Linear regression. Least squares estimator. Asymptotic properties. Gauss-Markov theorem

Jakub Mućk
SGH Warsaw School of Economics

About course

Topics I

SGH

1. Linear regression. Least squares estimator. Asymptotic properties. Gauss-Markov theorem
2. Testing economic hypotheses. Multiple hypothesis testing. Linear and non-linear hypotheses. Confidence intervals. Delta method.
3. Verifying key assumptions: normality, colinearity and functional form. Godness-of-fit.
4. Heteroskedasticity and serial correlation. Generalized least squares estimator. Weighted least squares. Robust and clustered standard errors.
5. Endogeneity. Instrumental variables estimation. Properties of instrumental variables.
6. Simultaneous equations model. Parameter identification problem. Estimation method for SEM.
7. Time series. Stationarity, spurious regression and cointegration.
8. Autoregressive distributed lags models. Vector Autoregression (VAR) models. Structural VAR.
9. Panel data. Between and within variation. Random and fixed effects models. Between regression. Hausman-Taylor estimator.

Topics II

10. Limited dependent variable. Models for binary and multinomial outcome variable. ML estimator. Panel data and limited dependent variable.
11. Count data models. Tobit regression.
12. Generalized method of moments. Selected applications
13. Dynamic panel data models. Nickell's Bias. Anderson-Hsiao estimator. Arellano-Bond estimator. System GMM estimator.
14. Estimating treatment effect. Difference-in-difference.
15. Regression discontinuity design

Econometrics textbooks:

1. Wooldridge J. M., *Econometric Analysis of Cross Section and Panel Data*.
2. Pesaran M. H., *Time Series and Panel Data Econometrics*.
3. Greene W. H., *Econometric Analysis*.
4. Hall R. C., Griffiths W. E., Lim G. C., *Principles of Econometrics*.
5. Wooldridge J. M., *Introductory Econometrics: A Modern Approach*.

Software

- Stata.
- R.

- Exam
- Homework ($\times$ 2-3) and classroom activity.

Introduction to Econometrics

What is econometrics about

Econometrics

is an application of statistical techniques to economics in the study of problems, the analysis of data, and the development and testing of theories and models.

Econometrics

is an application of statistical techniques to economics in the study of problems, the analysis of data, and the development and testing of theories and models.

We use econometrics

- to estimate economic parameters (e.g. elasticities),
- to forecast economic outcomes,
- to verify economic hypotheses.

- **Economic model** represents quantitative relationships between set of economic variables. For instance, the marginal propensity to consume:

$$C = \beta_0 + \beta_1 Y \tag{1}$$

- **Economic model** represents quantitative relationships between set of economic variables. For instance, the marginal propensity to consume:

$$C = \beta_0 + \beta_1 Y \tag{1}$$

where C is the consumption expenditures and Y is the disposable income.

- **Econometric model** is additionally extended by stochastic component:

$$C = \beta_0 + \beta_1 Y + \varepsilon, \tag{2}$$

where ε is the (disturbance) error term.

The general single-equation linear econometric model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_1 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad i = 1, 2, \ldots, N \tag{3}$$

where

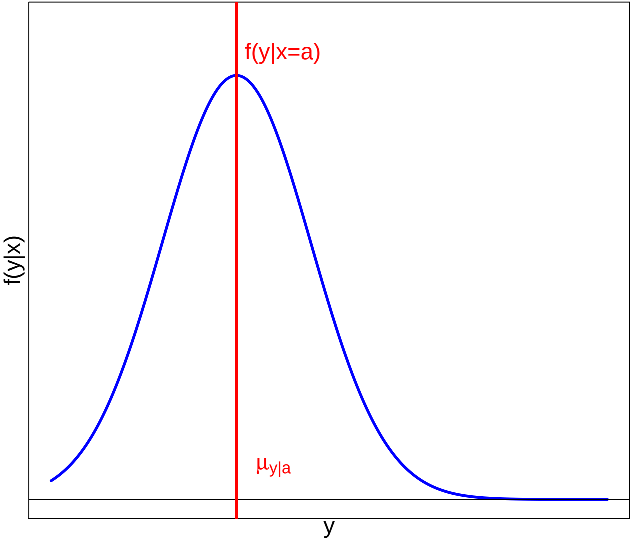
- y_i a dependent (outcome) variable,
- $x_{1i}, \ldots, x_{ki}$ is a set of k explanatory (independent) variables,
- $\beta_0, \beta_1, \ldots, \beta_k$ are the model parameters,
- ε_i is error term,
- i is an index of observation.

- **Cross-section data** are collected across sample units (individuals) in a particular time period.
  y_i where
  $i \in \{1, \ldots, N\}$.
- **Time series:** are collected over discrete intervals of time:
  y_t where
  $t \in \{1, \ldots, T\}$.
- **Panel or longitudinal data** is are collected across individual units over time:
  y_{it} where
  $i \in \{1, \ldots, N\}$
  $t \in \{1, \ldots, T\}$.

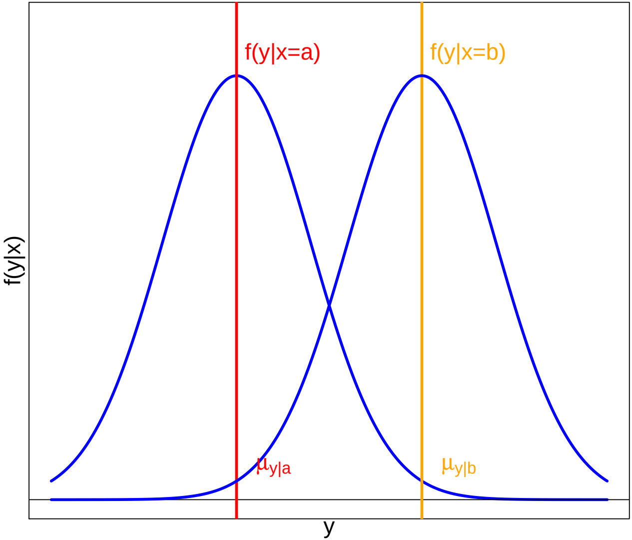
- Experimental data
- Microeconomic data
- Macroeconomic data
- Financial data

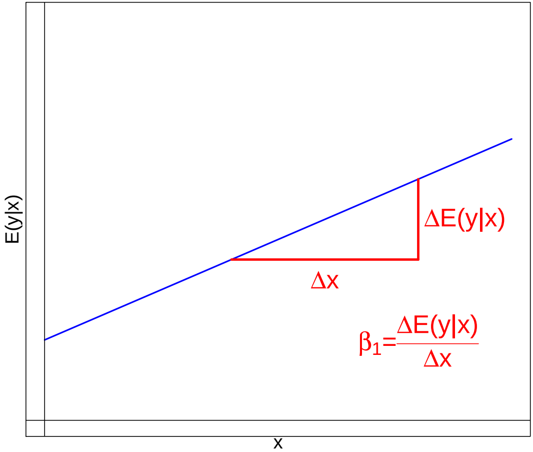
1. **Formulation of research hypotheses**. Based on research problem and a literature review.
2. **Choice of economic model** that leads to econometric model. This includes choosing the functional form as well as set of explanatory variables.
3. **Data collection**. Obtain sample and select method that allow to apply statistical interference.
4. **Estimating parameters**.
5. **Model diagnostics**. Check the validity of assumptions.

Simply Linear Regression Model


f(y|x=a)
$\mu_{y|a}$
f(y|x)
y

The starting point – conditional distribution of Y given X.





Simply Regression:

$$\mathbb{E}\left(y|x\right) = \beta_0 + \beta_1 x$$

- **Simply Linear Regression Model** :

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{4}$$

where
  - y is the (outcome) dependent variable;
  - x is independent variable;
  - ε is the error term.

- The dependent variable is explained with the components that vary with the **the dependent variable** and **the error term**.
- β_0 is the intercept.
- β_1 is the coefficient (slope) on x.

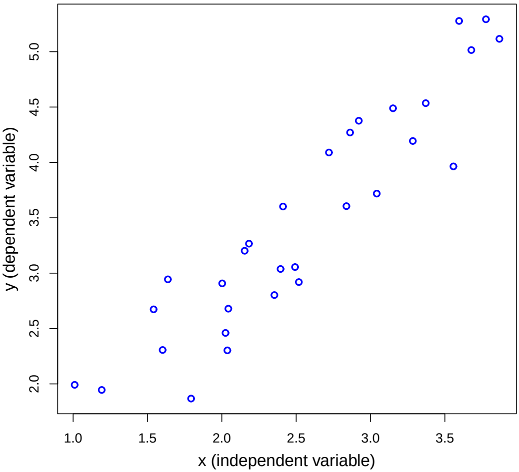
- **Simply Linear Regression Model** :

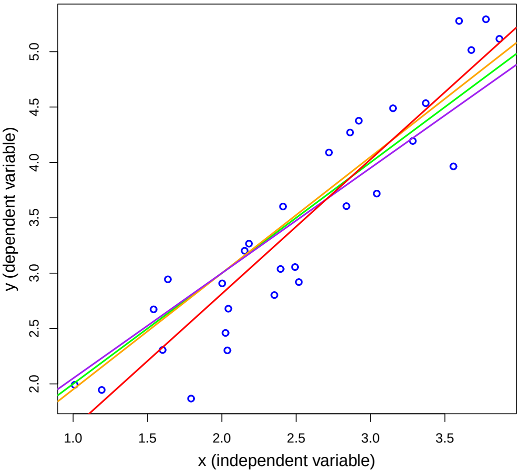
$$y = \beta_0 + \beta_1 x + \varepsilon \tag{4}$$

where
  - y is the (outcome) dependent variable;
  - x is independent variable;
  - ε is the error term.
- The dependent variable is explained with the components that vary with the **the dependent variable** and **the error term**.
- β_0 is the intercept.
- β_1 is the coefficient (slope) on x.

β_1 measures the effect of change in x upon the expected value of y (*ceteris paribus*).

The least squares (LS) estimator


y (dependent variable)
x (independent variable)
1.0
1.5
2.0
2.5
3.0
3.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


y (dependent variable)
2.0
2.5
3.0
3.5
4.0
4.5
5.0
1.0
1.5
2.0
2.5
3.0
3.5
x (independent variable)

Assumptions of the least squares estimators

- **Assumption #1**: true DGP (data generating process):

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{5}$$

- **Assumption #2**: the expected value of the error term is zero:

$$\mathbb{E}\left(\varepsilon\right) = 0, \tag{6}$$

  and this implies that $\mathbb{E}\left(y\right) = \beta_0 + \beta_1 x$.

- **Assumption #3**: the constant variance of the error term and zero covariance between observations. In particular:
  - the variance of the error term equals σ:

$$var\left(\varepsilon\right) = \sigma^2 = var\left(y\right). \tag{7}$$

  - the covariance between any pair of ε_i and ε_j is zero"

$$cov\left(\varepsilon_i, \varepsilon_j\right) = 0. \tag{8}$$

- **Assumption #4**: **Exogeneity.** The independent variable is **not random** and therefore it is not correlated with the error term.
- **Assumption #5**: the independent variable takes at least two values.
- **Assumption #6 (optional)**: the normally distributed error term:

$$\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right). \tag{9}$$

- **The fitted values of dependent variable ($\hat{y}_i$)**:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{10}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of intercept and slope, respectively.

- **The residuals ($\hat{e}_i$)**:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \tag{11}$$

are residuals between observed (empirical) and fitted values of dependent variable.

- **The sum of squared residuals (SSE)**:

$$SSE = \sum_{i}^{N} \hat{e}_i^2 = \sum_{i}^{N} (y_i - \hat{y}_i)^2 . \tag{12}$$

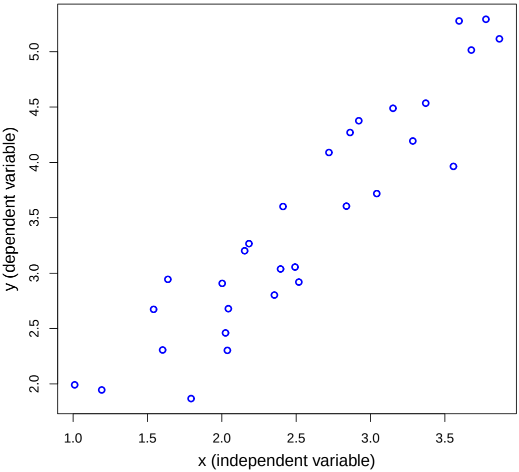
- The SSE can be expressed as function of the parameters β_0 and β_1:

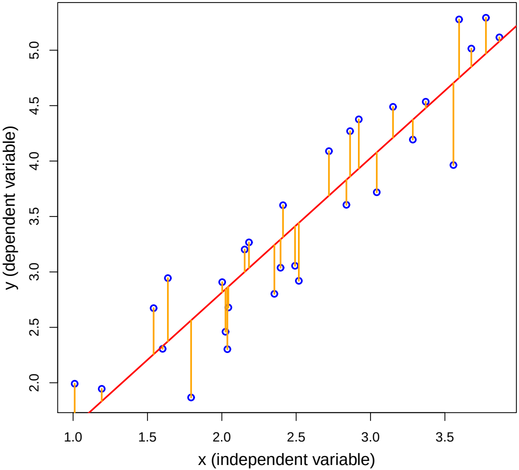
$$SSE(\beta_0, \beta_1) = \sum_{i}^{N} \hat{e}_i^2 = \sum_{i}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 . \tag{13}$$

- **The least squares principle** is a method of the parameter selection that provides the lowest SSE:

$$\min_{\beta_0, \beta_1} \sum_{i}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 . \tag{14}$$

In other words, the least squares principle minimizes the SSE.


y (dependent variable)
2.0
2.5
3.0
3.5
4.0
4.5
5.0
1.0
1.5
2.0
2.5
3.0
3.5
x (independent variable)



The LS estimators minimizes the sum of squared residuals (SSE).

- The least squares estimator for the simple regression model:

$$\hat{\beta}_0^{LS} = \bar{y} - \hat{\beta}_1^{LS}\bar{x}, \tag{15}$$

$$\hat{\beta}_1^{LS} = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^N (x_i - \bar{x})^2}. \tag{16}$$

where $\bar{y}$ and $\bar{x}$ are the sample averages of dependent and independent variables, respectively.

Gauss-Markov Theorem

Under the assumptions A#1-A#5 of the simple linear regression model, the least squares estimators $\hat{\beta}_0^{LS}$ and $\hat{\beta}_1^{LS}$ have the smallest variance of all linear and unbiased estimators of β_0 and β_1.

$\hat{\beta}_0^{LS}$ and $\hat{\beta}_1^{LS}$ are the Best Linear Unbiased Estimators (BLUE) of β_0 and β_1.

1. The estimators $\hat{\beta}_0^{LS}$ and $\hat{\beta}_1^{LS}$ are best when **compared to linear and unbiased estimators**.
   Based on the Gauss-Markov theorem we cannot claim that the estimators $\hat{\beta}_0^{LS}$ and $\hat{\beta}_1^{LS}$ are the **best of all possible estimators**.
2. Why the estimators $\hat{\beta}_0^{LS}$ and $\hat{\beta}_1^{LS}$ are *best*?
   Because they have the minimum variance.
3. The Gauss-Markov theorem holds if assumptions A#1-A#5 are satisfied.
   If not, then $\hat{\beta}_0^{LS}$ and $\hat{\beta}_1^{LS}$ are **not BLUE**.
4. The Gauss-Markov theorem does not require the assumption of normality (A#6)
5. Apart from that, the least squares estimator is consistent if assumptions A#1-A#5 are satisfied.

- The least squares estimator of β_1:

$$\hat{\beta}_1^{LS} = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^N (x_i - \bar{x})^2} \tag{17}$$

can be rewritten as:

$$\hat{\beta}_1^{LS} = \sum_{i=1}^{N} w_i y_i, \tag{18}$$

where $w_i = (x_i - \bar{x}) / \sum (x_i - \bar{x})^2$.

- After manipulation we get:

$$\hat{\beta}_1^{LS} = \beta_1 + \sum_{i=1}^{N} w_i \varepsilon_i. \tag{19}$$

Since the w_i are known this is linear function of random variable (ε).

Unbiasedness

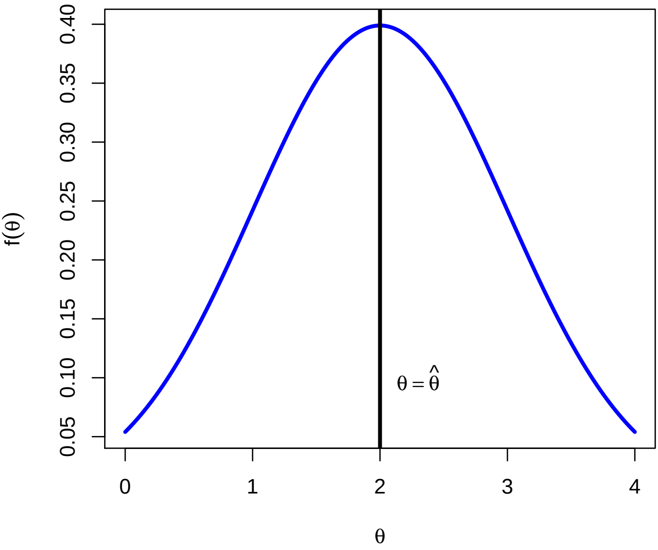
- The estimator is unbiased if its expected value equals the true value, i.e.,

$$\mathbb{E}\left(\hat{\beta}\right) = \beta. \tag{20}$$

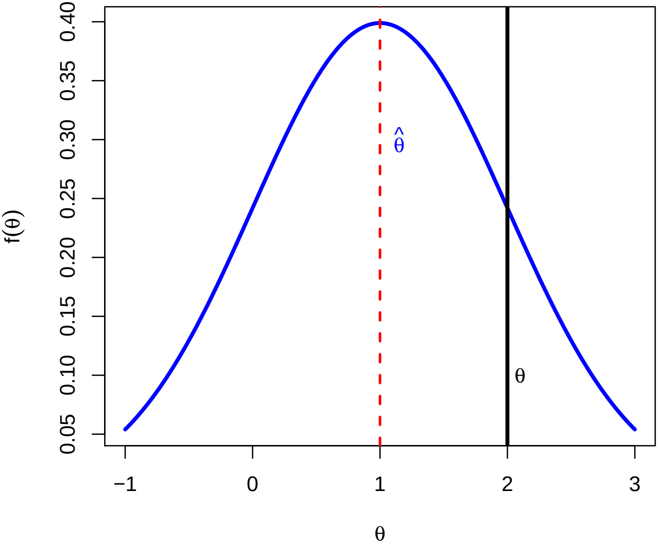
- For the least squares estimator:

$$\begin{aligned}\mathbb{E}\left(\hat{\beta}_1^{LS}\right) &= \mathbb{E}\left(\beta_1 + \sum_{i=1}^{N} w_i \varepsilon_i\right) = \mathbb{E}\left(\beta_1\right) + \mathbb{E}\left(\sum_{i=1}^{N} w_i \varepsilon_i\right) \\ &= \beta_1 + \sum_{i=1}^{N} w_i \mathbb{E}\left(\varepsilon_i\right) = \beta_1.\end{aligned}$$

- In the above manipulation, we take the advantage of two assumption: (i) $\mathbb{E}(\varepsilon_i) = 0$, and (ii) $\mathbb{E}(w_i\varepsilon_i) = w_i\mathbb{E}(\varepsilon_i)$. The latter assumption is equivalent the exogeneity of the independent variable.
- The unbiasedness is mostly about the average of our estimates from many samples (drawn form the same population).


f(θ)
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
1
2
3
4
θ
θ = θ̂

SGH



- In general, variance measures efficiency.
- If the assumption A#1-A#5 are satisfied then:

$$\begin{aligned} var\left(\hat{\beta}_0^{LS}\right) &= \sigma^2\left[\frac{\sum_{i=1}^N x_i^2}{N\sum_{i=1}^N\left(x_i-\bar{x}\right)^2}\right] \\ var\left(\hat{\beta}_1^{LS}\right) &= \frac{\sigma^2}{\sum_{i=1}^N\left(x_i-\bar{x}\right)^2} \\ cov\left(\hat{\beta}_0^{LS},\beta_1^{LS}\right) &= \sigma^2\left[\frac{-\bar{x}}{\left(x_i-\bar{x}\right)^2}\right] \end{aligned}$$

- The greater the variance of the error term (σ^2), i.e., the larger role of the error term, the larger variance and covariance of estimates.
- The larger variability of the dependent variable $\sum_{i=1}^N\left(x_i-\bar{x}\right)^2$, the smaller variance of the least squares estimators.
- The larger sample size (N) the smaller variance of the least squares estimators.
- The larger $\sum_{i=1}^N x_i^2$ the greater variance of the intercept estimator
- The covariance of estimator has a sign opposite to that of $\bar{x}$ and if $\bar{x}$ is larger then the covariance is greater.

- If the assumption of normality is satisfied then:

$$\hat{\beta}_0^{LS} \sim \mathcal{N}\left(\hat{\beta}_0^{LS}, var(\hat{\beta}_0^{LS})\right) \tag{21}$$

$$\hat{\beta}_1^{LS} \sim \mathcal{N}\left(\hat{\beta}_1^{LS}, var(\hat{\beta}_1^{LS})\right) \tag{22}$$

- What if the assumption of normality does not hold?
  If assumptions A#1-A#5 are satisfied and if the sample (N) is sufficiently large, the least squares estimators, i.e., β_0^{LS} and β_1^{LS}, have distribution that approximates the normal distributions described above.

- The variance of the error term:

$$var(\varepsilon_i) = \sigma^2 = \mathbb{E}\left[\varepsilon_i - \mathbb{E}(\varepsilon_i)\right]^2 = \mathbb{E}(\varepsilon_i)^2 \tag{23}$$

  since we have assumed that $\mathbb{E}(\varepsilon_i) = 0$.
- The estimates of the error term variance based on the residuals:

$$\hat{\sigma}^2 = \frac{1}{N-2}\sum_{i=1}^{N} \hat{e}_i^2. \tag{24}$$

  where $\hat{e}_i = y - \hat{y}_i$.
- The $\hat{\sigma}^2$ can be directly used to estimates the variance/covariance of the least squares estimator.

- To obtain estimates of the $var(\hat{\beta}_0^{LS})$ and $var(\hat{\beta}_1^{LS})$ the estimated variance of the error term is used ($\hat{\sigma}^2$):

$$\hat{var}\left(\hat{\beta}_0^{LS}\right) = \hat{\sigma}^2\left[\frac{\sum_{i=1}^N x_i^2}{N\sum_{i=1}^N (x_i-\bar{x})^2}\right]$$

$$\hat{var}\left(\hat{\beta}_1^{LS}\right) = \frac{\hat{\sigma}^2}{\sum_{i=1}^N (x_i-\bar{x})^2}$$

$$\hat{cov}\left(\hat{\beta}_0^{LS},\beta_1^{LS}\right) = \hat{\sigma}^2\left[\frac{-\bar{x}}{(x_i-\bar{x})^2}\right]$$

- Based on the variance we can calculate the standard errors are simply the standard deviation of the estimators:

$$\hat{se}\left(\hat{\beta}_0^{LS}\right) = \sqrt{\hat{var}\left(\hat{\beta}_0^{LS}\right)} \quad \text{and} \quad \hat{se}\left(\hat{\beta}_1^{LS}\right) = \sqrt{\hat{var}\left(\hat{\beta}_1^{LS}\right)}. \tag{25}$$

Multiple regression

- **Multiple regression** :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon \tag{26}$$

where
  - y is the (outcome) dependent variable;
  - $x_1, x_2, \ldots, x_K$ is the set of independent variables;
  - ε is the error term.
- The dependent variable is explained with the components that vary with the **the dependent variable** and **the error term**.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_K$ are the coefficients (slopes) on $x_1, x_2, \ldots, x_K$.

- **Multiple regression** :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon \quad (26)$$

where
  - y is the (outcome) dependent variable;
  - $x_1, x_2, \ldots, x_K$ is the set of independent variables;
  - ε is the error term.
- The dependent variable is explained with the components that vary with the **the dependent variable** and **the error term**.
- β_0 is the intercept.
- $\beta_1, \beta_2, \ldots, \beta_K$ are the coefficients (slopes) on $x_1, x_2, \ldots, x_K$.

$\beta_1, \beta_2, \ldots, \beta_K$ measure the effect of change in $x_1, x_2, \ldots, x_K$ upon the expected value of y (*ceteris paribus*).

General form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon. \tag{27}$$

Matrix form:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{28}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,K} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \ldots & x_{N,k} \end{bmatrix}_{N\times(K+1)},$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}_{(K+1)\times 1}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}_{N\times 1},$$

K – the number of explanatory variables; N – the number of observations.

- **Assumption #1**: true DGP (data generating process):

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \tag{29}$$

- **Assumption #2**: the expected value of the error term is zero:

$$\mathbb{E}\left(\varepsilon\right) = 0, \tag{30}$$

and this implies that $\mathbb{E}\left(y\right) = \mathbf{X}\beta$.

- **Assumption #3**: Spherical variance-covariance error matrix.

$$var(\varepsilon) = \mathbb{E}(\varepsilon\varepsilon') = I\sigma^2 \tag{31}$$

. In particular:
  - the variance of the error term equals σ:

$$var\left(\varepsilon\right) = \sigma^2 = var\left(y\right). \tag{32}$$

  - the covariance between any pair of ε_i and ε_j is zero"

$$cov\left(\varepsilon_i, \varepsilon_j\right) = 0. \tag{33}$$

- **Assumption #4**: **Exogeneity.** The independent variable are **not random** and therefore they are not correlated with the error term.

$$\mathbb{E}(\mathbf{X}\varepsilon) = 0. \tag{34}$$

- **Assumption #5**: the full rank of matrix of explanatory variables (there is no so-called collinearity):

$$rank(\mathbf{X}) = K + 1 \leq N. \tag{35}$$

- **Assumption #6 (optional)**: the normally distributed error term:

$$\varepsilon \sim \mathcal{N}\left(0, \sigma^2\right). \tag{36}$$

- The starting point is the DGP:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{37}$$

- As previously, the least square estimator is obtained by minimizing the sum of squared residuals:

$$\hat{\beta}^{OLS} = \arg\min_{\beta} \mathbf{e}'\mathbf{e}, \tag{38}$$

where $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta}$.

- The SSE can be expressed as a function of unknown parameters:

$$SSE(\beta) = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta), \tag{39}$$

- After manipulating (39) we get:

$$SSE(\beta) = \mathbf{y}\mathbf{y}' - 2\mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta, \tag{40}$$

- The FOC (first order condition) for (40):

$$\frac{\partial SSE(\beta)}{\partial \beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta, \tag{41}$$

Derivation of the least squares estimator II

- after manipulations

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\beta, \tag{42}$$

Finally, using assumption about full rank of $\mathbf{X}$ we get:

$$\hat{\beta}^{OLS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}. \tag{43}$$

- General variance of the least square estimator ($\hat{\beta}^{OLS}$):

$$Var(\hat{\beta}^{OLS}) = \mathbb{E}\left[\left(\hat{\beta}^{OLS} - \beta\right)\left(\hat{\beta}^{OLS} - \beta\right)'\right]. \tag{44}$$

- Let rewrite the least square estimator:

$$\hat{\beta}^{OLS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{X}\beta + \varepsilon\right) = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon, \tag{45}$$

- then

$$\begin{aligned} Var(\hat{\beta}^{OLS}) &= \mathbb{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\right)'\right] \\ &= \mathbb{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbb{E}\left[\varepsilon\varepsilon'\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \end{aligned}$$

- If the assumption #3 about **spherical variance-covariance error matrix**, i.e.. $\mathbb{E}(\varepsilon\varepsilon') = \sigma^2 I$, the above expression can be simplified and written as:

$$Var(\hat{\beta}^{OLS}) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{46}$$

- The variance of the OLS estimator can be calculated with the estimates of the variance of the error term ($\mathbb{S}^2_\varepsilon$):

$$Var(\hat{\beta}^{OLS}) = \mathbb{S}^2_\varepsilon(\mathbf{X}'\mathbf{X})^{-1}, \tag{47}$$

where

$$\mathbb{S}^2_\epsilon = \frac{\mathbf{e}'\mathbf{e}}{N-(K+1)} = \frac{SSE(\hat{\beta}^{OLS})}{df} \tag{48}$$

where $SSE(\hat{\beta}^{OLS})$ is the sum of squared residuals, and df stands for **degree of freedom**.

- Diagonal elements of the variance-covariance matrix (denotes as $\hat{d}_{ii}$) measure the variance of respective parameters. Then **standard error**:

$$\mathbb{S}(\hat{\beta}_i) = \sqrt{d_{ii}}. \tag{49}$$

- **Relative standard errors**

$$\left|\frac{\mathbb{S}(\hat{\beta}_i)}{\hat{\beta}_i}\right|. \tag{50}$$

Gauss-Markov Theorem

Under the assumptions A#1-A#5 of the multiple linear regression model, the least squares estimator $\hat{\beta}^{OLS}$ has the smallest variance of all linear and unbiased estimators of β.

$\hat{\beta}^{OLS}$ is the Best Linear Unbiased Estimators (BLUE) of β.

Asymptotic properties ..

are discussed in Appendix.

References

GREENE, W. (2003): *Econometric Analysis*. Pearson Education.

HILL, R. C., W. E. GRIFFITHS, G. C. LIM, AND M. A. LIM (2012): *Principles of econometrics*, vol. 5. Wiley Hoboken, NJ.

PESARAN, M. H. (2015): *Time Series and Panel Data Econometrics*. Oxford University Press.

WOOLDRIDGE, J. (2008): *Introductory Econometrics: A Modern Approach*. South-Western College Pub.

WOOLDRIDGE, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, no. 0262232588 in MIT Press Books. The MIT Press.

Appendix. Probability Primer

Random variable

is a variable whose value is unknown until it is observed

- **Discrete random variable** – takes only limited and countable numbers of values
- **Indicator random variable** – takes only the values 1 or 0.
- **Continuous random variable** – can take any value.

The probability density function

- **The probability density function (*pdf*)** of random variable summarizes the information concerning the possible outcomes of random variable and the corresponding probabilities.
- The *pdf* discrete random variable X :

$$f(x) = P(X = x), \tag{51}$$

and $\sum_i^k f(x_i) = 1$.

- Because for continuous random variables $P(X = x) = 0$ the *pdf* for continuous random variable can be expressed only for a range of values:

$$P(a \leq X \leq b) = \int_a^b f(x)dx, \tag{52}$$

and $\int_{-\infty}^{\infty} f(x)dx = 1$.

- **The cumulative distribution function (*cdf*)** is an alternative way to represent probabilities. For any value x the cdf:

$$F(x) = P(X \leq x). \tag{53}$$

  - For a discrete random variables, the *cdf* is obtained by summing the *pdf* over all values x_i.
  - For a continuous random variable, $F(x)$ is the area under the *pdf* to the left of the point x.
- The *cdf* is useful in calculating probabilities. For instance:

$$P(X > x) = 1 - P(X \leq x) = 1 - F(x), \tag{54}$$

$$P(a < X \leq b) = F(b) - F(a). \tag{55}$$

- For a set (at least two) of random variables it is useful to analyse **joint distribution**.
- **The joint probability density function** summarize the information concerning the possible outcomes of (at least two) random variables and the corresponding probabilities. For discrete random variables

$$f(x,y) = P(X = x, Y = y), \tag{56}$$

and for continuous random variables,

$$P(a \leq X \leq b, c \leq X \leq d) = \int_a^b \int_c^d f(x,y)dydx. \tag{57}$$

- **The marginal distribution** allows to get distribution of individual random variable:

$$f_X(x) = P(X = x) = \sum_y f(x,y). \tag{58}$$

- **The conditional distribution** is the probability distribution of Y when the value of X is known:

$$f(x|y) = P(Y = y|X = x) = \frac{f(x,y)}{f_X(x)}. \tag{59}$$

- Random variables are independent if and only if they joint *pdf* is the product of the individuals *pdfs*. For, two random variables:

$$f(x,y) = f_X(x) f_Y(y). \tag{60}$$

- **The expected value (expectation)** of a random variable X is a weighted (by probability density) average of all possible outcomes of X.
- For the discrete random variable X, the expected value can be expressed as:

$$\mathbb{E}(X) = \sum_{i=1}^{N} X_i f(x_i), \tag{61}$$

where $f(x)$ is the probability density function of X.

- The expected value can be called **the population mean** ($\neq$ sample average).
- For the continuous random variable X, $\mathbb{E}(X)$ could be defined as:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx. \tag{62}$$

- Properties of the expected value:
  1. For any constant c:
  $$\mathbb{E}(c) = c. \tag{63}$$
  2. For any constants a and b:
  $$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b. \tag{64}$$

3. For any constant $a_1, \ldots, a_k$ and random variables $X_1, \ldots X_k$:

$$\mathbb{E}\left(\sum_i^k a_i X_i\right) = \sum_i^k a_i \mathbb{E}\left(X_i\right), \tag{65}$$

and when $a_i = 1$ for all i then the expected value of the sum is exactly the sum of expected values.

4. For a function creating new random variable $g(\cdot)$:

$$\mathbb{E}\left(g(X)\right) = \sum_i^k g(x_i) f_x(x_i), \tag{66}$$

and for continuous random variable:

$$\mathbb{E}\left(g(X)\right) = \int_{-\infty}^{\infty} g(x_i) f_x(x_i), \tag{67}$$

- In the context of joint distribution, the **conditional expected value** is the expected value of X when the value of Y is known. For discrete random variables:

$$\mathbb{E}(X|Y = y) = \sum_{i=1}^k x_i f(x_i, y). \tag{68}$$

- **The variance** is one the measure of variability:

$$var(X) = \mathbb{E}\left[(X-\mu)^2\right], \tag{69}$$

where $\mathbb{E}(X) = \mu$.

- The variance is usually denoted σ^2 (or σ_X^2 for X).
- Alternatively, the variance can be expressed:

$$\begin{aligned} var(X) &= \mathbb{E}\left(X^2 - 2X\mu + \mu^2\right) = \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) - \mu^2 \\ &= \mathbb{E}\left(X^2\right) - \mu^2. \end{aligned}$$

- For any constants a and b:

$$var(aX+b) = a^2 var(X). \tag{70}$$

- **The standard deviation (sd)** is the square root of the variance is

$$sd(X) = \sqrt{var(X)} \tag{71}$$

- For any constants a and b:

$$var(aX+bY) = a^2 var(X) + b^2 var(Y) + cov(X,Y). \tag{72}$$

- **The covariance** is the measure of association between random variables. For random variables X and Y:

$$cov\left(X,Y\right)=\mathbb{E}\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right], \tag{73}$$

where $\mathbb{E}(Y)=\mu_Y$ and $\mathbb{E}(X)=\mu_X$.

- The covariance between X and Y is sometimes denoted σ_{XY}.
- The covariance can be further expressed as follows:

$$\begin{aligned} cov\left(X,Y\right) &= \mathbb{E}\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]\mathbb{E}\left[X\left(Y-\mu_Y\right)\right] \\ &= \mathbb{E}\left[\left(X-\mu_X\right)Y\right]=\mathbb{E}\left(XY\right)-\mu_X\mu_Y \end{aligned}$$

- Importantly, when X and Y are independent then $cov(X,Y)=0$ and $\mathbb{E}(XY)=\mathbb{E}(X)\mathbb{E}(Y)$.
- Based on the covariance it is hard to assess the magnitude of association between two random variables. However, **the correlation (ρ)** accounts for differences in variances:

$$\rho=\frac{cov(X,Y)}{\sqrt{var(X)}\sqrt{var(Y)}}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}, \tag{74}$$

and $\rho \in < 0,1 >$.

- If X is a normally distributed random variable with mean μ and variance σ^2, i.e. $X \sim \mathcal{N}(0, \sigma^2)$ then the *pdf* of X:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty. \tag{75}$$

- **A standard normal distribution** takes place if $\mu = 1$ and $\sigma^2 = 1$.
- **Standardization**. When $X \sim \mathcal{N}(0, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1). \tag{76}$$

- Example:

$$P(a \leq X \leq b) = F(b) - F(a) = \Phi\left(\frac{X-b}{\sigma}\right) - \Phi\left(\frac{X-a}{\sigma}\right), \tag{77}$$

where $\Phi(\cdot)$ is the *cdf* for standard normally distributed Z.

- Let $Z_1, \ldots, Z_k$ are independent, standard normal random variable. Then, the sum of squared random variable is χ^2 distributed

$$V = Z_1^2 + \ldots + Z_k^2 \quad \sim \chi^2(k), \tag{78}$$

where k denotes the number of degree of freedom. Importantly,

$$\begin{aligned} \mathbb{E}(V) &= k \\ var(V) &= 2k. \end{aligned}$$

- If Z is standard normal random variable and $V \sim \chi^2(k)$ then

$$t = \frac{Z}{\sqrt{V/k}} \sim t_m. \tag{79}$$

Appendix. Asymptotic properties of the OLS estimator

 SGH

- Unbiasedness of estimator:

$$\mathbb{E}\left(\hat{\beta}^{OLS}\right) = \beta. \tag{80}$$

- Using true DGP, i.e., $y = \mathbf{X}\beta + \varepsilon$ we can rewrite the least square estimator:

$$\hat{\beta}^{OLS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(X\beta + \varepsilon\right). \tag{81}$$

- Using $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'X = \mathbf{I}$:

$$\hat{\beta}^{OLS} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon. \tag{82}$$

- The expected value

$$\mathbb{E}\left(\hat{\beta}^{OLS}\right) = \mathbb{E}\left(\beta\right) + \mathbb{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\right], \tag{83}$$

- Using a fact that $\mathbb{E}\left(\beta\right) = \beta$ and assumption that explanatory variables are not random, i.e., $\mathbb{E}\left(\mathbf{X}\right) = X$:

$$\mathbb{E}\left(\hat{\beta}^{OLS}\right) = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbb{E}(\varepsilon), \tag{84}$$

under the assumption that $\mathbb{E}(\varepsilon\mathbf{X}) = 0$ we get

$$\mathbb{E}\left(\hat{\beta}^{OLS}\right) = \beta. \tag{85}$$

- Using previous derivations

$$\hat{\beta}^{OLS} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon. \tag{86}$$

- Let us assume that $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then

$$\hat{\beta}^{OLS} = \beta + \mathbf{A}\varepsilon, \tag{87}$$

- so $\hat{\beta}^{OLS}$ is the linear function of the error term which is the random variable. As a result, the estimator $\hat{\beta}^{OLS}$ is linear.

- Let us introduce with some unbiased linear estimator β, eg. $\hat{\mathcal{B}} = C\mathbf{y}$. Then

$$\mathbb{E}(\hat{\mathcal{B}}) = \mathbb{E}\left(C\mathbf{X}\beta + C\varepsilon\right) = \beta \tag{88}$$

- It can be observed that $C\mathbf{X} = \mathbf{I}$.
- Variance of the estimator $\hat{\mathcal{B}}$:

$$Var(\hat{\mathcal{B}}) = \sigma^2 CC'. \tag{89}$$

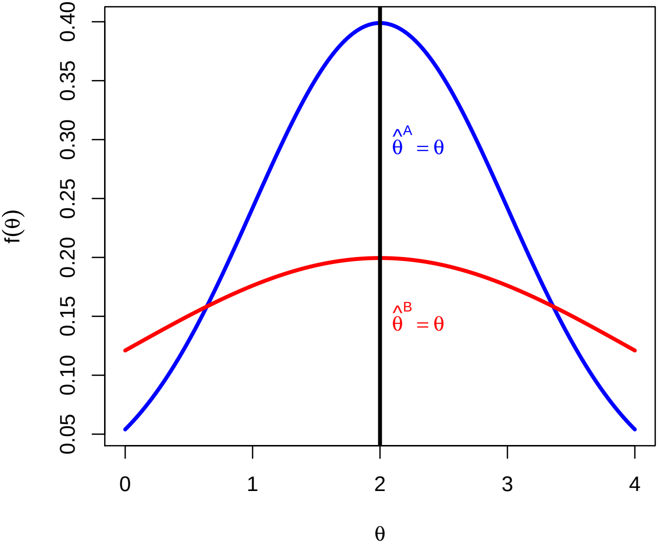
- Let us introduce $D = C - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$.
- Variance of the estimator $\hat{\mathcal{B}}$:

$$Var(\hat{\mathcal{B}}) = \sigma^2\left[\left(D + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\left(D + \left(\mathbf{X}'\mathbf{X}\right)^{-1}X'\right)'\right], \tag{90}$$

- $D\mathbf{X} = 0$ since
  $C\mathbf{X} = \mathbf{I} = D\mathbf{X} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{X}\right)$. The variance can be written as:

$$Var(\hat{\mathcal{B}}) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} + \sigma^2 DD' = Var(\hat{\beta}^{OLS}) + \sigma^2 DD'. \tag{91}$$

- If D is zero matrix then the variance of $\hat{\mathcal{B}}$ is the lowest. But then we consider the least square estimator.


f(θ)
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
1
2
3
4
θ
$\hat{\theta}^A = \theta$
$\hat{\theta}^B = \theta$

Consistency of the OLS estimator I

- Consistent estimator converges in probability to the true value of the unknown parameter.
- Consistency of the OLS estimator

$$\text{plim}_{N\to\infty}\hat{\beta}^{OLS} = \beta \tag{92}$$

- Using previous derivations ($\mathbb{E}\left(\hat{\beta}^{OLS}\right) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\varepsilon)$):

$$\text{plim}_{N\to\infty}\hat{\beta}^{OLS} = \beta + \text{plim}_{N\to\infty}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\varepsilon). \tag{93}$$

- Let us multiply by one, i.e., $1 = 1/N \times N$:

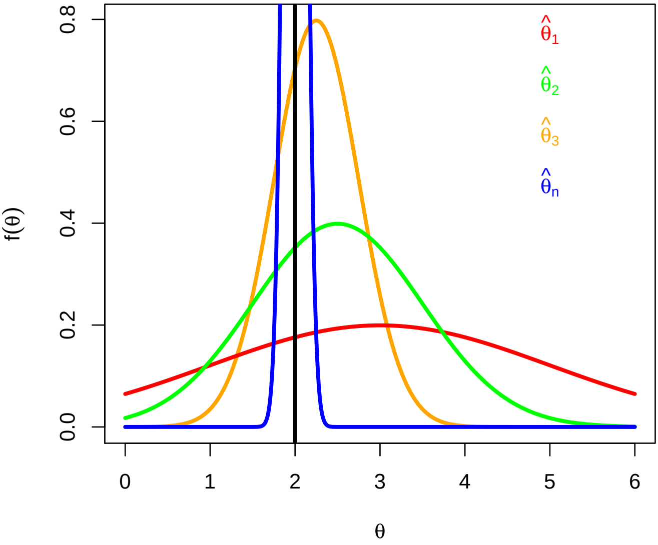
$$\text{plim}_{N\to\infty}\hat{\beta}^{OLS} = \beta + \text{plim}_{N\to\infty}\left(\frac{1}{N}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{N}\mathbf{X}'\mathbb{E}(\varepsilon). \tag{94}$$

- Using the assumption on exogeneity:

$$\text{plim}_{N\to\infty}\frac{1}{N}\mathbf{X}'\mathbb{E}(\varepsilon) = 0, \tag{95}$$

- It is hard to limit the expression $\text{plim}_{N\to\infty}(\mathbf{X}'\mathbf{X})$ but the expression $\text{plim}_{N\to\infty}(1/N\mathbf{X}'\mathbf{X})$ can be limited by some $\mathcal{C}$. Then:

$$\text{plim}_{N\to\infty}\hat{\beta}^{OLS} = \beta + \mathcal{C}\times 0 = \beta. \tag{96}$$


0.0
0.2
0.4
0.6
0.8
f(θ)
0
1
2
3
4
5
6
θ
$\hat{\theta}_1$
$\hat{\theta}_2$
$\hat{\theta}_3$
$\hat{\theta}_n$