TI: Econometrics I 2021/2022
Assignment I

Instructions

- You are supposed to make the assignments individually or in a group of two. Of course, you may discuss the questions and your general ideas with your fellow students, but the actual answers should be given individually or in a group of two. Fraud, in the sense of copying answers, will be reported to the examination board and may have serious consequences.
- The grading of the assignments will be anonymous. Do not write any non-anonymous identifier on your assignment (only write student numbers but no names on your assignment).
- Before you start answering the questions, first read the complete exercise.
- In answering the questions, always state explicitly what you did, why you did it, and what your conclusions are. Be clear and concise in your statements.
- Submitting your answers in LaTeX is highly appreciated, though Word (converted to pdf) is also accepted.
- For all questions where R is used also provide the code that you used to generate your results.
- Only include relevant R output in your file. Either include your code in your answers, or send it in a separate file.
- For all tests, use a significance level of 5%, unless indicated otherwise.
- Due date: 2021 November 16, 23:59 at Canvas.

Question 1

Here, you will simulate some data to explore some of the properties of the OLS estimator. Please submit answers as a RStudio Notebook containing code. Compute all estimators using lm. At the beginning of the code, **set the random seed to 810 using set.seed()**. Failure to do so will be penalised.

(a) [1 point] Simulate 100,000 observations from the DGP $Y_i = \alpha + 0.3X_i + 0.1X_i^2 + \epsilon_i$, where: $\alpha = 5$, $X_i \sim N(6, 3)$, $\epsilon_i \sim N(0, 1)$. Plot histograms of $Y_i, X_i,$ and X_i^2, with 50 breaks in each case. Create a scatterplot of y_i (vertical axis) against X_i. Use this data throughout the rest of the question.

(b) [2 points] Break the data into 1000 datasets of 100 observations sequentially (i.e dataset 1 comprises observations 1-100 from your simulation, dataset 2 comprises observations 101-200 and so on). Fit the equations $Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$ and $Y_i = \alpha + \beta_1 X_i + \epsilon_i$ to each of these datasets. Plot histograms of the $\hat{\beta}_1$ and $se(\hat{\beta}_1)$ for the estimator of β_1 in each case. Plot two histograms for the coefficient estimates, and the two histograms for the standard errors of the estimates. What do you see?

(c) [1 point] Imagine we had instead generated $X_i = c \; \forall i$ and tried to perform our simulation above. This would fail because we would violate one of the necessary conditions for computing the least squares estimator. Which one, and how?

(d) [2 points] Simulate another dataset as in a), but with 1,000,000 observations. Repeat part b), but now using 1000 observations per model. Only fit the equation $Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$ this time. Plot the histograms of the estimate of $\hat{\beta}_1$ and $se(\hat{\beta}_1)$ and compare them to the estimates from the same model in b). What happens to the distribution of the coefficient estimates and standard errors when we increase the sample size per model?

(e) [1 point] Now, generate a new variable C_i that is correlated with X_i. Do this by creating a vector of observations drawn from $N(1, 2)$ and adding them to X_i. Add C_i to the dataset from the previous questions (i.e with 100,000 observations total). Re-run the regressions as $Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 C_i + \epsilon_i$ and plot the distribution of these coefficient estimates and standard errors for β_1 next to the coefficient estimates from the second part. What happens to the coefficient estimates, and why?

Question 2

You are a labor economist trying to estimate the gender wage gap within occupations for women with children - that is, the effect of gender on wages given occupational choice. You have access to a dataset containing a set of wages, W_i, a gender dummy D_i, a set of occupational dummies O_{ij} and hours spent on childcare C_i for a sample of men and women with children. Assume that there is a positive covariance between each of your regressors, some covariance between each of the regressors and wages, and that gender at least partially determines occupational choice and hours spent on childcare.

(a) [3 points] Derive the expected value of the least-squares estimator for the coefficient on the gender dummy without controlling for either occupation or hours spent on childcare. From this, derive its variance as compared to the variance

of the estimator where we control for both variables. Do both in terms of two projection matrices P_O, P_C.

(b) [3 points] A friend who has taken an undergraduate econometrics course suggests including both the occupational dummies and hours spent on childcare as control variables, to remove omitted variable bias. Now imagine we control for both of these in our least-squares regression. Derive the coefficient estimate and the variance of the estimator.

(c) [2 points] Think about what you are trying to estimate. Why would including hours spent on childcare as a control be incorrect, despite your result above? To try to sharpen your thinking you might want to draw a causal diagram for the data-generating process (if you do not know what these are, see `https://mixtape.scunning.com/dag.html` - you do not need to worry about the formalisms), thinking about how each variable determines each other.

Question 3

A researcher is studying the service life of n machines. The researcher assumes that the service life of the machine has an exponential distribution. So the probability density function for the service life of machine i where $i = 1, 2, \cdots, n$ is given by

$$f(x_i; \alpha) = \begin{cases} \alpha e^{-\alpha x_i}, x_i \geq 0 \\ 0, x_i < 0 \end{cases}$$

(a) [1 point] Derive the log-likelihood function for the service life of n machines.

(b) [1 point] Derive $\hat{\alpha}$, the maximum likelihood estimator for α. Please check the second-order condition to ensure your result indeed maximizes the log-likelihood function.

(c) [2 points] After reading more papers, the researcher thinks the Weibull distribution might work too. The researcher now assumes the density function is

$$g(x_i; \lambda, \beta) = \begin{cases} \frac{\beta}{\lambda}(\frac{x_i}{\lambda})^{\beta-1} e^{-(\frac{x_i}{\lambda})^{\beta}}, x_i \geq 0 \\ 0, x_i < 0 \end{cases}$$

Derive the log-likelihood function.

(d) [1 point] The researcher further assumes that β is known but λ is unknown. Derive $\hat{\lambda}$, the maximum likelihood estimator for λ. You do not need to verify the second-order condition in this question.

(e) [2 points] Which distribution specification do you prefer and why? You might consider the two hints as follows, though you do not have to use any of the hints. (Hint 1: it could be shown that the Exponential distribution is a special case of the Weibull distribution; Hint 2: consider what happens if β is unknown)

Question 4

In this question you need to use the data set named **DataAS1**. The description of variables are as follows:

- *birthweight*: birth weight of infant (in grams)
- *smoker*: indicator equal to one if the mother smoked during pregnancy and zero, otherwise
- *age*: age of the mother
- *educ*: years of educational attainment of the mother (more than 16 years coded as 17)
- *unmarried*: indicator =1 if the mother is unmarried
- *alcohol*: indicator =1 if mother drank alcohol during pregnancy
- *drinks*: number of drinks per week of the mother during pregnancy

(a) [2 points] Estimate the two following models. Interpret the results of the two models and also compare the results.
model 1: Use a linear model to explain *birthweight* using *age*, *smoker*, *alcohol*, *drink*, and a constant term.

model 2: add additional independent variables *unmarried* and *educ*.

(b) [1 point] Test the null hypothesis H_0: $\beta_{unmarried} = \beta_{educ} = 0$. Please include the name of the test, the value of the test statistics, the p-value and the conclusion in your answer.

(c) [1 point] Are the residuals of model 2 normally distributed? Provide a histogram for the residuals and also perform a formal test.

(d) [2 points] There could be nonlinear relations between the dependent variable and the independent variables. Perform a test to investigate the possible non-linear relation. What is your conclusion?

(e) [2 points] Use the OLS method to estimate a new model (i.e. model 3) whose dependent variable is the log of *birthweight* and the independent variables are the same as those in model 2. Then use the maximum likelihood method to estimate model 3 again. Provide a table to show the coefficients of all regressors for two methods. Explain why the coefficients of regressors for the two methods are similar or why they are different.