Model Diagnostics

→ Introduction

In the world of statistical modeling, the accuracy of our conclusions depends heavily on the assumptions we make about errors in our model. To ensure the reliability of our results, it is essential to confirm that our model align with these assumptions.

→ Assumptions of a regression model

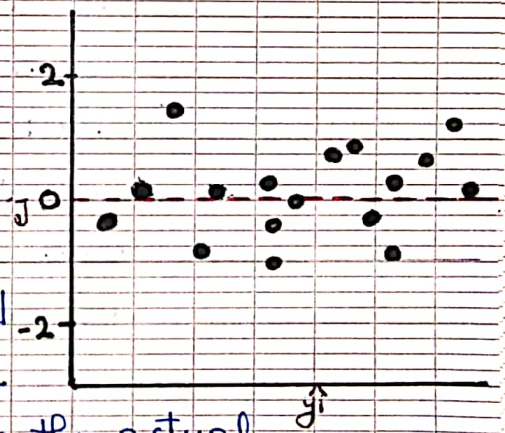
- Linearity of the model
- Independence of errors
- Homoscedasticity of errors (constant variance)
- Normality of errors $\varepsilon_i \rightsquigarrow N(0, \sigma^2)$

→ Graphs to analyse residues

1) Residual Plot

↳ x-axis : estimated values ($\hat{y}_i$)

↳ y-axis : residuals (e_i)



The distance of each point on the plot from the horizontal line at 0, corresponds to the error or the discrepancy between the actual and predicted value for that observation

Points above the line indicates overpredictions, while points below the line indicate underpredictions

↳ Importance of Residual Plots in Linear Regression

Residual plots play a crucial role in the evaluation and improvement of linear regression models by helping to identify potential issues. Some of the key reasons for using residual plots include:

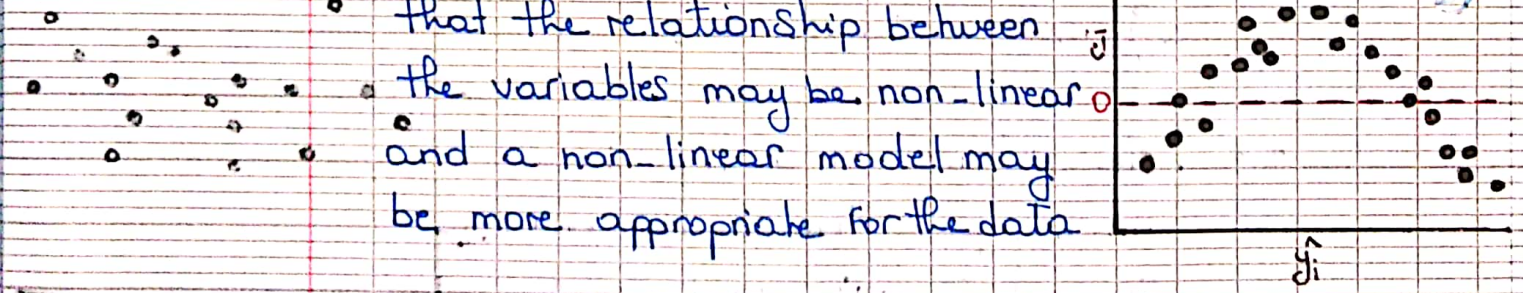
1) Accessing Model Fit:

A good residual plot should show a random, horizontal pattern, without any clear trends or structures.

Residuals should randomnly scattered around the horizontal axis.

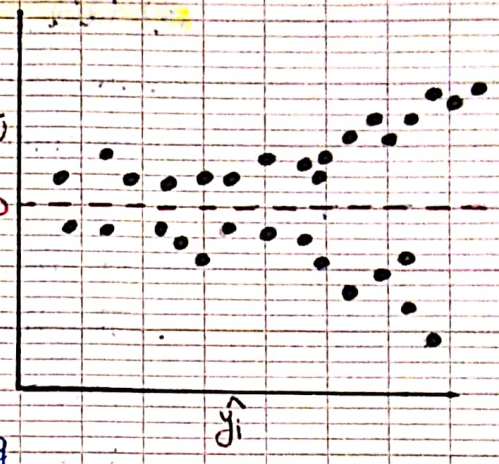
2) Detecting Non-Linearity

If the residual plot shows a non-random pattern, such as a curve, it indicates that the relationship between the variables may be non-linear and a non-linear model may be more appropriate for the data



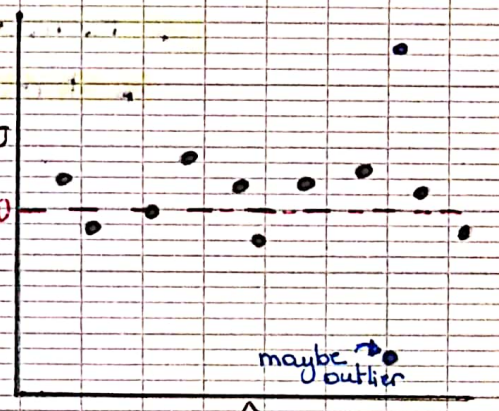
3) Identifying Heteroscedastic

Heteroscedasticity occures when the variance of the residuals is not constant across the range of predicted values. A residual plot can help identify heteroscedasticity by showing a pattern of increasing or decreasing spread in the residuals as the predicted values change



4) Spotting Outliers

Residual plot can identify outliers or influential data points that may have a significant impact on the regression model. Outliers appear as points that are far away from the majority of the other residuals in the plot and may require further investigation or removal from the analysis



2) Residuals vs Explanatory Variables x_{ij}

→ x-axis: independent variables (x_{ij})

→ y-axis: residuals (e_i)

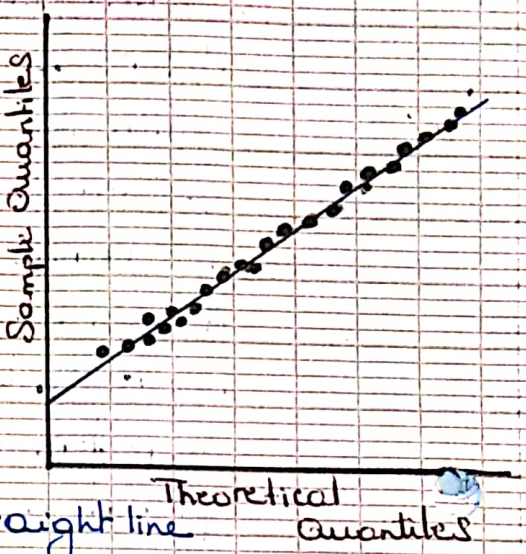
↳ It allow us to choose the correct form to use for each of the variables considered

↳ A horizontal behavior of the residues is expected to consider the model as satisfactory

3) QQ-plot of residuals

. In order to test the normality of the residuals, we can draw the QQ-plot of the residuals

. In QQ-plot, if the residuals (e_i) are normally distributed the points on the graph should roughly follow a straight line



→ Remedies

• Problem 1: Non-linearity of the model

→ detection: can be achieved by examining residual plots (where non-random patterns may indicate non-linear relationship)

→ Remedy: Transformation of variable

. Variable transformation involves applying a mathematical operation to one or more variables to alter their distribution or relationship with the response variable

. Common transformations include taking logarithms $\log(x)$, square roots ($\sqrt{x}$), power (x^2), or other mathematical functions of the variables

. We can transform the I.V., the D.V. or both

Choosing the best transformation involves exploring various transformations and assessing the transformatio that maximizes linearity and meets statistical assumptions

Problem 2: Heteroscedasticity of errors

→ detection: examine the residuals scatter plot, and if the spread of residuals varies systematically with the predicted values, it indicates the presence of heteroscedasticity

→ Redemy: Weighted Least Squares Method

The weighted Least Square method is employed as a remedy when dealing with heteroscedasticity. This method recognizes and addresses the non-constant variance of error by assigning different weights to observations based on their predicted values

→ $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $V(\varepsilon_i) = \sigma^2$

instead of having for every $i = 1, \dots, n$: $V(\varepsilon_i) = \sigma^2$

we will have:

$V(\varepsilon_i) = \frac{\sigma^2}{w_i}$ where $w_i > 0$ are the weight for each i

doing the math, we have the equivalent model:

$Y_w = X_w \beta + \varepsilon_w$

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 : $s^2 = \frac{\sum w_i (y_i - \hat{y}_i)^2}{n - p - 1}$

problem 3: Outliers

→ Outliers: extreme value of residual (dep. var) observation that does not resemble to the rest of the data.

→ Remedy: reject this observation / impose a detailed investigation.

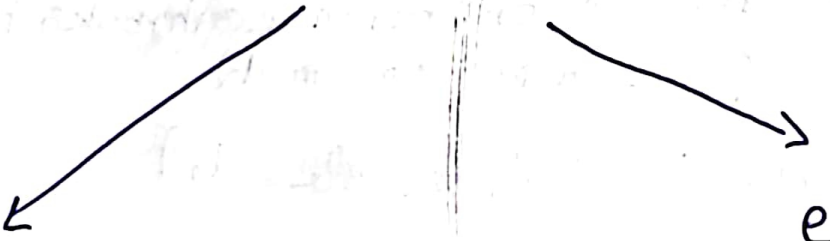
→ outlier have strong influence on slope ⇒ can modify the value of the correlation.

the can decrease the value of a legitimate correlation.

$|e_{outlier}| > |e_i|$

↓ To identify outliers (the are not necessarily influential)

Studentization of Residues:



Internal

the standard deviation is calculated including the point for which the residual is being calculated.

residual

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

↓

follows a student's distribution $t(n-p-1)$

↓ Leverage for ith obs.

i will be suspect if

$|r_i| > t((1-\alpha/2, (n-p-1))$

(quantile of Student's law $(\frac{1-\alpha}{2}, n-p-1)$

external

residuals that are scaled by their estimated st dev. calculated by all points except the point for which the residual is calculated

$r(-i) = \frac{e_i}{s(-i)\sqrt{1-h_{ii}}}$

,,

//

Relationship between them:

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

Hat Matrix: X' (transpose)

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

diagonal element of the matrix:

$$0 \leq h_{ii} = x_i(X'X)^{-1}x_i' \leq 1$$

↑ Leverage value

how much the i^{th} observation contributes to fitted values of the regression model.

$$-1 \leq h_{ij} \leq 1 \quad \text{(all the elements)}$$

$$\sum_{i=1}^{n} h_{ij} = 1$$

$$\sum_{j=1}^{n} h_{ij} = 1$$

diagonal elements $= p + 1$

$$\text{Trace}(H) = p + 1$$

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

↑ sum of square of all elements in the matrix.

$$e = y - \hat{y}$$

$$= (I_n - H)y$$

$$V(e) = \sigma^2(I_n - H) \qquad V(e_i) = \sigma^2(1 - h_{ii})$$

$$S^2(e_i) = S^2(1 - h_{ii})$$

Point of Leverage: → extreme value of independent variable

→ A Lever represents → the influence of obs "i" on $\hat{y}_i$ because of x_i

→ if levers were all equal the common value $\frac{p+1}{n}$ (Sum of levers / Sum of diagonal of hat matrix)

leverage $> 2\frac{p+1}{n}$ is a suspect

Cook's distance:

measure the influence of an observation on all forecast by taking into account leverage and importance of residuals.

for ith observation:

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

1- identifying atypical points by comparing C_i with 1

if $C_i > 1 \Longrightarrow$ atypical obs

$\Longrightarrow$ explaining this influence by consider for these obs. their residue as well as their leverage effect.

* C_i high

if r_i^2 high $\Rightarrow$ aberrant data

if $\frac{h_{ii}}{1-h_{ii}}$ " $\Rightarrow$ data having leverage effect

either both