Lecture 1 : Atypical points and model validation

M2-Modèles pour la régression

K. Meziani

Info sur ce cours : faire la démonstration des théorèmes, ils tomberont à l'exam.

Dauphine | PSL
UNIVERSITÉ PARIS

Outliers

Outliers :For some atypical observations, the response variable Y and/or predictors X_j appear to behave differently from the majority of observations.

Outliers may occur for various reasons:

- obvious cases: measurement errors, data transcription errors,... *Example: a customer recorded to be 400 years old; a 1 year old baby running the 100m in 10 seconds...*
- Adversarial error to scuttle the analysis.
- *outliers* sometimes reveal a particular phenomenon that may be different from the model followed by the majority of observations. *Example: gene expression in Cancer patient as compared to healthy individuals.*

Outliers

Goal of a model: explain as well as possible a general phenomenon but it can have its own limitations (too simple).

Presence of *outliers* can suggest to build more elaborate models:

- missing regressor
- feature engineering

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.

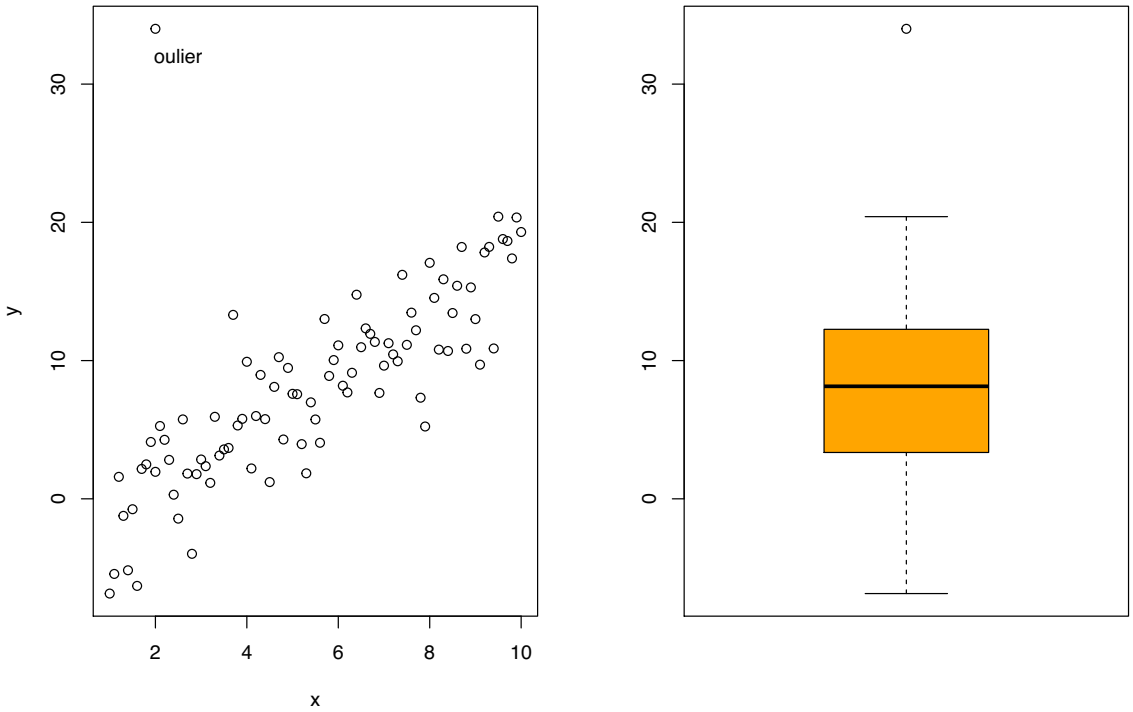
Section 1

1. Outlier in the Y-direction

{Outlier scenario: Outlier in the Y-direction but not in the predictors X_{ij}}

Consider linear regression.

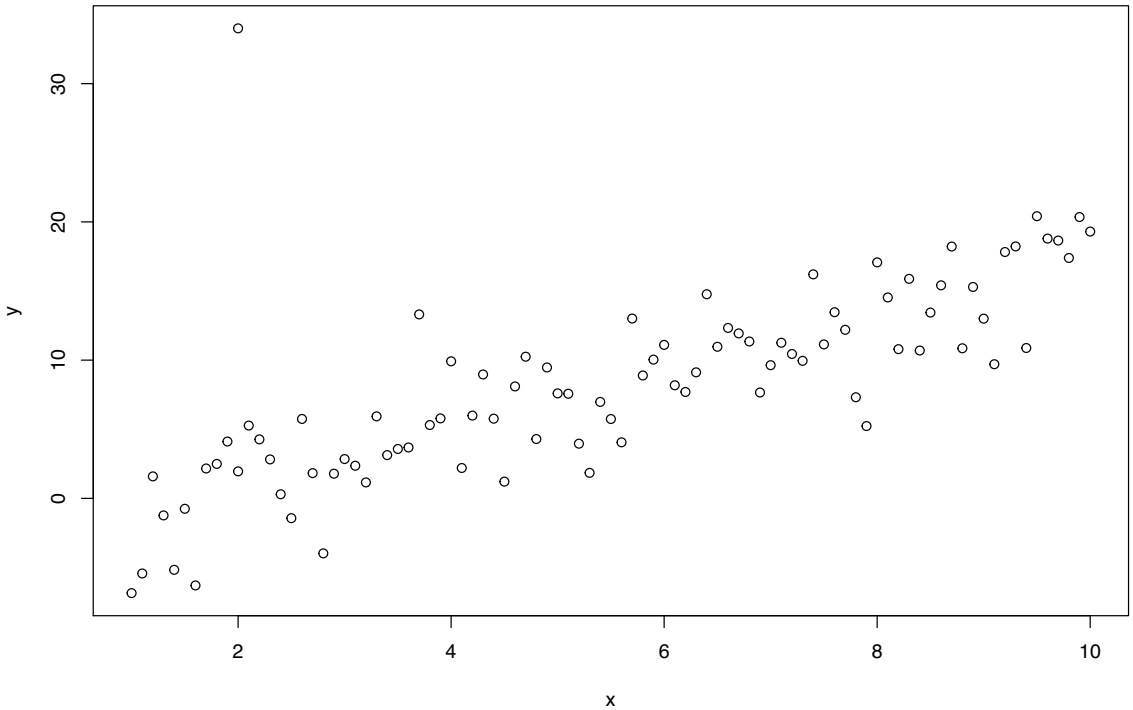
Example: Scatter plot of the toy dataset $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$. Boxplot reveals the *outliers*.



{Outlier scenario: Outlier in the Y-direction but not in the predictors X_{ij}}

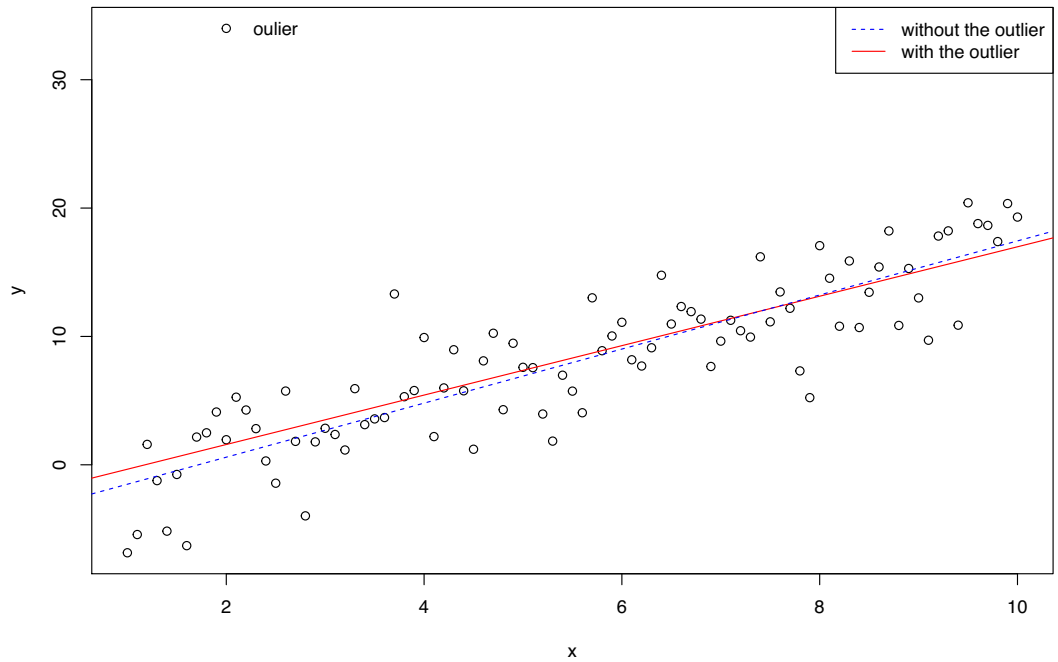
According to this scatter plot, omitting the outlier, a linear model can be considered:

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



{Outlier scenario: Outlier in the Y-direction but not in the predictors X_{ij}}

We plot the two LSE regression lines with ($y = 1.9254378x - 2.2701754$) and without ($y = 2.1061473x - 3.6203342$) the outlier



Question: what do you think of the presence of the outlier?

Comments

- For THIS scenario, 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.**

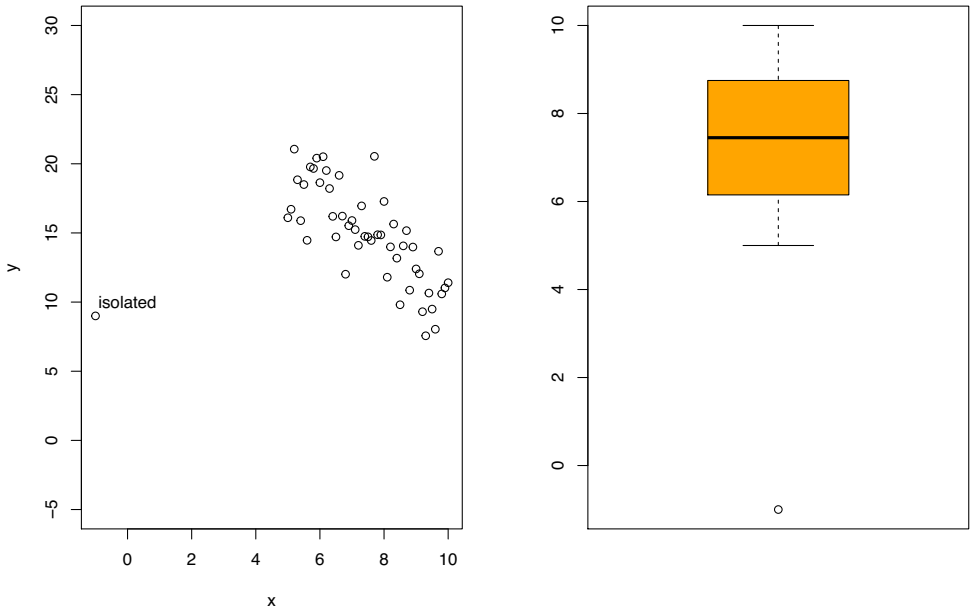
Section 2

2. Isolated observations

{Outlier scenario: out of domain point}

Isolated observation has atypical values in the predictors X_{ij}. It means that the values $(X_{ij})_j$ of the observation i are relatively far from all the value $(X_{i'j})_j$ of the other observations $i' \neq i$. We say this point is out of domain.

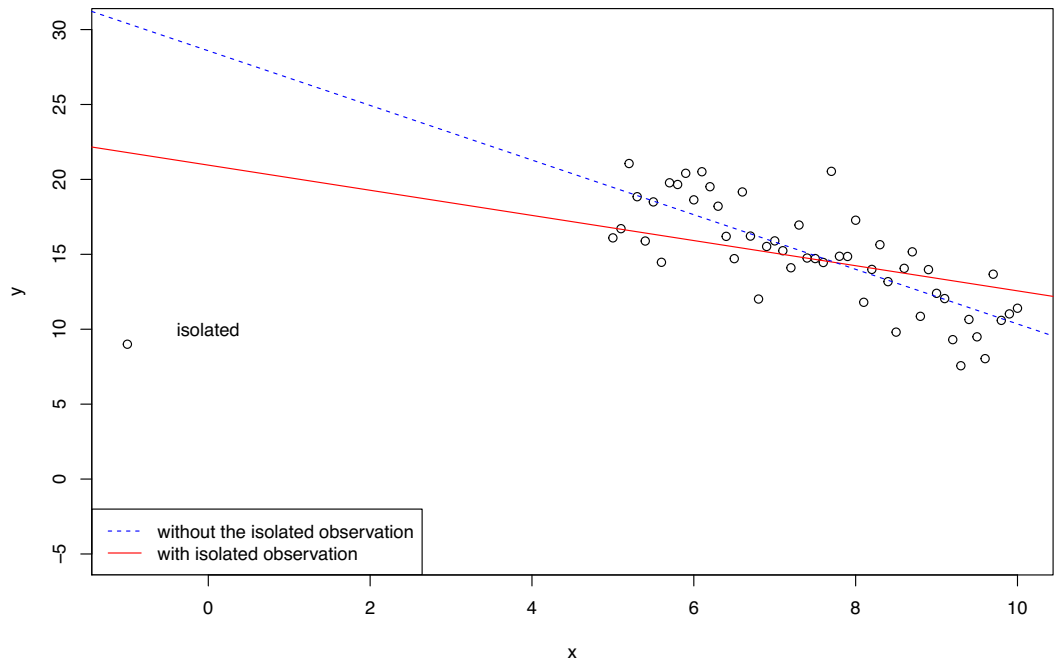
Let us consider an other toy example. According to the scatter plot, a linear model can be considered. $Y_i = ax_i + \varepsilon_i, \qquad \varepsilon_i$ i.i.d. $\mathcal{N}(0, \sigma^2)$.



{Outlier scenario: out of domain point}

Question: what is the impact of this outlier on this regression model?

We plot the two LSE regression lines with ($y = -0.8391409x20.9530211$) and without ($y = -1.8231533x + 28.5839412$) the outlier



Comments

- For THIS scenario, the *isolated observation* (*out of domain*) has a significant impact on the estimation of β. Indeed, removing this point significantly changes the LSE regression line.
- The *isolated observation* (X_i, Y_i) is quite far from the regression line. It does not follow the general linear trend of the majority of observations.
- Such of points are called ***high leverage point***. Statisticians are always wary of such points. Sometimes they do not significantly change the estimation of β and sometimes they do.
- ***Leverage points*** **can be detected by a multivariate detection study of the "leverage effect".**

Section 3

3. Leverage effect

{Atypical points can be high *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 be done for each observation in the dataset.

Estimation of β without the i-observation (x_i, Y_i)

- The index "$(-i)$" means "without the i-observation". For example, the matrix $X_{(-i)}$ is the $(n-1) \times p$ matrix corresponding to the matrix X without the i-*th* line.
- Denote by $\widehat{\beta}_{(-i)}$ the LSE computed from the dataset without the i-observation:

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

{Prediction without the i-observation (x_i, Y_i)}

LSE prediction for the i-observation : $\widehat{Y}_i^P = x_i^\top \widehat{\beta}_{(-i)}$.

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

☛ If the i-observation is not too influential, we expect that the residues in both cases to be close:

$$Y_i - \widehat{Y}_i \approx Y_i - \widehat{Y}_i^P.$$

☛ If not, the i-observation deserves special attention.

✍ These two quantities are related to the projector P_X

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

Recall about projector

Consider the orthogonal projector P_X onto $[X]$: $P_X = X(X^\top X)^{-1}X^\top$.

Proposition 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 $h_{ij} = 0$.

(revoir les propriétés d'un projecteur)

Chaque point en X plus les hii proche de 1 : fort impact sur sa propre est (> 0,5)

Hat matrix is $X(X^TX)^{-1}X = P_X$

$\hat{y} = P_X y$

Impact of the i-observation on its own estimation

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

poids de l'obs i sur sa propre

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

le poids de l'obs j sur l'est de i

Theorem Assume X is full rank and **[P1]–[P4]**. Then 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.

l'erreur d'estimation est proportionnelle à l'erreur de prédiction

- $\widehat{Y}_i$ is entirely determined by Y_i as soon as $h_{ii} = 1$. $h_{ii} = 1 \Rightarrow Y_i = \widehat{Y}_i$
- If $h_{ii} = 0$, Y_i has no influence on $\widehat{Y}_i$.
- The prediction error $(Y_i - \widehat{Y}_i)$ and the associated prediction error $(Y_i - \widehat{Y}_i^P)$ are equal for $h_{ii} = 0$.

Leverage effect definition

Definition 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).

☛ If an observation is such that $h_{ii} > s$, it influences its own estimate. But it does not necessarily affect the overall model, that is, the estimate of β.

Consulter ces articles?

Comments

- h_{ii} "corresponds" in a way to the distance of x_i from the gravity center $\bar{x}$ of the scatter plot x_i. **The Leverage h_{ii}'s tell us which observations are isolated from the rest of the sample**.
- Without necessarily being an *regression outlier* (residuals analysis), leverage points are atypical points in explanatory variables.

- We may not systematically eliminate them,
- It is important to detect and analyze them: do they come from measurement errors or from a population of a different nature?

Question: Do they impact the estimation of β (cook distance) ?

Section 4

4. Residuals analysis

Estimated residuals/Standardized residuals

- Recall that $\varepsilon = Y - X\beta$ is the vector of the theoritical errors/residuals such that

$$\mathbb{E}_\beta[\varepsilon] = 0_n, \quad \mathbb{V}\mathrm{ar}_\beta[\varepsilon] = \sigma^2 \boldsymbol{I}_n.$$

$P_{X^\perp} Y = P_{X^\perp}(X\beta + \varepsilon)$
$= P_{X^\perp} X\beta + P_{X^\perp}\varepsilon$
$= 0$

- Define the **estimated residuals** by

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

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

$P_{X^\perp} = 1 - P_X$
$= 1 - [h_{ij}]_{ij}$

$V[\hat{\varepsilon}_i] = \sigma^2(1 - h_{ii})$

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

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

normalization

Standardized residuals/studentized residuals

- The **standardized residuals** do not satisfy **[P3]**, il y a de la corrélation ...

We introduce the **studentized residuals** $t^* = (t_1^*, \cdots, t_n^*)^\top$ 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*):

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

Studentized residuals

Theorem Under **[P1]–[P4]**, if $rank(X_{(-i)}) = p$, then the **studentized residuals** satisfy

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

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

Proof : The demonstration is left as exercise. □

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.

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

$|Y_i - \hat{Y}_i| = |\hat{\varepsilon}_i| > c$ → outliers

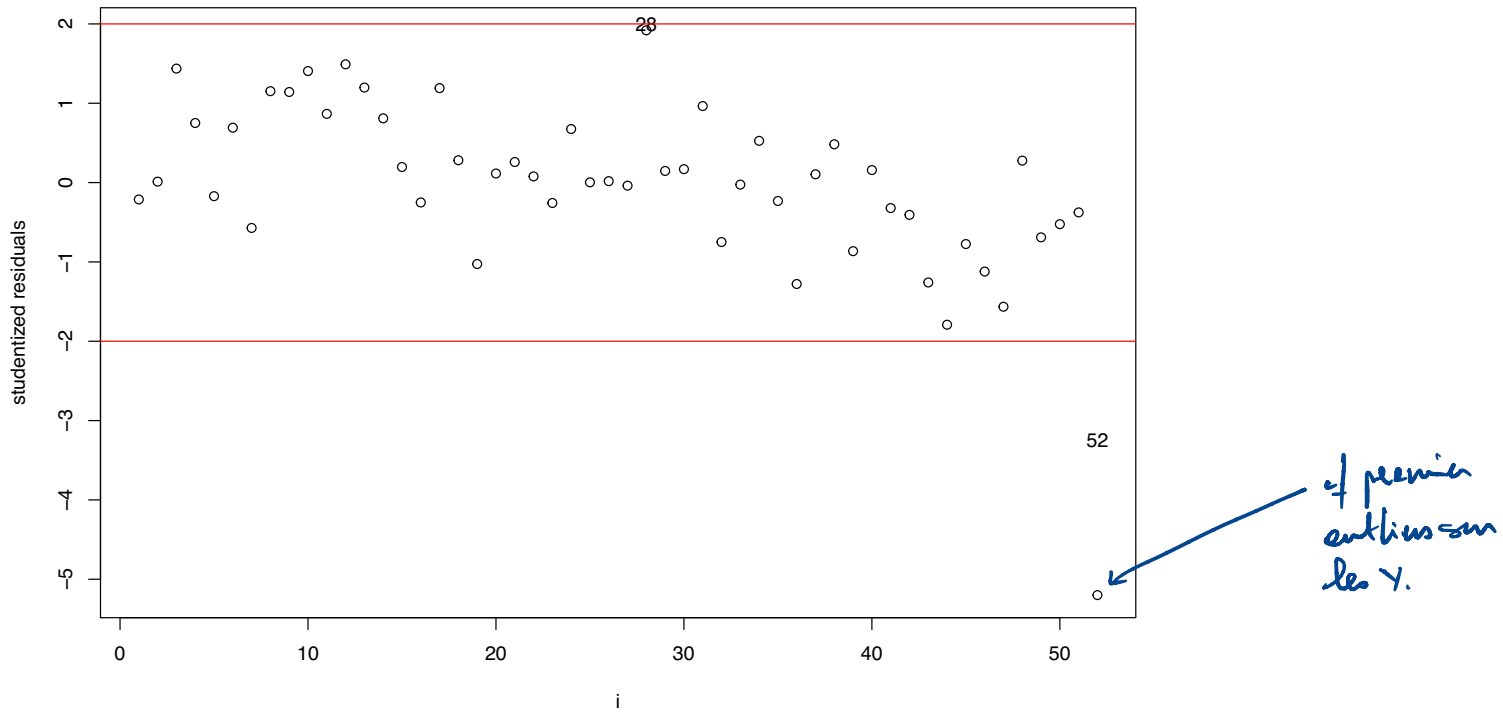
$$|t_i^*| > t_{n-p-1,1-\alpha/2}.$$

Comments:

- If its standardized residual (or studentized residual) is large, then the observation is a *regression outlier*.
- 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.

Residuals analysis for outliers detection

- Only the point "52" is a regression outlier.



Residuals analysis for outliers detection : comments

- 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 $\widehat{\sigma}^2$.
- We will see in a next chapter how to identify and to deal with such of point in pratice. (On a real dataset)

Comments

- Residuals analysis : identify atypical values related to the explained variable Y;
- The analysis of P_X : detect atypical values related to predictors X_j.
- **Cook's distance** combines these two analyzes. It is essentially a standardized distance measure that describes the change in the β estimator when we remove the observation i.

2 façons d'avoir du leverage → fait impact sur sa propre estimation $\hat{Y}_i$ $(h_i > 0{,}5)$
→ fait impact sur le modèle $\hat{\beta}_i$

Il faudrait pouvoir évaluer dist $(\hat{\beta}, \hat{\beta}_{(-i)})$ → DISTANCE DE COOKE

Cook's distance

☛ The Cook's distance can be seen as a criterion measuring the leverage effect of the i-observation on the model (so on β) : distance betewwen the 2 models (with and without the i-observation).

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

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

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

Cook's distance : proposition

☛ The Cook's distance can be seen as a criterion measuring both the *regression outlier* character of the i-observation (measured by its standardized residual).

Proposition 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

- The Cook's distance can be large if the standardized residues are large or if the levers are large (or if both are large).

A high value of Cook's distance suggests that observation i has a high influence (in practice compared to 1).

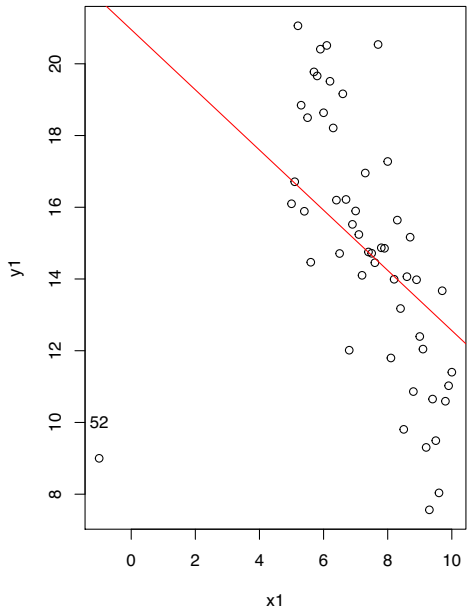
- $D_i < 1$ suggests small impact of i-observation.
- $D_i > 1$ suggests high impact of i-observation.

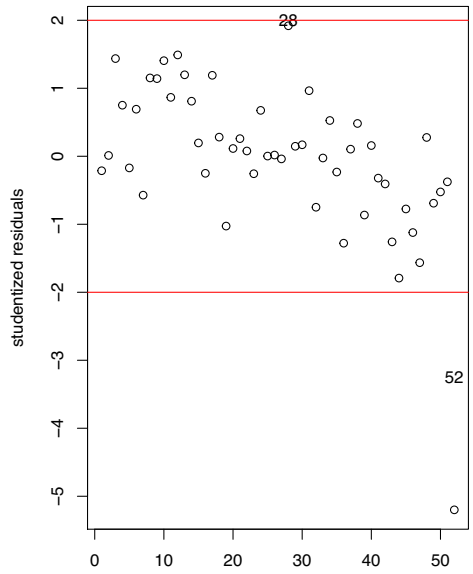
- It is strongly recommended to delete points with a large Cook distance. Nevertheless, if we want to keep these points, we have to make sure that they do not change too much the estimation of β and the interpretation.

Some examples

Some examples

{Toy example 1: the 52-*th* point $(-1, 9)$ is an *isolated point*}



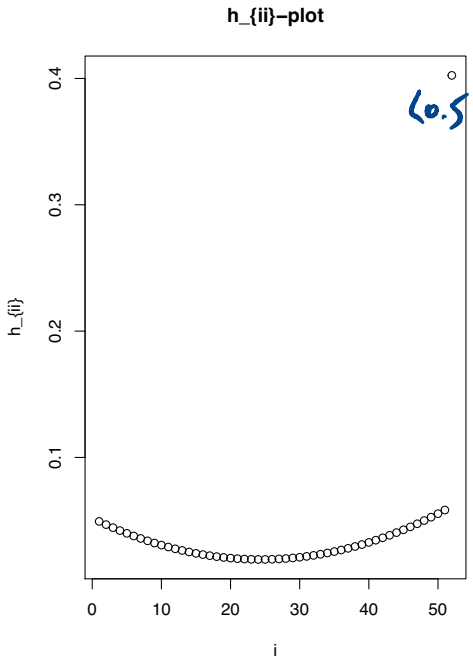


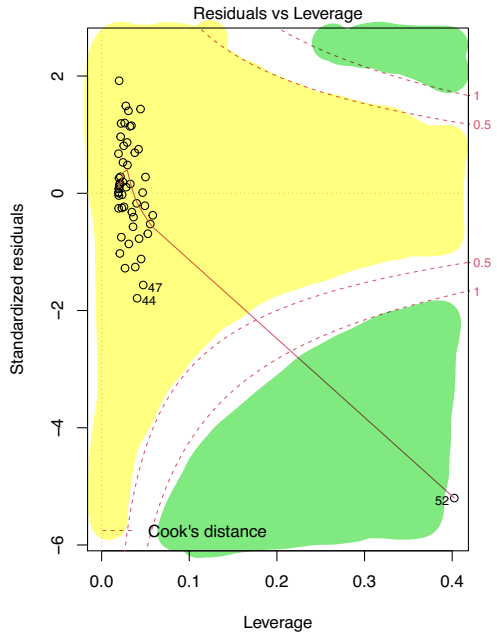
x- outlier

mal expliqué par la regression

- *Studentized residals*-plot: the 52-*th* point is a *regression outlier* as $t^*_{52} > 2$.
- Does this observation have *high leverage* ?

{Example 1 : the h_{ii}-plot and *Residuals vs leverage*-plot}



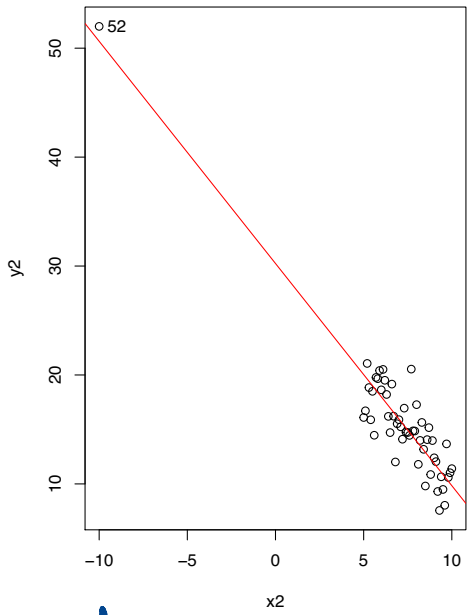


distance de cook < 0.5

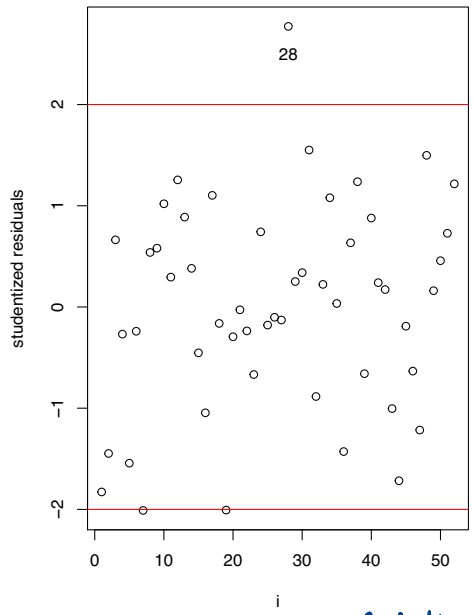
distance de cooke > 1

- The h_{ii}-plot : $h_{ii} < 0.5$, so no point is influent on its own estimation.
- According to the *Residuals vs leverage*-plot, the 52-*th* point $D_i > 1$. It has a large impact on the estimation of β, this point **may** be removed.

{Toy example 2: the 52-*th* point $(-10, 50)$: *isolated point* and outlier in Y}



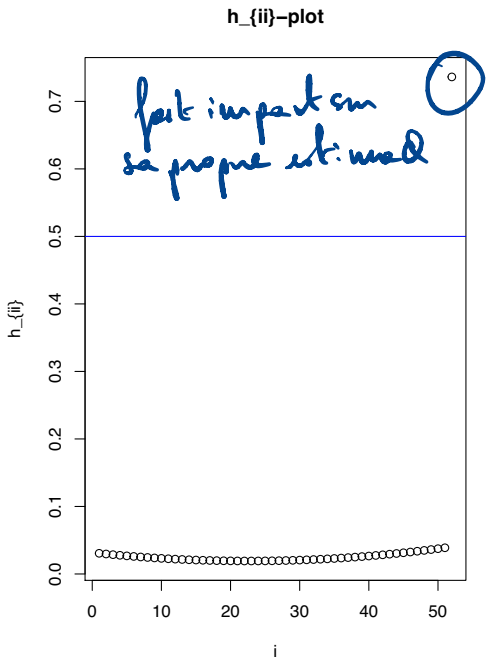
outlier en x et en y.

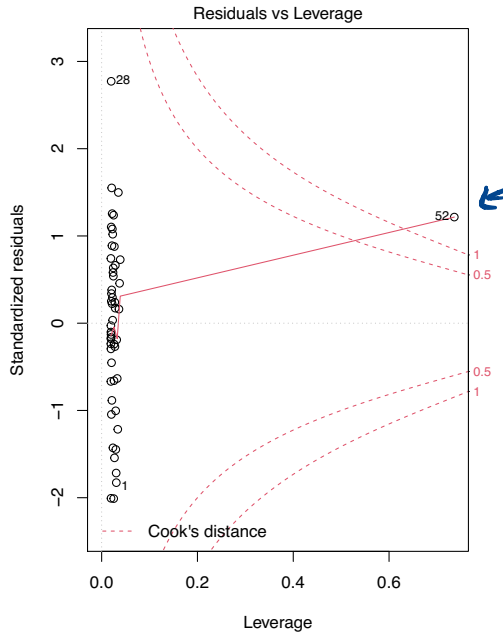


pas un outlier pour la régression

- Note that the point follows the model as it is close to the least square line.
- 52-*th* observation is not a *regression outlier* as $t^*_{52} < 2$ but now 28-*th* observation is an *regression outlier* as $t^*_{28} > 2$.

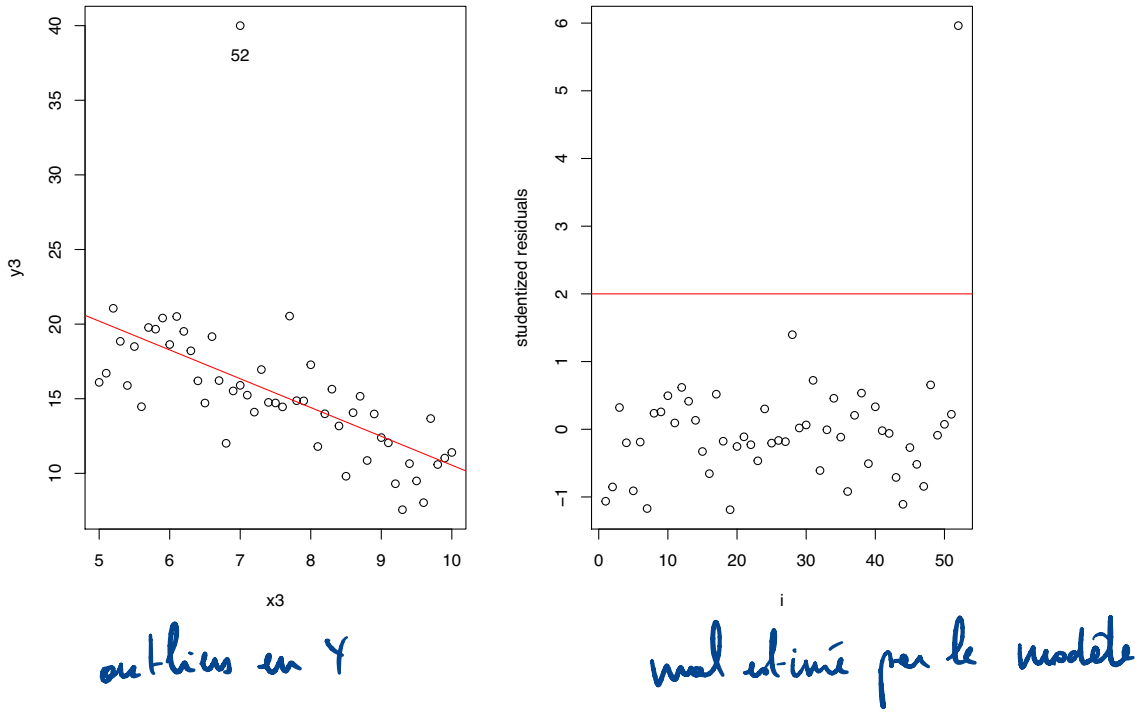
{Example 2 : the h_{ii}-plot and *Residuals vs leverage*-plot}





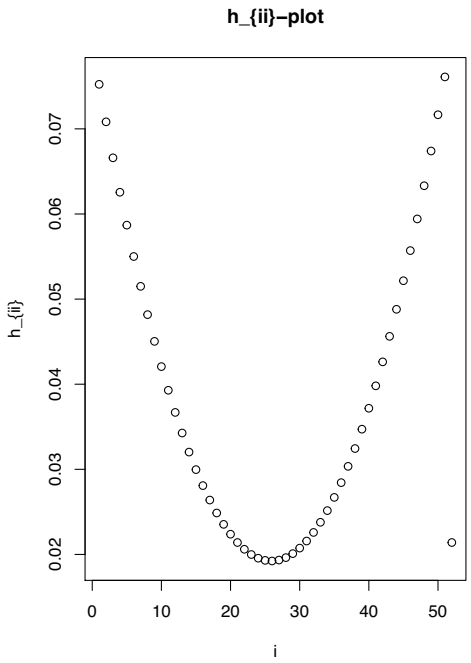
- The h_{ii}-plot: only *leverage point* is the $52-th$ point as $h_{ii} > 0.5$.
- According to the *Residuals vs leverage*-plot: the 52-*th* point has a $D_i > 1$.
- It has a large impact on the estimation of β, this point is a *leverage point and a regression outlier*, it may be removed.

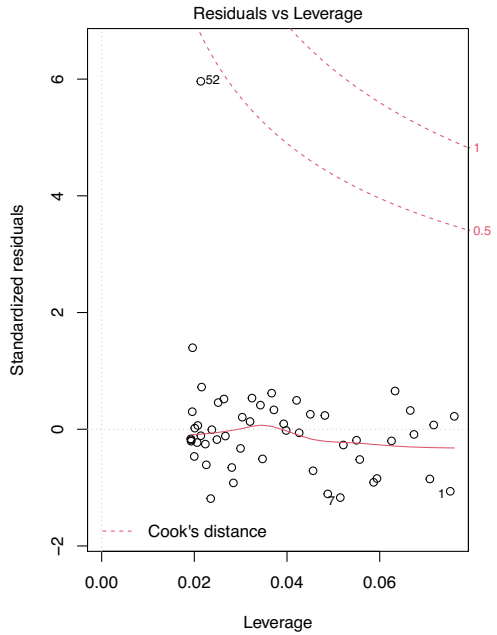
{Toy example 3: the 52-*th* point $(7, 40)$ is an is an *outlier* in Y}



- The *Studentized residals*-plot indicates that this point is a *regression outlier* as $t^*_{52} > 2$.

{Example 3 : the h_{ii}-plot and *Residuals vs leverage*-plot}





no impact on own est

Cook's distance not high.

- The h_{ii}-plot : no point has a high impact of its own estimation as $h_{ii} < 0.5$.
- According to the *Residuals vs leverage*-plot, the 52-*th* point : $D_i < 1$.
- Not abig influence on the estimation of β, this point is a *regression outlier* but not a *leverage point*. We may keep it.

Section 5

5. Model Validation

Gaussian linear regression model

Condiser the linear model $Y = X\beta + \varepsilon$, with full rank X and the postulates

[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 : $\forall i = 1, \cdots, n \quad \mathbb{V}\text{ar}_\beta[\varepsilon_i] = \sigma^2 > 0$.

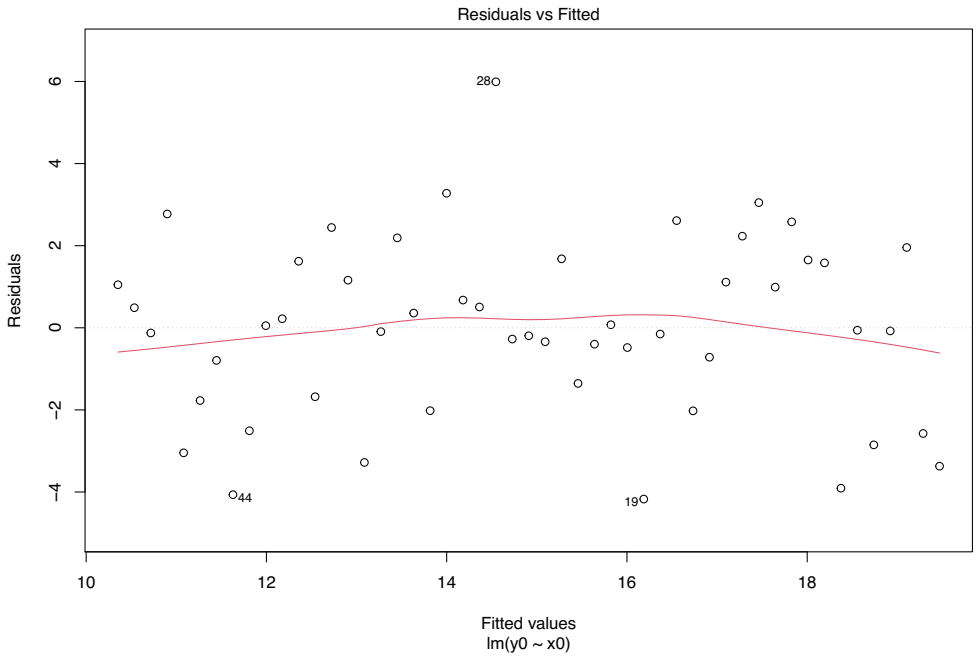
[P3] Errors are uncorrelated: $\forall i \neq j \quad \mathbb{C}\text{ov}(\varepsilon_i, \varepsilon_j) = 0$.

[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 .

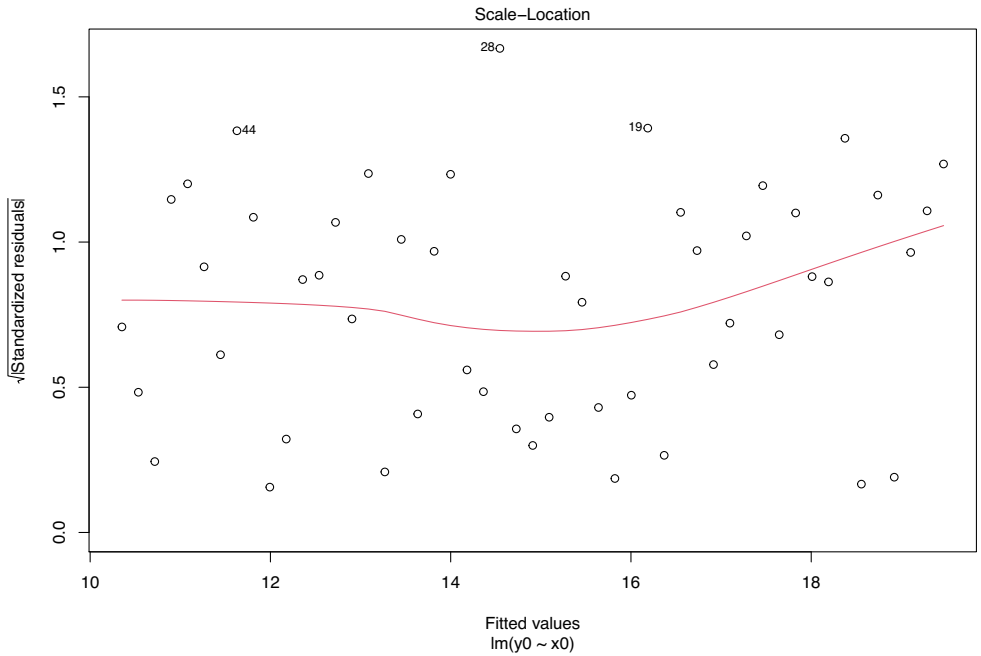
[P1]: Errors are centered

- It can be checked by inspecting the *Residuals vs Fitted*-plot (or $(\widehat{Y}_i, t_i^*)$ plot).
- Ideally, we should observe no particular pattern. That is, the red line should be approximately horizontal at zero. The presence of a pattern may indicate a problem with some aspects of the linear model.



[P2]: homosedascticity

- 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 for all observations. It's good if you see a horizontal line with equally spread points.



courbe rouge doit être plate proche de 1

[P2]: homosedascticity

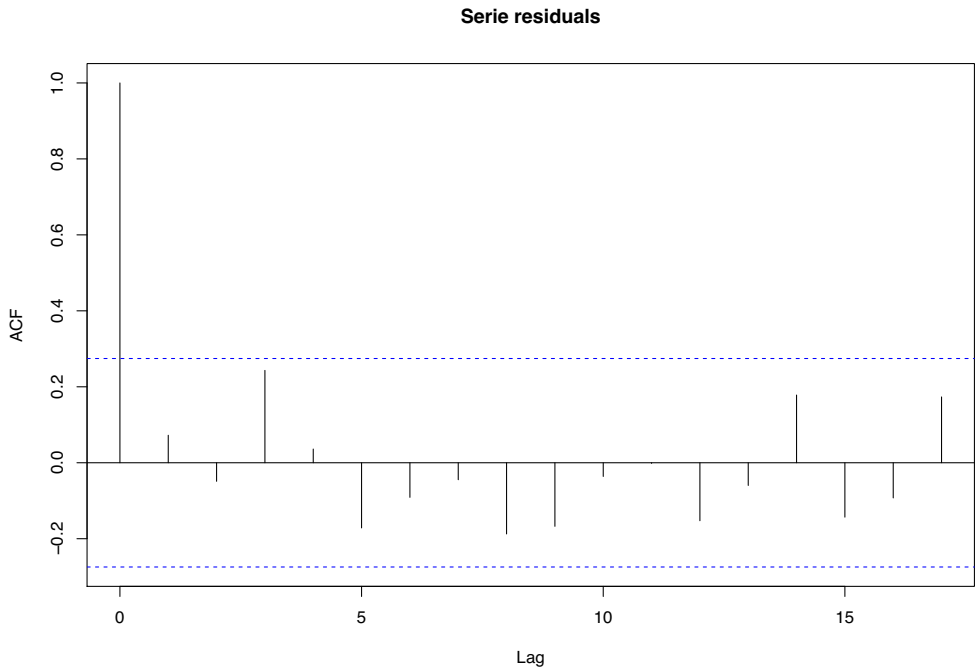
- If there is a doubt of heteroscedasticity, we advise to make a test a **Breush-Pagan test ($\mathcal{H}_O$: homoscedasticity)** to assess it.
- A possible solution to reduce the heteroscedasticity problem is to use a log or square root transformation of the outcome variable Y.

The command for the Breush-Pagan test is `ncvTest`. The homoscedasticity is rejected if the *p-value* is less than 0.05. Here, *p-value*$= 0.63576 > 0.05$, thus we can't reject $\mathcal{H}_0$, the posutlate is validated.

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.2243309, Df = 1, p = 0.63576
```

[P3]: Errors are uncorrelated

- Plot the auto-correlation of the residuals using the command `acf()`.
- Its interpretation is simple. If a bar, exept the first one, exceeds the dashed thresholds, the postulate is violated. Here, **[P3]** is validated



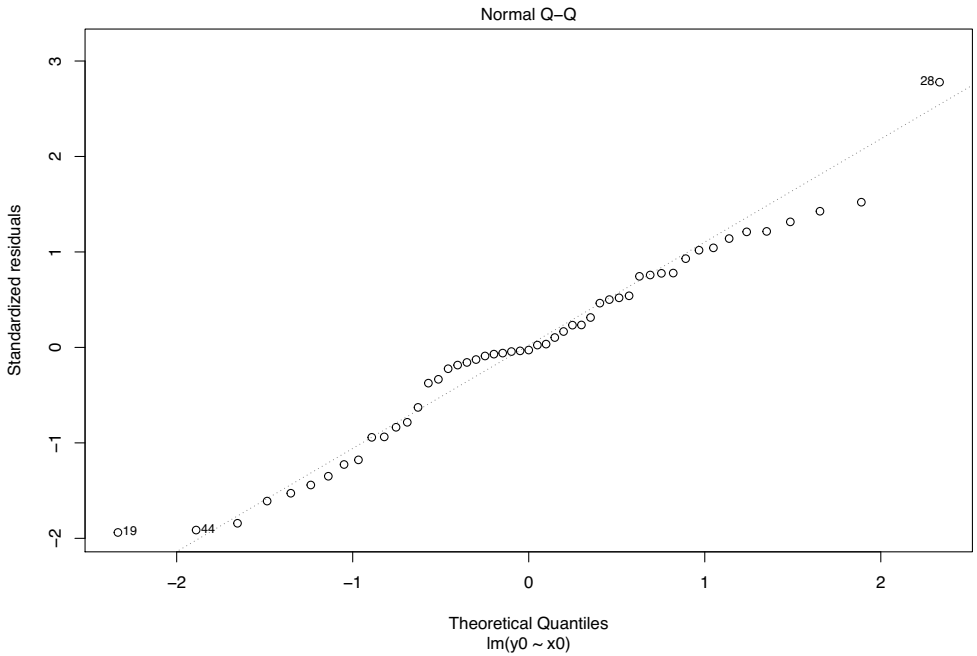
[P3]: Errors are uncorrelated

- The **Durbin-Watson test ($\mathcal{H}_O$: uncorrelation)** can be also used to validate this assumption. The command is `durbinWatsonTest`.
- Here, the *p-value*$= 0.376 > 0.05$ thus we can't reject $\mathcal{H}_0$, the posutlate is validated.

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.07242181       1.80157   0.376
##  Alternative hypothesis: rho != 0
```

[P4]: Errors are gaussian

- Q-Q plot: It consists in comparing the quantiles of the standardized residues denoted t_i to the theoritical quantiles of the standard normal (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.



[P4]: Errors are gaussian

- 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.
- The **Shapiro-Wilk test ($\mathcal{H}_O$: gaussian)** can also be used to assess the normality of residuals. Here, the *p-value*$= 0.4061 > 0.05$ thus we can't reject $\mathcal{H}_0$, the posutlate is validated.

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals((mod0))
## W = 0.9766, p-value = 0.4061
```

⚠ doesn't work with #sample >5000.