Matematické Metody v Ekonometrii 5.

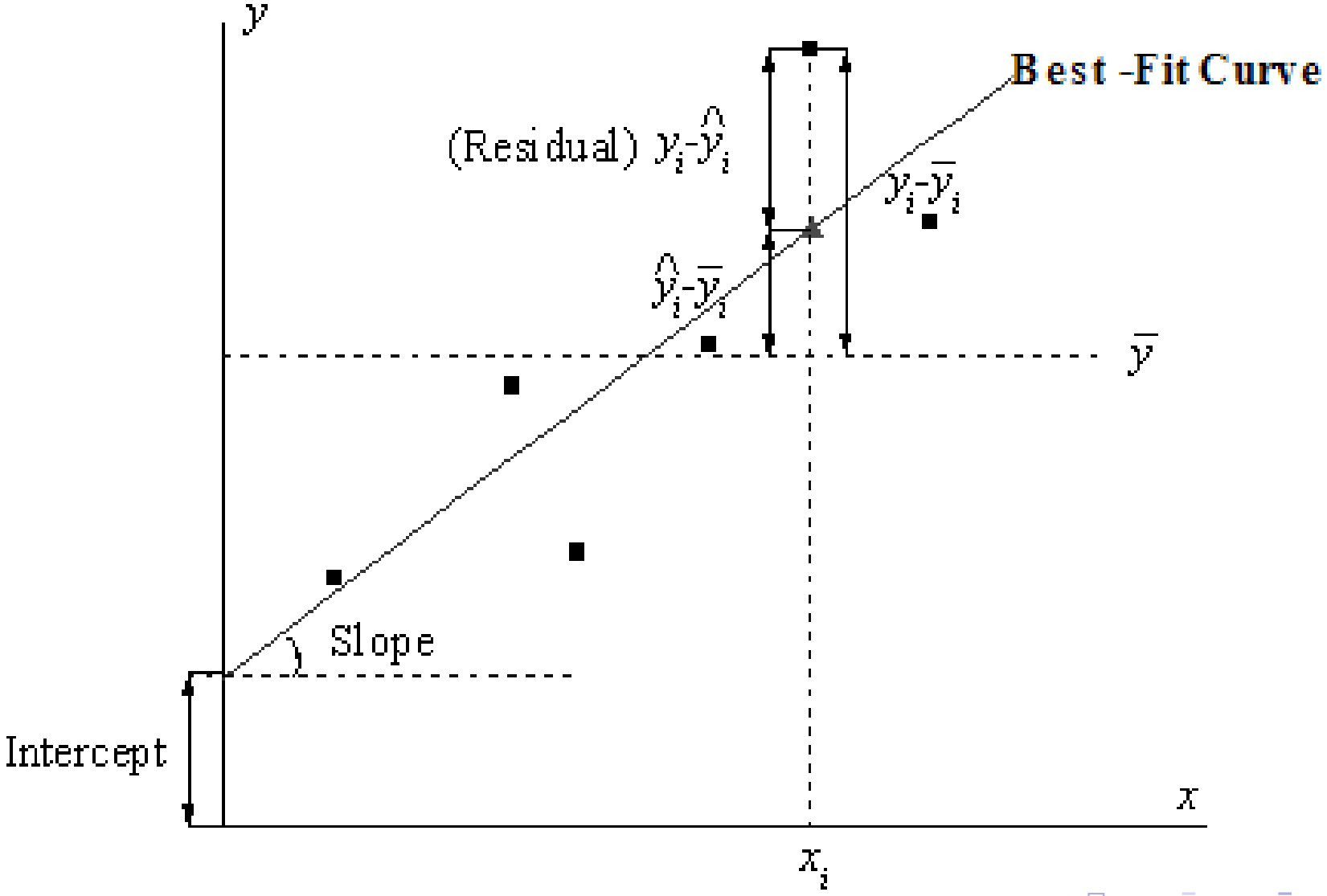
Multiple Regression Analysis - goodness of fit

Blanka Šedivá

KMA

zimní semestr 2016/2017

Illustration of the least square method for simple linear regression line



Goodness of fit and coefficient of determination for simple linear regression model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

- $(y_i - \overline{y}) = (\widehat{y_i} - \overline{y}) + (y_i - \widehat{y_i})$
- $\sum_i (y_i - \overline{y})^2 = \sum_i (\widehat{y_i} - \overline{y})^2 + \sum_i (y_i - \widehat{y_i})^2$
  (hold only for model with absolute part)

  Total sum of squares $TSS = S_y = \sum_i (y_i - \overline{y})^2$
  Explained sum of squares $ESS = S_{\widehat{y}} = \sum_i (\widehat{y_i} - \overline{y})^2$
  Residual sum of squares $RSS = \sum_i (y_i - \widehat{y_i})^2$
- $TSS = ESS + RSS$
- coefficient of determination

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{RSS}{\sum_i (y_i - \overline{y})^2}$$

- R^2 measure the proportion or percentage of total variation in Y explained by the regression model
- for sample line linear regression $\sqrt{R^2} = \widehat{\mathrm{cor}}(X, Y)$

Adjusted coefficient of determination

Mean Total sum of squares

$$MTSS = TSS/(n-1) = s_y = \sum_i (y_i - \overline{y})^2 / (n-1)$$

Mean Explained sum of squares

$$MESS = ESS/(p-1) = s_{\widehat{y}} = \sum_i (\widehat{y}_i - \overline{y})^2 / (p-1)$$

Mean Residual sum of squares

$$MRSS = RSS/(n-p) = s_R = \sum_i (y_i - \widehat{y}_i)^2 / (n-p)$$

- $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{(n-p)s_R^2}{(n-1)s_y^2}$
- $R^2_{adj} = 1 - \frac{s_R^2}{s_y^2}$
- $R^2_{adj} = 1 - \left(1 - R^2\right) \frac{n-1}{n-p}$

Regression through the Origin (RTO)

- $Y_i = \beta x_i + \varepsilon_i$
- the RTO residuals will usually have a nonzero mean, because forcing the regression line through the origin is generally inconsistent with the best fit
- equation $\sum_i (y_i - \overline{y})^2 = \sum_i (\widehat{y_i} - \overline{y})^2 + \sum_i (y_i - \widehat{y_i})^2$ is not valid
- if the RTO model provides a sufficiently poor fit, the data may exhibit more variation around the regression line than around y, in which case $\sum_i (y_i - \widehat{y_i})^2 > \sum_i (y_i - \overline{y})^2$
- $(y_i - 0) = (\widehat{y_i} - 0) + (y_i - \widehat{y_i})$
- $\sum_i (y_i)^2 = \sum_i (\widehat{y_i})^2 + \sum_i (y_i - \widehat{y_i})^2$
- $R^2_{RTO} = \frac{\sum_i \widehat{y_i}^2}{\sum_i y_i^2}$

Wald statistics for Testing Multiple Hypothesis

- Normal linear regression model $\boldsymbol{Y} \sim N\left(\boldsymbol{X}\beta, s^2\boldsymbol{I}\right)$
- Let $\boldsymbol{R}$ be a $q \times p$ matrix with $q \leq p$ and restriction of vektor can by describe be equation $\boldsymbol{R}\beta = \boldsymbol{r}$
- $W = (\boldsymbol{Rb} - \boldsymbol{r})^T \left(\boldsymbol{R}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{R}^T\right)^{-1}(\boldsymbol{Rb} - \boldsymbol{r})/s^2 \overset{assymp.}{\sim} \chi^2_q$
- If $s^2 = RSS/(n-p)$ than $W/q \sim F_{q,n-p}$
- We can test multiple hypotheses test or a joint hypotheses test $H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$ or $H_0 : \beta_2 = \beta_4, \beta_1 = \beta_5$ or $H_0 : \beta_2 + \beta_3 = 1$ or

Testing the overall significance of regression

- $x_1, x_2, \ldots, x_n$ do not help to explain y
- $H_0 : \beta_1 = 0; \beta_2 = 0; \ldots; \beta_k = 0$, all slope coefficients are simultaneously zero
- $H_1 = nonH_0$ exist $\beta_j \neq 0$
- for normally distributed disturbances the variable $F = \frac{ESS/(p-1)}{RSS/(n-p)}$ is distributed as F distribution with $\nu_1 = p - 1$ a $\nu_2 = n - p$
- $F = \frac{R^2}{1-R^2} \frac{n-p}{p-1}$
- if the F value exceeds the critical F value from F table at the α percent level of significance, we reject H_0
- otherwise we do not reject it
- joint testing of hypotheses

Testing model with restriction

- RSS_r sum of square residuals from the restricted model
- RSS_{ur} sum of square residuals from the unrestricted model
- $$F = \frac{(RSS_r - RSS_{ur})/q}{RSS_{ur}/(n-p)}$$
- using the fact $RSS = TSS(1 - R^2)$ we can obtain R square form of statistics F
$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-p)}$$

Testing linear equality restrictions - example

- $H_0 : \beta_1 + \beta_2 = 1$
- One restriction$q = 1$
- Testing statistic

$$F = \frac{b_1 + b_2 - (\beta_1 + \beta_2)}{\sqrt{\mathrm{var}\, b_1 + \mathrm{var}\, b_2 + 2\mathrm{cov}\,(b_1, b_2)}}$$

- If the F value exceeds the critical F value with degrees of freedom with $\nu_1 = q$ a $\nu_2 = n - p$ from F table at the α percent level of significance, we reject H_0 otherwise we do not reject it.
- For this restriction

$$F_{\nu_1=1,\nu_2=n-2} = t_{\nu=n-2}$$

Example I

- $\boldsymbol{y} = (1, 2, 3, 4)^T$ a vysvětlující proměnné $\boldsymbol{x} = (-2, -1, 1, 2)^T$ maticově $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- $\boldsymbol{X} = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$ $\qquad \boldsymbol{Y} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$
- $\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_1 \end{bmatrix}$
- $\boldsymbol{b} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$
  - $\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 4 & 0 \\ 0 & 10 \end{bmatrix}$ $\qquad \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/10 \end{bmatrix}$
  - $\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} 10 \\ 7 \end{bmatrix}$

Example II

- $\boldsymbol{b} = \begin{bmatrix} 5/2 \\ 7/10 \end{bmatrix}$
- $\widehat{\boldsymbol{y}} = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 5/2 \\ 7/10 \end{bmatrix} = \begin{bmatrix} 11/10 \\ 18/10 \\ 32/10 \\ 39/10 \end{bmatrix}$
- residua $\boldsymbol{e} = \begin{bmatrix} 1/10 \\ -2/10 \\ 2/10 \\ -1/10 \end{bmatrix}$
- $RSE = \boldsymbol{e}^T \boldsymbol{e} = 10/100 = 1/10$
- $s_R^2 = RSS/2 = 1/20 = 0.05$
- $\boldsymbol{s}_b = \left(\sqrt{\frac{1}{20}} \sqrt{\frac{1}{4}}, \sqrt{\frac{1}{20}} \sqrt{\frac{1}{10}} \right) = (0.1118, 0.0707)$

Example III

- confidence interval $\boldsymbol{\beta}$
  - for $\alpha = 5\%$ and df $= 2$ is according tables $t_{0.975} = 4.3027$
  - $\mathcal{P}\left(2.0189 < \beta_1 < 2.9811\right) = 0.95$
  - $\mathcal{P}\left(0.3958 < \beta_2 < 1.0042\right) = 0.95$
- $H_0 : \beta_1 = 0$
  statistics $t = \frac{2.5}{\sqrt{1/80}} = 22.3607$
  the critical region for $\alpha = 5\%$ is $W = (-\infty, -4.3027) \cup (4.3027, \infty)$
  $t \in W \Rightarrow$ reject H_0
- $H_0 : \beta_2 = 0$
  $t = \frac{0.7}{\sqrt{1/200}} = 9.8995$
  the critical region for $\alpha = 5\%$ is
  $W = (-\infty, -4.3027) \cup (4.3027, \infty)$ $t \in W \Rightarrow$ reject H_0
- total sum of squares
  $TSS = (1 - 2.5)^2 + (2 - 2.5)^2 + (3 - 2.5)^2 + (4 - 2.5)^2 = 5$

Example IV

- $R^2 = 1 - \frac{0.1}{5} = 0.98$
- $H_0 : \beta_2 = 0$

  $F = \frac{0.98}{0.02} \frac{2}{1} = 98$

  the critical region for $\alpha = 5\%$ is $W = (18.513, \infty)$
  $t \in W \Rightarrow$ reject H_0