MOEC0021
Empirical Methods
University of Zurich — HS 2019
PROF: Greg Crawford / TA: E. Dicarlo, M.R. Greco, S. Bagagli, A. Jenni — Handout 1

Problem Set 1

This problem set is due on the **21st of October** at **23:59**.
Solutions should be turned in via email to **emanuele.dicarlo@econ.uzh.ch** in PDF form. Please follow the following steps when submitting your solution:

1. Email Title: MOEC0021 Problem Set 1 Solutions
2. Attachment Title: GroupName_PS1.pdf
   For example, if my group was called 'DataJedi' I would name the attachment DataJedi_PS1.pdf

Remember, your goal is to communicate. Full credit will be given only to the correct solution which is described clearly. Convoluted and obtuse descriptions might receive low marks, even when they are correct. Also, aim for concise solutions, as it will save you time spent on write-ups, and also help you conceptualize the key idea of the problem.

1 Theory

1. Consider the simple linear regression model

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

   (a) Suppose that the *unconditional* expectation $E(\epsilon_i) = \mu_\epsilon \neq 0$. Using the formulas for $\hat{\beta}_1$ and $\hat{\beta}_2$ on slide 4 of the lecture notes, evaluate $E(\hat{\beta}_1)$ and $E(\hat{\beta}_2)$. What does your answer tell you about the robustness of OLS estimation to ϵ_i having a non-zero error?

   (b) Suppose you decided to measure all of your X variables in different units such that your new X variable, call it $\tilde{X}$, is exactly double your old one, i.e. $\tilde{X} = 2X$. Suppose you run the regression of y on $\tilde{X}$; call the resulting estimate $\tilde{\beta}$. What is the relationship between $\tilde{\beta}$ and $\hat{\beta}$, the regular OLS estimator from the regression of y on X?

   (c) Suppose you decided now to measure y in different units such that your new y variable, call it y^*, is exactly double your old one, i.e. $y^* = 2y$. Suppose you run the regression of y^* on X; call the resulting estimate β^*. What is the relationship between β^* and $\hat{\beta}$?

   (d) Given your answers to the last two questions, how meaningful are the units in which X and y are measured for the conclusions you draw from an OLS regression?

(e) Return to the scenario in part (1b) above with $\tilde{X} = 2X$. Calculate $V(\tilde{\beta})$. What is the relationship between $V(\tilde{\beta})$ and $V(\hat{\beta})$?

2 Empirical Application

1. In this empirical exercise, we will try to understand more about people's attitude toward smoking. In particular we will try to understand how smoking relates to a person's education and age.

   (a) Download the data from OLAT or from this link [1] and import them into Stata or R. How many observations are there? For each observation, i, there are 10 variables in the dataset, listed here:

   - *educ*: i's years of schooling
   - *cigpric*: the average cigarette price (in cents/pack) in i's state
   - *white*: a dummy variable=1 if i is white
   - *age*: i's age, measured in years
   - *income*: i's annual income,
   - *cigs*: the number of cigarettes smoked by i per day
   - *restaurn*: a dummy variable =1 if the restaurants in i's state restrict smoking
   - *lincome*: the log of income
   - *agesq*: the square of age
   - *lcigpric*: the log of the average cigarette price

   (b) Provide a table of summary statistics for the variables *cigs*, *educ*, *age*, *income*, *white*, *restaurn*. Briefly describe patterns you find particularly interesting (if any).

   (c) We want to estimate the relationship between number of cigarettes smoked and education, measured as i's years of schooling.

   $$cigs_i = \beta_1 + \beta_2 educ_i + \epsilon_i$$

   i. Compute β_1 and β_0 (easier in this order) using the formulas on either slide 4 and/or slide 16 of the lecture notes.

   ii. Run the regression in equation (1c) How do the computer's estimates of β_0 and β_1 relate to the ones you have just computed?

   iii. Suppose that Assumption 2 (Mean-zero error) is satisfied. How do you interpret the coefficient of *educ*? Is this a big or small effect?

   iv. Using your estimates, predict the number of cigarettes consumed by i and denote this $\widehat{cigs}$. In a graph, display both the scatterplot of cigarettes smoked against education and your regression line.

   v. Now regress cigarettes on education **without** including a constant. Generate predicted values and add the new regression line to the previous graph. What changes compared to the earlier regression line? Do you think you should include a constant or not?

[1] http://fmwww.bc.edu/ec-p/data/wooldridge/smoke.dta

(d) Now regress $cigs$ on $educ$, age, age^2, $white$ and $restaurant$ and assume again that Assumption 2 (Mean-zero error) is satisfied

i. What are the coefficients of race (i.e. white) and of the dummy restaurant? How would you interpret them?

ii. Calculate the marginal effect of age on cigarette consumption. What is the value of this marginal effect at age 20? At age 40? At age 60?

iii. Predict the residuals from your model, $e_i = cigs_i - \hat{cigs}_i$, where $\hat{cigs}_i$ is the fitted value of $cigs_i$ from your regression.

A. Construct a scatter plot of these residuals against age. What does this tell you about the likely validity of our Assumption 3?

B. Calculate the correlation of these residuals across individuals. What does this tell you about the likely validity of our Assumption 4?

C. Plot the density of the residuals together with the density of a normal distribution. What does this tell you about the likely validity of our Assumption 5?