Linear model

MS-C2128 Prediction and Time Series Analysis

Fall term 2020

Week 1: Linear model

1. Definition and assumptions
2. Matrix representation of linear models
3. Least squares estimation
4. Fitted values and residuals
5. Variance decomposition and coefficient of determination
6. Significance testing
7. Confidence intervals
8. Prediction interval
9. Random explanatory variables
10. On optimality of least squares estimators

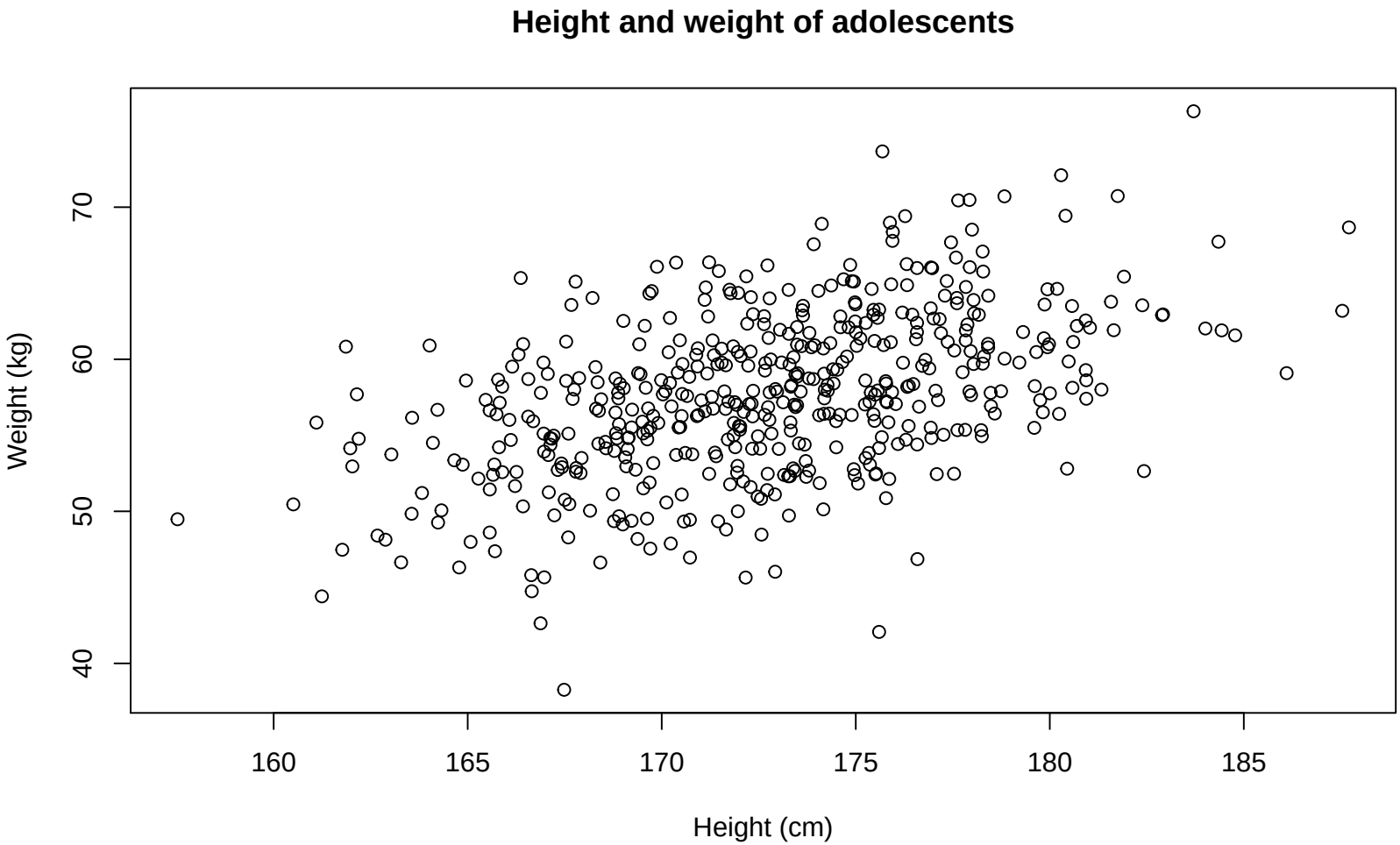
Content

1. Definition and assumptions
2. Matrix representation of linear models
3. Least squares estimation
4. Fitted values and residuals
5. Variance decomposition and coefficient of determination
6. Significance testing
7. Confidence intervals
8. Prediction interval
9. Random explanatory variables
10. On optimality of least squares estimators

Response variable and explanatory variables

- Variation in a *response variable* Y is explained by variation in *explanatory variables* $X_1, ..., X_k$.
  - $y_1, ..., y_n$ are the observed values of the response variable Y.
  - $x_{1j}, ..., x_{nj}$ are the observed values of the explanatory variable X_j.

Response variable and explanatory variable



Definition

- Multiple linear model is given by

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik} + \epsilon_i, \quad i = 1, ..., n,$$

where

- The variables Y_i are random and the variables $X_{i1}, ..., X_{ik}$ are non-random.
- The residual (error term) ϵ_i are random.
- The coefficients (regression parameters) $\beta_0, \beta_1, ..., \beta_k$ are constants.

Remark

For simplicity, we assumed that the variables $X_{i1}, ..., X_{ik}$ are non-random. Then all the randomness of Y_i comes from the residual ϵ_i.

Linear model

Remark

Conventional notation

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \quad i = 1, ..., n,$$

is used even though some of the variables are random and some are not.

Standard assumptions

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \quad i = 1, ..., n,$$

(i) The explanatory variables are non-random.

(ii) The explanatory variables are linearly independent.

(iii) $\mathrm{E}[\epsilon_i] = 0$ for all $i = 1, ..., n$

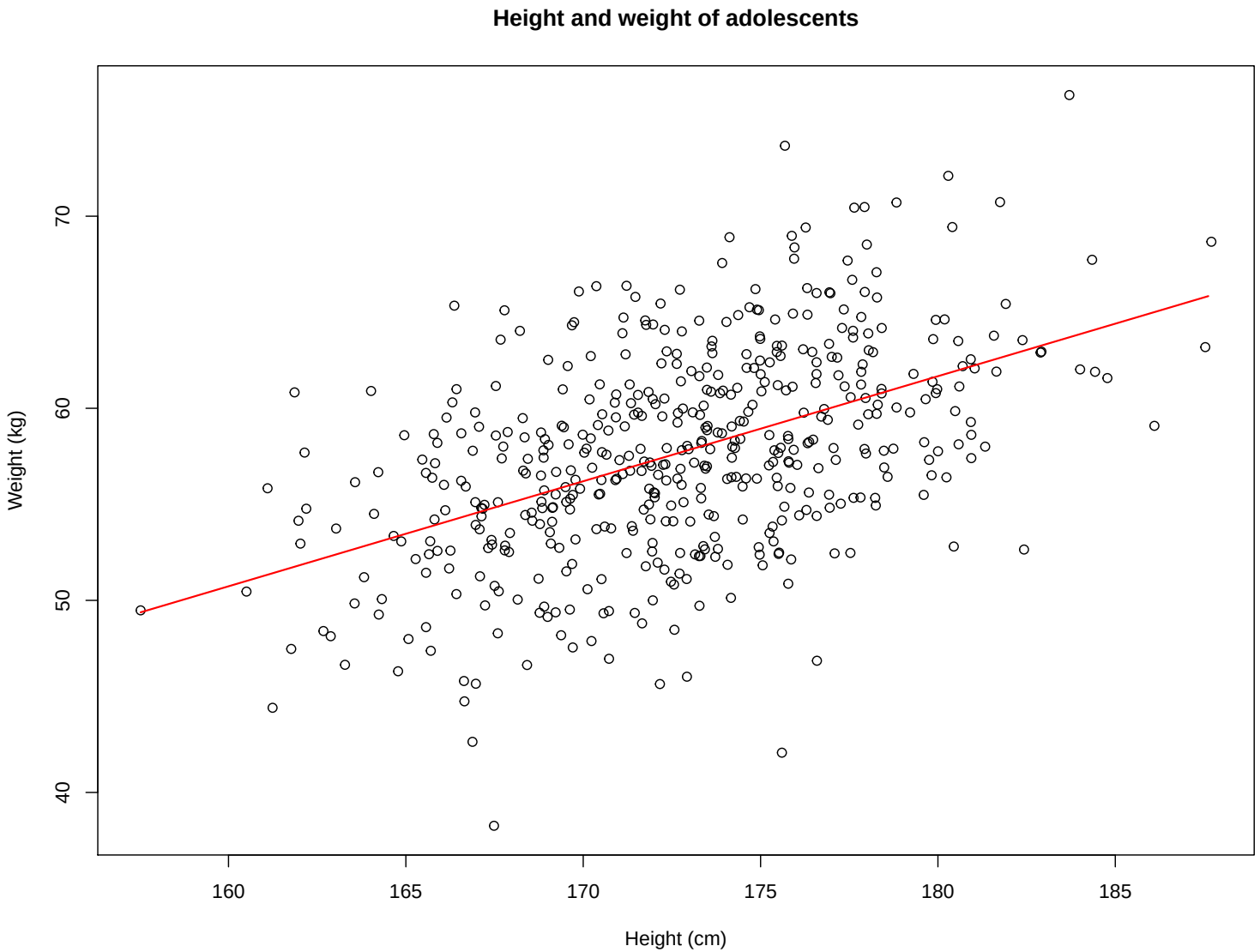
(iv) $\mathrm{var}(\epsilon_i) = \sigma^2$ for all $i = 1, ..., n$

(v) $\mathrm{cor}(\epsilon_i, \epsilon_l) = 0$, for $i \neq l$

Remark

In addition to the standard assumptions, it is customary to assume that the residuals are independent and identically distributed. Moreover, it is often assumed that the residuals are normally distributed, but this is not a necessary assumption.

Response variable and explanatory variable



Standard assumption (i)

(i) The variables X_j are non-random.

- This is a strong assumption.
- However, linear regression analysis works perfectly well also in the case when the variables X_j are assumed to be random. In that case, the notations are a bit more complicated. (One has to consider conditional expected values instead of expected values.) Estimation procedures are exactly the same in the case when the X_j are assumed to be non-random and in the case when the X_j are assumed to be random.

Standard assumptions (ii) and (iii):

(ii) The explanatory variables are linearly independent.

- Multicollinearity makes it difficult to assess the effect of the different explanatory variables separately.
- If X_j can be given as a linear combination of the other explanatory variables, it can be removed from the model.
- Assumption (*ii*) guarantees that the so called least square estimators for the parameters $\beta_0, \beta_1, ..., \beta_k$ can be given in closed form.

(iii) $\mathbf{E}[\epsilon_i] = 0$ for all $i = 1, ..., n$.

- Assumption (*ii*) guarantees that there is no systematic bias. That is,

$$\begin{aligned} \mathrm{E}[y_i] &= \mathrm{E}\left[\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i\right] \\ &= \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \mathrm{E}\left[\epsilon_i\right] \\ &= \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}. \end{aligned}$$

Standard assumption (iv):

(iv) $\mathbf{var}(\epsilon_i) = \sigma^2$ for all $i \neq I$.

- The parameter σ^2 is called the residual variance.
- If Assumption (iv) holds, the residuals ϵ_i are said to be homoscedastic.
- If Assumption (iv) does not hold, the residuals ϵ_i are said to be heteroscedastic.
- Heteroscedasticity makes standard least squares estimators unstable.
  - In the presence of heteroscedasticity one could apply the so called generalized least squares estimators. (This is beyond the scope of this course.)
- Homoscedasticity can be tested. (See Week 2, regression diagnostics.)

Standard assumption (v):

(v) $\mathbf{cor}(\epsilon_i, \epsilon_l) = 0$ for all $i = 1, \ldots, n$.

- The residuals are uncorrelated.
- If Assumption (v) does not hold, the residuals ϵ_i are correlated.
- Correlatedness can make the regression parameter estimators inefficient and even biased.
  - In the presence of correlated residuals, one could apply dynamic regression models.
- Uncorrelatedness can be tested. (See Week 2, regression diagnostics.)

Linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \quad i = 1, ..., n,$$

If Assumptions (i)-(vi) hold, then

(iii)' $\mathrm{E}[y_i] = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}$, $i = 1, ..., n$

(iv)' $\mathrm{var}(y_i) = \sigma^2$, $i = 1, ..., n$

(v)' $\mathrm{cor}(y_i, y_j) = 0$, $i \neq l$

Remark

If the residuals are assumed to be normally distributed, then

$$y_i \sim N\big(\mathrm{E}[y_i], \sigma^2\big), \qquad i = 1, ..., n.$$

Model part and random part

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \quad i = 1, ..., n,$$

- Linear model can be given as a sum of two parts:

$$y_i = \mathrm{E}[y_i] + \epsilon_i, \quad i = 1, ..., n$$

  - The expected value $\mathrm{E}[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$ is called the systematic part or the model part. This part depends on the x_j.
  - The residual ϵ_i forms the random part. This part is independent of x_j.

Regression plane and regression parameters

- The systematic part $\mathrm{E}[y_i]$ defines the regression plane

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

in $\mathbb{R}^{k+1}$.

  - The variance σ^2 measures the variation of $(x_{i1}, \ldots, x_{ik}, y_i) \in \mathbb{R}^{k+1}$ around the regression plane.

- The regression parameters β_j can be interpreted as follows:

  - Assume that the value of the explanatory variable x_j grows by one ($x_j \longrightarrow x_j + 1$) and assume that the values of all the other explanatory variables remain unchanged. The parameter β_j models the change in the expected value of the response variable y as the value of x_j changes by one unit:

$$\mathrm{E}[y] \longrightarrow \mathrm{E}[y] + \beta_j.$$

Content

1. Definition and assumptions
2. Matrix representation of linear models
3. Least squares estimation
4. Fitted values and residuals
5. Variance decomposition and coefficient of determination
6. Significance testing
7. Confidence intervals
8. Prediction interval
9. Random explanatory variables
10. On optimality of least squares estimators

Matrix representation

A linear model can be represented as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\boldsymbol{y} = (y_1, y_2, ..., y_n)^\top = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{\beta} = (\beta_0, ..., \beta_k)^\top = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad \boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^\top = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Standard assumptions: Matrix representation

A linear model can be represented as $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$

(i) The elements of $\boldsymbol{X}$ are non-random.

(ii) The columns of $\boldsymbol{X}$ are linearly independent.

(iii) $\mathrm{E}[\epsilon] = \mathbf{0}$.

(iv)-(v) $\mathrm{cov}(\epsilon) = \sigma^2 \boldsymbol{I}$.

Remark

- Let $\boldsymbol{z} = (z_1, z_2, ..., z_p)^\top$ be a random vector, where the elements $z_1, z_2, ..., z_p$ are random.
- The expected value $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{z}]$ of a vector $\boldsymbol{z}$ is taken componentwise

$$\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{z}] = \big(\mathrm{E}[z_1], \mathrm{E}[z_2], ..., \mathrm{E}[z_p]\big)^\top \in \mathbb{R}^p$$

- The covariance $\Sigma = \mathrm{cov}(\boldsymbol{z})$ of $\boldsymbol{z}$ refers to the matrix

$$\Sigma = \mathrm{cov}(\boldsymbol{z}) = \mathrm{E}\left[(\boldsymbol{z} - \mathrm{E}[\boldsymbol{z}])(\boldsymbol{z} - \mathrm{E}[\boldsymbol{z}])^\top\right]$$

$$= \begin{bmatrix} \mathrm{var}(z_1) & \mathrm{cov}(z_1, z_2) & \mathrm{cov}(z_1, z_3) & \cdots & \mathrm{cov}(z_1, z_p) \\ \mathrm{cov}(z_2, z_1) & \mathrm{var}(z_2) & \mathrm{cov}(z_2, z_3) & \cdots & \mathrm{cov}(z_2, z_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(z_p, z_1) & \mathrm{cov}(z_p, z_2) & \mathrm{cov}(z_p, z_3) & \cdots & \mathrm{var}(z_p) \end{bmatrix} \in \mathbb{R}^{p \times p}$$

Content

1. Definition and assumptions
2. Matrix representation of linear models
3. Least squares estimation
4. Fitted values and residuals
5. Variance decomposition and coefficient of determination
6. Significance testing
7. Confidence intervals
8. Prediction interval
9. Random explanatory variables
10. On optimality of least squares estimators

Least squares estimation

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \quad i = 1, ..., n,$$

The parameters $\beta_0, \beta_1, ..., \beta_k$ are usually estimated using the least squares estimation method:

- The sum of the squared residuals

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_k x_{ik}\right)^2$$

  is minimized with respect to the parameters $\beta_0, \beta_1, ..., \beta_k$:
- The partial derivatives with respect to $\beta_0, \beta_1, ..., \beta_k$ are calculated.
- $k + 1$ equations are obtained by setting the derivatives equal to zero.
- If the explanatory variable matrix $\boldsymbol{X}$ is of full rank, there exist a unique solution.
- The least squares estimators b_j for β_j are obtained.

Estimator vs estimate

- The *random* object b_j is called an estimator.
- If we calculate the numerical value for b_j from an observed sample, we obtain an estimate, that is a non-random realization of the corresponding estimator.

Remark

An estimator and the corresponding estimate are often denoted by the same symbol. The interpretation depends on the context.

Least squares estimator

- Let $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$ and let the matrix $\boldsymbol{X}$ be of full rank. Then the least squares estimator of β is

$$\boldsymbol{b} = (b_0, ..., b_k)^\top = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

- If the standard assumptions (i)-(v) hold, then

$$\mathrm{E}[\boldsymbol{b}] = \beta \quad \text{and} \quad \mathrm{cov}(\boldsymbol{b}) = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}.$$

  - As the expected value $\mathrm{E}[\boldsymbol{b}] = \beta$, the least squares estimator $\boldsymbol{b}$ is an unbiased estimator of β.
- If the residuals are normally distributed, then

$$\boldsymbol{b} \sim N_{k+1}\left(\beta, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\right)$$

Content

1. Definition and assumptions
2. Matrix representation of linear models
3. Least squares estimation
4. Fitted values and residuals
5. Variance decomposition and coefficient of determination
6. Significance testing
7. Confidence intervals
8. Prediction interval
9. Random explanatory variables
10. On optimality of least squares estimators

Fitted values and residuals

- The fitted values of the estimated model are given as follows:

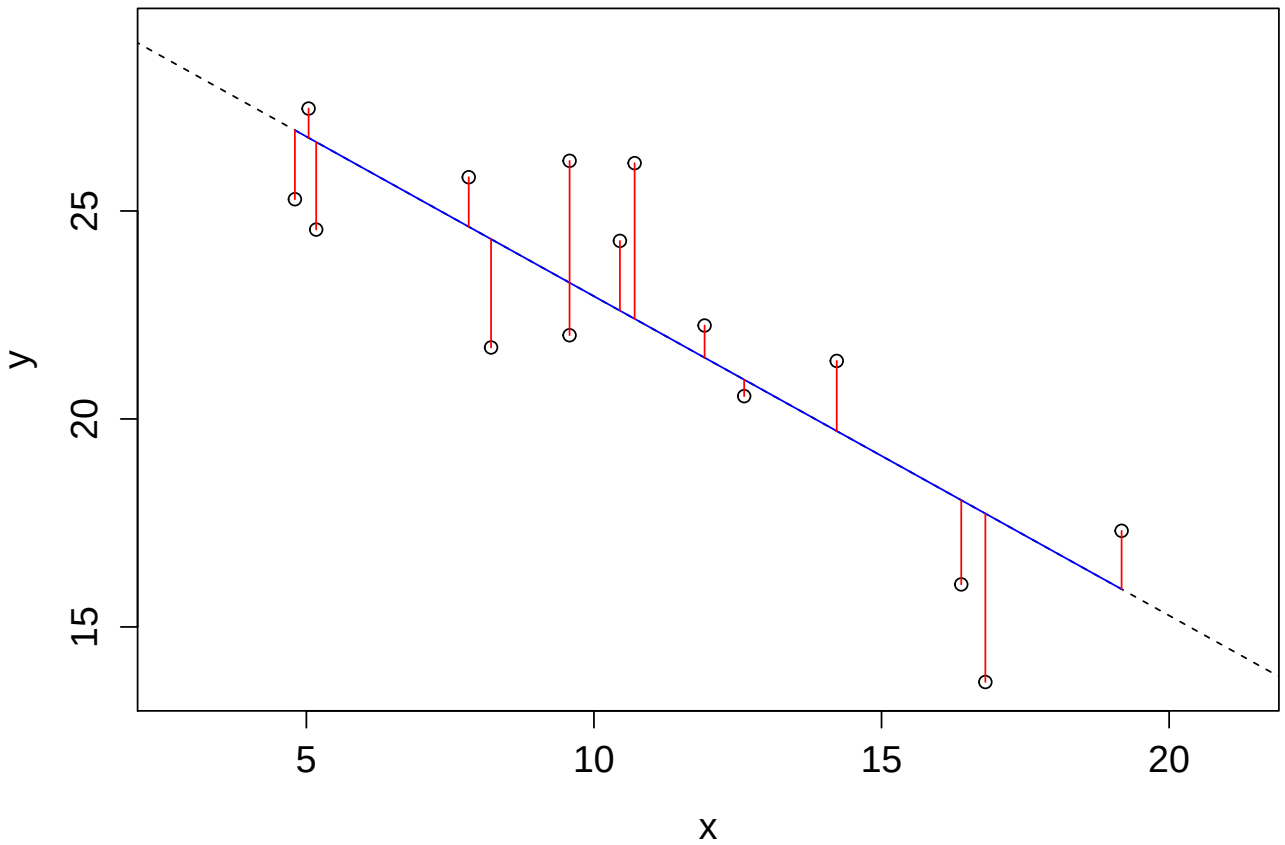
$$\hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_k x_{ik}$$

  - This gives the estimated values of the response variable y in points i.
- The estimated residuals:

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_{i1} - \ldots - b_k x_{ik}$$

  - The residuals are the differences between the observed values of the response variables and the estimated values of the response variables.
- The model explains the variation in y the better the smaller are the estimated residuals e_i.

Fitted values and residuals



Fitted values and residuals

Under the standard assumptions (i)-(v), we have that:

- For the fitted values:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} \\ \mathrm{E}[\hat{\boldsymbol{y}}] &= \boldsymbol{X}\beta \\ \mathrm{cov}(\hat{\boldsymbol{y}}) &= \sigma^2 \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top = \sigma^2 \boldsymbol{P}. \end{aligned}$$

- For the residuals:

$$\begin{aligned} \boldsymbol{e} &= \boldsymbol{y} - \hat{\boldsymbol{y}} = \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right)\boldsymbol{y} = \left(\boldsymbol{I} - \boldsymbol{P}\right)\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y} \\ \mathrm{E}[\boldsymbol{e}] &= \boldsymbol{0} \\ \mathrm{cov}(\boldsymbol{e}) &= \sigma^2\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right) = \sigma^2\left(\boldsymbol{I} - \boldsymbol{P}\right) = \sigma^2 \boldsymbol{M}. \end{aligned}$$

Fitted values and residuals

- The $n \times n$-matrices

$$\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$$
$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$$

  are symmetric ja idempotent:

$$\boldsymbol{P}^\top = \boldsymbol{P} \qquad \boldsymbol{P}^2 = \boldsymbol{P}$$
$$\boldsymbol{M}^\top = \boldsymbol{M} \qquad \boldsymbol{M}^2 = \boldsymbol{M}$$

- Moreover: $\boldsymbol{PM} = \boldsymbol{MP} = \boldsymbol{0}$
- The above mentioned properties of the matrices $\boldsymbol{P}$ and $\boldsymbol{M}$ play a crucial role when distributions related to estimating and testing the linear regression parameters are derived.

Content

1. Definition and assumptions
2. Matrix representation of linear models
3. Least squares estimation
4. Fitted values and residuals
5. Variance decomposition and coefficient of determination
6. Significance testing
7. Confidence intervals
8. Prediction interval
9. Random explanatory variables
10. On optimality of least squares estimators

Residual variance

- If the standard assumptions (i)-(v) hold, then an unbiased estimator for $\text{var}(\epsilon_i) = \sigma^2$ can be given as

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2,$$

where $e_i = y_i - \hat{y}_i$, k is the number of the explanatory variables and where n is the number of the observations.

- This is actually simply the sample variance of the e_i as there are $k+1$ estimated parameters and as

$$\sum_{i=1}^{n} e_i = 0 \implies \bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0 \qquad \text{and}$$

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} (e_i - \bar{e})^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2$$

Variance decomposition

The goal in regression analysis is to explain the variation in the response variable by the variation in the explanatory variables. Success in this can be assessed using the so called variance decomposition.

Variance decomposition: Definition

- The total sum of squares $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  - measures the variation of the observations y_i
  - the sample variance of the y_i, $s_y^2 = SST/(n-1)$
- The error sum of squares $SSE = \sum_{i=1}^{n} e_i^2$
  - measures the variation of the estimated residuals e_i
  - the sample variance of the e_i, $s^2 = SSE/(n-k-1)$
- The model sum of squares $SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
  - measures the part of the variation of the observations y_i that is explained by the estimated regression model.
- The variance decomposition: $SST = SSM + SSE$

Coefficient of determination: Definition

- The variance decomposition $SST = SSM + SSE$ tells about the goodness of the estimated regression model.
  - The larger the proportion of the model sum of squares SSM (that is, the smaller the proportion of the error sum of squares SSE) is of the total sum of squares SST, the better the estimated model explains the variation in the response variable.
- This motivates the use of the coefficient of determination

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST} \in [0, 1]$$

as a measure of goodness of a regression model.

  - The coefficient of determination measures the proportion of the variance in the response variable that is predictable from the explanatory variables.

Coefficient of determination: Properties

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST} \in [0, 1]$$

- The following conditions are equivalent:
  1. $R^2 = 1$.
  2. $e_i = 0$ for all $i = 1, 2, ..., n$.
  3. all the observed $(x_{i1}, x_{i2}, ..., x_{ik}, y_i)$ lie on the same plane in $\mathbb{R}^{k+1}$.
  4. "The model explains the variation in the response variable perfectly."
- The following conditions are equivalent:
  1. $R^2 = 0$.
  2. $b_1 = b_2 = ... = b_k = 0$.
  3. "The model does not explain the variation in the response variable at all."
- The coefficient of determination is equal to the square of the sample Pearson correlation coefficient of the observed values y_i and the fitted values $\hat{y}_i$.

Content

1. Definition and assumptions
2. Matrix representation of linear models
3. Least squares estimation
4. Fitted values and residuals
5. Variance decomposition and coefficient of determination
6. Significance testing
7. Confidence intervals
8. Prediction interval
9. Random explanatory variables
10. On optimality of least squares estimators

Significance testing

Let the standard assumptions (i-v) hold. Now, the expected value and the covariance matrix of the least squares estimator $\boldsymbol{b}$ are

$$\mathrm{E}[\boldsymbol{b}] = \boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$$
$$D^2(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

An unbiased estimator for the covariance matrix $D^2(\boldsymbol{b})$ is

$$\hat{D}^2(\boldsymbol{b}) = s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

where s^2 is the unbiased estimator of the residual variance σ^2,

$$s^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} e_i^2.$$

Significance testing

- The coefficient of determination measures how big part of the variation in the response variable is explained by the estimated regression model. However, it does not tell whether the coefficient of determination differs significantly from zero or not.
- The significance of the coefficient of determination can be tested by examining how likely it is to obtain as large value as was obtained, under the assumption that the residuals explain all of the variation in the response variable.
- Note that if the residuals explain all of the variation in the response variable, then $\beta_j = 0$ for all $j = 1, ..., k$.

Significance testing

The significance of the coefficient of determination can be tested by applying the permutation test as follows:

1. The null hypothesis is $H_0 : \beta_j = 0$ for all $j = 1, ..., k$ and the alternative hypothesis is $H_1 : \beta_j \neq 0$ for at least one j.
2. Calculate the value R^2 from the original observations $(y_i, \boldsymbol{x_i})$, $\boldsymbol{x_i} = (x_{i1}, ..., x_{ik})$.
3. Form n new pairs $(y_i, \boldsymbol{x_I})$, such that each original explanatory variable vector $\boldsymbol{x_I}$ (and each original response variables y_i) is used exactly once in the new sample. Construct all the possible $n!$ permuted samples.
4. For each sample p, calculate the corresponding coefficient of determination, R_p^2. You obtain $n!$ values.
5. Order the values R_p^2 from the smallest to the largest and calculate the empirical $(1 - \alpha) \cdot 100$th percentile from the sample. If the original R^2 is larger than the calculated percentile, the coefficient of determination is considered significant on level α.

Significance of one single regression parameter

Permutation test can be applied also in testing the significance of one single regression parameter β_j.

1. The null hypothesis is $H_0 : \beta_j = 0$ and the alternative hypothesis is $H_1 : \beta_j \neq 0$.
2. Calculate the value R^2 from the original observations $(y_i, \boldsymbol{x_i})$, $\boldsymbol{x_i} = (x_{i1}, ..., x_{ik})$.
3. Pair each y_i with the original $\boldsymbol{x_i} = (x_{i1}, ..., x_{ik})$, $i = 1, ..., n$, but permute the components (and only the components) $x_{\cdot j}$ that correspond to the parameter β_j. You can construct $n!$ permuted samples. Each sample size is n.
4. For each sample p, calculate the corresponding coefficient of determination, R_p^2. You obtain $n!$ values.
5. Order the values R_p^2 from the smallest to the largest and calculate the empirical $(1 - \alpha) \cdot 100$th percentile from the sample. If the original R^2 is larger than the calculated percentile, the null hypothesis is rejected.

Permutation test

Remark

If the sample size $n \geq 10$ it is not reasonable to construct all the $n!$ permuted samples. In that case, instead of considering all the possible permutations, one can take for example 1000 or 10000 randomly chosen permuted samples and base the test on using these.

Significance testing, $\epsilon \sim N(0, \sigma^2)$

Let the standard assumptions hold. Assume also that the residuals are normally distributed. Now, the significance of the regression model can be tested using the F test statistic.

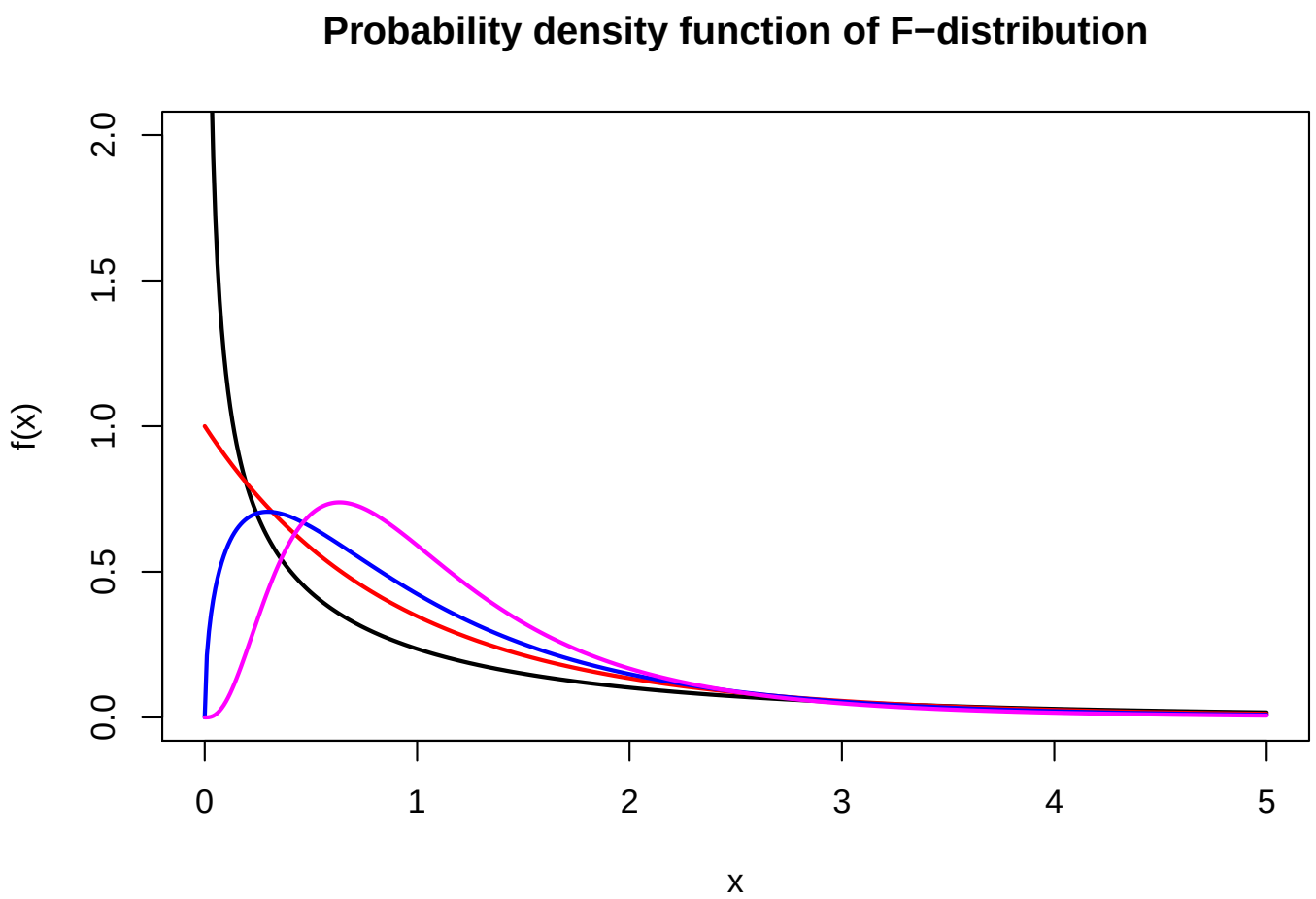
1. The null hypothesis is $H_0 : \beta_j = 0$ for all $j = 1, ..., k$ and the alternative hypothesis is $H_1 : \beta_j \neq 0$ for at least one j.
2. The F test statistic:

$$F = \frac{n-k-1}{k} \frac{R^2}{1-R^2} = \frac{n-k-1}{k} \frac{SSM}{SSE}$$

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

Note that this test statistic is reliable only when the residuals are normally distributed!

F distribution



Probability density functions of the $F(k, n-k-1)$ distribution for $n = 20$ and $k = 1$ (black), $k = 2$ (red), $k = 3$ (blue) and $k = 8$ (purple).

Significance of one single regression parameter, $\epsilon \sim N(0, \sigma^2)$

Let the standard assumptions hold. Assume also that the residuals are normally distributed.

- Now
$$\boldsymbol{b} \sim N_{k+1}\left(\boldsymbol{\beta}, D^2(b)\right),$$
where $D^2(b) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.
- Thus
$$\frac{b_j - \beta_j}{s_{bj}} \sim t(n-k-1),$$
where $t(n-k-1)$ denotes the t distribution with $n-k-1$ degrees of freedom and $s_{bj}^2 = \left[\hat{D}^2(b)\right]_{jj}$ is the estimated variance of β_j. Here $\left[\hat{D}^2(b)\right]_{jj}$ denotes the jj element of the estimated covariance matrix $\hat{D}^2(b) = s^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

Significance of one single regression parameter, $\epsilon \sim N(0, \sigma^2)$

Let the standard assumptions hold. Assume also that the residuals are normally distributed. Now, the significance of the regression parameter β_j can be tested using the t test statistic.

1. The null hypothesis is $H_0 : \beta_j = 0$ and the alternative hypothesis is $H_1 : \beta_j \neq 0$.
2. The t test statistic:

$$t = \frac{b_j}{s_{bj}}, \quad j = 0, 1, 2, ..., k,$$

follows, under the null hypothesis, the t distribution with $n - k - 1$ degrees of freedom.

Note that this test statistic is reliable only when the residuals are normally distributed!

Content

1 Definition and assumptions

2 Matrix representation of linear models

3 Least squares estimation

4 Fitted values and residuals

5 Variance decomposition and coefficient of determination

6 Significance testing

7 Confidence intervals

8 Prediction interval

9 Random explanatory variables

10 On optimality of least squares estimators

Confidence intervals

Fact

A level $(1-\alpha)$ confidence interval for a parameter θ is a random interval that contains the true (non-random) parameter value θ with probability $(1-\alpha)$.

If one calculates a $(1-\alpha)$ confidence interval for a parameter θ from 500 independent samples of i.i.d. observations, then approximately $500 \times (1-\alpha)$ of the intervals do contain the true value θ.

Bootstrap confidence interval for β_j

A $(1-\alpha)$ bootstrap confidence interval for the regression parameter β_j, $j = 0, 1, ..., k$ can be obtained as follows.

1. Select n data points randomly with replacement from the original observations $(y_i, \boldsymbol{x_i})$, $i = 1, ..., n$. Each data point $(y_i, \boldsymbol{x_i})$ can be selected once, multiple times, or not at all.
2. Calculate a new parameter vector estimate $\boldsymbol{b}_b$ from the new sample formed in the previous step.
3. Repeat the previous steps $m-1$ times. (For example 999 times.)
4. Order all the obtained estimates $b_j = (\boldsymbol{b}_b)_j$ from the smallest to the largest. Include also the original estimate.
5. Set the lower end of the bootstrap confidence interval to be smaller than or equal to the $[\frac{\alpha}{2} \times m]$th ordered estimate and set the upper end of the bootstrap confidence interval to be larger than or equal to the $[(1-\frac{\alpha}{2}) \times m]$th ordered estimate.

Bootstrap confidence interval for β_j

Some remarks:

- In regression settings, one can calculate bootstrap confidence intervals for β_j also by bootstrapping the residuals.
- Bootstrap confidence intervals are distribution free.
- The estimate is the better the larger the original sample size is and the larger the number (m) of the bootstrap samples is.

Confidence interval for β_j, $\epsilon \sim N(0, \sigma^2)$

Let the standard assumptions hold. Assume also that the residuals are normally distributed.

- Now, as noted above,

$$\frac{b_j - \beta_j}{s_{bj}} \sim t(n-k-1),$$

where $t(n-k-1)$ is the t distribution with $n-k-1$ degrees of freedom and s_{bj}^2 is the estimated variance of β_j.

Confidence interval for β_j, $\epsilon \sim N(0, \sigma^2)$

Let the standard assumptions hold. Assume also that the residuals are normally distributed. Under these assumptions, a level $(1 - \alpha)$ confidence interval for β_j can be given as

$$\left(b_j - t_{1-\alpha/2} s_{bj},\ b_j + t_{1-\alpha/2} s_{bj}\right),$$

where $t_{1-\alpha/2}$ is the $(1 - \alpha/2) \cdot 100$th percentile of the $t(n - k - 1)$ distribution.

Content

1 Definition and assumptions

2 Matrix representation of linear models

3 Least squares estimation

4 Fitted values and residuals

5 Variance decomposition and coefficient of determination

6 Significance testing

7 Confidence intervals

8 Prediction interval

9 Random explanatory variables

10 On optimality of least squares estimators

Prediction interval

The predicted value of y for a fixed $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_k)$ is

$$\hat{y} = b_0 + b_1\tilde{x}_1 + b_2\tilde{x}_2 + ... + b_k\tilde{x}_k.$$

This prediction is done under uncertainties as

1. the regression parameters have been estimated and
2. the residuals explain part of the variation in y.

$\Longrightarrow$ We wish to obtain an interval estimator that contains the true value with probability $(1 - \alpha)$.

Prediction interval

One can apply bootstrapping in constructing interval estimates for the predicted value of y, for fixed $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_k)$, as follows.

1. Calculate one bootstrap estimate $\boldsymbol{b}_b$ as described above.
2. Select a residual e_b randomly from the original estimated residuals e_i and calculate a bootstrap prediction $y_b = \tilde{\boldsymbol{x}}_*^\top \boldsymbol{b}_b + e_b$, where $\tilde{\boldsymbol{x}}_* = (1, \tilde{x}_1, ..., \tilde{x}_k)^\top$.
3. Repeat the previous steps $m - 1$ times. (For example 999 times.)
4. Order all the obtained predictions y_b from the smallest to the largest. Include also the original prediction.
5. Set the lower end of the bootstrap prediction interval to be smaller than or equal to the $[\frac{\alpha}{2} \times m]$th ordered bootstrap prediction and set the upper end of the bootstrap prediction interval to be larger than or equal to the $[(1 - \frac{\alpha}{2}) \times m]$th ordered bootstrap prediction.

Prediction interval, $\epsilon \sim N(0, \sigma^2)$

Let the standard assumptions hold. Assume also that the residuals are normally distributed. Now, a prediction interval for y conditioned on $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_k)$ can be given as

$$\tilde{\boldsymbol{x}}_*^\top \boldsymbol{b} \pm t_{1-\alpha/2} s \left[1 + \tilde{\boldsymbol{x}}_*^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \tilde{\boldsymbol{x}}_*\right]^{\frac{1}{2}},$$

where s^2 is the estimated residual variance and $t_{1-\alpha/2}$ is the $(1 - \alpha/2) \cdot 100$th percentile of the $t(n - k - 1)$ distribution.

Content

1. Definition and assumptions
2. Matrix representation of linear models
3. Least squares estimation
4. Fitted values and residuals
5. Variance decomposition and coefficient of determination
6. Significance testing
7. Confidence intervals
8. Prediction interval
9. Random explanatory variables
10. On optimality of least squares estimators

Random explanatory variables

Everything above can be applied directly also when the matrix $\boldsymbol{X}$ is random.

Random explanatory variables

Standard assumptions, when the explanatory variables are random are given as follows.

(i) The explanatory variables are random.

(ii) The explanatory variables are linearly independent.

(iii) $\mathrm{E}[\epsilon \mid \boldsymbol{X}] = \mathbf{0}$ for all $i = 1, ..., n$

(iv)-(v) $\mathrm{cov}(\epsilon \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}$.

Content

1. Definition and assumptions
2. Matrix representation of linear models
3. Least squares estimation
4. Fitted values and residuals
5. Variance decomposition and coefficient of determination
6. Significance testing
7. Confidence intervals
8. Prediction interval
9. Random explanatory variables
10. On optimality of least squares estimators

Positive definite matrices

A $d \times d$ matrix $\boldsymbol{C}$ (with real valued or complex valued elements) is positive definite, if $\boldsymbol{a}^* \boldsymbol{C}\boldsymbol{a} > 0$, for all $\boldsymbol{a} \in \mathbb{C}^d \setminus \{\mathbf{0}\}$.

A $d \times d$ matrix $\boldsymbol{C}$ (with real valued or complex valued elements) is positive semidefinite, if $\boldsymbol{a}^* \boldsymbol{C}\boldsymbol{a} \geq 0$, for all $\boldsymbol{a} \in \mathbb{C}^d \setminus \{\mathbf{0}\}$.

Positive definite matrices

It follows from the definition of positive definiteness that all positive definite matrices are Hermite symmetric. That is, if a square matrix $\boldsymbol{C}$ is positive definite, then $\boldsymbol{C} = \boldsymbol{C}^*$. For real square matrices, positive definiteness can be defined as follows.

A $d \times d$ matrix $\boldsymbol{M}$ with real valued elements is positive definite if it is symmetric and if $\boldsymbol{a}^\top \boldsymbol{M}\boldsymbol{a} > 0$ for all $\boldsymbol{a} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$.

A $d \times d$ matrix $\boldsymbol{M}$ with real valued elements is positive semidefinite if it is symmetric and if $\boldsymbol{a}^\top \boldsymbol{M}\boldsymbol{a} \geq 0$ for all $\boldsymbol{a} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$.

Optimality

If the standard assumptions hold, then the least squares estimator $\boldsymbol{b}$ is

1. unbiased ($\mathrm{E}[\boldsymbol{b}] = \beta$)
2. efficient in the sense that if $\boldsymbol{b}^*$ is another unbiased linear estimator for β, then the matrix $\boldsymbol{C} := \mathrm{cov}(\boldsymbol{b}^*) - \mathrm{cov}(\boldsymbol{b})$ is positive semidefinite.
3. consistent.

Optimality

Standard least squares estimator is not always efficient.

- Standard least squares estimator is not necessarily efficient if the standard assumption (iv) homoscedasticity and/or the standard assumption (v) uncorrelatedness do not hold.
- If the standard assumptions hold, but the vector β has linear restrictions, then the corresponding restricted least squares estimator is efficient.

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:

- Regression diagnostics
  - Regression graphics
  - Outlying observations
  - Constant/Non-constant regression parameters
  - Multicollinearity
  - Heteroscedasticity
  - Normality/Non-normality
  - Prediction capability
- Model selection
  - Model selection tests and strategies
  - Model selection criteria
  - Linearization