Advanced Applied Econometrics
Exercises

Exercise 1. Let assume that the random variable X_i is drawn from the normal distribution with mean μ and variance σ^2. Consider the following estimator of the mean of variable:

$$\hat{\theta}_N = \frac{1}{N} + \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{1}$$

(i) Check whether the above estimator is unbiased.

(ii) Check whether the above estimator is consistent.

Exercise 2. Let assume that the random variable X_i is drawn from the normal distribution with mean μ and variance σ^2. Conisder the following estimator of the mean of variable X:

$$\hat{\theta}_N = X_N. \tag{2}$$

(i) Check whether the above estimator is unbiased.

(ii) Check whether the above estimator is consistent.

Exercise 3. Consider the following model explaining wages (w):

$$w = \beta_0 + \beta_1 educ + \varepsilon, \tag{3}$$

where the *educ* is the number of years of education attainment, ε is the error term and β_0 and β_1 are the unknown parameters.

(i) What signs would you expect on the coefficients β_0 and β_1.

(ii) Analyze the conditional empirical distribution (or histogram) of wages for low- and high-educated workers. Are there any differences? Can you explain why?

(iii) Using dataset `cps.dta` plot wages versus education attainment. Do you observe any empirical pattern?

(iv) Calculate correlation and covariance between variables of interest, i.e., wages and educational attainment. If it is possible interpret obtained numbers.

(v) Based on the previous calculations and respective sample means and variances calculate the least square estimators of β_0 and β_1. Are they in line with your expectations?

(vi) Use an appropriate built-in command to get least squares estimates of β_0 and β_1. Interpret them. Plot the regression line.

(vii) Consider now the extended model:

$$w = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 female + \varepsilon, \tag{4}$$

where *exper* is the number of years of professional experience and $female$ is the dummy variable for women. What signs would you expect on the coefficients β_2 and β_3.

(viii) Estimate the underlying parameters of the extended model and interpret them.

Exercise 4. Consider the following data generating process for the variable y:

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

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

(i) Provide the following simulation study.

- Assume that $\beta_0 = \beta_1 = 1$. Besides, consider the case when $\sigma = .5$ while $x \sim \mathcal{N}(2, 1)$. Based on the definition of the DGP simulate data ($N = 100$) and estimate the underlying parameters, i.e., β_0, β_1.
- Repeat previous step 1000 times.
- Plot and discuss the obtained distribution of estimates on β_1.

(ii) Calculate variance of the distribution. Try to explain analytically obtained number.

Exercise 5. Let assume that the temperature in July is random variable that is normally distributed with the mean 20 and the variance 4.

(i) Test whether the temperature in July is higher than 30 degrees.

(ii) What is the probability that temperature in July equals exactly 21 degrees.

(iii) Calculate probability of the temperature between 19 and 23 degrees.

(iv) Provide 99% confidence intervals for the temperature in July.

Exercise 6. Consider the following estimated production function:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 k + \hat{\beta}_2 l, \tag{6}$$

where y is the logged output, k is the logged capital input and l is the logged labor. The underlying estimates are as follows:

$$\hat{\beta} = \begin{bmatrix} 0.1 \\ 0.21 \\ 0.62 \end{bmatrix} \quad \text{and} \quad var\left(\hat{\beta}\right) = \begin{bmatrix} 0.01 & -0.002 & -0.003 \\ -0.002 & 0.01 & -0.004 \\ -0.003 & -0.004 & 0.025 \end{bmatrix}. \tag{7}$$

In addition, it is known that the sample consist of 44 observations.

(i) Interpret all estimates.

(ii) Test significance of the above estimates. Provide also the p-values.

(iii) Calculated a 95% confidence intervals for β_2.

(iv) Test whether the parameter β_2 is larger than 0.5. Calculate the corresponding p-value.

(v) Test whether $\beta_1 = \beta_2$.

(vi) Based on the above estimates try to assess whether returns to scale are constant, diminishing or rising.

Exercise 7. Consider the following model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon, \tag{8}$$

where ε is the error term.

(i) Is the relationship between y and x monotonic?

(ii) Find the extremum point of the above relationship.

(iii) Assume that estimated variance of the parameters is known. Using the delta method calculate the asymptotic variance for the derived extremum point from previous point.

Exercise 8. Consider the following model explaining the logged wages ($\log w$):

$$\log w = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female + \varepsilon, \tag{9}$$

where the *educ* is the number of years of education attainment, *exper* is the number of years of professional experience and *female* is the dummy variable for women and ε is the error term.

(i) Using dataset `cps.dta` estimate underlying parameters and interpret them.

(ii) Test significance of estimates.

(iii) Are returns to education above 10%.

(iv) Calculate 99% confidence intervals for gender wage gap.

(v) Test whether tertiary education can mitigate effect of gender wage gap.

(vi) Test whether experience moves wages.

(vii) Provide confidence intervals for number of year of experience at which wages are the highest or the lowest.

Exercise 9. Consider the following model explaining the logged wages ($\log w$):

$$\log w = \beta_0 + \beta_1 educ + \varepsilon, \tag{10}$$

where the *educ* is the number of years of education attainment and ε is the error term.

(i) Using dataset `cps.dta` draw with replacement 4733 observations. Estimate the underlying parameters. Store the obtained estimate on β_1.

(ii) Repeat the above step 10000 times. Count in how many cases estimates are below 0.1.

(iii) Using dataset `cps.dta` estimate (without sampling/resampling) underlying parameters. Calculate the probability value of the null that β_1 is less than 0.1 Compare the results with numbers from previous point.

Exercise 10. Consider simply regression model. Show that the R^2 equals squared correlation coefficient.

Exercise 11. Consider the following linear model:

$$y_{\mathcal{I}} = \beta_{0,\mathcal{I}} + \beta_{1,\mathcal{I}} x_{\mathcal{I}} + \varepsilon_{\mathcal{I}} \tag{11}$$

where $\mathcal{I} \in \{1, 2, 3, 4\}$.

(i) Using the dataset `Anscombe.dta` estimate the linear regression (11) for each $\mathcal{I}$. Interpret estimated coefficients and its significance.

(ii) Interpret the coefficient of determination (R^2) in all model. Compare the R^2 between models. Which model is *the best*?

(iii) Use RESET to test for misspecification for all considered models.

(iv) Plot residuals versus dependent variable in all cases. Can you explain the result of the RESET test?

(v) For all considered models plot the fitted regression line along with the data scatter. Try to explain once again the results of the RESET test. Would you change your opinion (that based on the R^2) from the point (ii)?

Exercise 12. Consider the following models explaining the wages (w):

$$w = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female + \varepsilon, \tag{12}$$

or the logged wages ($\log w$):

$$\log w = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female + \varepsilon, \tag{13}$$

where the *educ* is the number of years of education attainment, *exper* is the number of years of professional experience and *female* is the dummy variable for women and ε is the error term.

(i) Using dataset `cps.dta` estimate the underlying parameters of both models. Interpret the implied marginal effects and elasticities.

(ii) Plot histograms for both considered models. In which case the assumption about normality of the error term is not satisfied?

(iii) Test the normality of the error-term.

(iv) Plot histograms of outcome variable in both considered models. Discuss the results. Based on that results try to explain previous points.

(v) Test collinearity of explanatory variables. Explain source of collinearity in this case. Are implications of collinearity observable here?

(vi) Extend the log-linear model by interaction between *educ* and *female*, i.e,

$$\log w = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female + \beta_5 female \times educ + \varepsilon, \tag{14}$$

and estimate the underlying parameters. Calculate the marginal effect of education and female. Interpret the obtained values.

(vii) Repeat the above step for the interaction between education and experience.

Exercise 13. Consider the following model explaining the logged wages (w):

$$\log w = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female + \varepsilon, \tag{15}$$

where the *educ* is the number of years of education attainment, *exper* is the number of years of professional experience and *female* is the dummy variable for women and ε is the error term.

(i) Do you expect presence of heteroskedasticity of the error term in the above model?

(ii) Using dataset `cps.dta` estimate the underlying parameters of both models. Calculate the squared residuals and plot them versus each explanatory variable. Do you observe heteroskedasticity in this case? Please provide economic interpretation.

(iii) Use the Goldfeld-Quandt test to test heteroskedasticity of the error term. Group sample into females and males.

(iv) Repeat previous point but divide sample in the educational group.

(v) Use the White to test heteroskedasticity of the error term.

(vi) Apply (15) to the least squares estimator with the HC/White robust standard errors. Discuss differences/ a lack of differences.

(vii) Assume that the variance of the error term depends on education attaitment and estimate the underlying parameters with the weighted least squares estimator.

(viii) Apply the feasible least squares estimator to (15) assuming that the variance of the error term can be explained by the same set of explanatory variable. Compare all results and discuss differences.

Exercise 14. Consider the relationship between the inflation (π_t) and the unemployment rate ($\mathcal{U}_t$):

$$\pi_t = \beta_0 + \beta_1 \mathcal{U}_t + \varepsilon_t, \tag{16}$$

where ε is the error term.

(i) What signs would you expect on the coefficients β_0 and β_1.

(ii) Using dataset `USPhillipsCurve.dta` estimate the underlying parameters. Interpret estimates. Discuss the significance of obtained estimates.

(iii) Interpret the R^2 coefficient of determination.

(iv) Plot residuals and discuss whether they are serially correlated.

(v) Using various statistical test test whether residuals are serially correlated.

(vi) Plot dependent variable and try to explain your previous findings about serial correlation of the error term.

(vii) Use the Newey-West HAC standard errors and compare obtained results with the baseline estimates.

(viii) Use appropriate FGLS estimator and compare the results with previous estimates.

(ix) Discuss credibility of all estimates. What is the (empirical) relationship between the inflation and unemployment rate?

Exercise 15. Consider the following model explaining the logged wages (w):

$$\log w = \beta_0 + \beta_1 educ + \varepsilon, \tag{17}$$

where the *educ* is the number of years of education attainment, *exper* is the number of years of professional experience and *female* is the dummy variable for women and ε is the error term.

(i) Using dataset `mroz.dta` estimate the return to education. Discuss the reliability of this estimate. Do you think that the estimated return to education is underestimated or overestimated?

(ii) Explain why *educ* might be endogenous variable in this case.

(iii) Use the mother's education level as IV (instrumental variable). Discuss whether this IV might be relevant and exogenous. Estimate the first stage regression and test whether mother's education level is irrelevant IV.

(iv) Run the second stage regression and discuss the differences between IV and OLS estimates.

(v) Test endogeneity of the OLS estimates.

(vi) Extend set of instrumental variables by the father's education. Reestimate the parameters. Are they different from the previous 2SLS estimates? Re-run the weak instruments test and Hausman test and compare their results.

(vii) Run the Sargan test and discuss the results.

(viii) Extend structural equation (17) by experience and its squares. Reestimate parameter with the least squares. Discuss the reliability of the estimated return to education.

(ix) Estimate extended regression with the 2SLS using mother's and father's education as IV. Discuss differences with previous points.

(x) Test endogeneity of education in extended model and overidentifying restrictions.

(xi) Check whether the error term from previous point is heteroskedastic. Reestimate extended regression with robust (to the heteroskedasticity of the error term) standard error and run the diagnostic test for IV estimation.

(xii) Consider now the number of siblings as IV. Discuss whether this IV is relevant and exogenous. Test its relevance.

Exercise 16. Consider demand on cigarettes:

$$\ln C = \beta_0 + \beta_1 \ln P + \beta_2 \ln Y + \varepsilon \tag{18}$$

Where $\ln C$ is the logged consumption of cigarettes, $\ln P$ is the logged price of cigarettes, $\ln Y$ is the logged income and ε is the error term.

(i) What signs would you expect on the coefficients β_1 and β_2.

(ii) Using dataset `cig85_95.dta` calculate the variables of interest. Using ten year changes estimate the demand function. Interpret estimates.

(iii) Discuss the reliability of the above estimates.

(iv) Explain why price might be endogenous variable in this case.

(v) Consider the taxes on cigarettes as instrumental variables for price. Discuss why the tax could be relevant and exogenous IV here.

(vi) Reestimate parameters for the demand function. Compare results and discuss differences.

(vii) Test endogeneity of the initial OLS estimates and relevance of used instrumental variable.

(viii) Consider now the sales taxes as instrumental variable. Discuss why this variable could be relevant and exogenous IV here.

(ix) Reestimate parameters for the demand function. Compare results and discuss differences. Test endogeneity of the initial OLS estimates and relevance of used instrumental variable.

(x) Use both instrumental variables. Reestimate parameters. Compare results and discuss differences. Test endogeneity of the initial OLS estimates, relevance of used instrumental variable and over-identifying restrictions.

Exercise 17. Consider the following equations describing the international trade flows:

$$\begin{aligned} im_t &= \beta_0 + \beta_1 dd_t^{PL} + \beta_2 reer_t + \varepsilon_t, \quad (19)\\ ex_t &= \gamma_0 + \gamma_1 dd_t^{EU} + \gamma_2 reer_t + \eta_t, \quad (20)\end{aligned}$$

where im_t is the logged imports in Poland, dd_t^{PL} is the logged domestic demand in Poland, $reer_t$ is the logged real effective exchange rated deflated by CPI and an increase in the $reer_t$ is related to appreciation, ex_t is the logged exports, dd_t^{EU} is the logged domestic demand in the EU countries while ε_t and η_t are the error terms.

(i) Start with the demand function for imported goods (19). What signs would you expect on the coefficients β_1 and β_2. Why?

(ii) Using dataset `InternationalTradePoland.dta` estimate the underlying parameters in (19). Interpret obtained values. Are they in line with you expectations. Discuss their reliability.

(iii) Move to the demand function for exported goods (20). What signs would you expect on the coefficients γ_1 and γ_2. Why?

(iv) Estimate the underlying parameters in (20). Interpret obtained values. Are they in line with you expectations. Discuss their reliability.

(v) Discuss whether fluctuations in real exchange rate have stronger impact on imports or exports?

(vi) Combine (19) and (20) in system of equations and test whether fluctuations in real exchange rate have symmetric effect on trade flows in Poland.

(vii) Consider now the following extension:

$$\begin{aligned} im_t &= \beta_0 + \beta_1 dd_t^{PL} + \beta_2 reer_t + \beta_3 ex_t + \varepsilon_t, \quad (21)\\ ex_t &= \gamma_0 + \gamma_1 dd_t^{EU} + \gamma_2 reer_t + \gamma_3 im_t + \eta_t, \quad (22)\end{aligned}$$

What signs would you expect on the coefficients γ_3 and β_3. Why?

(viii) Why the parameters of the system described by (21)-(22) cannot be estimated credibly with the OLS estimator?

(ix) Are parameters in (21)-(22) identified?

(x) Estimate underlying parameters of (21)-(22) with the 2SLS estimator. Compare results with previous estimates. Try to explain differences.

(xi) Test once again whether fluctuations in real exchange rate affect symmetrically trade flows.

(xii) Estimate underlying parameters of (21)-(22) with the 3SLS estimator. Are there any reasons to use the 3SLS estimator instead of 2SLS estimator? Compare results with previous estimates. Try to explain differences.

Exercise 18. Use dataset `InternationalTradePoland.dta` and answer the following questions.

(i) Start with the logged series of imports. Discuss whether this variable could be stationary.

(ii) Plot the logged imports and judge whether this series is stationary or has a unit root.

(iii) Run the ADF test and interpret its result.

(iv) Assess the order of integration of the logged imports.

(v) Is the logged imports trend-stationary (stationary around deterministic trend)?

(vi) Regress the logged imports on linear trend and analyze the residuals. Are they mean-reverting?

(vii) Consider now the logged real effective exchange rate. Based on you economic knowledge discuss whether this varriable should be stationary.

(viii) Plot the logged real effective exchange rate and discuss whether this macroeconomic variable looks like stationary series.

(ix) Perform the ADF test for the logged real effective exchange rate. Interpret the results.

(x) Trim the sample and consider period after 2000. Plot once again series and run the ADF test. Are the results different from ones that you obtained in previous point?

(xi) Consider now the relationship between the logged imports and the logged domestic demand in Poland. Are these variables cointegrated?

(xii) Turn to the export demand. Is there long-run relationship between the logged exports and the logged domestic demand in the EU?

Exercise 19. Consider the quarterly inflation rate inf_t:

$$\pi_t = \ln CPI_t - \ln CPI_{t-1}, \tag{23}$$

where CPI_t is the consumer price index.

(i) Discuss whether the inflation rate should be stationary. Plot this series and try to judge whether it is stationary. Run the ADF test.

(ii) Consider now the autoregressive model of order P for inflation rate:

$$\pi_t = \mu + \sum_{i=1}^{P} \rho_i \pi_{t-i}, \tag{24}$$

and start with $P = 1$. Using the dataset estimate the underlying parameters. Interpret the implied half-life. Plot impulse response function (IRF) for the inflation. Test the serial correlation of the error term.

(iii) Consider now the AR(2) model for inflation. Estimate the parameters of (24). Plot IRF and test serial correlation.

(iv) Consider now the AR(3) and AR(4) model. Estimate the parameters, test serial correlation and plot IRF. Compare results with the previous points.

Exercise 20. Consider the relationship between the inflation (π_t) and the unemployment rate ($\mathcal{U}_t$):

$$\pi_t = \mu + \beta_0 \mathcal{U}_t + \varepsilon_t, \tag{25}$$

where ε is the error term.

(i) Using dataset `USPhillipsCurve.dta` estimate the underlying parameters. Interpret estimates. Discuss the significance of obtained estimates. And test serial correlation of the error term.

(ii) Consider now the ADL model:

$$\pi_t = \mu + \sum_{i=1}^{P} \rho_i \pi_{t-i} + \sum_{j=1}^{D} \beta_j \mathcal{U}_{t-j} + \varepsilon_t, \tag{26}$$

and estimate the parameters if $P = 1$ and $D = 0$. Estimate and interpret the short- and long-run multipliers. Test serial correlation.

(iii) Extend subsequently the autoregressive order P, i.e., $P \in \{2, 3, 4\}$. Estimate the underlying parameters. Estimate and interpret the short- and long-run implied multipliers. Each time test serial correlation. Which estimates are the most credible?

(iv) Consider now the cases when $P = 4$ while $D \in \{1, 2\}$. Once again estimate the underlying parameters and calculate the short- and long-run multipliers. Compare the results.

Exercise 21. Consider the error correction model (ECM) for the first difference of logged imports (Δim_t):

$$\Delta im_t = \delta e_{t-1} + \varepsilon_t, \tag{27}$$

where e_{t-1} is the error correction term. Assume that there is a long-run relationship between the logged imports and logged domestic demand in Poland.

(i) Using dataset `InternationalTradePoland.dta` estimate the parameter δ. Calculate and interpret half-life. Test the serial correlation of the error term. Use both methods of estimating the ECM model. Calculate and interpret the long-run elasticity of imports to changes in domestic demand.

(ii) To eliminate the serial correlation add subsequently next lags of dependent variable in (27). Include also Δdd_t. Test serial correlation. And compare the robustness the error correction mechanism. Compare the long-run elasticites.

(iii) Turn to the demand for exported goods. Assume the long-run relationship between logged exports (ex_t) and the logged domestic demand in the EU economies. Estimate the following error correction model:

$$\Delta ex_t = \mu + \phi_1 ex_{t-1} + \phi_2 dd_{t-1}^{EU} + \varepsilon_t, \tag{28}$$

and calculate and interpret half-life. Test the serial correlation of the error term. Calculate and interpret the long-run elasticity of exports to changes in domestic demand in the EU economies.

(iv) Extend (28) by Δex_{t-1} and Δdd_t^{EU} and check whether your findings from previous points are still valid.