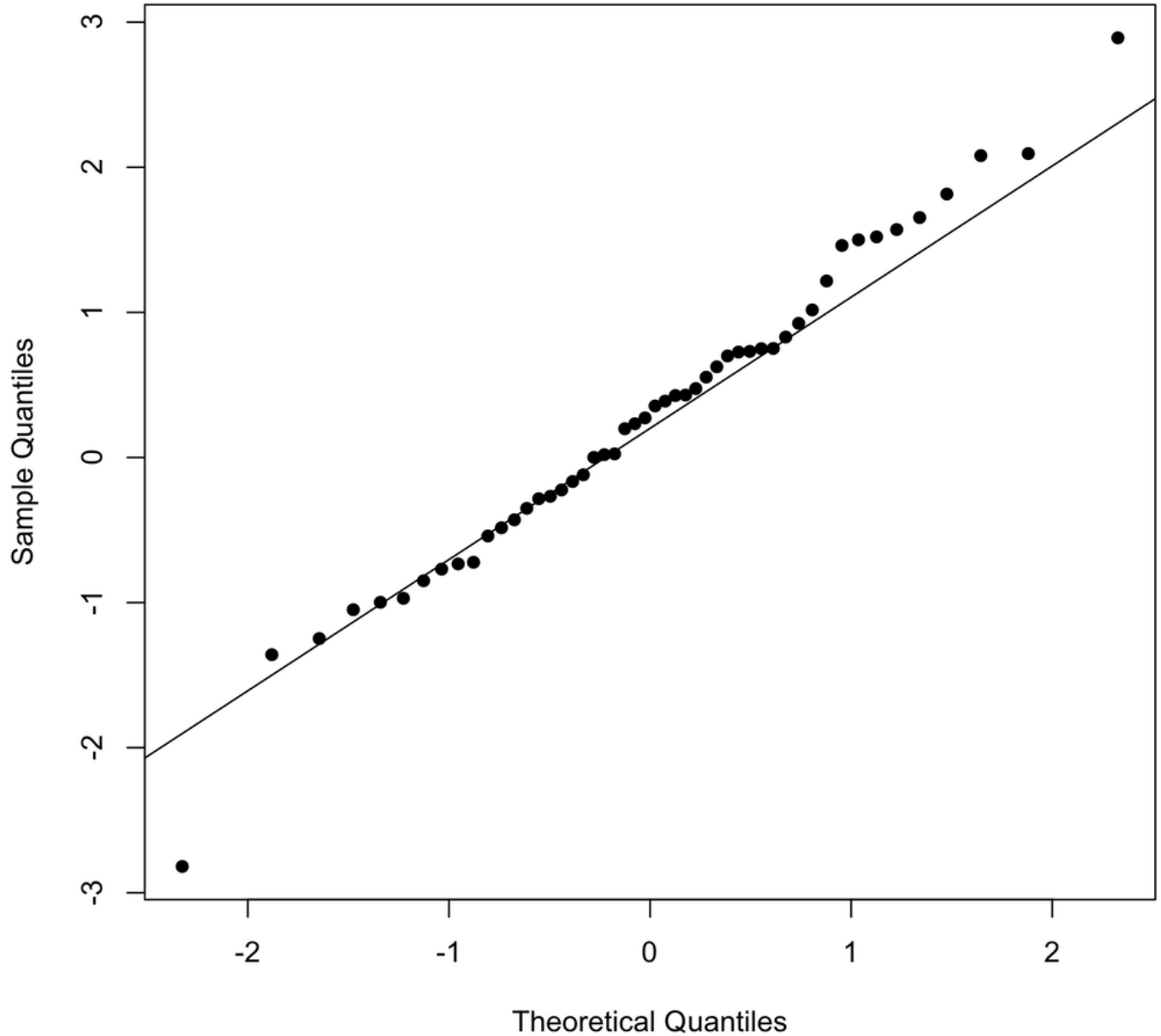
Plot of residuals versus time

- This type of graph is mainly used in the context of time series analysis (time is an explanatory variable)
- The residues are represented as a function of time.



QQ-plot of residuals

- In order to test the normality of the residuals, we can draw the QQ-plot of the residuals.
- To draw the QQ-plot, we order the e_i in ascending order and then we associate for each e_i, the quantile q_i of a reduced centered normal distribution.
- We plot the residuals e_i as a function of the quantiles q_i.
- If the errors ε_i are normally distributed, the points on the graph should be approximately aligned with the equation line $e_i = q_i$.


Sample Quantiles
Theoretical Quantiles
3
2
1
0
-1
-2
-3
-2
-1
0
1
2

Remedies

- **Problem 1: Heteroscedasticity of errors**: The errors relating to observations with large $\widehat{y_i}$ have a greater variance than the errors to observations with small $\widehat{y_i}$.
- **Remedy: Weighted Least Squares Method**: This method is used when the error variance is not constant.

- Instead of having for everything $i = 1, \cdots, n$: $V(\varepsilon_i) = \sigma^2$ we will have:

$V(\varepsilon_i) = \dfrac{\sigma^2}{w_i}$ où $w_i > 0$ are the weights for each i

$$V(\varepsilon) = \sigma^2 \mathbf{V} = \sigma^2 \begin{pmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1/w_n \end{pmatrix}$$

On note:

- $Y_w = WY$
- $X_w = WX$
- $\varepsilon_w = W\varepsilon$

avec: $W = \begin{pmatrix} \sqrt{w_1} & 0 & \cdots & 0 \\ 0 & \sqrt{w_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sqrt{w_n} \end{pmatrix}$ tel que $W'W = \mathbf{V}^{-1}$.

- The equivalent model is:

$$Y_w = X_w\beta + \varepsilon_w$$

$$\text{où } V(\varepsilon_w) = V(W\varepsilon) = WV(\varepsilon)W' = \sigma^2 WVW' = \sigma^2 I_n.$$

- We then apply the least squares method since the variance of the errors is constant in this new model.
- The vector of estimators is: $\hat{\beta}_w = (X'V^{-1}X)^{-1}X'V^{-1}Y$
- The vector of estimated values: $\hat{Y}_w = X\hat{\beta}_w$
- The variance σ^2 is estimated by: $s_w^2 = \dfrac{\sum w_i(y_i - \hat{y}_i)^2}{n - p - 1}$

Remedies

- **Problem 2: Non-linearity of the model:** The model is not linear.
- **Remedy: Transformation of variables:** A possible remedy is to transform one or more variables of the model in order to make it linear.

- We must replace the linear model

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

by another regression model which is not linear.

For example:

$$y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i \quad (1)$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \quad (2)$$

$$y_i = \beta_0 . \beta_1^{x_i} . \varepsilon_i \quad (3)$$

$$y_i = \beta_0 . x_i^{\beta_1} . \varepsilon_i \quad (4)$$

- For models (1) and (2), we apply a transformation under the form $z_i = x_i^2$ and models (1) and (2) become respectively:

$$y_i = \beta_0 + \beta_1 z_i + \varepsilon_i \qquad (5)$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i \qquad (6)$$

- For models (3) and (4), we apply a transformation in logarithm form and models (3) and (4) become respectively:

$$log(y_i) = log(\beta_0) + log(\beta_1) x_i + log(\varepsilon_i) \qquad (7)$$

$$log(y_i) = log(\beta_0) + \beta_1 log(x_i) + log(\varepsilon_i) \qquad (8)$$

Remedies

- **Problem 3: Outliers:** The regression model has an outlier data point on one of the variables. An outlier is an observation that does not resemble the rest of the data.
- **Remedy: Detection of outliers:** A possible remedy is to reject this observation or to impose a detailed investigation into these observations.

- Outlier (or atypical) points have a strong influence on the slope of the regression line and consequently on the value of the correlation coefficient. A single outlier can considerably modify the slope of the regression line and therefore the value of the correlation.

- Needless to say, outliers can not only artificially increase the value of a correlation coefficient, but they can also decrease the value of a “legitimate” correlation.

- An outlier observation is an observation whose residual (in absolute value) is much higher than the others.
- It is possible that this observation is a measurement or data entry error, or it is simply wrong, in which case it should be removed.
- (As an example, let's take a dataset containing water temperature measurements. If we find a temperature value at 150°C, we will say that this is impossible. In fact, water evaporates at 100°C. C. This value at 150°C is therefore meaningless and this observation will have to be deleted.)

- It is possible that what appears to be just a few outliers is actually a skewed distribution. You must transform the variable if one of your variables has an asymmetric distribution (i.e. it does not have a bell shape).
- If this is indeed a legitimate outlier, the impact of the outlier must be assessed.

Outlier detection

- Most graphical and statistical methods for detecting outliers use the notion of studentized residuals.
- Studentization of residues:
    1. Internal studentized residues
    2. External studentized residues
- Point of Levier

Hat Matrix

- We note “Hat matrix” the matrix defined by:

$$H = X(X'X)^{-1}X'.$$

- This matrix is a square matrix, n × n and symmetric:
  1. $H^2 = H$
  2. $H' = H$
  3. $HX = X$
  4. $(I_n - H)X = 0$
  5. $H(I_n - H) = 0$
  6. $(I_n - H)^2 = (I_n - H)$
  7. $\hat{Y} = HY$

- The diagonal element of the matrix H is given by $h_{ii} = x_i(X'X)^{-1}x_i'$

Properties

- All elements of the matrix H are between -1 and +1.

$$-1 \leq h_{ij} \leq 1$$

- The elements of the diagonal are between 0 and 1.

$$0 \leq h_{ii} \leq 1$$

- The sum of the elements of each column and each row is equal to 1

$$\sum_{i=1}^{n} h_{ij} = 1, j = 1, \cdots, n \quad \sum_{j=1}^{n} h_{ij} = 1, i = 1, \cdots, n$$

Properties

- The sum of the elements of the diagonal is equal to p + 1

$$Trace(H) = p + 1$$

- The sum of the squares of all elements of the matrix H is equal to p + 1

$$\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}^2 = p + 1$$

Variance of residuals

- We have $e = Y - \hat{Y}$
- We can express e in terms of H: $e = (I_n - H)Y$
- We then have: $\mathbb{V}(e) = \sigma^2(I_n - H)$ et $\mathbb{V}(e_i) = \sigma^2(1 - h_{ii})$
- We estimate the variances by $s^2(e_i) = s^2(1 - h_{ii})$

Internal studentized residues

- The internal studentized residues (also called standardized residue) are defined by:

$$r_i = \frac{e_i}{s.\sqrt{1 - h_{ii}}}$$

$\mathrm{r_i}$ follows approximately a Student's distribution $t(n - p - 1)$.

- **Rule**

i will be suspect if $|r_i| > t_{(1-\alpha/2,(n-p-1))}$ (quantile of Student's law (1 − α/2, n − p − 1), for the threshold α = 5%, with the approximation by a normal law if n large).

External studentized residues

- We replace in the expression of r_i the estimate of s by $s_{(-i)}$ estimate of s by redoing the adjustment of the model without observation i, which makes e_i independent of $s_{(-i)}$.
- The external studentized residues (called RSTUDENT) are then defined by:

$$r_{(-i)} = \frac{e_i}{s_{(-i)} \cdot \sqrt{1 - h_{ii}}}$$

- r_i follows approximately a Student's distribution $t(n - p - 1)$.

- **Rule**

  i will be suspect if $|r_{(-i)}| > t_{(1-\alpha/2,(n-p-1))}$ (quantile of Student's law (1 − α/2, n − p − 1), for the threshold α = 5%, with the approximation by a normal law if n large).

Relationship between internal and external studentized residues

- We have the following relationship between the internal and external studentized residues

$$r_{(-i)} = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}}$$

- The advantage of this relationship is that we will not need to restart a estimation procedure by ordinary least squares by removing the i^{th} observation.

Point of Leverage

- The levers of the observations are the n values h_{ii}.
- A lever represents the influence of observation i on the adjusted value $\widehat{y_i}$, because of the values x_i taken by the variables in i.

- **Rule**

  If the levers were all equal, the common value would be $(\mathrm{p}+1)/\mathrm{n}$. A leverage greater than $2^{(\mathrm{p}+1)/\mathrm{n}}$ is suspect.

Cook Distance

- Cook's distance measures the influence of an observation on all forecasts by taking into account leverage and importance of residuals.
- The Cook distance for the i^{th} observation is defined by:

$$C_i = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}}r_i^2.$$

- The detection strategy most often consists of identifying atypical points by comparing the Cook distances with the value 1 (i.e. $C_i > 1 \Rightarrow$ atypical observations) then explaining this influence by considering, for these observations, their residue as well as their leverage effect.

Interpretation

Interpretation of a high C_i value observed:

- Let r_i^2 high : aberrant data
- Let $\frac{h_{ii}}{1-h_{ii}}$ high : data having a leverage effect,
- either both