Course 02403 Introduction to Mathematical Statistics

Lecture 8: Linear regression

DTU Compute
Technical University of Denmark
2800 Lyngby – Denmark

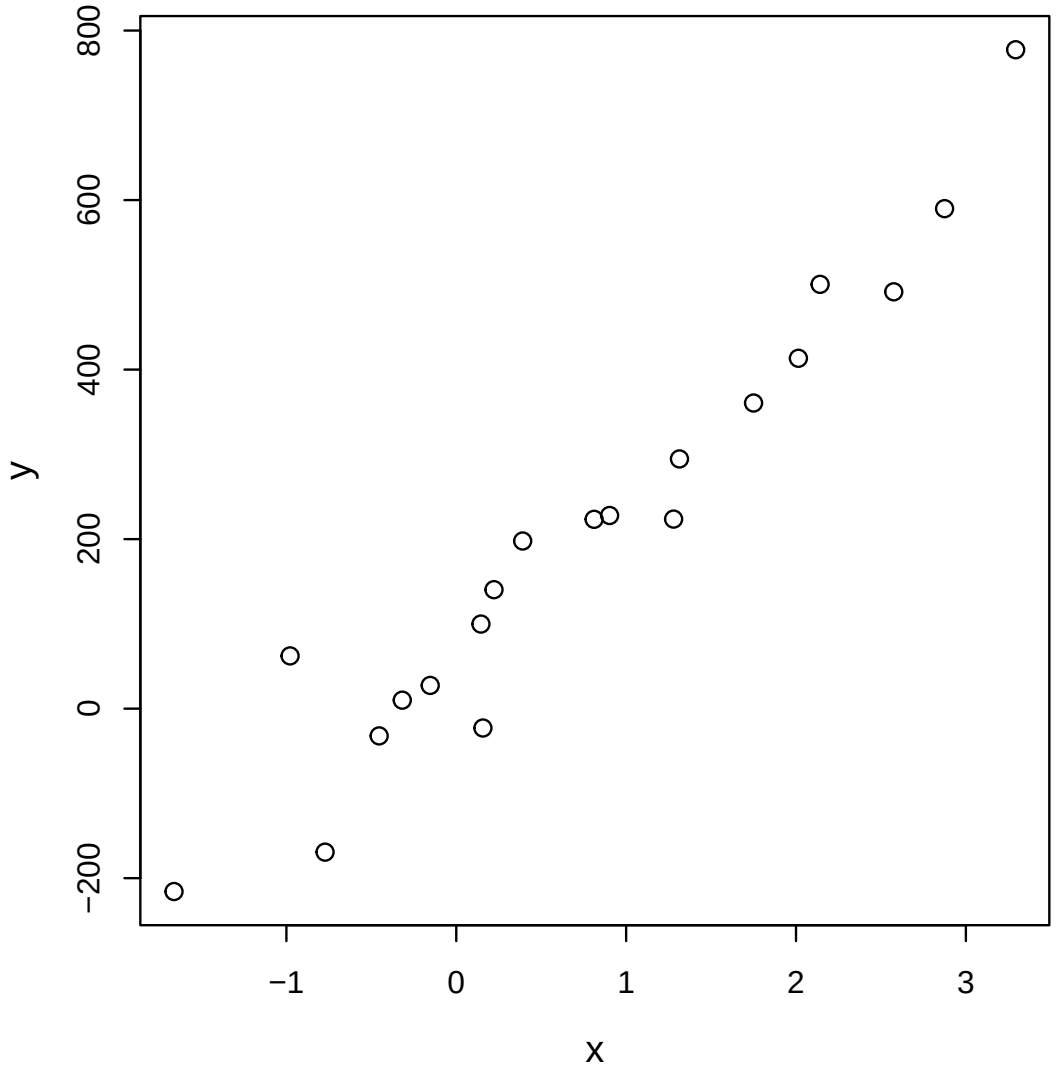
Overview

1. Linear regression models
   - Least squares method
2. Statistics and linear regression
   - Hypothesis tests and confidence intervals for β_0 and β_1
   - Confidence and prediction intervals
   - Correlation
3. Multiple linear regression
4. Model selection
5. Model validation - Analysis of residuals

Overview

A scatterplot

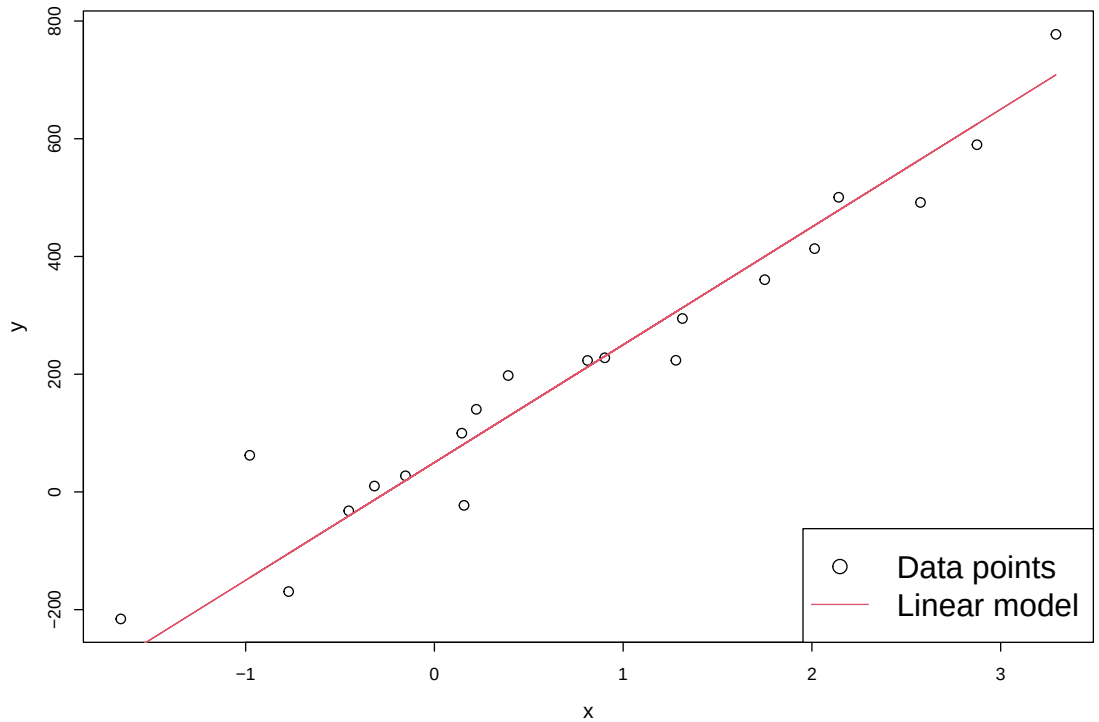
- We have n pairs of data points (x_i, y_i).



A linear model

If the data points lie on a straight line, the relationship between x and y values can be described by the equation:

$$y_i = \beta_0 + \beta_1 x_i.$$



- We need a description of the *random variation*.

The simple linear regression model

- The *linear regression model*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \ldots, n).$$

or

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

- Y_i is the *dependent variable*.
- x_i is the *explanatory variable*.
- ε_i is the deviation (residual).
- We assume $\varepsilon_i \sim N(0, \sigma^2)$ (and i.i.d.).

Consider: *What kind of distribution does Y_i follow? Are the Y_is identically distributed?*

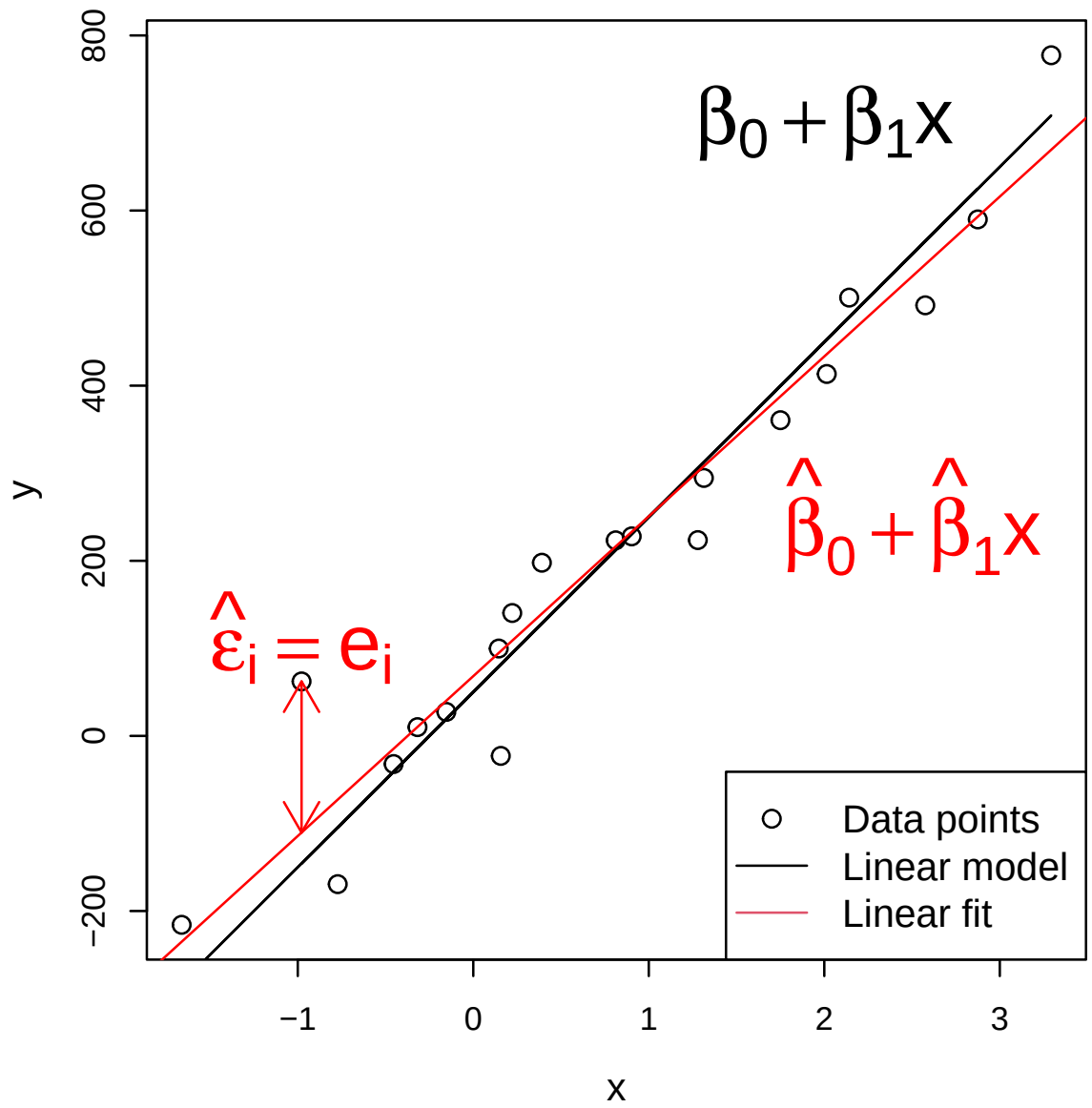
Least squares method

- We want to estimate the parameters β_0 and β_1.
- Good idea: Let's minimize the variance of the residuals (σ^2).
- We minimize the sum of the squared residuals (Residual Sum of Squares, RSS):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

That is, we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that they minimize RSS.

Illustration of model, data, and fit



'Least squares' estimators

Theorem 5.4 (for estimators)

'Least squares' estimators for β_0 and β_1 are given by:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}},$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

or

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$

'Least squares' estimators – how?

$$\begin{aligned} RSS(\boldsymbol{\beta}) &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

and

$$\nabla RSS(\boldsymbol{\beta}) = -2\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

Example: Skive fjord

Formulate a model of $\log(chla)$ as a function of water temperature.

Overview

Variation in parameter estimates

There is variation in the parameters!

A new sample leads to new realizations of the estimators, i.e., new estimates.

What are the distributions of the parameter estimators?

We need to know them to create confidence intervals, etc.

Standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed with variances:

Theorem 5.8 (first part)

$$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}},$$

$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}},$$

$$Cov[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}\sigma^2}{S_{xx}}.$$

or

$$V[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Variance estimator

Theorem 5.8 (second part)

Since σ^2 is unknown, we use the *central estimate for* σ^2:

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

Why divide by $n-2$?

- $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$
- $Tr(\boldsymbol{I}-\boldsymbol{H}) = n-2 = n-Rank(\boldsymbol{X})$
- $\frac{1}{\sigma^2}\sum_{i=1}^n e_i^2 = \frac{1}{\sigma^2}\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} \sim \chi^2(n-2)$

Standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$

Thus, we estimate the variance (standard deviation) for the error and thereby also the variances (standard deviations) of the estimators. We denote these $\hat{\sigma}^2_{\beta_0}$ and $\hat{\sigma}^2_{\beta_1}$.
We obtain the following estimates of the standard deviations (standard errors) for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}, \quad \hat{\sigma}_{\beta_1} = \hat{\sigma}\sqrt{\frac{1}{S_{xx}}}.$$

or

$$[\hat{\sigma}_{\beta_0}, \hat{\sigma}_{\beta_1}] = \hat{\sigma}\sqrt{\mathrm{diag}((\boldsymbol{X}^T\boldsymbol{X})^{-1})}$$

Hypothesis tests for β_0 and β_1

We can conduct hypothesis tests for the parameters in a linear regression model:

$$\begin{aligned} H_{0,i}: &\quad \beta_i = \beta_{0,i}, \\ H_{1,i}: &\quad \beta_i \neq \beta_{0,i}. \end{aligned}$$

Theorem 5.12

Under the null hypotheses ($\beta_0 = \beta_{0,0}$ and $\beta_1 = \beta_{0,1}$), the test statistics are

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}, \quad T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}},$$

t-distributed with $n-2$ degrees of freedom.

Confidence intervals for β_0 and β_1

Method 5.15

$(1-\alpha)$ confidence intervals for β_0 and β_1 are given by:

$$\hat{\beta}_0 \pm t_{1-\alpha/2}\,\hat{\sigma}_{\beta_0},$$
$$\hat{\beta}_1 \pm t_{1-\alpha/2}\,\hat{\sigma}_{\beta_1},$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the t-distribution with $n-2$ degrees of freedom.

- In Python, $\hat{\sigma}_{\beta_0}$ and $\hat{\sigma}_{\beta_1}$ can be found under "std err".

Example: Skive Fjord

Go to Python and make a table that include parameters estimates, standard errors, the t-test, and confidence interval.

Method 5.18: Confidence interval for the regression line

A simple linear regression model can be written as

$$Y_i \sim N(\mu(x_i), \sigma^2),$$

where $\mu(x_i) = \beta_0 + \beta_1 x_i$.

For a new observation x_{new}, we can find a confidence interval for $\mu(x_{new}) = \beta_0 + \beta_1 x_{new}$.

The $(1-\alpha)$ confidence interval for the regression line at $x = x_{new}$ (for $\mu(x_{new})$) can be found by:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) \pm t_{\alpha/2}(n-2) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}.$$

Method 5.18: Prediction interval for a new observation

- We want a prediction interval for a new observation Y_{new} at $x = x_{new}$.
- The $(1-\alpha)$ prediction interval for a new observation Y_{new} at $x = x_{new}$ can be found by:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}.$$

- The prediction interval will contain the observed y_{new} in $100(1-\alpha)\%$ of cases.
- For fixed α, the prediction interval is larger than the confidence interval.

Confidence and prediction interval for a new observation

Set $\boldsymbol{x}_{new} = [1, \quad x_{new}]^T$, then the confidence interval is

$$\boldsymbol{x}_{new}^T \hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2} \hat{\sigma} \sqrt{\boldsymbol{x}_{new}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_{new}}$$

and the prediction interval is

$$\boldsymbol{x}_{new}^T \hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2} \hat{\sigma} \sqrt{1 + \boldsymbol{x}_{new}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_{new}}.$$

Example: Skive Fjord

Go to Python and add confidence and prediction lines to the plot.

Explained variance and correlation

- The explained variance in a model is R^2 (R-squared).
- Calculated with

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

  where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- The proportion of the total variance explained by the model.

Explained variance and correlation

- The correlation ρ is a measure of *linear relationship* between two stochastic variables.
- The estimated (i.e., empirical) correlation satisfies

$$\hat{\rho} = \sqrt{R^2}\, sign(\hat{\beta}_1),$$

where $sign(\hat{\beta}_1)$ is -1 for $\hat{\beta}_1 \leq 0$ and 1 for $\hat{\beta}_1 > 0$

Explained variance and correlation

- The correlation ρ is a measure of *linear relationship* between two stochastic variables.
- The estimated (i.e., empirical) correlation satisfies

$$\hat{\rho} = \sqrt{R^2}\, sign(\hat{\beta}_1),$$

  where $sign(\hat{\beta}_1)$ is -1 for $\hat{\beta}_1 \leq 0$ and 1 for $\hat{\beta}_1 > 0$
- Thus:
  - Positive correlation with positive slope.
  - Negative correlation with negative slope.

Test for significant correlation

- Test for significant correlation (linear relationship) between two variables:

$$H_0 : \rho = 0,$$
$$H_1 : \rho \neq 0,$$

is equivalent to

$$H_0 : \beta_1 = 0,$$
$$H_1 : \beta_1 \neq 0,$$

where β_1 is the slope in the simple linear regression model.

Example: Skive Fjord

Go to Python and make the analysis using `smf.ols`.

Overview

Multiple linear regression

We can of course imagine more than one explanatory variable, corresponding to the model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \text{ og i.i.d.}$$

or

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \varepsilon_i \sim N(0, \sigma^2)$$
$$= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

Parameter estimetes

All results from the simple linear regression carry over (with minor adjustment). The estimators of the parameters in the simple multiple regression model are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$$

and the covariance matrix of the estimates is

$$V[\hat{\boldsymbol{\beta}}] = \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}$$

and central estimate for the residual variance is

$$\hat{\sigma}^2 = \frac{RSS}{n - (p+1)}$$

Hypotese test (partial t-test)

The estimate of the parameters in the simple linear regression model are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

and the covariance matrix of the estimates is

$$\hat{\boldsymbol{\Sigma}}_\beta = \hat{\sigma}^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}$$

The obeserved t-statistic for the hypothesis: $H_0 : \beta_i = \beta_{i,0}$ is

$$t_{obs,i} = \frac{\hat{\beta}_i - \beta_{i,0}}{\sqrt{(\hat{\boldsymbol{\Sigma}}_\beta)_{ii}}}.$$

Should be compared with a t-distribution with $n - (p+1)$ degrees of freedom.

Confidence and prediction intervals

$\boldsymbol{x}_{new} = [1, x_{1,new}, \ldots, x_{p,new}]^T$:

Variance of the mean estimator

$$\begin{aligned} V(\hat{Y}_{new}) &= V(\boldsymbol{x}_{new}^T \hat{\boldsymbol{\beta}}) \\ &= \sigma^2 \boldsymbol{x}_{new}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_{new}. \end{aligned}$$

Prediction variance

$$\begin{aligned} V(Y_{new}) &= V(\boldsymbol{x}_{new} \hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}_{new}) \\ &= \sigma^2 (1 + \boldsymbol{x}_{new}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_{new}) \end{aligned}$$

in practice replace σ^2 with its estimate ($\hat{\sigma}^2$), and hence use quantiles of the appropriate t-distribution.

Example: Skive Fjord

Go to Python expand the model with global radiation.

Overview

Model extension (forward selection)

- *Not included in the book*
- Start with a *simple linear regression model* with one significant explanatory variable
- *Extend the model* with other explanatory variables one at a time
- *Stop* when there are no more significant extensions

Model reduction (backward selection)

- *Described in the book under section 6.5*
- Start with the full model
- Remove the "least significant" variable
- Stop when all remaining parameters are significant

Model selection

- There is no certain method to find the best model!
- Selecting a model requires subjective decisions.
- Different procedures, either forward or backward selection (or both), depend on the circumstances.
- Statistical methods and tests exist to compare models.

Overview

Residual analysis

Method 5.28

- Check the normality assumption with a QQ-plot.
- Check for any systematic deviations by plotting the residuals (e_i) as a function of the fitted values $(\hat{y}_i)$.

(Method 5.29)

- Is the independence assumption reasonable?

Agenda

1. Linear regression models
   - Least squares method
2. Statistics and linear regression
   - Hypothesis tests and confidence intervals for β_0 and β_1
   - Confidence and prediction intervals
   - Correlation
3. Multiple linear regression
4. Model selection
5. Model validation - Analysis of residuals