MOEC0021
Empirical Methods
University of Zurich HS 2020
PROF: Emanuele Dicarlo / TA: A. Jenni, M.R. Greco, S. Bagagli, L. Schubert

Problem Set 1 – The Classical Linear Regression Model

This problem set is due on the **19th of October** at **23:59**. Solutions should be turned in on OLAT in a single PDF file. Please name your file as GroupName_PS1.pdf. Include any code you wrote to answer the questions in the file.
One of your goals is to communicate efficiently. Please keep your answers succinct. Lengthy answers will be marked down.

1. Theory – Using the CLRM to Make Predictions

Consider the following regression model

$$y_i = x_i'\beta + \varepsilon_i, \qquad i = 1, 2, \dots, n, n+1$$

where x_i and β are vectors of dimensions $K \times 1$, and y_i and ε_i are scalars. Suppose that you observe the regressors for all observations, $x_1, x_2, \dots, x_n, x_{n+1}$, and the outcome variable only for the first n observations, $y_1, y_2, \dots, y_n$. You want to use your model to predict the unobserved outcome value y_{n+1}. Let your prediction be

$$\hat{y}_{n+1} = x_{n+1}'\hat{\beta}_n,$$

where $\hat{\beta}_n$ is the OLS estimator computed using the n observations for which y_i is observed.

(a) This may feel very abstract. Can you think of a economic context where this framework may be applied? Provide an example, specifying what i, y_i, x_i and ε_i would be in this context.

Assume for the rest of this exercise that the CLRM assumptions hold. In particular,
$\varepsilon|\mathbf{X} \sim \mathcal{N}(0, \sigma^2 I_{n+1})$, where we define $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{n+1})'$ and $\mathbf{X}' = (x_1, x_2, \dots, x_{n+1})$.

(b) Derive the conditional expectation function of y_i given x_i, $\mathbb{E}(y_i|x_i)$, and the conditional variance of y_i given x_i, $\text{Var}(y_i|x_i)$.

(c) Suppose the CLRM assumptions hold in the example you provided in part (a). Briefly interpret your results for $\mathbb{E}(y_i|x_i)$ and $\text{Var}(y_i|x_i)$ in this context.

(d) Define your prediction error for observation $n+1$ as $\hat{e}_{n+1} = y_{n+1} - \hat{y}_{n+1}$. We say that a prediction is unbiased if $\mathbb{E}(\hat{e}_{n+1}|\mathbf{X}) = 0$. Is you prediction $\hat{y}_{n+1}$ unbiased? Explain why in your own terms.

(e) What is the conditional variance of your prediction error for observation $n+1$, $\text{Var}(\hat{e}_{n+1}|\mathbf{X})$? Is it larger or smaller than $\text{Var}(y_i|x_i)$ derived in part (b)? Explain.

(f) What happens to $\text{Var}(y_i|x_i)$ and $\text{Var}(\hat{e}_{n+1}|\mathbf{X})$ as n, the size of the estimating sample, increases? Comment briefly.

2. Empirical Application – No Risk, No Steak? Interpreting Regressions in the CLRM

The goal of this question—besides learning some cool econometrics—is to deepen our understanding of how the American people prefer their steak: rare, medium rare, medium, medium well, or well done? After wrangling with the data for a little bit, we will run some regressions to better understand which traits are associated with which preferences.

(a) Install the R package `fivethirtyeight`. We will use the data set `steak_survey` that is included in *FiveThirtyEight*'s package. There should be 15 variables and 550 observations. For each observation, each variable captures an individual's response to a survey question. (To find out what the survey questions were, check out the original data set on GitHub. *https://github.com/fivethirtyeight/data/blob/master/steak-survey/steak-risk-survey.csv*)

Unfortunately, many variables are currently still in a non-readable format. Hence we would like you to **generate** the following variables:

- **cooking_temp**: This variable takes on the values 120, 130, 135, 140, or 150 degrees Fahrenheit whenever someone prefers their steak rare, medium rare, medium, medium well done, or well done respectively.
- **cheat**: This variable should be 1 if someone has cheated and 0 if not.
- **riskaverse**: This variable should be 1 if someone prefers lottery B and 0 for lottery A.
- **yrs_ed**: This variable should be 8, 12, 14, 16, 18 for less than high school, high school, associate's, bachelor's, or master's respectively.
- **rand_age**: Let this variable be a random number drawn from the age interval a respondent claims to be in. **Please use** the command `sample()` and `set.seed(123)` for reproducibility.

(b) Provide a table of summary statics for the variables: *cooking_temp, cheat, riskaverse, yrs_ed, rand_age*. Furthermore, depict the distribution of *steak_prep* in a histogram.

Hint: Because *steak_prep* contains text, you will need to a create a numeric categorical variable first. Make sure to add meaningful labels to the x-axis values though.

(c) We would like to study if there is an association between someone's risk aversion and liking their steak well done. What is your prior regarding the sign of the coefficient? Briefly explain your reasoning in one sentence.

To check, we consider the following regression:

$$cooking\_temp_i = \beta_1 + \beta_2 riskaverse_i + \epsilon_i \tag{1}$$

Hint: You ought to restrict your data set to contain only those observations that are "steak eaters".

- Compute the regression coefficients by hand using the formulas on slide 4 of the lecture notes.
- Run the regression using R (or Stata) and present your results in a table (that is—cheesy hint—just as beautiful as the stars you might gaze at at night). How do the computer's results compare to your own?
- Suppose the mean-zero-error assumption holds. How do you interpret the coefficient of risk aversion?

(d) Do you believe it is important to include a constant in the above regressions? Why or why not? Please explain your reasoning in 80 words or less.

(e) Let us consider a more elaborate regression model:

$$\begin{aligned} cooking\_temp_i = \beta_1 + \beta_2 riskaverse_i + &\beta_3 ln(yrs\_ed_i) + \beta_4 rand\_age_i + \\ &+ \beta_5 rand\_age_i^2 + \beta_7 cheat_i + \epsilon_i \end{aligned} \tag{2}$$

Please run the regression, present your results in a clean table, and interpret the following marginal effects on cooking temperature preferences:

- 1 additional year of education
- Having previously cheated on a spouse
- 10 additional years of age (compute the ME at means)

(f) What is the predicted preferred cooking temperature when all explanatory variables are at their mean? Why is this **not** an informative number to look at?

(g) You might also consider including both the estimated age and the categorical age variable. Is this a good idea? Name two reasons why (or why not).

(h) Regardless of your sincere attempts, it seems like all your models yield insignificant results regarding the relationship between risk aversion and one's preferred steak cooking temperature. Name two reasons why this could be the case.

(i) Using the model specified in 2, predict the residuals.

- Construct a scatter plot of these residuals against age. What does this tell you about the validity of our Assumption 3?
- Plot the density of the residuals against the density of a normal distribution. What does this tell you about the validity of our Assumption 5?

  Hint: You can randomly draw N=403 observations from a standard normal distribution and plot that series against the residuals using stacked data sets.