Statistical and Econometric Models

Lectures: Lennart Oelschläger · Tutorials: Sebastian Büscher ✉

Problem Set – Week 4 (for tutorial on May 13, 2024)
Solutions

Violations of MLR assumptions

Problem 4.1:

We want to check the behaviour of the OLS estimator when the assumptions MLR.1 - MLR.6 are not exactly matched. For this, we will simulate data from the following underlying process:

$$y_i = 3 + 0.5\, x_{1,i} + 7\, x_{2,i} + 0.003\, x_{2,i}^3 + 0.001\, x_{3,i} + u_i, \quad u_i \sim \mathcal{N}(0, \sigma_i^2), \tag{1}$$

with $\sigma_i^2 = 0.002\, x_1$.
We will, however, assume to not know the true data generating process and fit the model

$$y_i = \beta_0 + \beta_1\, x_{1,i} + \beta_2\, x_{2,i} + v_i, \quad v_i \sim \mathcal{N}(0, \sigma^2). \tag{2}$$

(a) Which assumptions of the classical linear model are violated by this model?

Solution:

- MLR.1 is fulfilled. The omitted terms are included in the error.
- MLR.2 is fulfilled, since we have drawn the data independently from the data generating process.
- MLR.3 is fulfilled, since we do not have multicollinearity:

$$\begin{aligned}\mathbb{E}\boldsymbol{x}\boldsymbol{x}' &= \mathbb{E}\left[\begin{pmatrix}1\\ \boldsymbol{x}_1\\ \boldsymbol{x}_2\end{pmatrix}(1 \quad \boldsymbol{x}_1 \quad \boldsymbol{x}_2)\right] = \mathbb{E}\left[\begin{pmatrix}1 & \boldsymbol{x}_1 & \boldsymbol{x}_2\\ \boldsymbol{x}_1 & \boldsymbol{x}_1^2 & \boldsymbol{x}_1\boldsymbol{x}_2\\ \boldsymbol{x}_2 & \boldsymbol{x}_1\boldsymbol{x}_2 & \boldsymbol{x}_2^2\end{pmatrix}\right]\\ &= \begin{pmatrix}1 & \mathbb{E}\boldsymbol{x}_1 & \mathbb{E}\boldsymbol{x}_2\\ \mathbb{E}\boldsymbol{x}_1 & \mathbb{E}\boldsymbol{x}_1^2 & \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_2)\\ \mathbb{E}\boldsymbol{x}_2 & \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_2) & \mathbb{E}\boldsymbol{x}_2^2\end{pmatrix} = \begin{pmatrix}1 & 2 & 0\\ 2 & 13/3 & 0\\ 0 & 0 & 9\end{pmatrix}\end{aligned}$$

  This is positive definite (you can check the eigenvalues by hand or in R), so $\mathbb{E}\boldsymbol{x}\boldsymbol{x}' > 0$.
- MLR.4 is violated, because the error in model (2) will depend on $\boldsymbol{x}_2^3$, so one can expect $\mathbb{E}(\boldsymbol{v} \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = 0.003\boldsymbol{x}_2^3$.
- MLR.5 is violated, because $\text{Var}(\boldsymbol{v} \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \sigma^2\boldsymbol{x}_1$, so it varies with the regressor $\boldsymbol{x}_1$.
- MLR.6 is violated (see reasons for violation of MLR.4 and MLR.5).

(b) Draw 500 times from the data generating process (1). For this, draw the values of the regressors randomly, such that x_1 is drawn from a uniform distribution on the interval $[1, 3]$, x_2 is drawn from a normal distribution with mean zero and standard deviation three, and x_3 is drawn from a normal distribution with mean zero and standard deviation one.

(c) Estimate model (2) on the simulated data to obtain the estimates of the parameters and the standard deviations of the estimates.

(d) Repeat (b) and (c) 10000 times and save the estimates you obtain every time. Compare the mean on the 10000 estimates with the true parameters. Compare the standard deviation of the 10000 estimates of each parameter with the standard deviation of the estimators you calculated in (c).

(e) Perform a RESET-Test with $p = 3$ on the model estimated in part (c). Does it detect the misspecification of the model?

(f) Add now x_2^2 and x_2^3 to the model (2) and fit it to the data.

(g) Repeat (d) for this new model and save the standard deviations of the 10000 obtained estimates.

(h) Test the model obtained in (e) for heteroskedasticity. You can use a Breusch-Pagan test or a White test.

(i) Based on the test you used to detect heteroskedasticity, calculate the Feasible Generalised Least Squares estimator. To do so, calculate the matrix $\boldsymbol{L}$ with $\boldsymbol{L}^2 = \texttt{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_N^2)$ and estimate the rewritten model

$$\boldsymbol{L}^{-1}\boldsymbol{y} = \boldsymbol{L}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{L}^{-1}\boldsymbol{u}.$$

(j) Calculate the Calculate the heteroskedasticity-consistent White estimator for the variance of the OLS estimator

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = (X'X)^{-1}X' \,\mathrm{diag}(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_N)X(X'X)^{-1}.$$

(k) Compare the estimates and their standard deviations of the FGLS estimation and the White estimator with those obtained in (f) and to the standard deviations of the 10000 estimates obtained in (g).

Instrumental Variables

Problem 4.2:

In the linear regression framework $y = X\beta + \varepsilon$, a central assumption states $E(\varepsilon \mid X) = 0$.

(a) Show that this assumption implies $\mathrm{Cov}(\varepsilon, X) = 0$ (this is referred to as "exogeneity" of X).

Solution: Using the law of iterated expectation:

$$\mathbb{E}(X'\varepsilon) = \mathbb{E}(\mathbb{E}(X'\varepsilon \mid X)) = \mathbb{E}(X'\mathbb{E}(\varepsilon \mid X)) = 0.$$

This implies

$$\mathbb{E}(\varepsilon) = \mathbb{E}(\mathbb{E}(\varepsilon \mid X)) = \mathbb{E}(\varepsilon) = 0,$$

and hence

$$\begin{aligned}\text{Cov}(\varepsilon, X) &= \mathbb{E}\left((\varepsilon - \mathbb{E}(\varepsilon))(X - \mathbb{E}(X))\right) \\ &= \mathbb{E}(X'\varepsilon) - \mathbb{E}(X')\mathbb{E}(\varepsilon) = 0.\end{aligned}$$

(b) In the opposite case ("endogeneity"), X is correlated with ε. In this case, the OLS estimator $\hat{\beta}_{\text{OLS}}$ is biased. As a remedy, a technique called "instrumental variables estimation" was introduced in the lecture. What are "instrumental variables"?

Solution: Instrumental variables Z (a matrix with the same dimension as X) are observable regressors that fulfil two properties:

- Relevance: $\text{Cov}(X, Z) \neq 0$
- Exogeneity: $\text{Cov}(\varepsilon, Z) = 0$

(c) The instrumental variables estimator for β is defined as

$$\hat{\beta}_{\text{IV}} = (Z'X)^{-1}Z'y.$$

Derive this equation by multiplying the equation $y = X\beta + \varepsilon$ by Z' on both sides, assuming that in the DGP it holds that $Z'\varepsilon = 0$ and $Z'X$ is invertible.

Solution:

$$\begin{aligned} & y = X\beta + \varepsilon \\ \Rightarrow\ & Z'y = Z'X\beta + Z'\varepsilon \approx Z'X\beta && \mid Z'X \text{ invertible} \\ \Rightarrow\ & \hat{\beta}_{\text{IV}} = (Z'X)^{-1}Z'y \end{aligned}$$

(d) Consider the model

$$\log(y) = x'\beta + \varepsilon,$$

where y is the increase in sales of a company compared to last year and x are their expenditures for marketing. One could argue that this is a case of endogeneity. Why? Would the number of employees be a valid instrument? Is there a problem in using the age of the CEO as an instrument?

Solution: Usually, $\log(y)$ is influenced by additional factors that would be contained in ε. An example could be the sector in which the company participates, say solar systems. But the sector in general is correlated with x (more marketing in certain sectors than in others). Hence, $\text{Cor}(\varepsilon, x) \neq 0$ most likely.

The number of employees could be a valid instrument, because it can be assumed to not have a direct influence on $\log(y)$ but on x (companies with more employees usually have a larger budget for marketing).

The age of the CEO would be a bad instrument, because it is not relevant.