Regression diagnostics and model selection

MS-C2128 Prediction and Time Series Analysis

Fall term 2020

Week 2: Regression diagnostics and model selection

1. Regression diagnostics
2. Model selection

Content

1 Regression diagnostics

2 Model selection

Regression diagnostics

- Questions:
  - Does the model describe the dependence between the response variable and the explanatory variables well
    1. contextually?
    2. statistically?
- A good model describes the dependencies as well as possible.
  Assessment of the goodness of a regression model is called regression diagnostics.
- Regression diagnostics tools:
  - graphics
  - diagnostic statistics
  - diagnostic tests

Regression model selection

In regression modeling, one has to select

1. the response variable(s) and the explanatory variables,
2. the functional form and the parameters of the model,
3. and the assumptions on the residuals.

Remark

The first two points are related to defining the model part and the last point is related to the residuals. These points are not independent of each other!

Problems in defining the model part

(i) Linear model is applied even though the dependence between the response variable and the explanatory variables is non-linear.

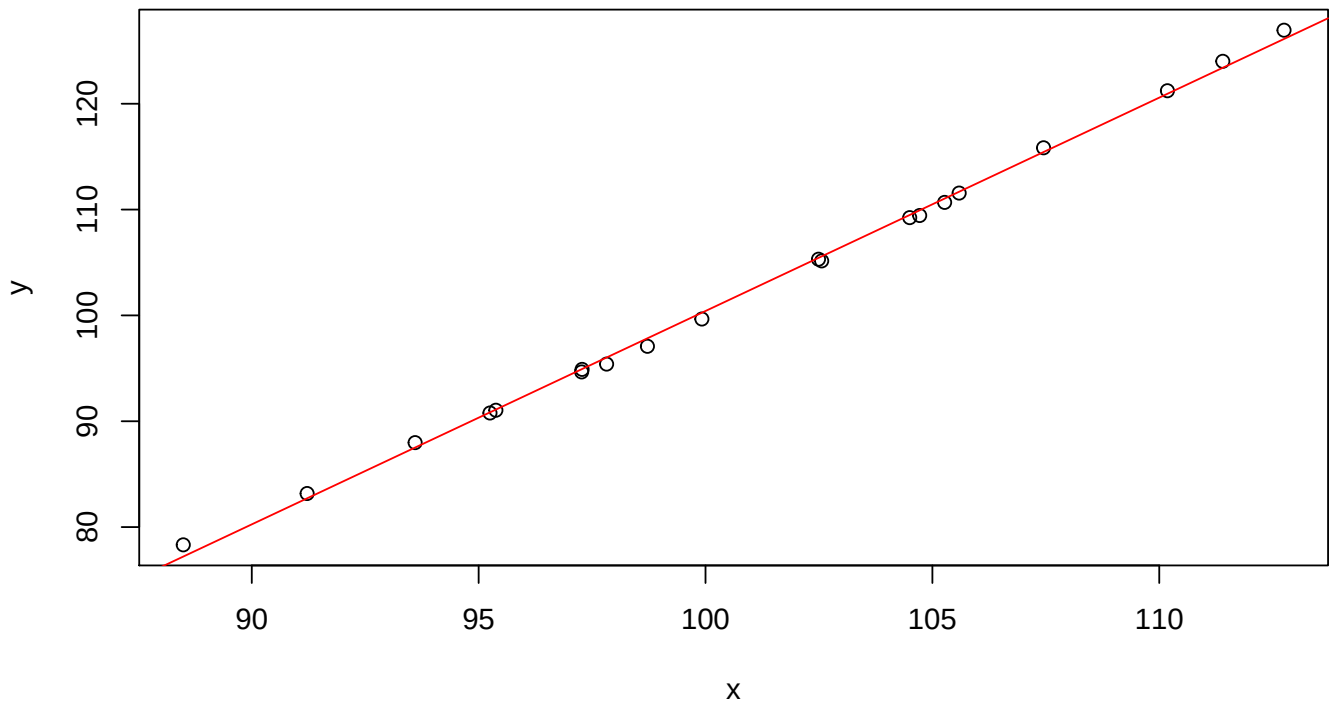
(ii) Too many or too few explanatory variables are chosen.

(iii) It is assumed that the model parameters are constants even though the assumption does not hold.

Remark

Fundamental errors in defining the model part can often be detected from the fitted values of the response variables.

Does a linear model fit to the observations (Data 1)?



Problems in assumptions on the residuals

(i) Homoscedasticity and/or uncorrelatedness is assumed even though the assumption does not hold.

(ii) Normality is assumed even though the residuals are not normally distributed.

Remark

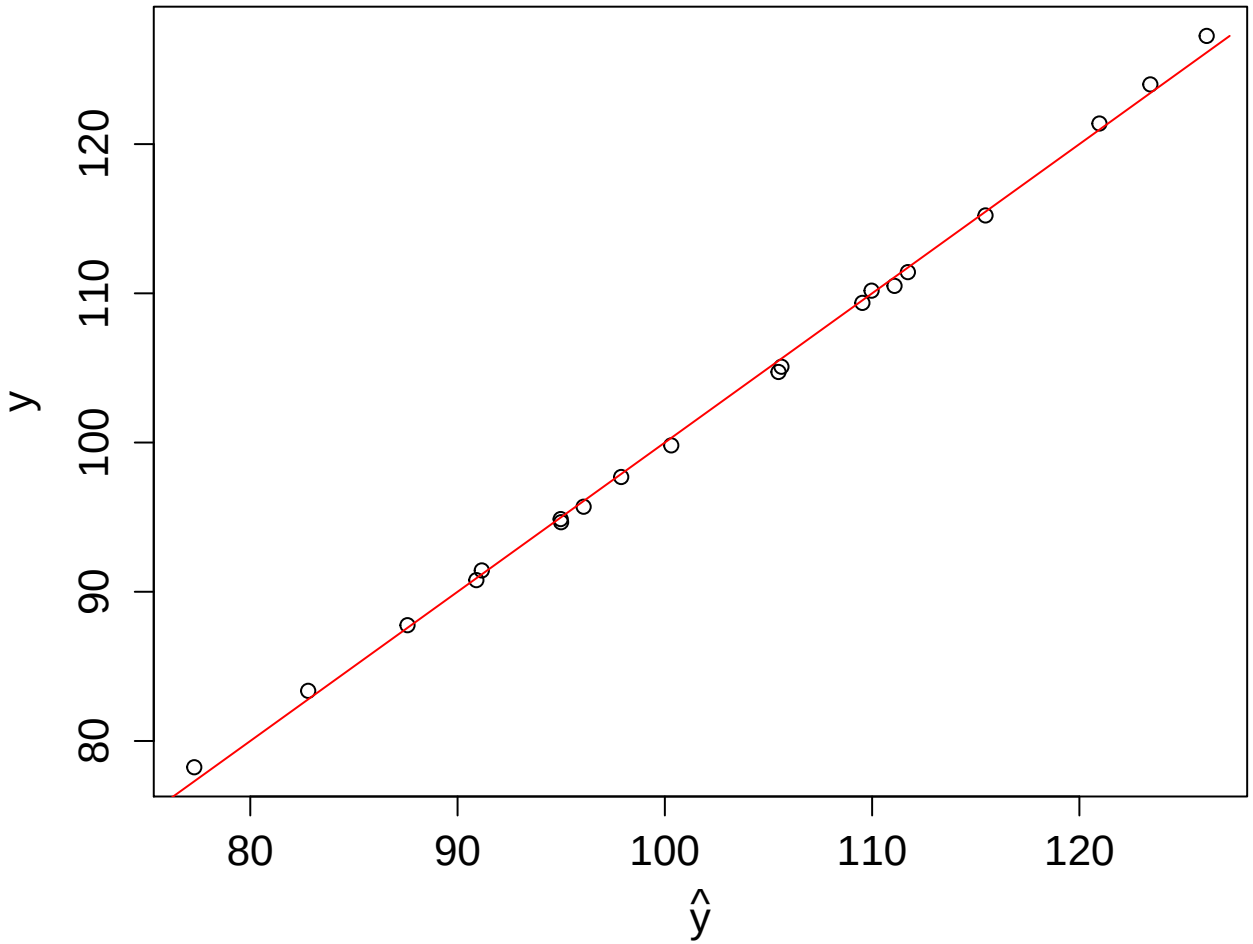
Fundamental errors in assumptions on the residuals can often be detected from the estimated residuals.

Linear regression: Diagnostic checks

- Are there outlying observations?
- Are the regression parameters constants?
- Are the explanatory variables linearly independent?
- Are the residuals homoscedastic and uncorrelated (and normally distributed)?

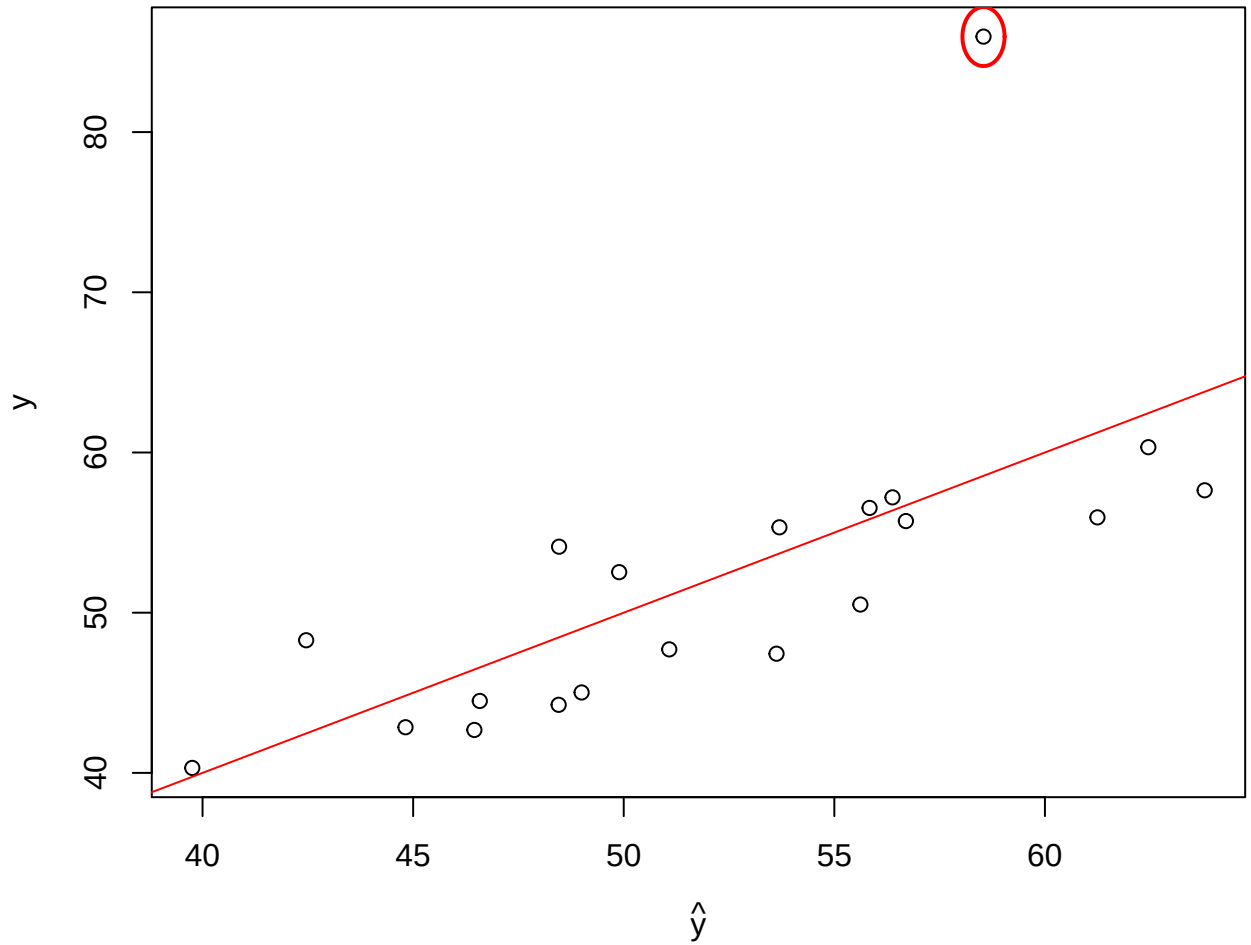
Graphics: Scatter plot

The scatter plot of the fits $\hat{y}$ and the observed values y (Data 1). Possibly some non-linear dependence can be detected.



Graphics: Scatter plot

An outlying observation can be detected (Data 2).

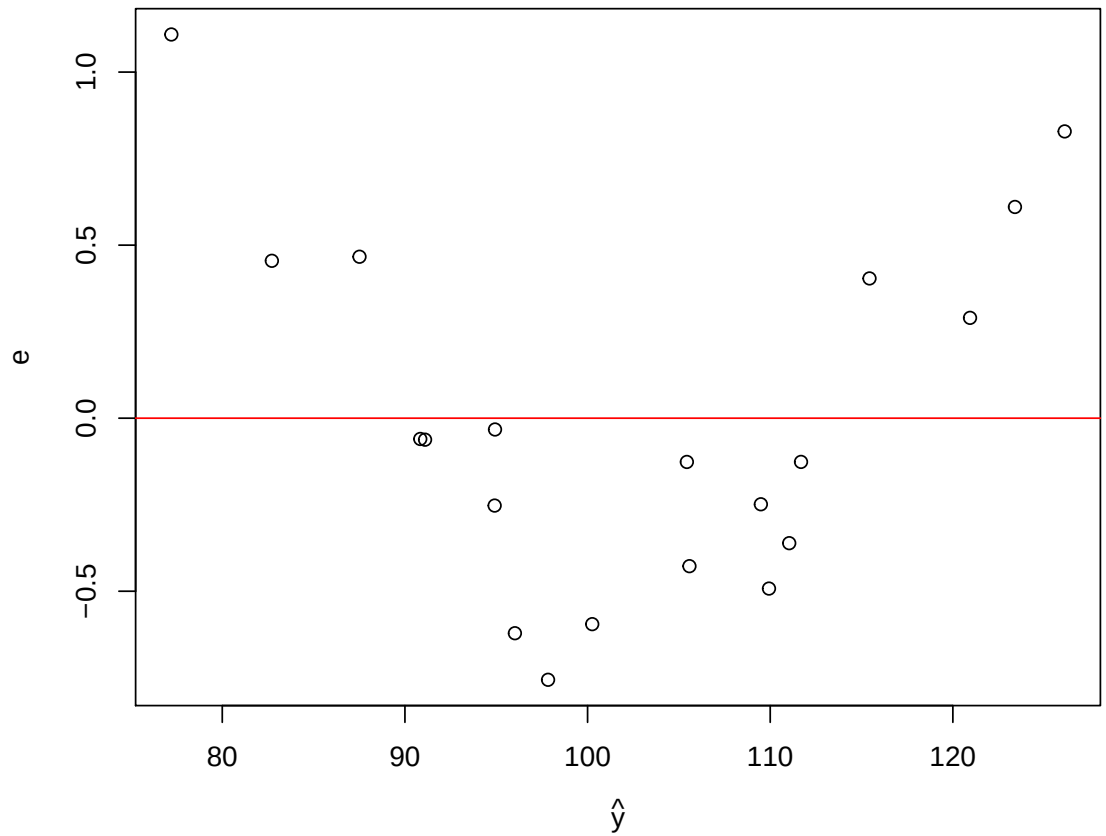


Graphics: Scatter plot, $(\hat{y}, y)$

- The model is the better the better the points follow a line with slope 1.
- Nonlinear shapes indicate that the functional form of the model part is not well selected.
- Outlying observations are typically far away from the line.

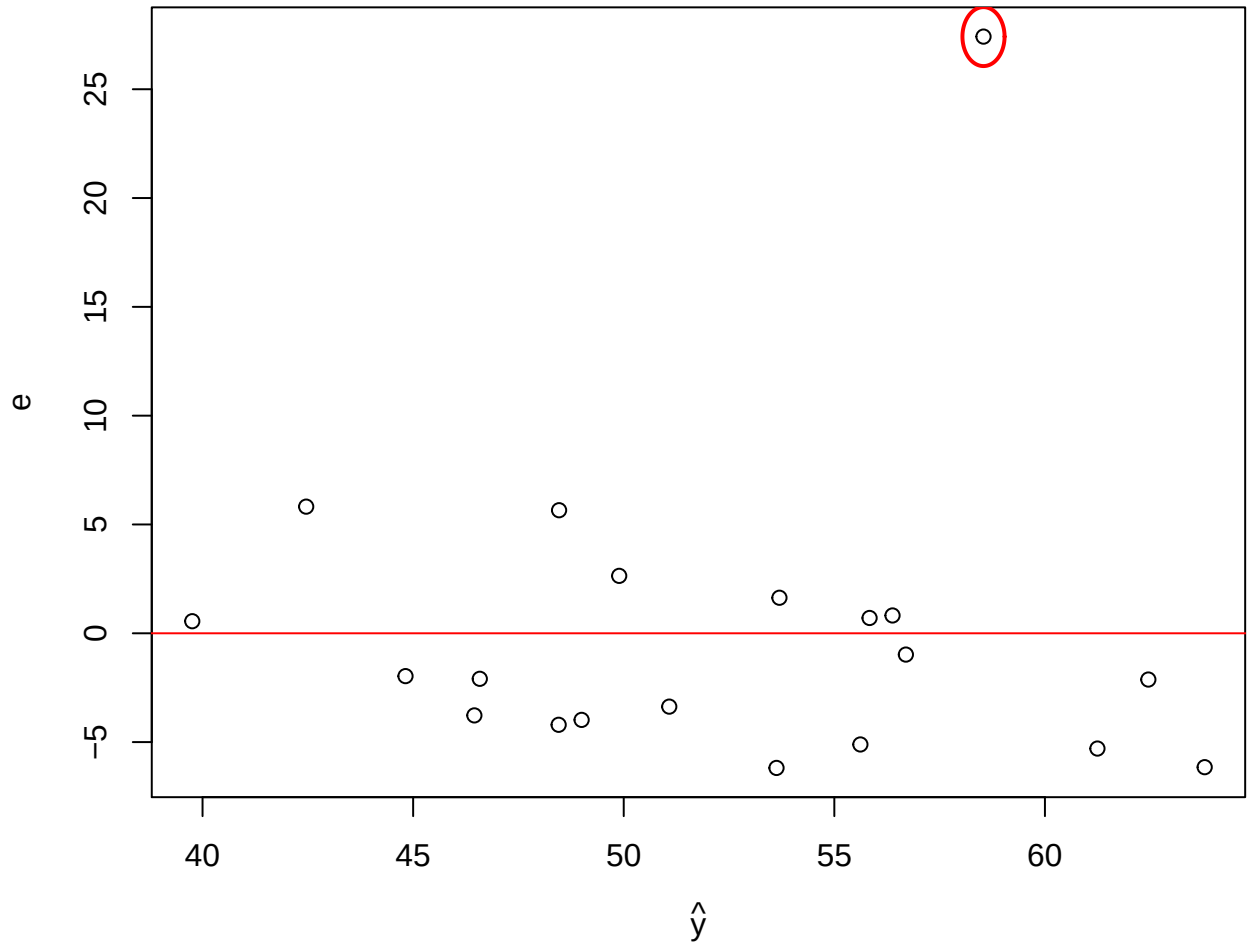
Graphics: Residual plot

A residual plot is a scatter plot of the fits $\hat{y}$ or of the explanatory variables x_j and the estimated residuals e. Nonlinear shape is clearly detected (Data 1).



Graphics: Residual plot

An outlying observation is clearly detected (Data 2).



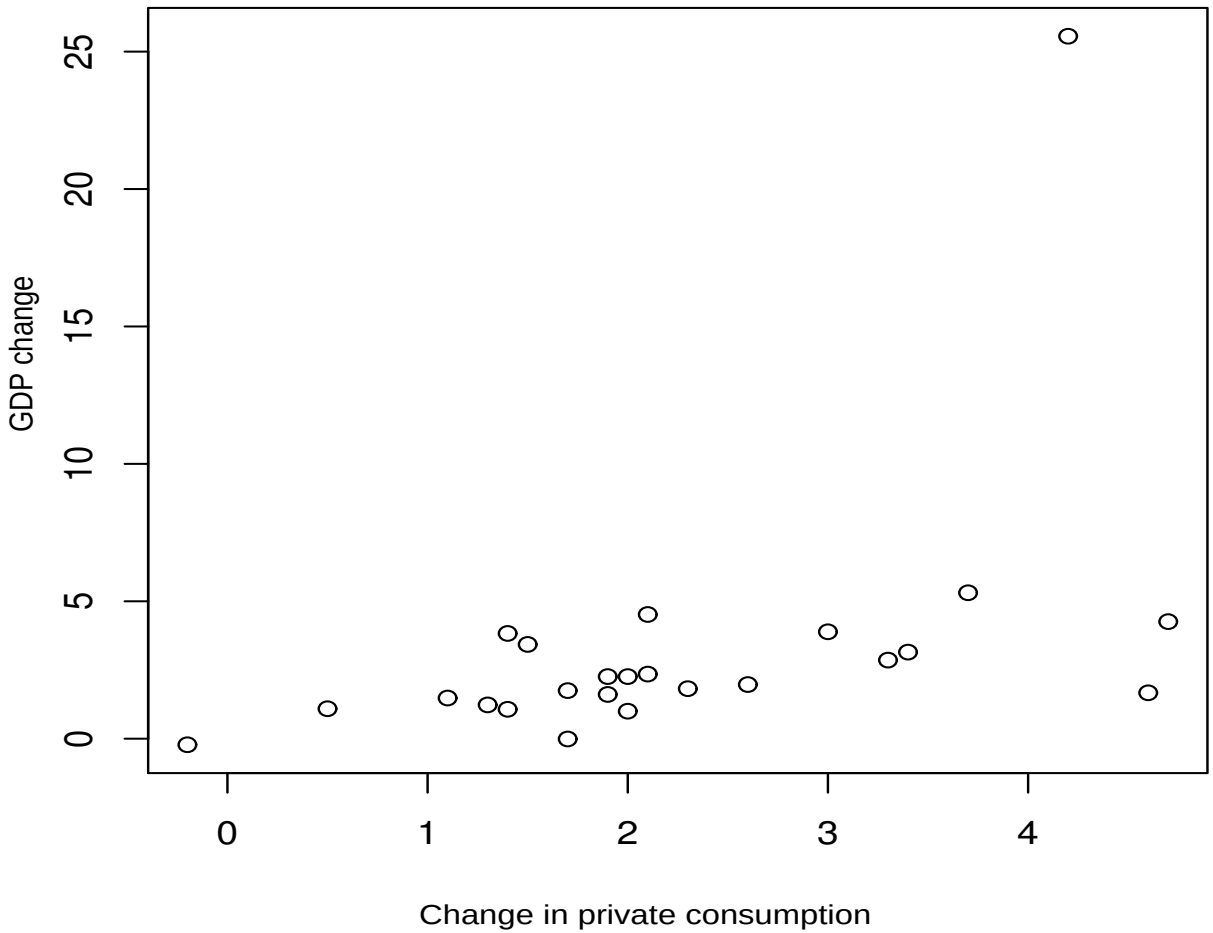
Graphics: Residual plot

- If the model is good, the residuals are approximately uniformly distributed on a rectangle. Outlying observations lie far away from the horizontal axis.
- If the shape of the scatter plot is for example quadratic (not approximately a rectangular), the functional form of the model part might be wrong.
- If the height of the scatter plot is not approximately the same everywhere, the residuals might be heteroscedastic or the functional form of the model part might be wrong.

Outlying observations

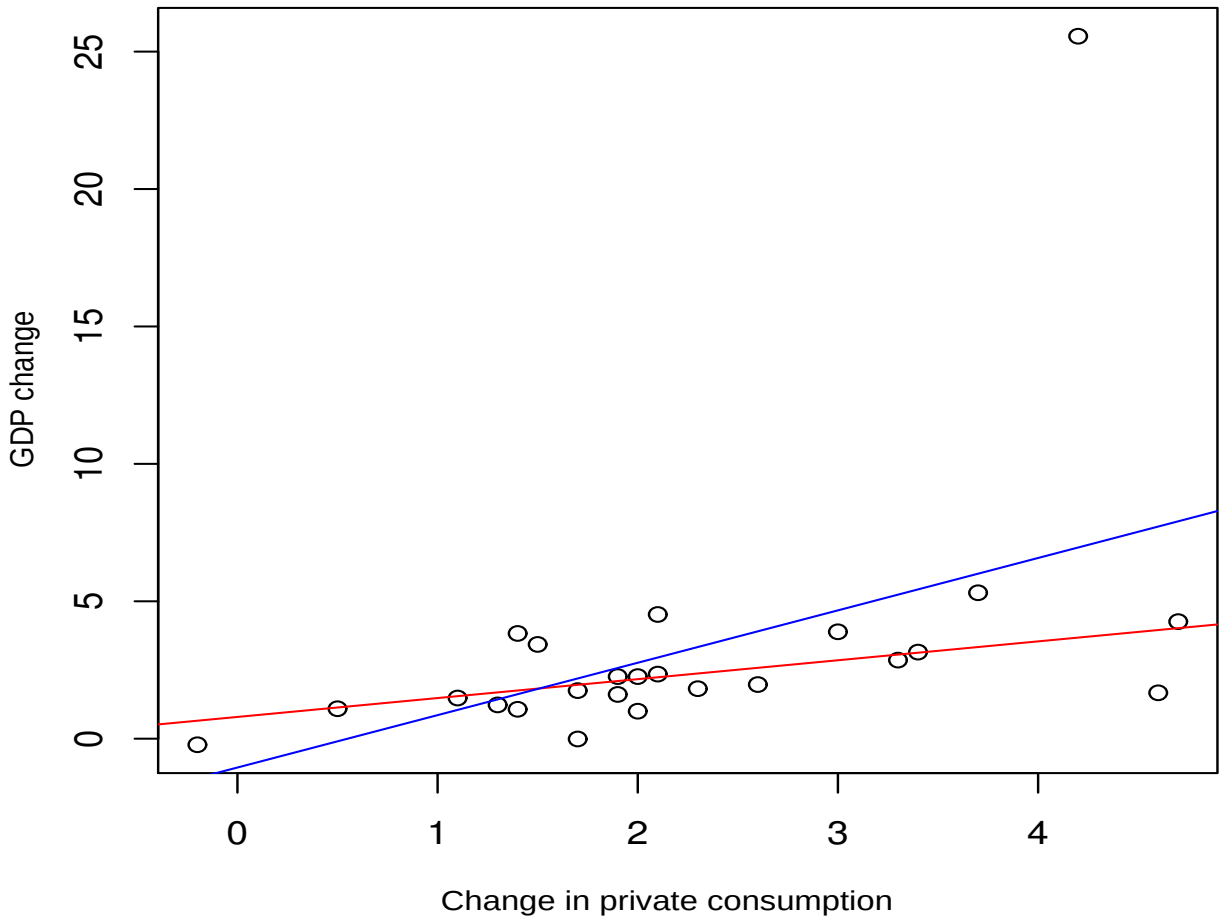
- Outlying observations are observations that, in some sense, differ significantly from the other observations.
  - In statistical analysis, an observation is outlying, if its effect is abnormally significant in the analysis.
    - If removing an observation changes significantly the results of an analysis, the observation is outlying.
  - Outlying observations may not be excluded from the analysis just because they do not fit to the model!
- In regression analysis, outlying observations may
  - complicate the model selection,
  - complicate estimation,
  - and lead to erroneous interpretations and inference.

Outlying observations: Example



GDP change and the change in private consumption (both compared to the previous year) in 2015 in 24 European countries. (Ireland is an outlying observation.) Source: OECD.com.

Outlying observations: Example



GDP change and the change in private consumption (both compared to the previous year) in 2015 in 24 European countries. (Ireland is an outlying observation.) Source: OECD.com.

Detecting outlying observations: Cook's distance

Cook's distance that corresponds to an observation y_i is given by

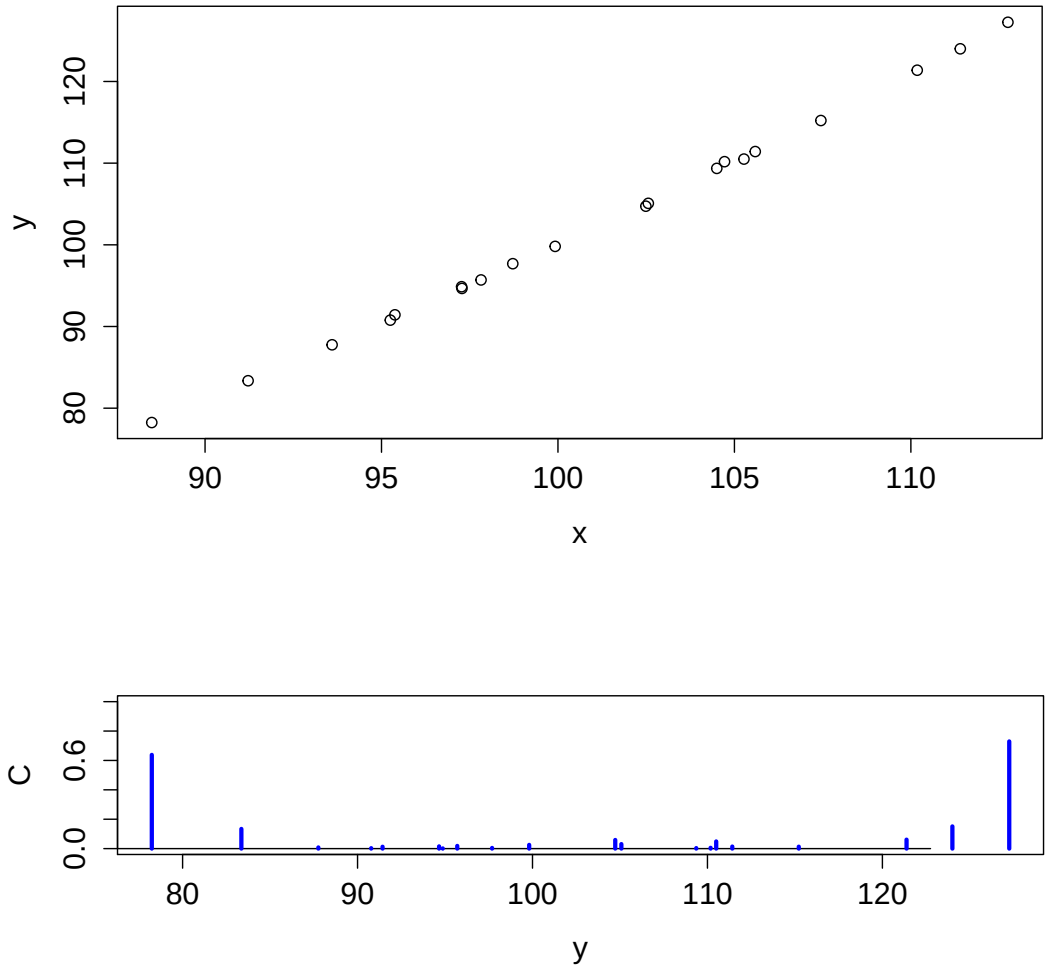
$$C_i = \frac{\sum_{l=1}^{n} (\hat{y}_l - \hat{y}_l^i)^2}{(k+1)s^2},$$

where the

- $(\hat{y}_1, ..., \hat{y}_n)$ are the fits obtained when the entire data is used and
- $(\hat{y}_1^i, ..., \hat{y}_n^i)$ are the fits obtained when all the data points except the observation i are used.

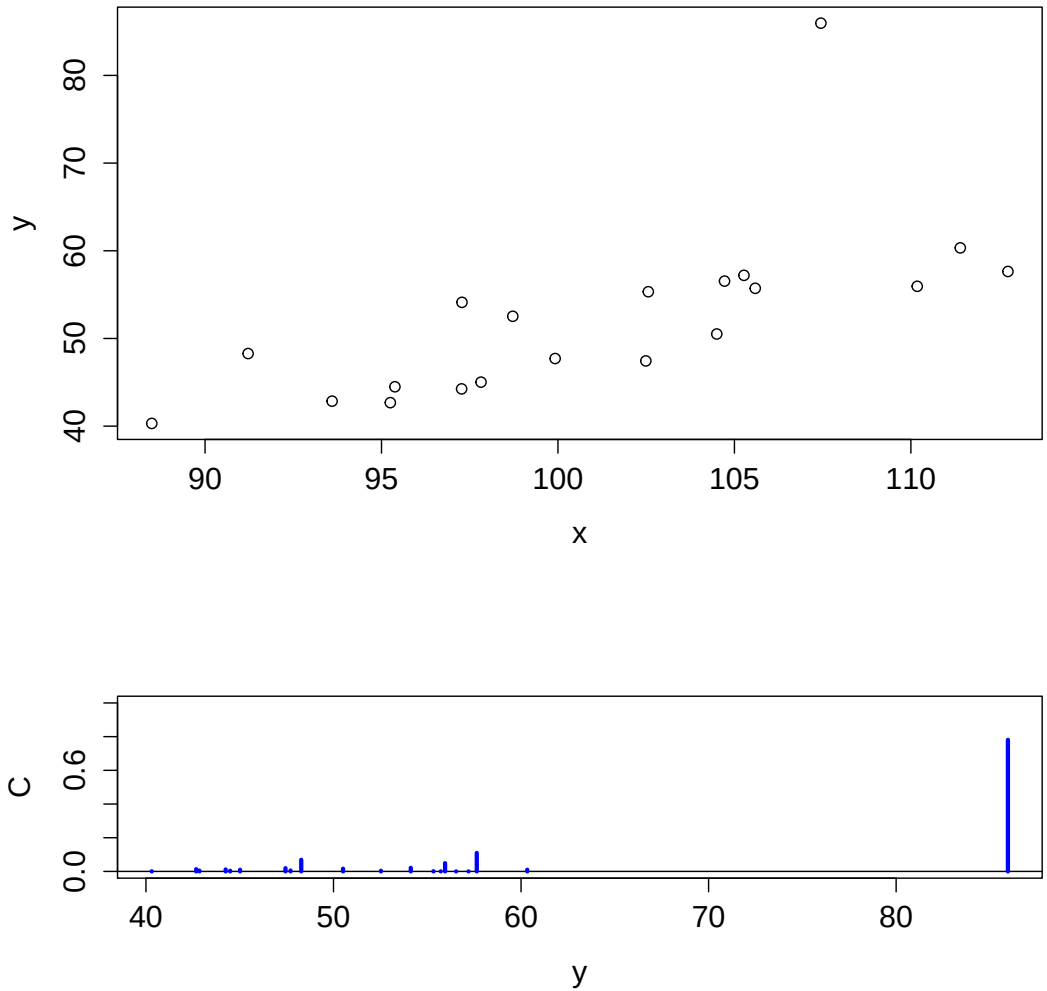
If the cook's distance C_i is significantly larger that the cook's distances of the other observations, then one should take a closer look at the observation y_i.

Cook's distances, Data 1



Note the directions of the *y*-axes.

Cook's distances, Data 2



Note the directions of the *y*-axes.

How to deal with outlying observations?

- If the outlying observations are clearly errors (for example, height of a man is 17.8 meters) one can correct or exclude the erroneous observation.
- What if the outlying observations are not erroneous?
  - Options:
    1. Ask a context expert if there is an explanation why the observation behaves differently than the other data points and should be analyzed separately.
    2. Apply a model that enables to split the model into separate parts.
    3. Apply some robust estimation procedure that is not too sensitive to outlying observations.
  - There are no general rules that would tell you what to do. However, one may not simply remove the data points that are unpleasant.
    - If it is well-justified to remove an observation, it still has to be reported and analyzed in detail.

Testing for parameter instability

If there is a reason to suspect that different linear models fit to different subgroups, one should consider testing for parameter instability. For example, it could be that the income level has a linear effect on the consumption of a certain good, but if the income level is high enough, the effect is smaller, or it could be that the effect is different for people that are under 40 years old and for people that are over 40 years old.

Testing for parameter instability

The purpose of the testing is to find out whether or not the parameters of the linear model are the same for two separate subgroups.

- The null hypothesis H_0: the model parameters are the same for both subgroups.
- The alternative hypothesis H_1: the model parameters for the two subgroups are not the same.

Testing can be based on considering the error sum of squares after fitting a linear model to the entire data and after fitting a linear model to the subgroups of the data.

Testing for parameter instability

Assume that the observations $(\boldsymbol{x}_1, y_1), ...(\boldsymbol{x}_n, y_n)$ have been divided into to two separate subgroups based on some criteria (for example under and over 40 year olds). Assume that the sample sizes of the subgroups are $h \geq k+1$ and $n-h \geq k+1$. (The k here is the number of the explanatory variables in the linear model.) Reorder the pairs such that the observations $(\boldsymbol{x}_1, y_1), ...(\boldsymbol{x}_h, y_h)$ form the first subgroup and the observations $(\boldsymbol{x}_{h+1}, y_{h+1}), ...(\boldsymbol{x}_n, y_n)$ form the second subgroup.

Testing for parameter instability

The null hypothesis H_0: the model parameters are the same for both subgroups, can be tested by applying the following permutation test:

1. Assume that the observations $(\boldsymbol{x}_i, y_i)_{i=1,\ldots,n}$ are split in to two subgroups:
   1. Subgroup 1: $(\boldsymbol{x}_i, y_i)_{i=1,\ldots,h}$ (sample size h)
   2. Subgroup 2: $(\boldsymbol{x}_i, y_i)_{i=h+1,\ldots,n}$ (sample size $n-h$)
2. Estimate the parameters of the linear model using the entire data and calculate the corresponding SSE.
3. Estimate the parameters of the linear model separately for the subgroup 1 and for the subgroup 2, and calculate the corresponding SSE_1 and SSE_2.
4. Calculate the statistic

$$Ch = \frac{\big(n - 2(k+1)\big)}{k+1} \, \frac{SSE - (SSE_1 + SSE_2)}{(SSE_1 + SSE_2)},$$

where k is the number of the explanatory variables.

Testing for parameter instability

5. Divide now the original entire sample randomly into two separate subgroups $p1$ and $p2$ that have sample sizes h and $n - h$.
6. Estimate the parameters of the linear model separately for the subgroup $p1$ and for the subgroup $p2$, and calculate the corresponding SSE_{p1} and SSE_{p2}.
7. Calculate the value

$$Ch_p = \frac{n - 2(k+1)}{k+1} \frac{SSE - (SSE_{p1} + SSE_{p2})}{(SSE_{p1} + SSE_{p2})}.$$

8. Repeat the steps 5, 6 and 7 m times.
9. Order the values Ch_p from the smallest to the largest and calculate the empirical $(1 - \alpha) \cdot 100$th percentile from the ordered sample. If the original statistic Ch is larger than the calculated percentile, then the null hypothesis is rejected (on significance level α).

Testing for parameter instability

Remark

In statistical testing one usually chooses the significance level α to be equal to 0.05 or 0.01 or 0.001.

Remark

If one wishes to divide n observations into two separate subgroups that have sample sizes h and $n - h$, there are $\binom{n}{k}$ such possible divisions. Usually it is impossible to consider all the possible divisions, but m should be chosen to be large enough (10000 or 20000).

Testing for parameter instability, $\epsilon \sim N(0, \sigma^2)$: Chow-test

If the residual are normally distributed, then one does not have to apply the permutation test in testing for parameter instability. Under the assumption of normally distributed residuals, the statistic

$$Ch = \frac{\big(n - 2(k+1)\big)}{k+1} \, \frac{SSE - (SSE_1 + SSE_2)}{(SSE_1 + SSE_2)}$$

follows, under the null hypothesis, the $F(k+1, n-2(k+1))$ distribution.

Non-constant regression parameters

What should one do, if the regression parameters are not constants?

- One can divide the data into two groups and analyze them separately.
- One can apply a (non-linear) model that allows non-constant parameters.

Multicollinearity

Multicollinearity is a phenomenon in which one explanatory variable in a regression model can be linearly predicted from the other explanatory variables with a substantial degree of accuracy. That is, it is a phenomenon in which the explanatory variables are linearly dependent. High degree of multicollinearity is harmful. Multicollinearity may complicate estimation, since it can make the matrix $\boldsymbol{X}$ to be singular. Even if the matrix $\boldsymbol{X}$ is of full rank, multicollinearity complicates examining the effect of one single explanatory variable and consequently it complicates prediction.

Multicollinearity: Variance inflation factor

The variance inflation factor of an explanatory variable x_j is given as

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, 2, ..., k$$

where R_j^2 is the coefficient of determination of the linear regression model where

- the variable x_j is the response variable, and
- the explanatory variables are all the other original explanatory variables x_l, $l \neq j$.

Multicollinearity: Variance inflation factor

- If $R_j^2 = 0$ and consequently $VIF_j = 1$ for all $j = 1, ..., k$, then
  - the explanatory variables $x_1, x_2, ..., x_k$ are orthogonal. (This is ideal.)
- If $R_j^2 = 1$ for some $j = 1, ..., k$, then
  - the explanatory variable x_j can be given as a linear combination of the other explanatory variables.
- If $VIF_j > 10$ for some $j = 1, ..., k$,
  - it indicates high degree multicollinearity.

Multicollinearity

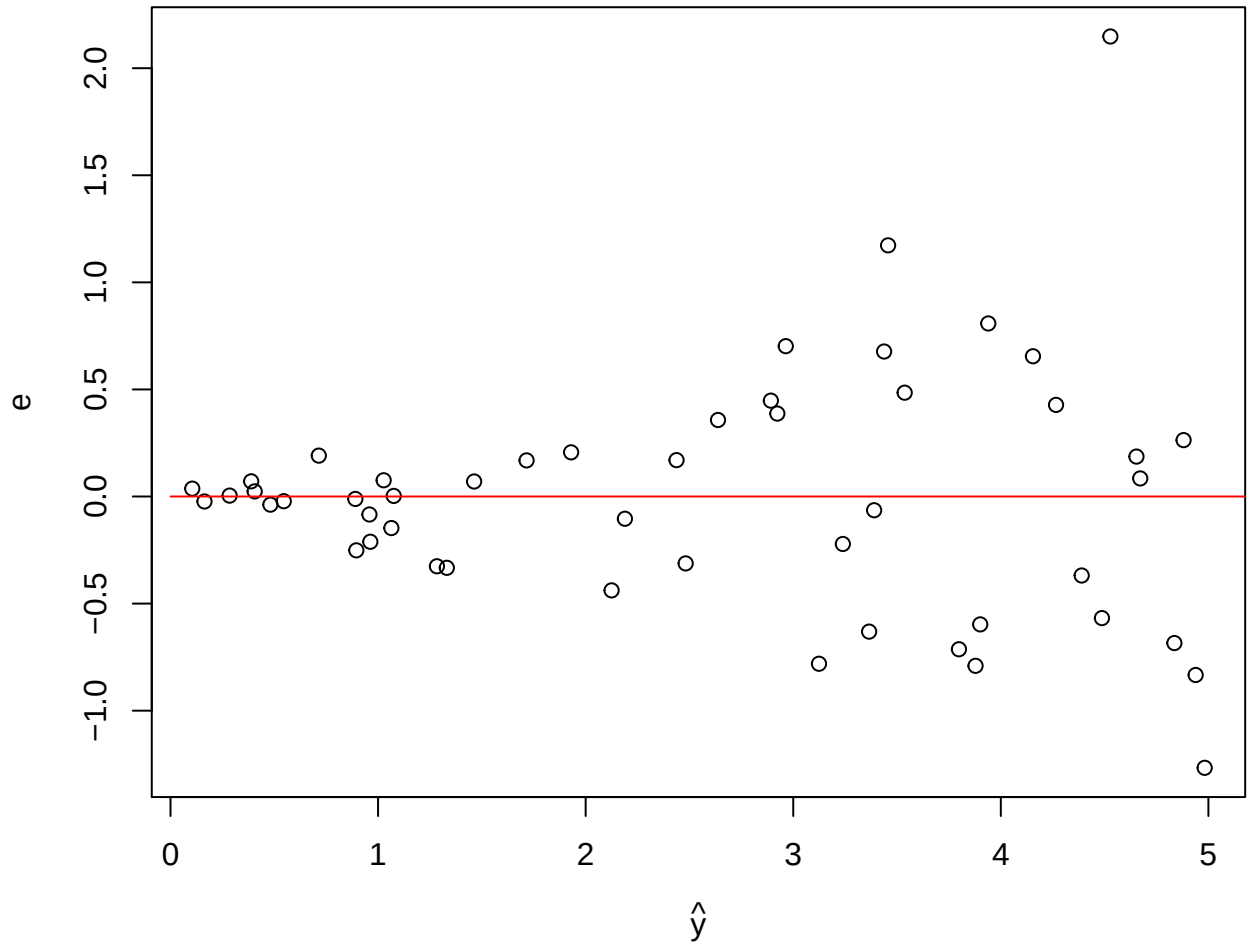
What should one do, if one detects high degree multicollinearity?

- One can remove some of the explanatory variables from the model.
- One can consider suitable variable transformations.

Heteroscedasticity

If the variances of the residuals ϵ_i are not the same for all the values of the explanatory variables, then the residuals are called heteroscedastic. The least squares estimators are unbiased also under heteroscedastic residuals, but heteroscedasticity makes the model unstable and the residual variance estimator may be biased. Heteroscedasticity can often be detected from the residual plot.

Heteroscedasticity, residual plot



Simple homoscedasticity test

- Calculate the coefficient of determination R_a^2 of the test model
$$e_i^2 = \alpha_0 + \alpha_1 \hat{y}_i + \delta_i.$$
- The null hypothesis H_0: The residuals ϵ_i are homoscedastic.
- Test statistic: nR_a^2.
- If the value of the test statistic differs significantly from zero, the null hypothesis is rejected.

White homoscedasticity test

White homoscedasticity test is almost like the simple homoscedasticity test but in addition to the explanatory variables in the simple test model, the White test model includes the squared explanatory variables and the products of the explanatory variables. The test statistic in White homoscedasticity test is the sample size times the coefficient of determination R_a^2 of the White test model. If the value of the test statistic nR_a^2 differs significantly from zero, the null hypothesis is rejected.

Remark

The p-value of the test statistic can be estimated by applying permutations. Moreover, if the residuals are normally distributed, then the test statistic follows, under the null, $\chi^2(p)$ distribution, where the degrees of freedom p is equal to the number of explanatory variables in the test model.

White homoscedasticity test, Example

Let the original regression model be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i.$$

The White test model is then

$$e_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i1}^2 + \gamma_4 x_{i2}^2 + \gamma_5 x_{i1} x_{i2} + \delta_i.$$

Heteroscedasticity

What should one do, if the residuals are heteroscedastic?

- One can consider variable transformations.
- One can consider methods (for example generalized least squares estimators) that are suitable for heteroscedastic residuals.

Testing distributional assumptions

If it is assumed that the variables follow some specific distribution (for example normal distribution), this assumption should be tested.

- One can visualize the data using a histogram.
- One can examine quantile-quantile plots, where the empirical quantiles are plotted against theoretical quantiles from the assumed distribution.
- One can apply the chi-square goodness of fit test.
- If normal distribution is assumed, one can apply for example the Shapiro-Wilk or Kolmogorov-Smirnov test.

Testing prediction ability

Assume that we observe the sample $(x_{i1}, ..., x_{ik}, y_i)$, $i = 1, ..., n + h$. For example, first part of the observations was measured in Finland, and the second part was measured in Uganda.

- Estimate the linear model from the observations $i = 1, ..., n$ $\longrightarrow$ least squares estimate $\boldsymbol{b}$, variance estimate s^2.
- Apply $\boldsymbol{b}$ to predict the values $y_{n+1}, ..., y_{n+h}$:

$$\hat{y}_i = \boldsymbol{x}_{*i}^\top b, \quad i = n + 1, ..., n + h,$$

  where $\boldsymbol{x}_{*i} = (1, x_{i1}, ..., x_{ik})^\top$.
- Calculate the prediction errors $u_i = y_i - \hat{y}_i$.

Testing prediction ability

The null hypothesis H_0: $\beta_1 = \beta_2$, $\sigma_1^2 = \sigma_2^2$.

- The parameter β_1 is related to the observations $1, ..., n$.
- The parameter β_2 is related to the observations $n+1, ..., n+h$.

The test statistic is

$$\chi^2 = \sum_{i=n+1}^{n+h} \frac{u_i^2}{s^2}.$$

Large values of the test statistic yield rejection of the null hypothesis. (This is similar to testing for instability of the model parameters.)

- The p-value can be estimated using permutations. If the residuals are normally distributed, the the test statistic follows, under the null, the $\chi^2(h)$ distribution.

Content

1 Regression diagnostics

2 Model selection

Model selection

In linear regression analysis, it is crucial to select good explanatory variable. Too many explanatory variables lead to inefficient model and unnecessary large variances for the regression parameter estimators. Missing explanatory variables lead to small coefficient of determination (and, in a sense, biased estimators).

Missing explanatory variables

Assume that the correct regression model is (1): $\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \epsilon$, but assume that we calculate the estimate vector $\boldsymbol{b}_1$ from the model (2): $\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \epsilon$. Then

- Part of the explanatory variables are missing, and the residual is of the form $\delta = \boldsymbol{X}_2\beta_2 + \epsilon$.
- The estimator $\boldsymbol{b}_1$ is

$$\boldsymbol{b}_1 = \beta_1 + \left(\boldsymbol{X}_1^\top \boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2\beta_2 + \left(\boldsymbol{X}_1^\top \boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1^\top \epsilon.$$

  This is unbiased if and only if $\beta_2 = 0$ or $\boldsymbol{X}_1^\top \boldsymbol{X}_2\beta_2 = 0$.

Note that the above mentioned estimator is biased only in the larger model (1). In the smaller model (2), it is an unbiased estimator.

Selecting explanatory variables step by step

In multiple regression models, the aim is to select variables such that the coefficient of determination is as high as possible and that the explanatory variables are significant and as independent of each other as possible. Significance of the parameters can be assessed by testing the null hypothesis H_0: $\beta_j = 0$. VIF (or some other measure of dependence) can be used in selecting variables that are not multicollinear. Variables can be added and removed one by one and the changes in significance, in VIF and in coefficient of determination can be tracked. Note that every time a variable is added or removed, the significance and VIF of the other variables may also change and everything has to be calculated again. Thus, the order in which the variables are tested and removed or added has an effect on the outcome.

Model selection criteria

- For a good model, the residual variance is small (and the coefficient of determination is large).
  - Error sum of squares SSE decreases (or at least does not increase) as explanatory variables are added.
  - Simply minimizing SSE (or maximizing R^2) leads to selecting all the possible explanatory variables to the model.
- Model selection criteria are usually based on minimizing SSE, but there is a penalty function that adds a penalty that is based on the number of the explanatory variables.
  - If an explanatory variable is added, the penalty term increases the value of a model selection criteria function unless the decrease in the error sum of squares is large enough.
- Principle of parsimony: If two models perform equally well, the simpler one is better.

Model selection criteria

Let $\boldsymbol{y} = \boldsymbol{X}\beta_p + \epsilon$ be a linear regression model.

1. The number of the estimated parameters is $p = k + 1$.
2. The least squares estimator $\boldsymbol{b}_p = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.
3. Error sum of squares

$$SSE_p = \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_p\right)^\top \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_p\right)$$

4. The residual variance $\hat{\sigma}_p^2 = SSE_p/(n-p)$.

Model selection criteria are often based on minimizing/maximizing a function that is of the form

$$C(p, \hat{\sigma}_p^2).$$

Model selection criteria

- Modified coefficient of determination is given as

$$\bar{R}_p^2 = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SST}, \quad SST = (n-1)s_y^2.$$

  - For a good model, the modified coefficient of determination is as large as possible.
- If the residuals are normally distributed, the model selection can be based on minimizing the *Akaike information criterion* (AIC)

$$C(p, \hat{\sigma}_p^2) = \log(\hat{\sigma}_p^2) + \frac{2p}{n},$$

where $\hat{\sigma}_p^2$ is the estimated residual variance.

Linearization

- If the response variable y depends on the explanatory variables in some non-linear way, then one usually has to build a non-linear regression model.
- However, sometimes one can obtain a linear model by applying suitable linearizing transformations.
  - We here consider linearizing transformations only in the case when we have one response variable and one explanatory variable.

Linearization

- Nonlinear dependence can be linearized if there exist bijective transformations f and g such that for $f(x_i),\ g(y_i)$, $i = 1, ..., n$ it holds that

$$f(y_i) = \beta_0 + \beta_1 g(x_i) + \epsilon_i, \quad i = 1, ..., n$$

  where the residuals ϵ_i satisfy the standard linear model assumptions.
  - Now standard linear model approach can be applied to this linearized model.
- In searching for suitable transformations f and g one can consider:
  - context knowledge (physics, economics,...),
  - graphics (scatter plots), well-known transformations

Linearizing transformations

- Nonlinear dependence of the variables y and x can usually be detected from the scatter plot of (x_i, y_i), $i = 1, ..., n$. The scatter plot can give hints on what type of functions could be applied in linearizing transformations.
- If the transformations f and g are successful in linearizing the dependence between y and x, then nonlinear dependencies are not present in the scatter plots and residual plots:

$$\begin{aligned}
&\big(g(x_i), f(y_i)\big), \quad i = 1, ..., n \\
&\big(f(y_i), e_i\big), \quad i = 1, ..., n \\
&\big(g(x_i), e_i\big), \quad i = 1, ..., n.
\end{aligned}$$

Linearizing transformations

$f(y)$	$g(x)$		
	x	$1/x$	$\log x$
y	$y = \beta_0 + \beta_1 x$	$y = \beta_0 + \beta_1/x$	$y = \beta_0 + \beta_1 \log x$
$1/y$	$1/y = \beta_0 + \beta_1 x$	$1/y = \beta_0 + \beta_1/x$	$1/y = \beta_0 + \beta_1 \log x$
$\log y$	$\log y = \beta_0 + \beta_1 x$	$\log y = \beta_0 + \beta_1/x$	$\log y = \beta_0 + \beta_1 \log x$

$f(y)$	$g(x)$		
	x	$1/x$	$\log x$
y	$y = \beta_0 + \beta_1 x$	$y = \beta_0 + \beta_1/x$	$y = \beta_0 + \beta_1 \log(x)$
$1/y$	$y = \frac{1}{\beta_1\left(x + \frac{\beta_0}{\beta_1}\right)}$	$y = \frac{1}{\beta_0} - \frac{\beta_1}{\beta_0^2} + \frac{1}{x + \frac{\beta_1}{\beta_0}}$	$y = \frac{1}{\beta_1\left(\log x + \frac{\beta_0}{\beta_1}\right)}$
$\log y$	$y = e^{\beta_0} e^{\beta_1 x}$	$y = e^{\beta_0} e^{\beta_1/x}$	$y = e^{\beta_0} x^{\beta_1}$

References:

1. J. S. Milton, J.S., Arnold, J.C. (1995): Introduction to Probability and Statistics, McGraw-Hill Inc
2. Hogg, R.V., McKean, J.W., Craig, A.T. (2005): Introduction to Mathematical Statistics, Pearson Education
3. Davison, A.C., Hinkley, D.V. (2009): Bootstrap Methods and their Applications, Cambridge University Press
4. Belsley, D.A., Kuh, E., Welsch, R.E. (2005): Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley
5. Harrell, F. E. Jr. (2015): Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, Springer

Next week:

1. Stationary stochastic processes
    1. Definition
    2. Autocorrelation function
    3. Partial autocorrelation function
    4. Lag and difference operators
    5. Difference stationarity
2. ARMA models
    1. Pure random process
    2. Different SARMA models
    3. Spectrum