Parametric Statistics

Week 9 - Linear Regression

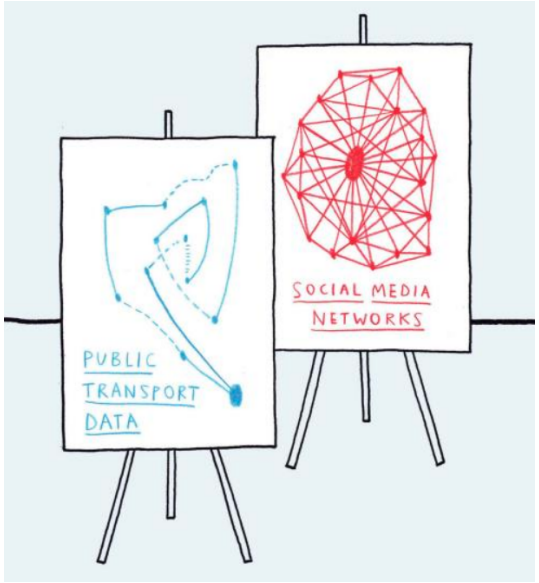
Juan Carlos Medina Serrano

Technische Universität München
Hochschule für Politik
Political Data Science

Munich, 08. January 2020

political data science

https://politicaldatascience.blogspot.de



From Correlation to Prediction

Until now, we were able to find the correlation between two variables.

We can go a step further and use these correlations to predict the value of one variable based on the value of another. Predict **outcome** variable **Y** from **predictor** variable **X**.

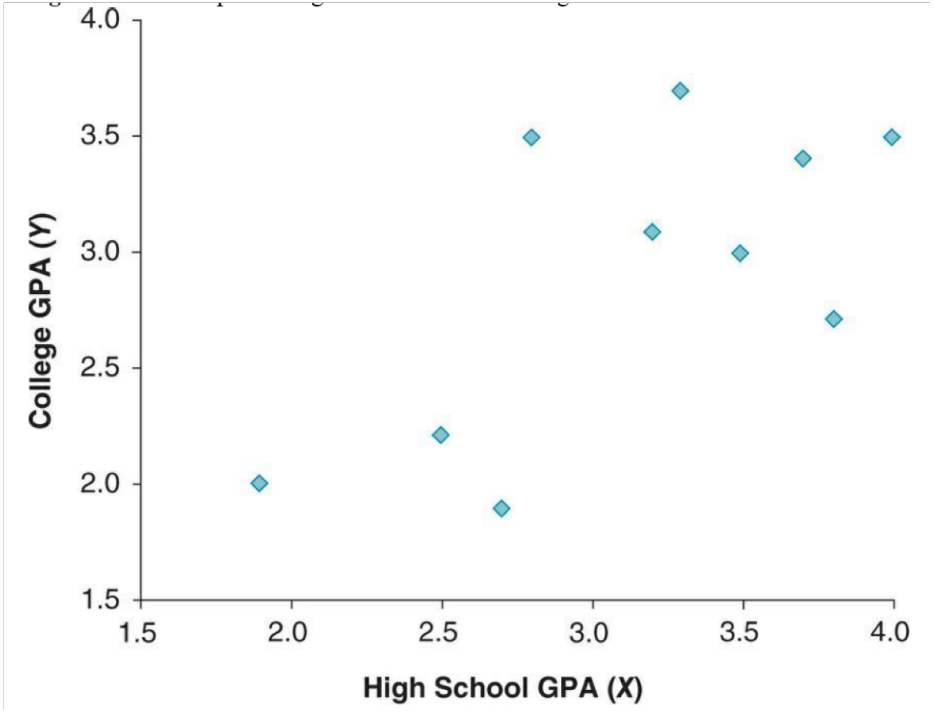
The easiest model that we learned at the beginning of the course was the mean. Here we did not take into account another variable.

We make the model better, now it will be a **linear** model.

Task: Find the regression line that express the relation between two variables at best.

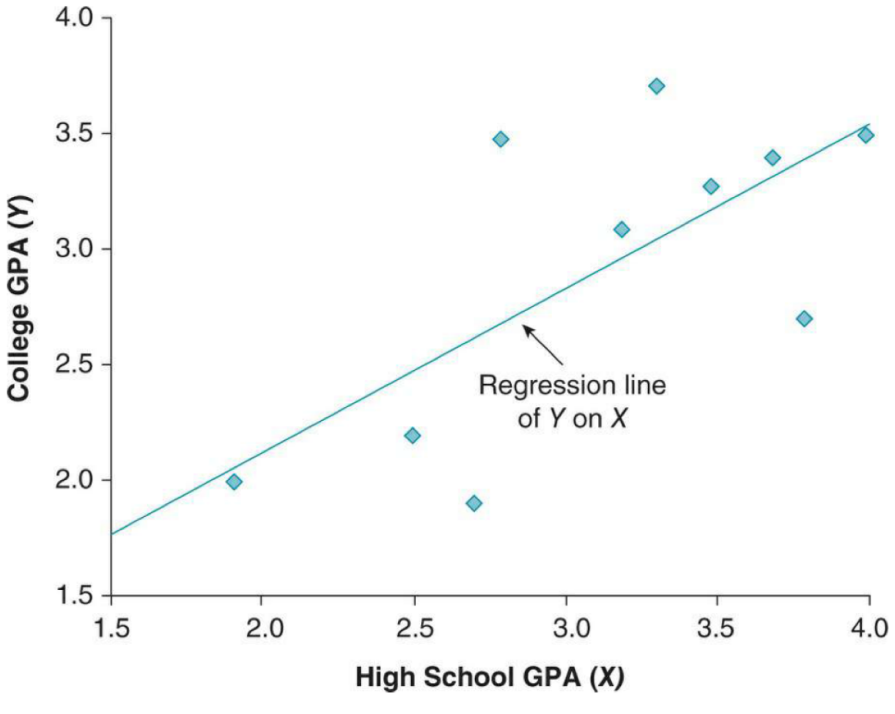
The Regression Line

Example Predicting college (first years of bachelor) exam grades using the high school exam grades.



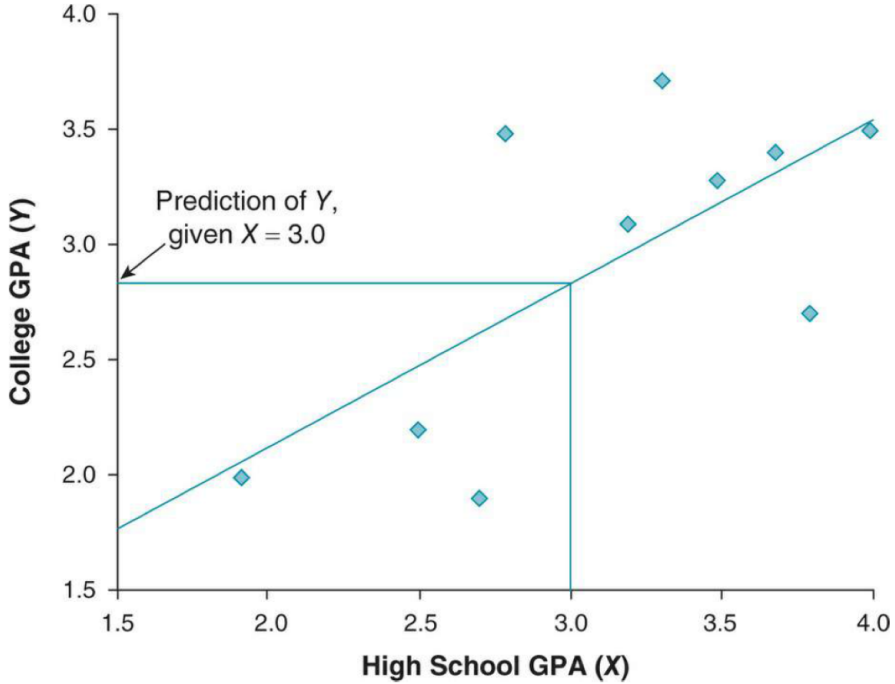
The Regression Line

Example Predicting college (first years of bachelor) exam grades using the high school exam grades.



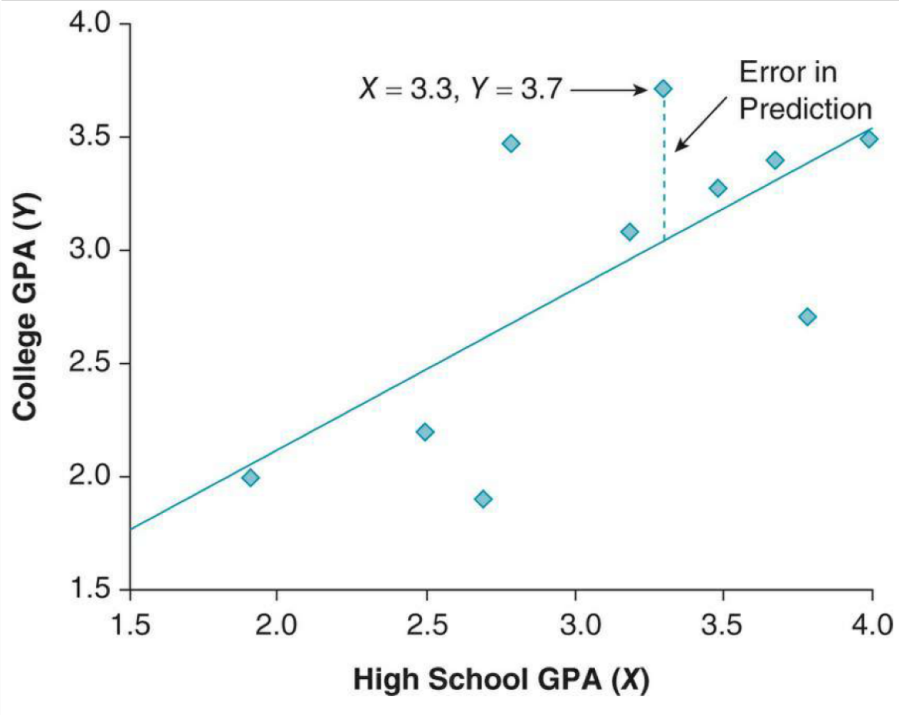
The Regression Line

Example Predicting college (first years of bachelor) exam grades using the high school exam grades.



The Regression Line

Example Predicting college (first years of bachelor) exam grades using the high school exam grades.



The Regression Line

Remember from math class the equation of the line?

$$y = mX + b$$

In statistics, we use the following notation:

$$Y = \beta_0 + \beta_1 X$$

Where the βs are called the regression coefficients, instead of slope (m) and intercept (b) from math class

Remember what we learned about models in week 2?

$$Outcome_i = (Model) + error_i$$

So our predictions for each data point are:

$$y_i = (\beta_0 + \beta_1 x_i) + error_i$$

A line that has a positive slope (β_1) describe a positive relationship, whereas a negative slope a negative relationship

The Regression Line

How to find the best line? = How to find the best β_0 and β_1?

With the coefficients that minimize the errors. (squared errors). We define the *residual sum of squares* as:

$$RSS = error_1^2 + error_2^2 + ... + error_N^2$$

or equivalent:

$$RSS = (y_1 - \beta_0 - \beta_1 x_1)^2 + (y_2 - \beta_0 - \beta_1 x_2)^2 + ... + (y_N - \beta_0 - \beta_1 x_N)^2$$

We want to minimize RSS! How? Using mathematics!

Note residual = error

The Regression Line

We learned last week that to maximize something we use the derivative and then equal to zero. This is the same to find the minimum!

Technique called: least squares.

Results:

$$\beta_1 = \frac{\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{N}(x_i - \overline{X})^2}$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

Where $\overline{X}$ is the mean of X, and $\overline{Y}$ is the mean of Y. **The regression line always passes through $\overline{X}$ and $\overline{Y}$.** Once again: there is no need to learn these formulas! Just try to understand what they represent.

After this calculations, replace the coefficients in the linear model (The hat is because we are estimating Y):

$$\hat{Y} = \beta_0 + \beta_1 X$$

The Regression Line

Example Predicting college (first years of bachelor) exam grades using the high school exam grades.

	X	Y	$(x-\overline{X})$	$(y-\overline{Y})$	$(x-\overline{X})*(y-\overline{Y})$	$(x-\overline{X})^2$
	3,5	3,3	0,4	0,4	0,1332	0,1
	2,5	2,2	-0,6	-0,7	0,4672	0,4
	4,0	3,5	0,9	0,6	0,4902	0,7
	3,8	2,7	0,7	-0,2	-0,1518	0,4
	2,8	3,5	-0,3	0,6	-0,1938	0,1
	1,9	2,0	-1,2	-0,9	1,1532	1,5
	3,2	3,1	0,1	0,2	0,0102	0,004
	3,7	3,4	0,6	0,5	0,2632	0,3
	2,7	1,9	-0,4	-1,0	0,4532	0,2
	3,3	3,7	0,2	0,8	0,1232	0,03
Mean	3,1	2,9				
Sum					2,748	3,904

$$\beta_1 = \frac{2,748}{3,904} = 0,704 \qquad \beta_0 = 2,9 - 0,704 * 3,1 = 0,719 \qquad \hat{Y} = 0,719 + 0,704X$$

How Good is the Linear Model?

There are two ways to asses how good the model is:

- Assessing the fit of the model
- Assessing the coefficients

Assessing the Fit

This section will remind you (hopefully) of one-way ANOVA, since its practically the same procedure...
We have again different sum of squares:

- Sum of Squares model: Calculate how much better the regression line is than just using the mean of the model. Differences between the mean and the predicted values on the line:

$$SS_{model} = \sum_{i=0}^{N} (\hat{y}_i - \overline{Y})^2$$

- Residual Sum of Squares: Difference between the predicted values and the real values. This we met before, the **RSS**

$$SS_{residual} = RSS = \sum_{i=0}^{N} (\hat{y}_i - y_i)^2$$

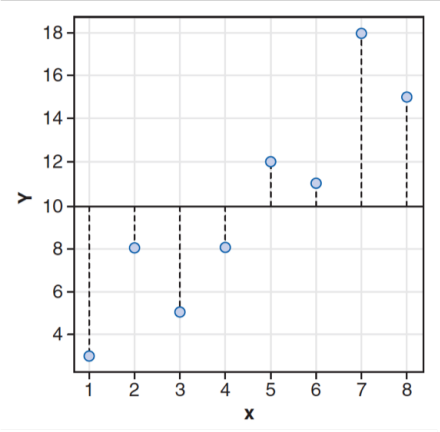
- Sum of Squares total: Difference between each observation and the most basic model, the mean:

$$SS_{total} = \sum_{i=0}^{N} (y_i - \overline{Y})^2$$

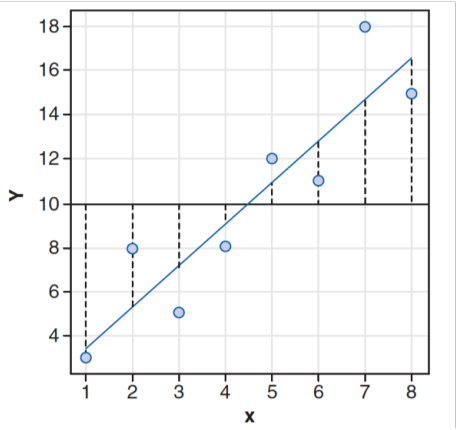
$$SS_{total} = SS_{model} + SS_{residual}$$

Assessing the Fit

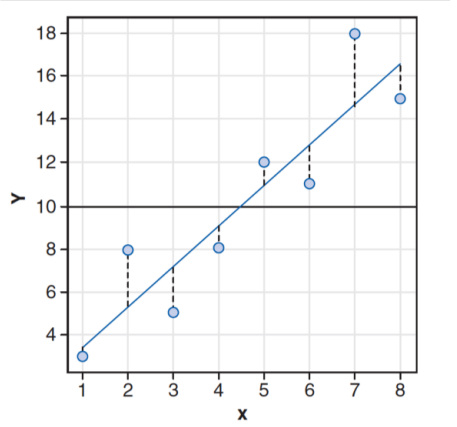
SS_{total}



SS_{model}



$SS_{residuals}$



Assessing the Fit

If the value of SS_{model} is large then the regression model is very different from using the mean to predict Y.

The R^2 measure of fit:

$$R^2 = \frac{SS_{model}}{SS_{total}}$$

represents the amount of variance in the outcome explained by the model (SS_{model}) relative to how much variation there was in the first place (SS_{total}).

This is the same R^2 as in correlation! The square root is the Pearson correlation coefficient (r) between variables!

R^2 measures the **proportion of variance in Y that can be explained using X**

Assessing the Fit

Time to test for significance of the lineal model! SAME test as one-way ANOVA :)

Degrees of freedom with N data points and p predictors (In linear regression we have $p = 1$):

- model - $df_{model} = p$
- residual - $df_{residual} = N - p - 1$

$$df_{total} = N - 1 = df_{model} + df_{residual}$$

Variances:

- variance explained by the model- $MS_{model} = \frac{SS_{model}}{df_{model}}$
- variance not explained by the model- $MS_{residuals} = \frac{SS_{residual}}{df_{residual}}$

The **F-statistic**:

$$F = \frac{variance\ explained}{variance\ not\ explained} = \frac{MS_{model}}{MS_{residuals}}$$

The higher the F value, the more statistical significance.

Assessing the coefficients

Another way to check model fit is to assess if the β_1 coefficient is significant.

Why β_1?

Because β_1 represents **the change in outcome resulting from a unit change in the predictor**

Example: If X = hours of study before an exam and Y= exam result and after asking the students from last semester, we obtain the linear model:

$$Y = 0.3X + 1$$

The 0.3 means that for every extra hour of study, the students obtained 0.3 points more in their exam result

Assessing the coefficients

If β_1 is 0, then there is no relation between variables, and no way to make predictions.

We need a significance test that tests if the value of β_1 is 0.

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

The t-test!

$$t = \frac{\beta_1 - 0}{SE(\beta_1)}$$

We compare with a t-distribution of degrees of freedom:

$$df = N - p - 1$$

Similar to the residual degrees of freedom. In case of p=1:

$$df = N - 2$$

Assessing the coefficients

Standard error of β_1:

$$SE(\beta_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \overline{X})^2}}$$

Where σ^2 is the variance of the residuals (errors)

We can use this standard error to create confidence intervals for β_1. The 95% confidence interval is:

$$[\beta_1 - 2 * SE(\beta_1), \quad \beta_1 + 2 * SE(\beta_1)]$$

Important is that the confidence interval does not include 0!

How Good is the Linear Model?

Should we use the F-statistic of the complete model or the t-statistic for the β_1 coefficient?

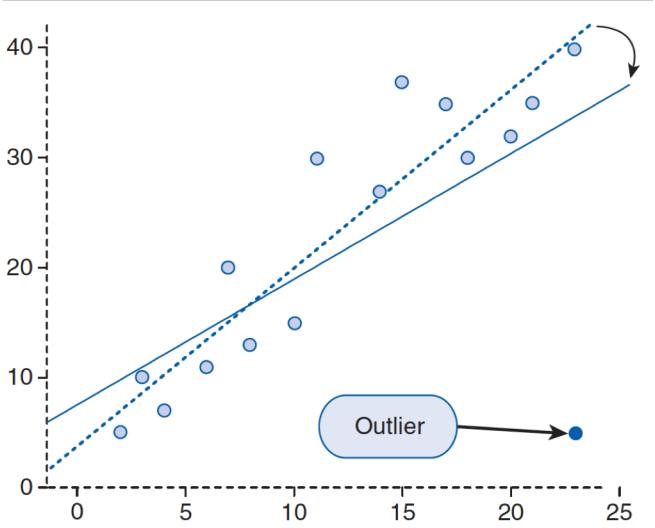
For linear regression, it does not matter which one you choose! They are equivalent!

$$\text{F-statistic} = \text{t-statistic}^2$$

This will change next week, when we have more than one predictors. We will then need BOTH tests for finding significance.

Outliers

One problem with linear regression is the presence of outliers!
The model can drastically change by including them:



Outliers

How to detect them?

Find the points with the biggest residuals! (distances to the line) Remember: Residual = Error

But how big is big?

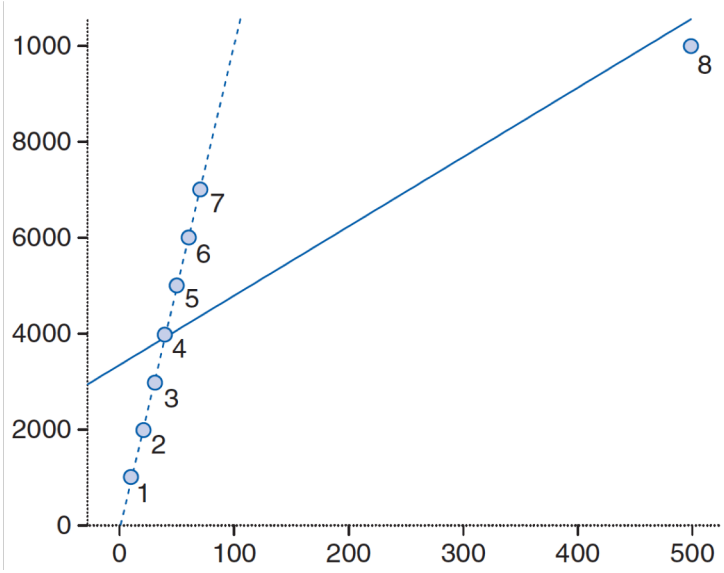
We can standardized the residuals:

Standardized residuals: Residuals divided by their standard deviation.

Observations with standardized residuals greater than **3** are most probable outliers!

Outliers

However, there are some outliers that have so much influence in the model, that they end up having a small residual:



We call these **influential** cases.

Outliers

How to detect them?

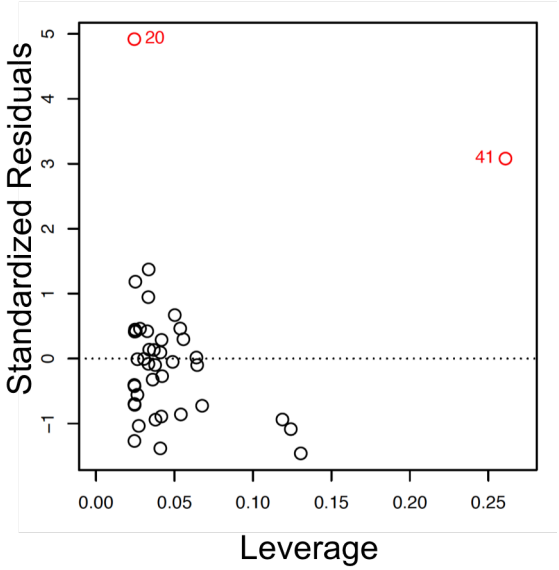
Two measures in statistics:

- **Cook's distance (D)**: Values greater than 1 are of concern.
- **Leverage (h)**: Can be between 0 and 1. The higher the leverage, the more influence the observation has.

Outliers

How to detect them at the same time?

Leverage vs. Standardized residuals plot:



Assumptions for Linear Regression

Assumptions

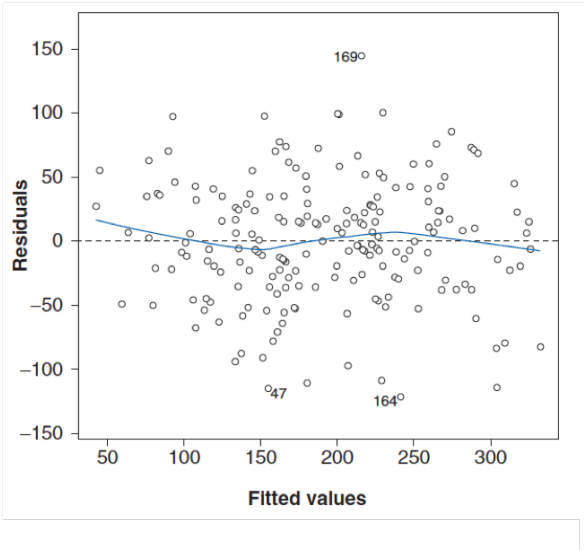
- **Linearity**: The relationship between variables is linear.
- **Homoscedasticity**: The variance of the residuals (errors) should be constant at each level of the predictor variable.
- **Normally distributed errors**: We asume that the errors are random. This means they should be normally distributed with mean value 0.
- **Independent errors**: The residuals should be uncorrelated.

How do we check if these assumptions are met?

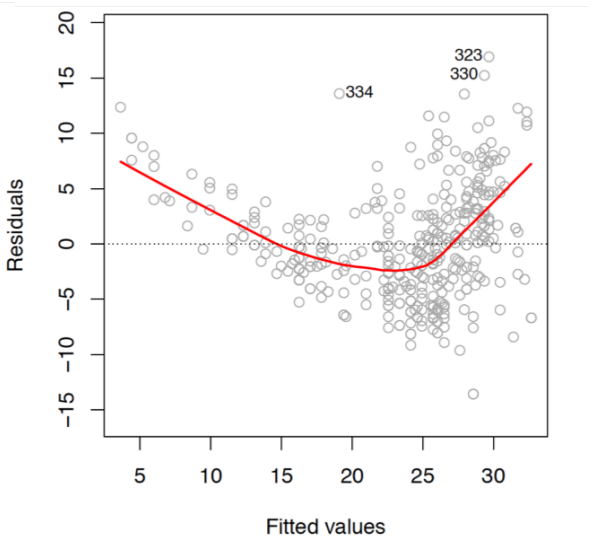
Linearity

The relationship between variables is linear.
We use a plot with the fitted values ($\hat{Y}$) and the residuals:

Linear

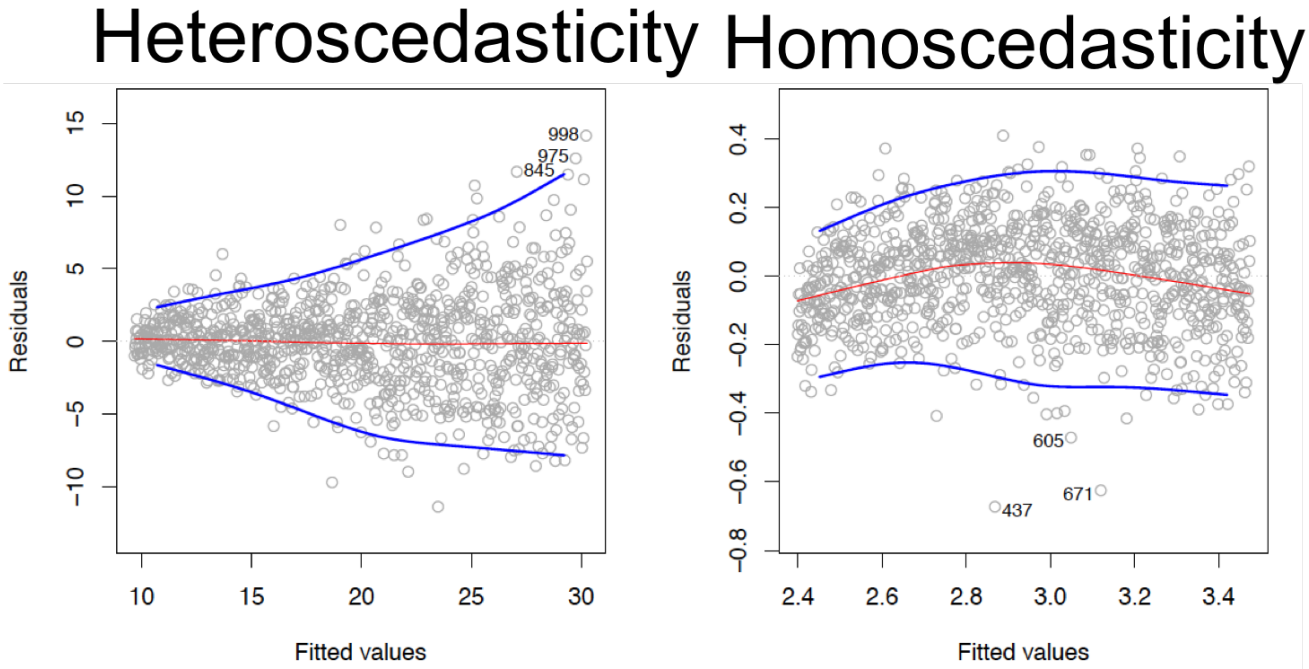


Non-Linear



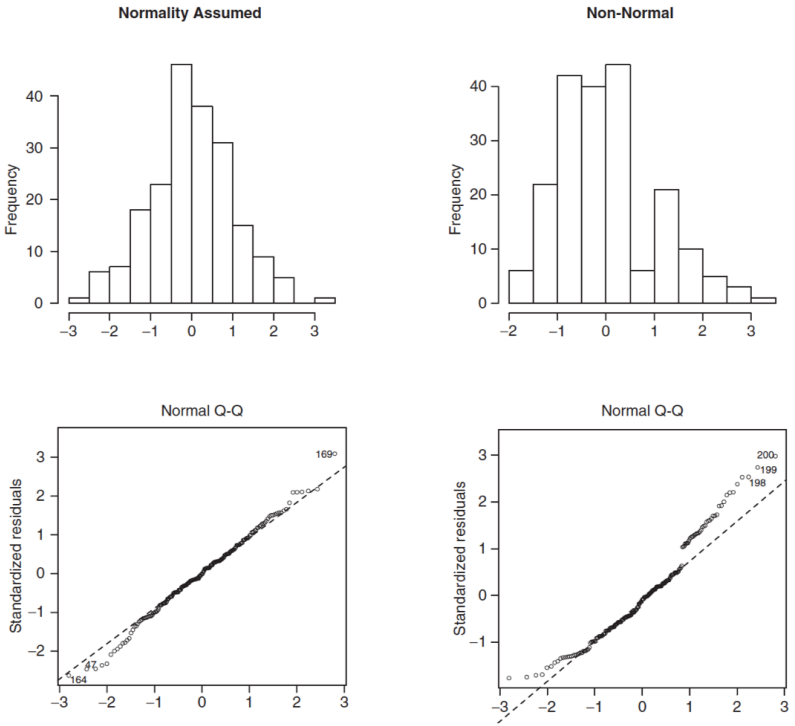
Homoscedasticity

The variance of the residuals (errors) should be constant at each level of the predictor variable.
We use again the plot with the fitted values ($\hat{Y}$) and the residuals:



Normally Distributed Errors

We already learned how to test for normality! We can use the Q-Q plots for example:

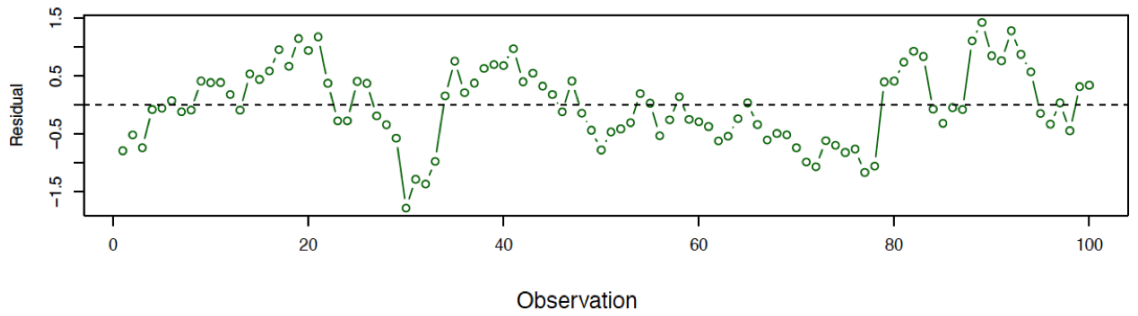


Independent errors

The residuals should be uncorrelated.

We use the **Durbin–Watson test**. Values less than 1 and bigger than 3 means correlation.

If they are correlated the most probable reason is that they are a time series and the residuals will correlate:



More on time series in three weeks...

Literature

- Statistics for people who hate statistics: Chapter 16
- Discovering Statistics using R: Chapter 7