22MO0010
Empirical Methods
University of Zurich HS 2024
PROF: Gregory S. Crawford
TAS: T. Friederiszick; T. Braun, A. Azzolini, J. Grigaitis

Problem Set 1 – The Classical Linear Regression Model

This problem set is due **October 21, 2024** at **23:59**. Solutions are to be handed in as a single PDF file on OLAT. **Please follow the 'Instructions for Submission', which you find on OLAT.**

---

1. Theory – Using the CLRM to Make Predictions

Consider the following regression model

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

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, $x_1, x_2, \ldots, x_n, x_{n+1}$, and the outcome variable only for the first n observations, $y_1, y_2, \ldots, y_n$. 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 an 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 $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{n+1})'$ and $\mathbf{X}' = (x_1, x_2, \ldots, x_{n+1})$.

(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, $\mathrm{Var}(y_i|x_i)$.

(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 $\mathrm{Var}(y_i|x_i)$ 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 $n+1$, $\mathrm{Var}(\hat{e}_{n+1}|\mathbf{X})$? Is it larger or smaller than $\mathrm{Var}(y_i|x_i)$ derived in part (b)? Explain.

2. Empirical Application – The Beauty and the Student. Interpreting Regressions in the CLRM

The goal of this question—besides learning some cool econometrics—is to investigate how university students evaluate the teaching performance of their professors. In particular, we will ask whether professors' looks have an effect on their overall teaching evaluation score. After calculating standard summary statistics and plotting the data, we will run some regressions to better understand which instructor characteristics are associated with high course evaluation ratings.[1]

(a) Install the R package `AER`. We will use the data set `TeachingRatings` that is included in the `AER` package. There should be 12 variables and 463 observations. Read the R Documentation of `TeachingRatings` and explain in one sentence: what is the unit of observation in this data set?

(b) Provide a table with the mean, standard deviation, minimum and maximum value for the variables: *eval, beauty, age, allstudents*. Furthermore, we want to make sure our data set is complete: count the number of missing values in each variable: *eval, beauty, age, allstudents, gender, minority*.

(c) Plot the joint distribution of *eval* and *beauty* using a scatterplot. Explain in one sentence why it is always a good idea to plot your data.

(d) We want to study the relationship between a course's overall teaching evaluation score and its instructor's physical appearance.

For this reason, we consider the following regression model:

$$eval_i = \beta_1 + \beta_2 beauty_i + \epsilon_i \tag{1}$$

- Compute the OLS coefficient estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ using only the following functions from the R 'base' package: `mean()`, `sum()`, `var()` and `cov()` (or the equivalent functions in your programming language of choice). (**Hint:** use the formulas from the lecture notes Topic 1b.)
- Suppose the CLRM Assumption 2 holds. How do you interpret the coefficient of instructor's physical appearance?

(e) 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.

[1] We use data from Hamermesh and Parker (2005): `https://doi.org/10.1016/j.econedurev.2004.07.013`.

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

$$\begin{aligned} eval_i = &\beta_1 + \beta_2 beauty_i + \beta_3 age_i + \beta_4 age_i^2 + \beta_5 ln(allstudents_i) + \\ &+ \beta_6 gender_i + \beta_7 minority_i + \beta_8 female\_minority_i + \epsilon_i \end{aligned} \tag{2}$$

where $female\_minority_i$ is 1 if the instructor of course i is female and belongs to a minority group, and 0 otherwise.

Please run the regression, present your results in a clean table, and calculate the marginal effects of

- an increase in the instructor's beauty rating by one standard deviation
- the instructor being a non-minority male as opposed to non-minority female
- the instructor being 60 as opposed to 50 years old

on a course's teaching evaluation score.

(g) You might also consider including a dummy variable $male\_minority_i$ in the above specified multivariate regression model. Why is this a good or bad idea?

(h) Suppose the CLRM Assumption 2 holds in the case of the regression model estimated in question (f). What do you conclude about the importance of professors' looks on their overall teaching evaluation score? (**Hint:** How large is the effect? Is the effect causal?)

(i) Provide a concrete scenario in which the CLRM Assumption 2 does not hold in the case of the regression model estimated in question (f).

(j) Using the model specified in question (f), calculate the predicted residuals $\hat{\epsilon}_i$.

- Calculate the covariance between $beauty_i$ and $\hat{\epsilon}_i$ (round your answer to 6 decimal places). What does this tell you about the validity of CLRM Assumption 2?
- Construct a scatter plot of the residuals against $ln(allstudents_i)$. What does this tell you about the validity of CLRM Assumption 3?
- Plot the density of the residuals over the density of a normal distribution. What does this tell you about the validity of CLRM Assumption 5?

  Hint: You can randomly draw N=463 observations from a standard normal distribution and plot that series against the residuals using a stacked data set.