MAP553 Regression

Chapter 3: Atypical points and model validation

Contents

For some atypical observations, the values of the response variable Y and/or predictors X_j appear to behave differently from the majority of observations, these points are called ***outliers***. Moreover, observations that do not follow the same linear regression model that most data are called ***regression outliers***.

Underline that it is important to first check the most obvious reasons of such of points : measurement errors, data transcription errors, and so on. For example : a boat passengers is written to be 400 years, a 1 year old baby running the 100m in 10 seconds...

Then, the remaining *outliers* are not necessarily wrong. Indeed, *outliers* sometimes reveal a particular phenomenon that may be different from the model followed by the majority of observations. Keep in mind that even if the aim of a model is to explain as well as possible a general phenomenon, it can have its own limits. Thus, *outliers* can suggest to track for more elaborate models (missing regressor,...).

In the regression setting, an typical values (*outliers*) can occur in three main ways :

- in the response Y but not in the predictors X_j,
- in the predictors X_j but not in the response variable Y,
- in both Y and X directions.

3.1. Atypical points

Let's place ourselves in the context of the Iinear regression and consider an *outlier* in the Y-direction (an atypical value in a response Y_i) but not in the predictors X_{ij}. We detect them easily by an univariate detection. Let us consider a toy example. A simple boxplot reveal the *outliers* (see figure 1, right).

According to the scatter plot (figure 1, left), a linear model can be considered. We set

$$Y_i = ax_i + \varepsilon_i, \qquad \varepsilon_i \text{ i.i.d. } \mathcal{N}(0, \sigma^2).$$

By simple calculation, the ordinary least square estimator (the same as the maximum likelihood estimator in our setting) is such that

$$\widehat{a} = \frac{s_{x,y}}{s_x^2} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2} \quad \text{and} \quad \widehat{b} = \overline{y} - \widehat{a}\,\overline{x} \quad \Rightarrow \quad y = \widehat{a}x + \widehat{b}.$$

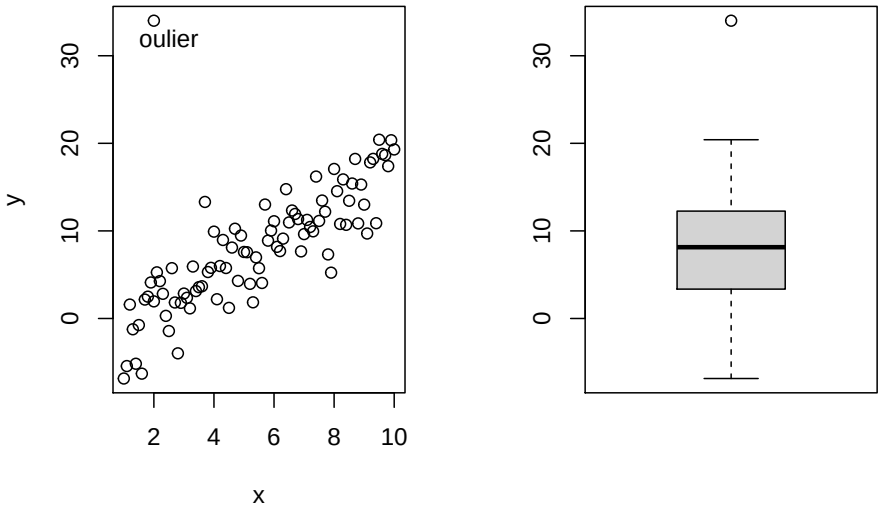

Figure 1: Scatter plot of the toy dataset/Boxplot of the toy dataset

Plot now (figure 2, right) the two least square lines with and without the outlier.

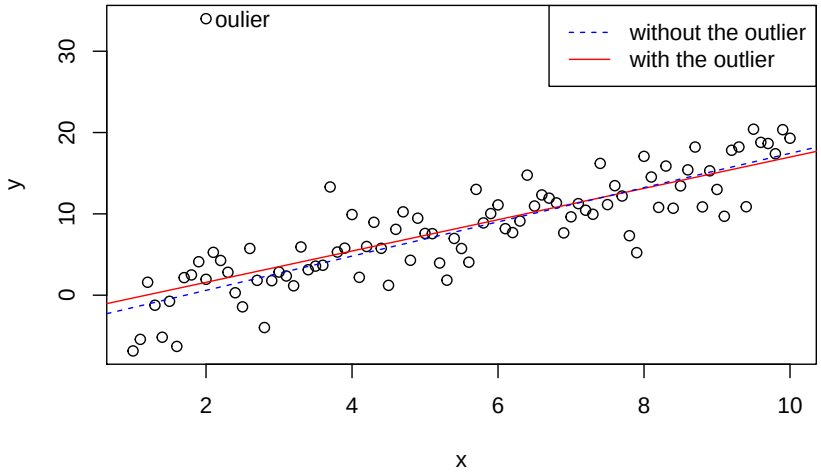

Figure 2: Scatter plot of the toy dataset/The least square lines with and without the outlier

Comments:

- Here, the outlier has only a small effect on the estimation. Indeed, removing this point slightly changes the regression line (least squares line).
- This type of atypical observations (*outliers*) has an impact on the estimation of σ^2 so on the residuals $\widehat{\varepsilon} = Y - \widehat{Y}$.
- **The *regression outliers* can be detected by a residuals analysis.**

3.2. Isolated observations

An *isolated observation* has atypical values in the predictors X_{ij}. It means that the values $\left(X_{ij}\right)_j$ of the observation i are relatively far from all the value $\left(X_{i'j}\right)_j$ of the other observations $i' \neq i$. Let us consider an other toy example. According to the scatter plot (figure~**??**), a linear model can be considered. We set

$$Y_i = ax_i + \varepsilon_i, \qquad \varepsilon_i \text{ i.i.d. } \mathcal{N}(0, \sigma^2).$$

By simple calculation, the ordinary least square estimator (the same as the maximum likelihood estimator in our setting) is such that

$$\widehat{a} = \frac{s_{x,y}}{s_x^2} = \frac{\overline{xy} - \overline{xy}}{\overline{x^2} - (\overline{x})^2} \quad \text{and} \quad \widehat{b} = \overline{y} - \widehat{a}\,\overline{x} \quad \Rightarrow \quad y = \widehat{a}x + \widehat{b}.$$

Plot now (figure~3) the two least square lines with and without the outlier.

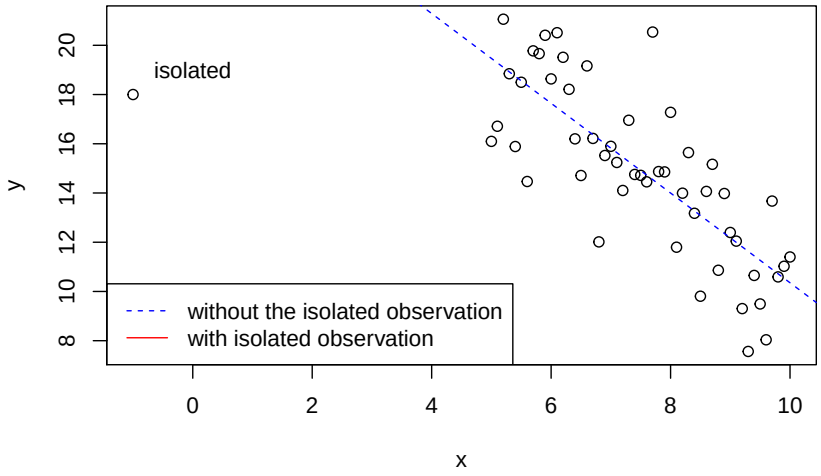

Figure 3: Scatter plot of the toy dataset/The least square lines with and without the isolated observation

Comments:

- Here, the *isolated observation* has a real impact on the estimation. Indeed, removing this point significantly changes the regression line (least squares line).
- This type of atypical observations (*isolated observations*) has an impact on the estimation of β. Here, the*isolated observation* influe on the estimation of β.
- Such of points which influe the estimation of β, are called *leverage point*.
- ***Leverage points*** **can be detected by a multivariate detection study of the "leverage effect".**
- Note that in this example the response Y_i of the *isolated observation* is quite far from the regression line. It does not follow the general linear trend of the majority of observations.

3.3. Leverage effect

Atypical points can be **leverage points** (and/or **regression outliers**). An analysis of the influence (leverage effect) of an observation is based on the idea of comparing the adjustment with and without this observation. Note that it should to be done for each of the observations in the dataset. To this end, let us introduce the calculation of the estimator of β without the observation $i : (x_i, y_i) \in \mathbb{R}^P \times \mathbb{R}$. The index "$(-i)$" means "without the observation i". For example, the matrix $X_{(-i)}$ is the $(n-1) \times p$ matrix corresponding to the matrix X without the *i-th* line. Therefore, calculating the least squares estimator without the observation (x_i^T, Y_i) gives:

$$\widehat{\beta}_{(-i)} = \left(X_{(-i)}^T X_{(-i)}\right)^{-1} X_{(-i)}^T Y_{(-i)}.$$

• Then, the predictive Y_i for the observation x_i in this setting is noted $\widehat{Y}_i^P = x_i^T \widehat{\beta}_{(-i)}$.

• The associated prediction error is $Y_i - \widehat{Y}_i^P$.

Intuitively, if the i observation is not too influential, the estimation error $(Y_i - \widehat{Y}_i)$ and prediction error $(Y_i - \widehat{Y}_i^P)$ will be relatively close. If not, the i observation deserves special attention. These two quantities are related to the projector P_X

$$P_X = X(X^T X)^{-1} X^T.$$

Recall that $\widehat{Y} = P_X Y$, then we deduce :

$$\widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j.$$

Proposition 1 *Note $h_{ij} = (P_X)_{ij}$, the entries of P_X. The trace of P_X is equal to :*

$$Tr(P_X) = \sum_{i=1}^{n} h_{ii} = p.$$

Moreover, for all $i = i, \cdots, n$ and for all $j \neq i$,

1. $0 \leq h_{ii} \leq 1, \quad -\frac{1}{2} \leq h_{ij} \leq \frac{1}{2}$.
2. *If $h_{ii} = 1$ then $h_{ij} = 0$.*

Theorem 1 *In the linear regression model, under the Rank assumption and under **[P1]–[P4]**, we have for all* $i = 1, \cdots, n$

$$Y_i - \widehat{Y}_i = (1 - h_{ii)}(Y_i - \widehat{Y}_i^P),$$

where h_{ii} *denote the i-th diagonal element of* P_X.

Comments:

- From the proposition 1, it follows the fact that $\widehat{Y}_i$ is entirely determined by Y_i as soon as $h_{ii} = 1$. If $h_{ii} = 0$, Y_i has no influence on $\widehat{Y}_i$.
- The theorem 1 suggests that the estimation error $(Y_i - \widehat{Y}_i)$ and prediction error $(Y_i - \widehat{Y}_i^P)$ are equal for $h_{ii} = 0$.

Definition 1 *An observation i is called a **leverage point** if* $h_{ii} > s$, *where*

- $s = 2p/n$ *according to Hoaglin & Welsch (1978),*
- $s = 3p/n$ *for* $p > 6$ *and* $(n - p) > 12$ *according to Velleman & Welsch (1981),*
- $s = 1/2$ *according to Huber& Welsch (1981).*

Comment:

- It is possible to prove that h_{ii} corresponds in a certain way to the distance of the point x_i to the gravity center $\bar{x}$ of the scatter plot x_i. **In other words, the h_{ii} tells us precisely which are the isolated observations of the sample**.
- If an observation is such that $h_{ii} > s$, influences its own estimate. But it does not necessarily affect the overall model, that is, the estimate of β.
- Without being necessarily *regression outlier* (residuals analysis), leverage points are atypical points in explanatory variables. Without systematically eliminating them, it is important to detect and analyze them: do they come from measurement errors or from a population of a different nature? Do they impact the estimation of β (cook distance) ?

3.4. Residuals analysis

3.4.1. The different residuals

Recall first that , the $\varepsilon = Y - X\beta$ is the vector of the theoritical errors/residuals such that $\mathbb{E}_\beta[\varepsilon] = 0_n$ and $\mathbb{V}\text{ar}_\beta[\varepsilon] = \sigma^2\mathbb{I}_n$. We give in definition~**??** a first estimation of the ε : the estimated residuals

$$\widehat{\varepsilon} = Y - \widehat{Y} = Y - X\widehat{\beta} = Y - P_X Y = (I - P_X)Y = P_{X^\perp}Y = P_{X^\perp}\varepsilon.$$

Moreover, according to proposition~**??**

$$\mathbb{E}_\beta[\widehat{\varepsilon}] = 0_n \text{ and } \mathbb{V}\text{ar}_\beta[\widehat{\varepsilon}] = \sigma^2 P_{X^\perp}.$$

The postulat **Postulat [P2]** is not satisfied by the estimated residuals. To fix it, we consider the standardized residuals $t = (t_1, \cdots, t_n)^T$ such that

$$t_i = \frac{\widehat{\varepsilon_i}}{\widehat{\sigma}\sqrt{1-h_{ii}}}.$$

But the standardized residuals do not satisfy the postulate **Postulat [P2]** on uncorrelation. Introduce then, the studentized residuals $t^* = (t_1^*, \cdots, t_n^*)^T$ such that

$$t_i^* = \frac{\widehat{\varepsilon_i}}{\widehat{\sigma}_{(-i)}\sqrt{1-h_{ii}}},$$

where $\widehat{\sigma}^2_{(-i)}$ is the estimation of σ^2 in the model deprived of the observation i (by *cross validation*):

$$\widehat{\sigma}^2_{(-i)} = \frac{\|Y_{(-i)} - X_{(-i)}\widehat{\beta}_{(-i)}\|^2}{n-1-p}$$

Theorem 2 *In the regression linear model, under **[P1]–[P4]**, if rank$(X_{(-i)} = p$ then the **studentized residuals** are such that*

$$t_i^* = \frac{\widehat{\varepsilon_i}}{\widehat{\sigma}_{(-i)}\sqrt{1-h_{ii}}} \sim t_{n-1-p},$$

where t_{n-1-p} denotes the student law of $(n-1-p)$ degrees of freedom.

Proof : The demonstration is left as exercise. □

3.4.2. Residuals analysis for outliers detection

To analyze the fit quality of an observation, that is, if the model explains the observation, we look at the associated residual $\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$. If its standardized residual (or studentized residual) is large, then the observation is a *regression outlier*.

Definition 2 *A **regression outlier** is an observation (x_i^T, Y_i) such that the associated studentized residual t_i^* is high :*

$$|t_i^*| > t_{n-p-1,1-\alpha/2}.$$

Comments:

- Note that in theory, $\alpha\%$ of the datas are outliers.
- In practice, we use $\alpha = 5\%$, then for a large enough sample (larger than $30+p$), $t_{n-p-1,1-\alpha/2} \approx 2$.
- We are actually looking for (x_i^T, Y_i) for which t_i^* is well outside the confidence band in the $i \longmapsto t_i^*$ plot. In figure 4, only the point "52" is an outlier.
- Explaining the presence of these outliers can be difficult. They can be caused by measurement errors or be the result of a population change. It is recommended to pay attention to these points and check if they do not have too much influence on the calculation of $\widehat{\beta}$ and $\widehat{\sigma}^2$.

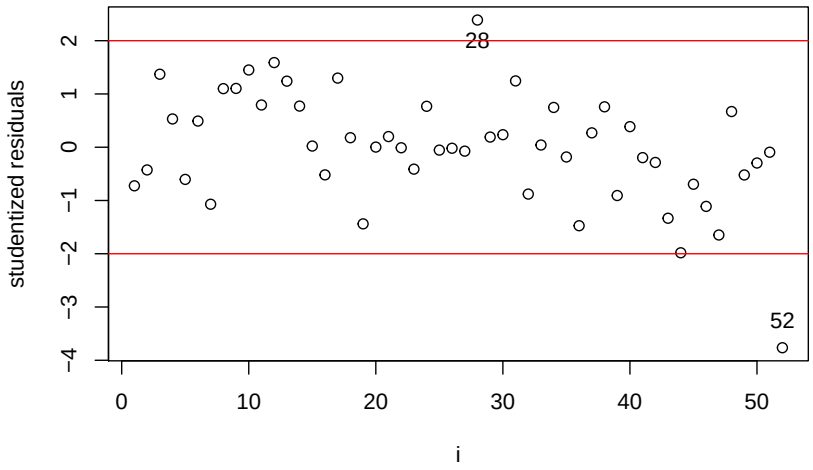

Figure 4: Scatter plot/Studentized residuals plot

3.4.2. Validation of the postulates

We recall that we assume in the regression linear model, the Rank assumption (easy to check) and the postulates **[P1]–[P4]** with

- **[P1]**: Errors are centered : $\forall i = 1, \cdots, n \quad \mathbb{E}_\beta[\varepsilon_i] = 0$. In practice, this means that the model is correct (the model is linear).
- **[P2]**: Errors have homosedasctic variance : $\forall i = 1, \cdots, n \quad \mathbb{V}\text{ar}_\beta[\varepsilon_i] = \sigma^2 > 0$.
- **[P3]**: Errors are uncorrelated: $\forall i \neq j \quad \mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j) = 0$.
- **[P4]**: Errors are gaussian : $\forall i = 1, \cdots, n \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The simplest way to validate postulates is graphically.

Validation of the postulate [P1]: Errors are centered

The linearity assumption (the centered postulat) can be checked by inspecting the *Residuals vs Fitted*-plot (or $(\widehat{Y}_i, t_i^*)$ plot). Ideally, figure~5 shows no fitted pattern. (figure~**??**) That is, the red line should be approximately horizontal at zero. The presence of a pattern may indicate a problem with some aspect of the linear model.

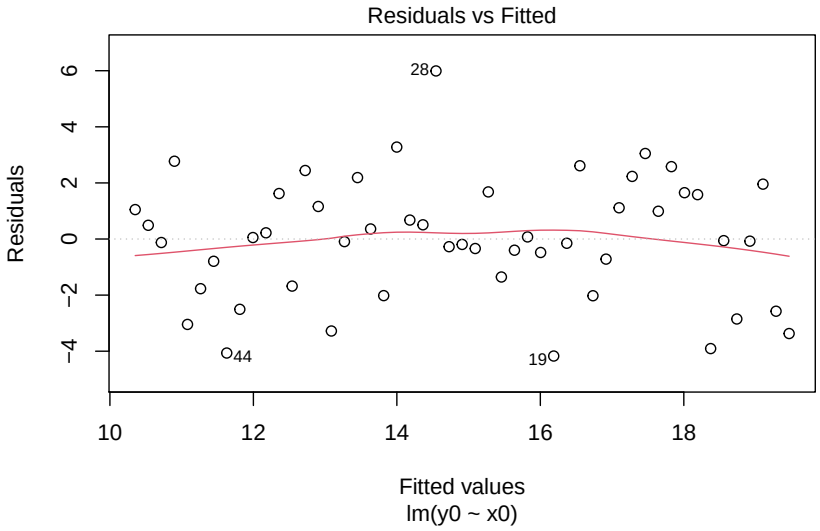

Figure 5: Residuals vs Fitted plot

Validation of the postulate [P2]: Errors have homosedasctic variance

This assumption can be checked by examining the *Residuals vs Fitted*-plot and the *Scale-location*-plot (plot of the points $(\widehat{Y}_i, \sqrt{t_i})$), also known as the *spread-location* plot. This last plot shows if residuals are spread equally along the ranges of predictors. It's good if you see a horizontal line with equally spread points. In our example (figure~6), this is the case.

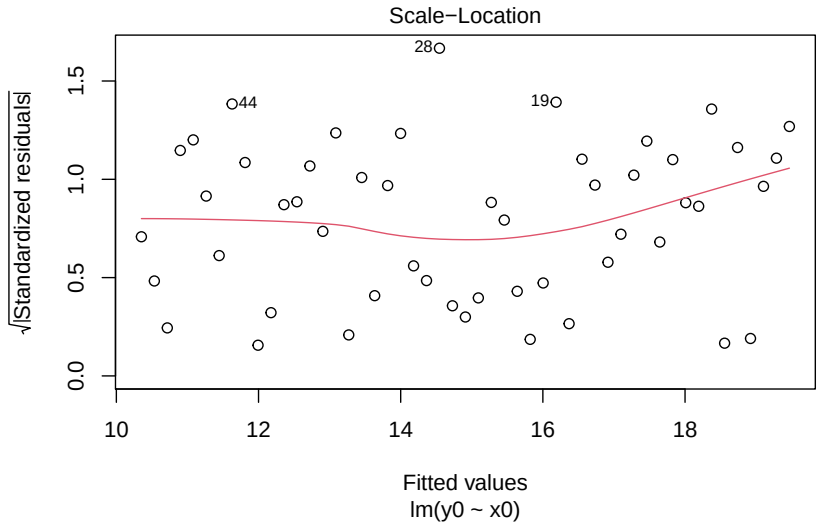

Figure 6: Scale-location plot

Comment:

- ☛ If there is a doubt of heteroscedasticity, we advise to make a test. A possible solution to reduce the heteroscedasticity problem is to use a log or square root transformation of the outcome variable Y.

Validation of the postulate [P3]: Errors are uncorrelated

Under **R**, we can represent the auto-correlation of the residuals using the command `acf()`. The vertical lines (igure~7) represent the correlation coefficients between the residues of each point and those of the points of the following line (lag = 1), or those separated by two lines (lag = 2) and so on. Its interpretation is simple. If a bar, exept the first one, exceeds dashed thresholds, uncorrelation isn't satified. In figure~7, the postulate is validated. We will see in the next chapter how to use it.

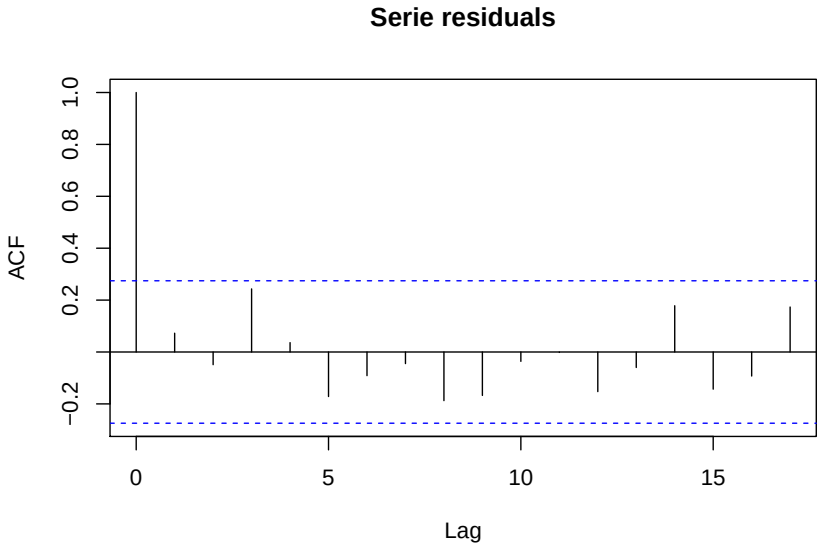

Figure 7: Autocorrelation-plot

Validation of the postulate [P4]: Errors are gaussian

To analyze the normality, we use the Q-Q plot. It consists in comparing the t_i to the theoritical quantiles of the reduced normal centered law (for n large enough, the standard normal is similar to the student law). If all the points fall approximately along this reference line, then the postulate is validated as in figure~8.

Comments:

- ☛ In general, it is often recognized that the normality assumption plays a minor role in regression analysis.
- ☛ The normality assumption is useful for inference purposes, especially for small samples. However, it should be noted that in the presence of small samples, non-normality may be particularly difficult to diagnose by residue examination.

3.5. Cook's distance

Residuals analysis allow to identify atypical values related to the explained variable ; the analysis of the orthogonal projector allows to detect atypical values related to predictors. In this section, we try to combine these two analyzes. For that, we introduce Cook's distance.

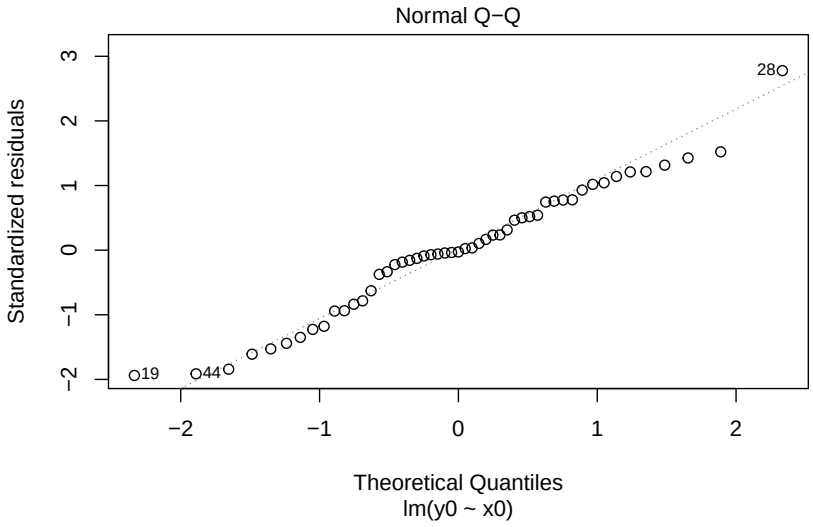

Figure 8: QQ-normal-plot

3.5.1 Definition

Definition 3 *For all i, the Cook's distance of the observation* (x_i^T, Y_i) *is given by the following formula :*

$$D_i = \frac{1}{p\widehat{\sigma^2}}(\widehat{\beta}_{(-i)} - \widehat{\beta})^T (X^T X)(\widehat{\beta}_{(-i)} - \widehat{\beta})$$

where $\widehat{\beta}_{(-i)}$ *is the estimation of* β *in the model without the i-th observation.*

Comments:

- ☛ The Cook distance is essentially a standardized distance measure that describes the change in the β estimator when we remove the observation i.
- ☛ A high value of Cook's distance suggests that observation i has a high influence. In practice, Cook's distances are often compared with 1. A value much lower than 1 suggests that the impact of observation i does not seem very important. In contrast, a Cook distance greater than one suggests that observation i has a large impact.
- ☛

Proposition 2 *The Cook's distance of the observation* (x_i^T, Y_i) *satisfies*

$$D_i = \frac{h_{ii}}{p\widehat{\sigma^2}(1-h_{ii})^2}(Y_i - \widehat{Y}_i)^2 = \frac{h_{ii}}{p(1-h_{ii})}t_i^2$$

where h_{ii} *is the i-th diagonal element of the orthogonal projector* P_X *and* t_i *is the standardized residual associated to the observation i.*

Proof : Let as exercice.

Comments:

- Recall that the standardized residuals measures the adequacy of the observation Y_i to the estimated model $\widehat{Y_i}$ while the quantity $\frac{h_{ii}}{1-h_{ii}}$ measure the sensitivity of the estimator $\widehat{\beta}$ to the observation i. Indeed, they are such that

$$\frac{h_{ii}}{1-h_{ii}} = \frac{\mathbb{V}\text{ar}_\beta(Y_i)}{\mathbb{V}\text{ar}_\beta(\widehat{\varepsilon_i})}.$$

  The $\frac{h_{ii}}{1-h_{ii}}$'s are called the **levers**.

- It can be seen that the Cook's distance for fixed p, can be large if the standardized residues are large or if the levers are large (or if both are large).

- Thus Cook's distance can be seen as a criterion measuring both the outlier (aberrant) character of an observation (measured by the standardized residual) and its leverage effect. Points with high Cook's distances (greater than 1) will be outliers, or levers, or both. It is strongly recommended to delete points with a high Cook distance. Nevertheless, if we want to keep these points, we have to make sure that they do not change too much the estimation of β and the interpretations.

3.5.2 Examples

Example 1 Consider here the toy example (figure~9 top/left) where the 52-*th* point $(-1, 18)$ is an *isolated point*. If we look at the *Studentized residals*-plot (figure~9 top/right), it comes out that the 52-*th* point is a *regression outlier* as $t^*_{52} > 2$. Moreover, the 28-*th* observation is also a *regression outlier* as $t^*_{28} > 2$. Are they *leverage points* ?

In the h_{ii}-plot (figure~9 bottom/left), we see that all the $h_{ii} < 0.5$, so none point is influent on its own estimation. Nevertheless, according to the *Residuals vs leverage*-plot (figure~9 bottom/right), it turns out that the 52-*th* point has a high Cook's distance (larger than 1). It has a large impact on the estimation of β, this point may be removed.

Example 2 Consider now, the toy example (figure~10 top/left) where the 52-*th* point $(-10, 50)$ is an *isolated point* and an *outlier*. Note that the point follows the model as it is close to the least square line. The *Studentized residals*-plot (figure~10 top/right) indicates that this point is not a *regression outlier* as $t^*_{52} < 2$. On the other hand, it appears that the 28-*th* observation is an *regression outlier* as $t^*_{28} > 2$.

In the h_{ii}-plot (figure~10 bottom/left), the only *leverage point* is the $52 - th$ point as its $h_{ii} > 0.5$. Moreover, according to the *Residuals vs leverage*-plot (figure~10 bottom/right), it turns out that the 52-*th* point has a Cook's distance larger than 1. It has a large impact on the estimation of β, this point is a *leverage point and a regression outlier*, it may be removed.

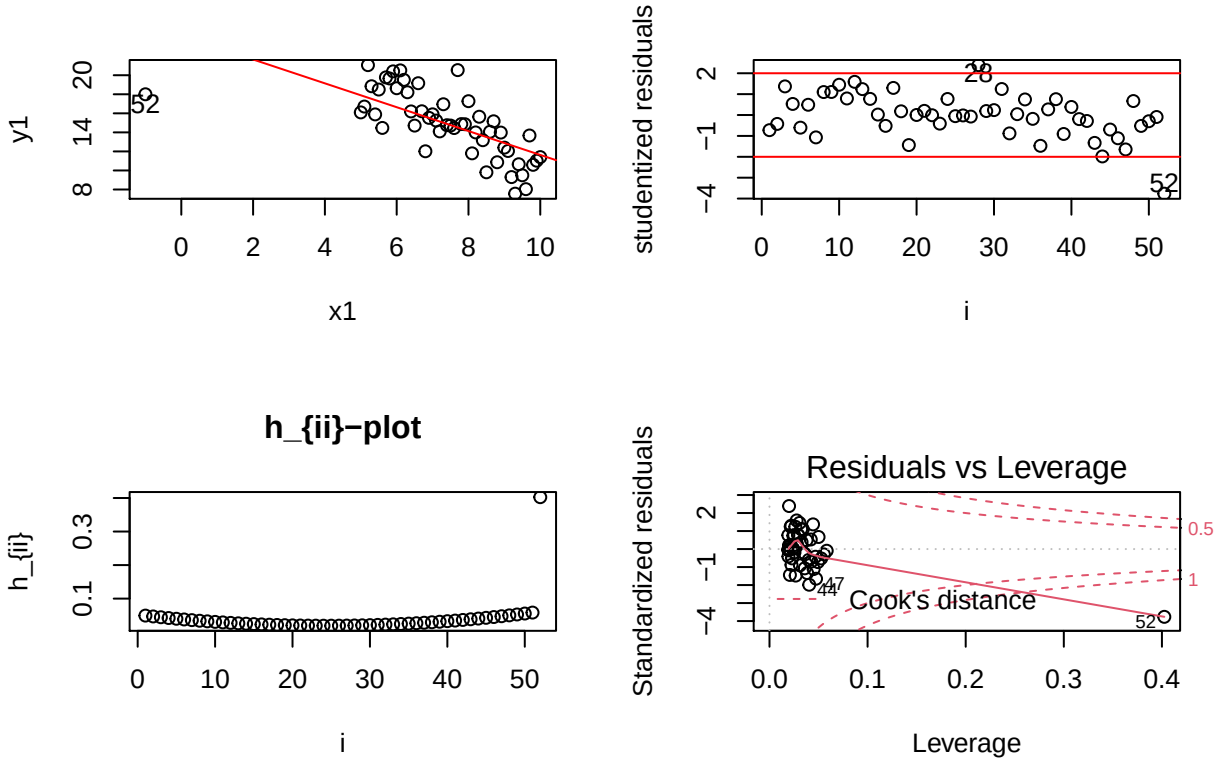

Figure 9: Some Plots for the toy dataset of example 1

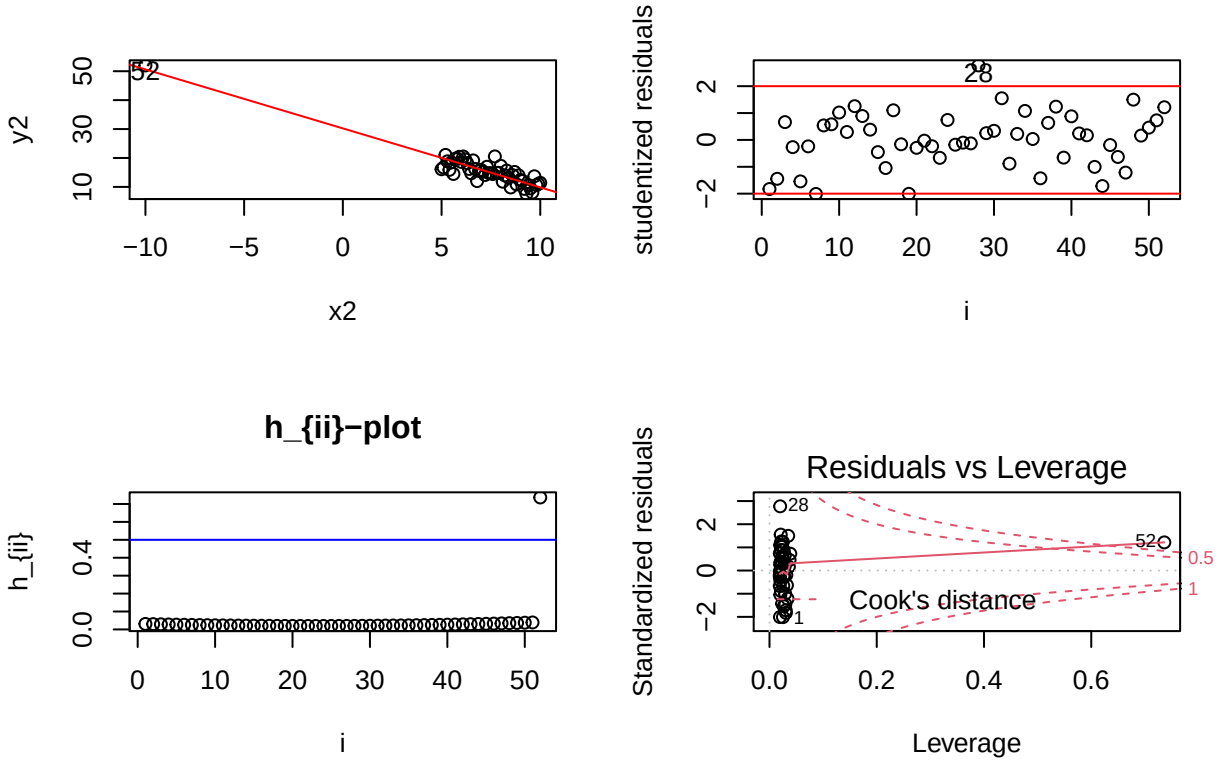

Figure 10: Some Plots for the toy dataset of example 2

Example 3 Consider here the toy example (figure~11 top/left) where the $52-th$ point $(7, 40)$ is an *outlier*.The *Studentized residals*-plot (figure~**??** top/right) indicates that this point is a *regression outlier* as $t^*_{52} > 2$.

In the h_{ii}-plot (figure~11 bottom/left), so none point is influent on its own estimation as for each observation $h_{ii} < 0.5$. Moreover, according to the *Residuals vs leverage*-plot (figure~11 bottom/right), it turns out that the $52-th$ point has a Cook's distance smaller than 1. It has not a big influence on the estimation of β, this point is a *regression outlier* but not a *leverage point*, it may be keeped.

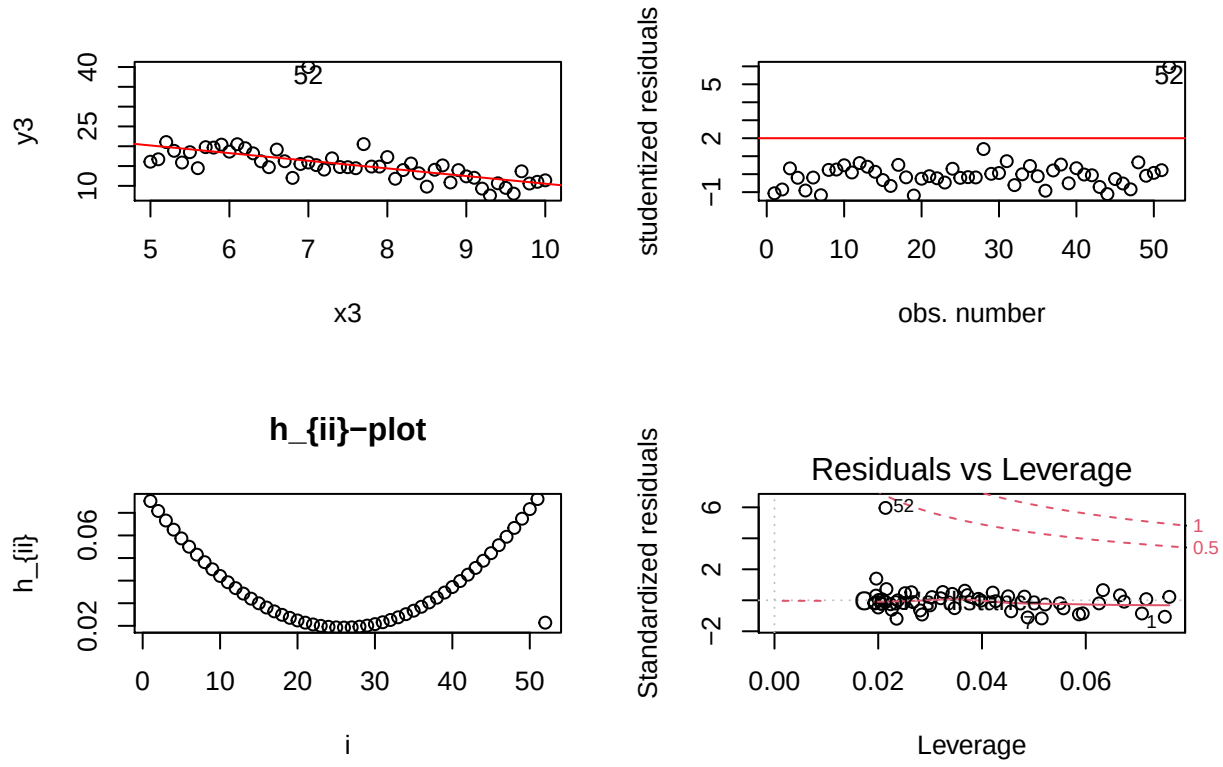

Figure 11: Some Plots for the toy dataset of example 3

3.6. Conclusion

The statistics can not in any case determine which variables have been forgotten,which can be one of the reasons why the postulate on homoscedsticity is not verified. On the other hand, it can determine the function f_j such that Y and $ f_j (X_j) $ has a linear relation. Most of the time, we will leave a model that includes a maximum of variables that can explain Y, even if we remove one using the Fisher or Student tests. In the next Chapter, we will go further into the issue of choice of models. First, start with an example under **R** with a real dataset.