ECOM-G314 Econometrics 1
Homework Assignment 1

This homework assignment will be discussed in the exercise session on Wednesday 15 November (groups at 10.15am and 12.15pm) in the seminar room 3–4 at Economicum. Please submit your solution **by 9.45 a.m. on Wednesday 15 November**.

Peer review and self-assessment should be done by **Monday 20 November at 6 p.m.** at the latest. Please note that the peer review and self-assessment are compulsory, and a prerequisite for gaining points from your submission.

Tutorials will be held on Mondays 6 November and 13 November at 2.15pm in the Economicum lecture hall. You can ask the TA for help with the the homework assignments and discuss the assignments with other students. If you have any questions, please contact the TA via email at heikki.korpela@helsinki.fi.

The share of each exercise of the maximum number of points from the assignment is given in brackets.

When asked for an interpretation of any results, a purely statistical interpretation is sufficient. You do not need to refer to economic theory or make statements about causality. Examples will be discussed in the tutorial session.

Please return your submission in Moodle as one PDF file. It is not strictly necessary to return the code used, but if there are errors in your results, the code may be helpful in deciding whether you've made a fundamental or a minor mistake.

The peer review is anonymous. For this reason, please do **not** include your name or student ID in your submission, in the filename or the file description.

1. Consider the following total cost function of firm i: [15%]

$$TC_i = \beta_1 + \beta_2 \log(Q_i) + \beta_3 \log(p_{i1}) + \beta_4 \log(p_{i2}) + \beta_5 \log(p_{i3}) + \varepsilon_i,$$

where TC_i is the total cost function and Q_i is the total output of firm i, and p_{i1}, p_{i2} and p_{i3} are the prices of labour, fuel and capital, respectively.

   (a) The file `NERLOVE116.txt` contains the observations of the 116 largest electric utility companies in 1955 of the following variables: total costs in millions of dollars (column 1), output in billions of kilowatt hours (column 2), price of labour (column 3), price of fuel (column 4), and price of capital (column 5); all prices are in levels. Estimate the model by OLS, test for the significance of each of the regression coefficients, and interpret the estimation and test results.

   References: code for "wages" (week 2), slides "Linear regression", "Testing hypotheses under the normality assumption" (week 1), "Interpretation of linear regression model" (week 2), or book chapters 2.2, 2.5.1, 3.1. Read the model specification carefully.

   (b) Linear homogeneity in factor prices implies that $\beta_3 + \beta_4 + \beta_5 = 1$. Test this restriction by (i) the F-test and (ii) the Wald test.

References: code for "wages" (week 2), slides "Testing hypotheses under the normality assumption" (week 1, for F-test) and "Asymptotic properties of the OLS estimator" (week 2, for Wald test), or book chapter 2.5.6. The `linearHypothesis` function from R's `car` package may be helpful.

2. (Verbeek, Exercise 3.2) For this exercise we use data on sales, size and other characteristics of 400 Dutch men's fashion stores in the file `clothing2.xlsx`. The goal is to explain sales per square metre (sales) from the characteristics of the shop. [35%]

   (a) Estimate a linear model (model A) that explains sales from total number of hours worked (hoursw), shop size (ssize) and a constant. Interpret the results.

   (b) Test whether the number of owners (nown) affects shop sales, conditional upon hoursw and ssize.

   (c) Test whether further including of the number of part-time workers (npart) improves the model from (b).

   (d) Estimate a linear model (model B) that explains sales from the number of owners, full-time workers (nfull), part-time workers and shop size. Interpret the results.

References (for a–d): same as 1(a).

   (e) Compare the original model A from (a) and model B on the basis of $\bar{R}^2$ and AIC. [Hint: The easiest way to compute the AIC in R is to use the `extractAIC()` function.]

References: slides "Specification of regression model" (week 2, see slide 4 in particular) and book chapter 3.2.2.

   (f) Perform a non-nested F-test of model A against model B. Perform a non-nested F-test of model B against model A. What do you conclude?

References: slides "Specification of regression model" (week 2, see slide 6 in particular) and book chapter 3.2.3.

   (g) Include the numbers of full-time and part-time workers in model A to obtain model C. Estimate this model. Interpret the result and perform a RESET test with $Q = 2$. Are you satisfied with this specification?

References: slides "Specification of regression model" (week 2, see slide 7 in particular) and book chapter 3.3.2.

3. (Adapted from Verbeek, Exercie 3.3) The file `housing2.xlsx` contains the sale prices and a number of characteristics of 546 houses sold in 1987 in Windsor, Canada. [25%]

(a) Regress the logarithmic sale price (logprice) on a constant, the log of lot size (lotsize), the number of bedrooms (bedrooms), the number of bathrooms (bathrooms) and the indicator variable airco that takes value 1 if the house has air conditioning and 0 otherwise. Interpret the result.

References: same as 1(a), plus book 2.1.3 (regarding dummy variables). Interpretation of both dummies and (partly) logarithmic models are discussed on the slides 5 and 7–8 of the slide set on "Interpretation of linear regression model" (week 2), and on pages 63–64 of chapter 3.1 of the book.

(b) Create four dummy variables relating to the number of bedrooms, corresponding to categories of having 2 or less, 3, 4 and 5 or more bedrooms. Estimate a model for log prices that includes log lot size, the number of bathrooms, the air conditioning dummy and three of the dummies (don't include the dummy for having 3 bedrooms). Interpret the result. Why is this model not nested in the specification in (a)?

References: See the notes on dummies in (a). Definition of nested models is given on slide 6 of slides "Specification of regression model" (week 2) and in book in chapter 3.2.3.

(c) Perform two non-nested F-tests to test these two specifications against each other. What do you conclude?

References: slides "Specification of regression model" (week 2, see slide 6 in particular) and book chapter 3.2.3.

(d) Include all four dummies in the model estimated in (b). What happens? Why?

References: book chapter 3.6.2, p. 90.

4. Consider the simple linear regression model [25%]

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Let $\beta_1 = \beta_2 = 1.0$. Assume that x_i is independently normally distributed with mean zero and variance σ_X^2, and the error term ε_i is independently normally distributed with mean zero and variance σ_ε^2.

(a) Let $\sigma_X^2 = \sigma_\varepsilon^2 = 1$. Consider three sample sizes, $N = 25$, $N = 100$ and $N = 1000$. For each sample size, generate $S = 1000$ samples of that size from the regression model, and for each generated sample, compute the OLS estimate of β_2. Compute the variance of the OLS estimates in each case. How does the estimation accuracy vary with the sample size? [Hint: If the model is estimated using the `lm()` function in R, the OLS estimate of β_2 for variable `x` is obtained with `coef(lm(y~x))['x']`.]

(b) Repeat (a), but set the sample size $N = 100$, and compare three values (0.1, 1.0, and 10.0) of the error variance σ_ε^2. How does the estimation accuracy vary with error variance?

(c) Repeat (a), but set the sample size $N = 100$, and compare three values (0.1, 1.0, and 10.0) of the variance of x_i, σ_X^2. How does the estimation accuracy vary with σ_X^2?

References (for a-c): code `simulation.R` (week 1). For understanding the results, slides "Testing hypotheses under the normality assumption" (week 1, see slide 2 in particular for the variance of estimator) might be helpful (but not required to finish the problem).