Lecture 9 | 14.04.2025

Statistical inference
in a linear model (asymptotics)

Overview

- **Normal linear regression model**
  - **Assumptions:** random sample $(Y_i, \boldsymbol{X}_i^\top)^\top$ for $i = 1, \dots, n$ from the joint distribution $F_{(Y,\boldsymbol{X})}$ such that $Y_i|\boldsymbol{X}_i \sim N(\boldsymbol{X}_i^\top \beta, \sigma^2)$
  - **Inference:** confidence intervals for β_j, confidence regions for β and linear combinations of the form $\mathbb{L}\beta$ (plus the corresponding statistical tests)

- **Linear regression model without normality**

  Assumptions (A1):
  - random sample $(Y_i, \boldsymbol{X}_i^\top)^\top$, $i = 1, \dots, n$ from the joint distribution $F_{(Y,\boldsymbol{X})}$
  - mean specification $E[Y_i|\boldsymbol{X}_i] = \boldsymbol{X}_i^\top \beta$, respectively $E[\boldsymbol{Y}|\mathbb{X}] = \mathbb{X}\beta$
  - thus, for errors $\varepsilon_i = Y_i - \boldsymbol{X}_i^\top \beta$ we have $E[\varepsilon_i|\boldsymbol{X}_i] = E[Y_i - \boldsymbol{X}_i^\top \beta|\boldsymbol{X}_i] = 0$ and $Var(\varepsilon_i|\boldsymbol{X}_i) = Var[Y_i - \boldsymbol{X}_i^\top \beta|\boldsymbol{X}_i] = Var[Y_i|\boldsymbol{X}_i] = \sigma^2(\boldsymbol{X}_i)$
  - and for unconditional expectations, $E[\varepsilon_i] = E[E[\varepsilon_i|\boldsymbol{X}_i]] = 0$ and $Var(\varepsilon_i) = Var(E[\varepsilon_i|\boldsymbol{X}_i]) + E[Var(\varepsilon_i|\boldsymbol{X}_i)] = Var(0) + E[\sigma^2(\boldsymbol{X}_i)] = E[\sigma^2(\boldsymbol{X}_i)]$

  Inference:
  - involves different confidence intervals statistical tests of hypotheses

Parameter estimation without normality

- In the normal regression model $\boldsymbol{Y} = \mathbb{X}\beta + \varepsilon$ one can simply use the distributional specification to formulate the likelihood (loglikelihood)
- In a general regression model $\boldsymbol{Y} = \mathbb{X}\beta + \varepsilon$ where $\varepsilon \sim (\mathbf{0}, \Sigma)$ the likelihood (loglikelihood resp.) can not be formulated (the distribution is missing)

- The most common approach in this case is based on the method of least squares (LSE), thus, the vector of the estimated parameters is given as

$$\widehat{\beta}_n = \underset{\beta \in \mathbb{R}^p}{\mathrm{Arg\,max}} \quad \sum_{i=1}^{n} \left[Y_i - \boldsymbol{X}_i^\top \beta \right]^2$$

- and the estimated vector of parameters can be given explicitly as

$$\widehat{\beta}_n \equiv \widehat{\beta} = \left(\mathbb{X}^\top \mathbb{X} \right)^{-1} \mathbb{X}^\top \boldsymbol{Y}$$

which is the **BLUE** estimate for $\beta \in \mathbb{R}^p$ but for the statistical inference we need to know its (asymptotic) distributional properties (how does this random quantity behave when $n \in \mathbb{N}$ tends to infinity, $n \to \infty$)

Some additional assumptions

The random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$ drawn from some joint distribution $F_{(Y,\boldsymbol{X})}$ of a generic $(p+1)$-dimensional random vector $(Y, \boldsymbol{X}^\top)^\top$. Let $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$. Let the following holds:

Assumptions (A2):

- $E|X_j X_k| < \infty$ for $j, k \in \{1, \ldots, p\}$
- $E\big(\boldsymbol{X}\boldsymbol{X}^\top\big) = \mathbb{W} \in \mathbb{R}^{p \times p}$ is a positive definite (regular) matrix
- $\mathbb{V} = \mathbb{W}^{-1}$

Note, that the assumptions stated above refer to the population model—the population properties

Empirical counterparts for $\mathbb{W}$ and $\mathbb{V}$

- Both matrices, $\mathbb{W} \in \mathbb{R}^{p \times p}$ and $\mathbb{V} \in \mathbb{R}^{p \times p}$ are theoretical (population) characteristics, the dimensions are fixed for any $n \in \mathbb{N}$, and they are typically not known in practical applications
- Both matrices can be however estimated using the empirical data—the observed random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$

- Define the following:
  - $\mathbb{W}_n = \mathbb{X}^\top \mathbb{X} = \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top$
  - $\mathbb{V}_n = \mathbb{W}_n^{-1}$ if it exists (eventually it will for $n \in \mathbb{N}$ large enough)

- Under the assumptions in (A1) and (A2)
  - $\frac{1}{n}\mathbb{W}_n \longrightarrow \mathbb{W}$ a.s. (in P) as $n \to \infty$
  - $n\mathbb{V}_n \longrightarrow \mathbb{V}$ a.s. (in P) as $n \to \infty$

It is also good to realize that $\left(\mathbb{X}^\top \mathbb{X}\right)^{-1}$ may not exist for any $n \in \mathbb{N}$ but as far as $\frac{1}{n}\left(\mathbb{X}^\top \mathbb{X}\right)$ converges almost surely (in probability) to the matrix $\mathbb{W}$ (positive definite) we also have that $P(rank(\mathbb{X}^\top \mathbb{X}) = p) \to 1$, for $n \to \infty$

Problems of the statistical inference

Analogously as in the normal linear model, the statistical inference concerns confidence sets and statistical tests about $\boldsymbol{\beta} \in \mathbb{R}^p$ and its linear combinations

- Statistical inference can be performed with respect to the parameters β and σ^2 but, it can be also of some interest to do inference about some (appropriate) linear combination(s) of $\boldsymbol{\beta}$
- From the practical point of view, we are interested in the parameter vector $\boldsymbol{\beta}$ itself but also linear combinations of the form $\boldsymbol{l}^\top \boldsymbol{\beta}$ or $\mathbb{L}\boldsymbol{\beta}$

The estimates for the unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma^2 > 0$ are

- $\widehat{\boldsymbol{\beta}}_n = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X} \boldsymbol{Y} = \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i Y_i \right)$ (LSE)
- $s_n^2 = \frac{1}{n-p} \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 = \frac{1}{n-p} \| \boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}} \|_2^2$, where $\widehat{Y}_i = \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}$ (MSe)

Both estimates, quantities $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{s_n^2}$, are random quantities (random vector and random variable) and, therefore, it is reasonable to investigate their statistical properties (e.g., mean, variance, distribution, etc.)

Homoscedastic vs. heteroscedastic model

Recall, that in the assumption in (A1) the conditional variance of ε_i depends on $\boldsymbol{X}_i$, which is reflected by the notation $Var(\varepsilon_i|\boldsymbol{X}_i) = \sigma^2(\boldsymbol{X}_i)$

Two situations are typically distinguished:

- **Homoscedastic model)** (Assumption A3a)
  $\sigma^2(\boldsymbol{X}) = Var(Y|\boldsymbol{X}) = \sigma^2 > 0$

- **Heteroscedastic model** (Assumption A3b)
  $\sigma^2(\boldsymbol{X}) = Var(Y|\boldsymbol{X})$ such that $E[\sigma^2(\boldsymbol{X})] < \infty$ and moreover, it also holds that $E[\sigma^2(\boldsymbol{X})X_jX_k] < \infty$ for $j,k \in \{1,\ldots,p\}$

Consistency of the LSE estimates

- In particular, we are interested in the following parameters:
  - $\boldsymbol{\beta} \in \mathbb{R}^p$
  - $\sigma^2 > 0$
  - $\theta = \boldsymbol{l}^\top \boldsymbol{\beta} \in \mathbb{R}$, for some nonzero vector $\boldsymbol{l} \in \mathbb{R}^p$
  - $\Theta = \mathbb{L}\boldsymbol{\beta} \in \mathbb{R}^m$, for some matrix $\mathbb{L} \in \mathbb{R}^{m \times p}$ with linearly independent rows

- The corresponding estimates are defined straightforwardly and it holds (under (A1), (A2), and (A3a/A3b)) that
  - $\widehat{\boldsymbol{\beta}}_n \longrightarrow \boldsymbol{\beta}$ a.s. (in P), for $n \to \infty$
  - $\widehat{\theta}_n = \boldsymbol{l}^\top \widehat{\boldsymbol{\beta}}_n \longrightarrow \theta$ a.s. (in P), for $n \to \infty$
  - $\widehat{\Theta}_n = \mathbb{L}\widehat{\boldsymbol{\beta}}_n \longrightarrow \Theta$, a.s. (in P), for $n \to \infty$

- Under the homoscedastic model ((A1), (A2), and (A3a)) it also holds
  - $\widehat{s_n^2} \longrightarrow \sigma^2$, a.s. (in P), for $n \to \infty$

Assymptotic normality

Under the assumptions stated in (A1), (A2), and (A3a) and, additionally, for $E[\varepsilon^2 X_j X_k] < \infty$ for $j, k = 1, \ldots, p$ the following holds:

- $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N_p(\boldsymbol{\beta}, \sigma^2 \mathbb{V})$ for $n \to \infty$
- $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} N(0, \sigma^2 \boldsymbol{l}^\top \mathbb{V} \boldsymbol{l})$, as $n \to \infty$
- $\sqrt{n}(\widehat{\Theta}_n - \Theta) \xrightarrow{\mathcal{D}} N_m(\boldsymbol{0}, \sigma^2 \mathbb{L} \mathbb{V} \mathbb{L}^\top)$, as $n \to \infty$

Statistical inference based on asymptotics

- Define the random variable

$$T_n = \frac{\boldsymbol{l}^\top \widehat{\beta}_n - \boldsymbol{l}^\top \beta}{\sqrt{MSe \cdot \boldsymbol{l}^\top (\mathbb{X}^\top \mathbb{X})^{-1} \boldsymbol{l}}} \qquad \left(= \frac{\sqrt{n}(\boldsymbol{l}^\top \widehat{\beta}_n - \boldsymbol{l}^\top \beta)}{\sqrt{\sigma^2 \boldsymbol{l}^\top \mathbb{V} \boldsymbol{l}}} \cdot \sqrt{\frac{\sigma^2 \boldsymbol{l}^\top \mathbb{V} \boldsymbol{l}}{MSe \cdot \boldsymbol{l}^\top \left[n(\mathbb{X}^\top \mathbb{X})^{-1} \right] \boldsymbol{l}}} \right)$$

$\hookrightarrow$ where it is easy to see that the first term in the brackets converges (in distribution) to $N(0,1)$ and the second term converges (in probability) to one (Cramér-Slutsky)

- Define the random variable

$$Q_n = \frac{(\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \mathbb{L}\boldsymbol{\beta})^\top \left[\mathbb{L}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{L}^\top \right]^{-1} (\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \mathbb{L}\boldsymbol{\beta})}{MSe}$$

$\hookrightarrow$ where $\sqrt{n}(\mathbb{L}\widehat{\boldsymbol{\beta}}_n - \mathbb{L}\boldsymbol{\beta}) \xrightarrow{\mathcal{D}} N(0, \sigma^2 \mathbb{L}\mathbb{V}\mathbb{L}^\top)$ for $n \to \infty$ and, also, $\left(MSe \cdot \left[\mathbb{L} n (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{L}^\top \right] \right)^{-1} \xrightarrow{\mathcal{D}} \sigma^2 \mathbb{L}\mathbb{V}\mathbb{L}^\top$ for $n \to \infty$ (Cramér-Slutsky)

- Then it holds that
  - $T_n \xrightarrow{\mathcal{D}} N(0,1)$ for $n \to \infty$
  - $Q_n \xrightarrow{\mathcal{D}} \chi^2_m$ for $n \to \infty$

$\hookrightarrow$ what is this good for in practical applications and inference?

Standard inference tools – summary

In general, the **statistical inference** is a (mathematical) process of using observed data (e.g., random sample) to make **valid and consistent conclusions** or predictions about an unknown (much larger) population. It involves (mainly) the hypotheses testing and confidence intervals construction.

- **Confidence intervals**
  - normal linear regression model (exact coverage)
  - linear regression model without normality (asymptotic coverage)

- **Statistical tests**
  - normal linear regression model (based on the exact distribution)
  - linear regression model without normality (asymptotic validity)