Lecture 7 | 14.04.2026

Statistical inference

in a linear model without normality (asymptotics)

Overview

- **Normal linear regression model**
  - **Assumptions:** Random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$ from the joint distribution $F_{(Y,\boldsymbol{X})}$, such that $Y_i|\boldsymbol{X}_i \sim N(\boldsymbol{X}_i^\top \beta, \sigma^2)$ while the marginal distribution of r.v. $\boldsymbol{X} \in \mathbb{R}^p$ does not depend on $\beta \in \mathbb{R}^p$ neither on $\sigma^2 > 0$
  - **Inference:** Confidence intervals/tests for $\beta_j \in \mathbb{R}$, confidence regions/tests for $\beta \in \mathbb{R}^p$ (or a subset of parameters), including also linear combinations (reparametrizations) of $\beta \in \mathbb{R}^p$ in a form $\ell^\top \beta \in \mathbb{R}$ or $\mathbb{L}\beta \in \mathbb{R}^m$

Overview

- **Normal linear regression model**
  - **Assumptions:** Random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$ from the joint distribution $F_{(Y,\boldsymbol{X})}$, such that $Y_i|\boldsymbol{X}_i \sim N(\boldsymbol{X}_i^\top \beta, \sigma^2)$ while the marginal distribution of r.v. $\boldsymbol{X} \in \mathbb{R}^p$ does not depend on $\beta \in \mathbb{R}^p$ neither on $\sigma^2 > 0$
  - **Inference:** Confidence intervals/tests for $\beta_j \in \mathbb{R}$, confidence regions/tests for $\beta \in \mathbb{R}^p$ (or a subset of parameters), including also linear combinations (reparametrizations) of $\beta \in \mathbb{R}^p$ in a form $\ell^\top \beta \in \mathbb{R}$ or $\mathbb{L}\beta \in \mathbb{R}^m$

- **Linear regression model without normality**

  Assumptions (A1):
  - random sample $(Y_i, \boldsymbol{X}_i^\top)^\top$, $i = 1, \ldots, n$ 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 $\varepsilon_i = Y_i - \boldsymbol{X}_i^\top \beta$ it holds $E[\varepsilon_i|\boldsymbol{X}_i] = E[Y_i - \boldsymbol{X}_i^\top \beta|\boldsymbol{X}_i] = 0$ and $Var(\varepsilon_i|\boldsymbol{X}_i) = Var[Y_i - \boldsymbol{X}_i^\top \beta|\boldsymbol{X}_i] = Var[Y_i|\boldsymbol{X}_i] = \sigma^2(\boldsymbol{X}_i)$

Overview

- **Normal linear regression model**
  - **Assumptions:** Random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$ from the joint distribution $F_{(Y,\boldsymbol{X})}$, such that $Y_i|\boldsymbol{X}_i \sim N(\boldsymbol{X}_i^\top \beta, \sigma^2)$ while the marginal distribution of r.v. $\boldsymbol{X} \in \mathbb{R}^p$ does not depend on $\beta \in \mathbb{R}^p$ neither on $\sigma^2 > 0$
  - **Inference:** Confidence intervals/tests for $\beta_j \in \mathbb{R}$, confidence regions/tests for $\beta \in \mathbb{R}^p$ (or a subset of parameters), including also linear combinations (reparametrizations) of $\beta \in \mathbb{R}^p$ in a form $\ell^\top \beta \in \mathbb{R}$ or $\mathbb{L}\beta \in \mathbb{R}^m$

- **Linear regression model without normality**

  Assumptions (A1):
  - random sample $(Y_i, \boldsymbol{X}_i^\top)^\top$, $i = 1, \ldots, n$ 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 $\varepsilon_i = Y_i - \boldsymbol{X}_i^\top \beta$ it holds $E[\varepsilon_i|\boldsymbol{X}_i] = E[Y_i - \boldsymbol{X}_i^\top \beta|\boldsymbol{X}_i] = 0$ and $Var(\varepsilon_i|\boldsymbol{X}_i) = Var[Y_i - \boldsymbol{X}_i^\top \beta|\boldsymbol{X}_i] = Var[Y_i|\boldsymbol{X}_i] = \sigma^2(\boldsymbol{X}_i)$
  - Unconditional characteristics: $E[\varepsilon_i] = E[E[\varepsilon_i|\boldsymbol{X}_i]] = 0$ and $Var(\varepsilon_i) = Var(E[\varepsilon_i|\boldsymbol{X}_i]) + E[Var(\varepsilon_i|\boldsymbol{X}_i)] = Var(0) + E[\sigma^2(\boldsymbol{X}_i)] = E[\sigma^2(\boldsymbol{X}_i)]$

Overview

- **Normal linear regression model**
  - **Assumptions:** Random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$ from the joint distribution $F_{(Y,\boldsymbol{X})}$, such that $Y_i|\boldsymbol{X}_i \sim N(\boldsymbol{X}_i^\top \beta, \sigma^2)$ while the marginal distribution of r.v. $\boldsymbol{X} \in \mathbb{R}^p$ does not depend on $\beta \in \mathbb{R}^p$ neither on $\sigma^2 > 0$
  - **Inference:** Confidence intervals/tests for $\beta_j \in \mathbb{R}$, confidence regions/tests for $\beta \in \mathbb{R}^p$ (or a subset of parameters), including also linear combinations (reparametrizations) of $\beta \in \mathbb{R}^p$ in a form $\ell^\top \beta \in \mathbb{R}$ or $\mathbb{L}\beta \in \mathbb{R}^m$

- **Linear regression model without normality**

  Assumptions (A1):
  - random sample $(Y_i, \boldsymbol{X}_i^\top)^\top$, $i = 1, \ldots, n$ 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 $\varepsilon_i = Y_i - \boldsymbol{X}_i^\top \beta$ it holds $E[\varepsilon_i|\boldsymbol{X}_i] = E[Y_i - \boldsymbol{X}_i^\top \beta|\boldsymbol{X}_i] = 0$ and $Var(\varepsilon_i|\boldsymbol{X}_i) = Var[Y_i - \boldsymbol{X}_i^\top \beta|\boldsymbol{X}_i] = Var[Y_i|\boldsymbol{X}_i] = \sigma^2(\boldsymbol{X}_i)$
  - Unconditional characteristics: $E[\varepsilon_i] = E[E[\varepsilon_i|\boldsymbol{X}_i]] = 0$ 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:
  - Again, involves confidence intervals/regions and statistical tests of hypotheses about the unknown parameters ($\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$)

Parameter estimation without normality

- In the normal regression model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, one can simply use the distributional specification to formulate the whole likelihood (loglikelihood)
- In a general regression model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbb{I})$, the likelihood (the loglikelihood resp.) can not be formulated as the underlying distribution is not specified

Parameter estimation without normality

- In the normal regression model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, one can simply use the distributional specification to formulate the whole likelihood (loglikelihood)
- In a general regression model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbb{I})$, the likelihood (the loglikelihood resp.) can not be formulated as the underlying distribution is not specified
- 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

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

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 whole likelihood (loglikelihood)
- In a general regression model $\boldsymbol{Y} = \mathbb{X}\beta + \varepsilon$ where $\varepsilon \sim (\mathbf{0}, \sigma^2\mathbb{I})$, the likelihood (the loglikelihood resp.) can not be formulated as the underlying distribution is not specified
- 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

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

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

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

Parameter estimation without normality

- In the normal regression model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, one can simply use the distributional specification to formulate the whole likelihood (loglikelihood)
- In a general regression model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbb{I})$, the likelihood (the loglikelihood resp.) can not be formulated as the underlying distribution is not specified
- 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

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

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

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

  which is the **BLUE** estimate for $\boldsymbol{\beta} \in \mathbb{R}^p$ but for the statistical inference we need to know its (asymptotic) distributional properties (how does this random quantity behave as the sample size increases, i.e., $n \to \infty$)

Additional assumptions for assymptotics

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

Assumptions (A2):

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

Additional assumptions for assymptotics

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

Assumptions (A2):

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

Note, that the assumptions stated above all refer to the population model—the population properties related to the random vector $\boldsymbol{X} \in \mathbb{R}^p$.

The first two items in Assumptions (A2) are, indeed, technical conditions that need to be satisfied (for asymptotics). The last item is more like an introduction of a notation.

Empirical counterparts for matrices $\mathbb{W}$ 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 both these matrices are typically not known in practical applications
- Both matrices can be, however, effectively estimated using the empirical data—the observed random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$

Empirical counterparts for matrices $\mathbb{W}$ 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 both these matrices are typically not known in practical applications
- Both matrices can be, however, effectively estimated using the empirical data—the observed random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$
- Define the following:
  - $\mathbb{W}_n = \mathbb{X}^\top \mathbb{X} = \sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\top$ and, also
  - $\mathbb{V}_n = \mathbb{W}_n^{-1}$ if it exists (eventually it will for $n \in \mathbb{N}$ large enough)

Empirical counterparts for matrices $\mathbb{W}$ 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 both these matrices are typically not known in practical applications
- Both matrices can be, however, effectively estimated using the empirical data—the observed random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$
- Define the following:
  - $\mathbb{W}_n = \mathbb{X}^\top\mathbb{X} = \sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top$ and, also
  - $\mathbb{V}_n = \mathbb{W}_n^{-1}$ if it exists (eventually it will for $n \in \mathbb{N}$ large enough)
- Under the assumptions in (A1) and (A2), the following hold:
  - $\frac{1}{n}\mathbb{W}_n \longrightarrow \mathbb{W}$ a.s. (in P) as $n \to \infty$
  - $n\mathbb{V}_n \longrightarrow \mathbb{V}$ a.s. (in P) as $n \to \infty$

Empirical counterparts for matrices $\mathbb{W}$ 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 both these matrices are typically not known in practical applications
- Both matrices can be, however, effectively estimated using the empirical data—the observed random sample $\{(Y_i, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, n\}$
- Define the following:
  - $\mathbb{W}_n = \mathbb{X}^\top\mathbb{X} = \sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top$ and, also
  - $\mathbb{V}_n = \mathbb{W}_n^{-1}$ if it exists (eventually it will for $n \in \mathbb{N}$ large enough)
- Under the assumptions in (A1) and (A2), the following hold:
  - $\frac{1}{n}\mathbb{W}_n \longrightarrow \mathbb{W}$ 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 some (rather small) values of $n \in \mathbb{N}$ but as far as $\frac{1}{n}\left(\mathbb{X}^\top\mathbb{X}\right)$ converges almost surely (in probability) to the matrix $\mathbb{W}$ (which is 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 $\beta \in \mathbb{R}^p$ and its linear combinations

- Statistical inference can be performed with respect to the unknown parameters $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$ but, it can be also of some interest to do inference about some (appropriate) linear combination(s) of β
- From the practical point of view, it may be often of some more relevant interest to say something about a specific linear combination—for instance, $\boldsymbol{l}^\top \beta \in \mathbb{R}$ or $\mathbb{L}\beta \in \mathbb{R}^m$ rather than $\beta \in \mathbb{R}^p$ itself

Problems of the statistical inference

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

- Statistical inference can be performed with respect to the unknown parameters $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$ but, it can be also of some interest to do inference about some (appropriate) linear combination(s) of β
- From the practical point of view, it may be often of some more relevant interest to say something about a specific linear combination—for instance, $\boldsymbol{l}^\top \beta \in \mathbb{R}$ or $\mathbb{L}\beta \in \mathbb{R}^m$ rather than $\beta \in \mathbb{R}^p$ itself

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

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

Problems of the statistical inference

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

- Statistical inference can be performed with respect to the unknown parameters $\beta \in \mathbb{R}^p$ and $\sigma^2 > 0$ but, it can be also of some interest to do inference about some (appropriate) linear combination(s) of β
- From the practical point of view, it may be often of some more relevant interest to say something about a specific linear combination—for instance, $\boldsymbol{l}^\top \beta \in \mathbb{R}$ or $\mathbb{L}\beta \in \mathbb{R}^m$ rather than $\beta \in \mathbb{R}^p$ itself

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

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

Both estimates, quantities $\widehat{\beta}_n$ 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 $Var(\varepsilon_i|\boldsymbol{X}_i) = \sigma^2(\boldsymbol{X}_i)$

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 $j, k \in \{1, \ldots, p\}$

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 $j, k \in \{1, \ldots, p\}$

The **homoscedastic model** (without the normality) can be generally expressed as

$$Y_i|\boldsymbol{X}_i \sim (\boldsymbol{X}_i^\top \beta, \sigma^2)$$

while the **heteroscedastic model** (again, without the normality assumption) as

$$Y_i|\boldsymbol{X}_i \sim (\boldsymbol{X}_i^\top \beta, \sigma^2(\boldsymbol{X}_i))$$

for some positive variance function $\sigma^2 : \mathbb{R}^p \to (0, \infty)$

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

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
- Under Assumptions (A1) and Assumptions (A2), mohoscedastic model implies the heteroscedastic one, i.e. (A3a) $\Longrightarrow$ (A3b)

Consistency of the LSE estimates

- In particular, we are interested in the following parameters:
  - $\beta \in \mathbb{R}^p$
  - $\sigma^2 > 0$
  - $\theta = \boldsymbol{l}^\top \beta \in \mathbb{R}$, for some nonzero vector $\boldsymbol{l} \in \mathbb{R}^p$
  - $\Theta = \mathbb{L}\beta \in \mathbb{R}^m$, for some matrix $\mathbb{L} \in \mathbb{R}^{m \times p}$ with linearly independent rows
- Under Assumptions (A1) and Assumptions (A2), mohoscedastic model implies the heteroscedastic one, i.e. (A3a) $\Longrightarrow$ (A3b)
- The corresponding estimates are defined straightforwardly and it holds (under (A1), (A2), and (A3a/A3b)) that
  - $\widehat{\beta}_n \longrightarrow \beta$ a.s. (in P), for $n \to \infty$
  - $\widehat{\theta}_n = \boldsymbol{l}^\top \widehat{\beta}_n \longrightarrow \theta$ a.s. (in P), for $n \to \infty$
  - $\widehat{\Theta}_n = \mathbb{L}\widehat{\beta}_n \longrightarrow \Theta$, a.s. (in P), for $n \to \infty$

Consistency of the LSE estimates

- In particular, we are interested in the following parameters:
  - $\beta \in \mathbb{R}^p$
  - $\sigma^2 > 0$
  - $\theta = \boldsymbol{l}^\top \beta \in \mathbb{R}$, for some nonzero vector $\boldsymbol{l} \in \mathbb{R}^p$
  - $\Theta = \mathbb{L}\beta \in \mathbb{R}^m$, for some matrix $\mathbb{L} \in \mathbb{R}^{m \times p}$ with linearly independent rows
- Under Assumptions (A1) and Assumptions (A2), mohoscedastic model implies the heteroscedastic one, i.e. (A3a) $\Longrightarrow$ (A3b)
- The corresponding estimates are defined straightforwardly and it holds (under (A1), (A2), and (A3a/A3b)) that
  - $\widehat{\beta}_n \longrightarrow \beta$ a.s. (in P), for $n \to \infty$
  - $\widehat{\theta}_n = \boldsymbol{l}^\top \widehat{\beta}_n \longrightarrow \theta$ a.s. (in P), for $n \to \infty$
  - $\widehat{\Theta}_n = \mathbb{L}\widehat{\beta}_n \longrightarrow \Theta$, 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 (homoscedastic model)

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:

Assymptotic normality (homoscedastic model)

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:

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

Assymptotic normality (homoscedastic model)

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:

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

Analogous (asymptotic normality) result will also hold under the assumptions (A1), (A2), and (A3b) – i.e., under the heteroscedastic model – however, the asymptotic variance term will be slightly different. In both caes, the proof follow the same lines and, thus, it will be given for the heteroscedastic model only.

The assymptotic normality above already provides all necessary inference tools!

Statistical inference based on asymptotics

- Define the random variable T_n as

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

$\hookrightarrow$ 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)

Statistical inference based on asymptotics

- Define the random variable T_n as

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

  $\hookrightarrow$ 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 (a quadratic form) D_n as

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

  $\hookrightarrow$ 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, $\left(MSe \cdot \left[\mathbb{L}n(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{L}^\top \right] \right)^{-1} \xrightarrow{\mathcal{P}} \sigma^2 \mathbb{L}\mathbb{V}\mathbb{L}^\top$ for $n \to \infty$ (Cramér-Slutsky)

Statistical inference based on asymptotics

- Define the random variable T_n as

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

  $\hookrightarrow$ 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 (a quadratic form) D_n as

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

  $\hookrightarrow$ 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, $\left(MSe \cdot \left[\mathbb{L}n(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{L}^\top \right] \right)^{-1} \xrightarrow{\mathcal{P}} \sigma^2 \mathbb{L}\mathbb{V}\mathbb{L}^\top$ for $n \to \infty$ (Cramér-Slutsky)

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

Standard inference tools – overview

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/prediction intervals/sets construction.

Standard inference tools – overview

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/prediction intervals/sets 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)