# Introductory Econometrics I – Spring 2024

## Problem Set 2 – Due date: Apr 7

Last updated: March 24, 2024

**Notes:** Please submit a single PDF file containing your answers to all questions on Web-learning. For empirical questions, original codes and complete results need to be attached.

1. Consider the following regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

Assumptions MLR.1-MLR.5 hold. In particular, Assumption MLR.4 means $\mathbb{E}[u|x_1, x_2] = 0$. In addition, assume $x_1$ is independent of $x_2$, and $\mathbb{E}[x_2] = 0$. Let $\sigma_u^2 = \mathbb{V}[u|x_1, x_2] = \mathbb{V}[u] > 0$ and $\sigma_2^2 = \mathbb{V}[x_2|x_1] = \mathbb{V}[x_2] > 0$. There is a random sample $\{(y_i, x_{i1}, x_{i2}) : 1 \leq i \leq n\}$. Consider the following two estimators:

- Run a simple regression of $y$ on 1 and $x_1$. Denote the coefficient on $x_1$ by $\tilde{\beta}_1$.
- Run a multiple regression of $y$ on 1, $x_1$ and $x_2$. Denote the coefficient on $x_1$ by $\hat{\beta}_1$.

(a) Show that $\mathbb{E}[\tilde{\beta}_1] = \beta_1$. [Hint: Define $e = \beta_2 x_2 + u$. Note that $x_1$ is independent of $x_2$ implies $\mathbb{E}[x_2|x_1] = \mathbb{E}[x_2] = 0$.]

(b) Show that

$$\mathbb{V}[\tilde{\beta}_1|x_1] = \frac{\sigma_u^2 + \beta_2^2 \sigma_2^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2},$$

where $\bar{x}_1 = \frac{1}{n}\sum_{i=1}^n x_{i1}$, and $\mathbb{V}[\tilde{\beta}_1|x_1]$ denotes the conditional variance of $\tilde{\beta}_1$ given the explanatory variables $(x_{11}, \cdots, x_{n1})$.

[Hint: Define an error term $e = \beta_2 x_2 + u$. Use the law of iterated expectation to show $\mathbb{E}[x_2 u|x_1] = 0$.]

(c) In class, we have discussed the variance of the multiple regression estimator $\hat{\beta}_1$:

$$\mathbb{V}[\hat{\beta}_1|x_1, x_2] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 (1 - R_1^2)},$$

where $R_1^2$ is the R-squared from regressing $x_1$ on 1 and $x_2$. In large samples (when the sample size $n$ gets large), what value do you think $R_1^2$ gets close to? You only need to explain your intuition.

[Hint: $x_1$ is independent of $x_2$ implies that $\mathbb{E}[x_1|x_2] = \mathbb{E}[x_1]$. Think about the population coefficient on $x_2$ in the simple regression of $x_1$ on 1 and $x_2$.]

(d) When the sample size $n$ is large, which estimator do you prefer in this special case, $\hat{\beta}_1$ or $\tilde{\beta}_1$?

2. Consider the following regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

Assumptions MLR.1-MLR.6 hold. We want to test

$$H_0 : \beta_1 - \beta_2 = 1, \quad H_1 : \beta_1 - \beta_2 \neq 1.$$

(a) Write $\mathbb{V}[\hat{\beta}_1 - \hat{\beta}_2]$ in terms of $\mathbb{V}[\hat{\beta}_1]$, $\mathbb{V}[\hat{\beta}_2]$ and $Cov[\hat{\beta}_1, \hat{\beta}_2]$.

(b) Write the formula for the $t$ statistic used to test $H_0$.

(c) Let $\theta_1 = \beta_1 - \beta_2$. Rewrite the regression model so that $\theta_1$ appears on one of the independent variables.

3. (**Data exercise**) We are interested in exploring factors affecting labor income and collect a data set with the following variables:

   - `id`: individual index
   - `age`: age of an individual
   - `agesq`: the square of age, `age`$^2$
   - `IQ`: IQ score
   - `lwage`: natural log of the monthly earnings
   - `educ`: year of education
   - `exper`: years of work experience

   Please answer the following questions using the dataset:

   (a) We first use the dataset to verify the FWL theorem.
      i. Run the multiple regression of log(wage) on `educ` and `IQ` (including the intercept), and obtain the slope coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.
      ii. Run a simple regression of `educ` on `IQ` (including the intercept) to obtain the residual, $\hat{r}$
      iii. Run the simple regression of `lwage` on $\hat{r}$, get the slope coefficient $\hat{\delta}$. Verify that $\hat{\delta} = \hat{\beta}_1$.

   (b) Suppose `lwage` is jointly determined by education and IQ, in the following way:

   $$\texttt{lwage} = \beta_0 + \beta_1 \texttt{educ} + \beta_2 \texttt{IQ} + u,$$

   where the model satisfies MLR.1-MLR.4. Instead, we estimate the following model:

   $$\texttt{lwage} = \alpha_0 + \alpha_1 \texttt{educ} + e.$$

   Do you expect to overestimate or underestimate the effects of education on log wage? Please state your intuition, and verify using the data.

   For the following questions, consider the regression model:

   $$\texttt{lwage} = \beta_0 + \beta_1 \texttt{educ} + \beta_2 \texttt{exper} + \beta_3 \texttt{IQ} + \beta_4 \texttt{age} + \beta_5 \texttt{age}^2 + u.$$

   Assume that MLR.1-MLR.5 holds for this model.

   (c) Estimate the model and report the results.

   (d) What's the $\hat{\beta}_1$? Interpret it causally.

   (e) Test the joint significance of `age` and `age`$^2$ at the 1% significance level. What's your conclusion?

   (f) State the null hypothesis that another year of general workforce experience has the same effect on `lwage` as another year of education.

   (g) Test the null hypothesis in the last question against a two-sided alternative at the 10% significance level, by constructing a 90% confidence interval. What do you conclude? [Hint: Consider your answer for question 2. Let $\theta = \beta_1 - \beta_2$, then $\beta_1 = \theta + \beta_2$. Plug this formula into the original regression equation. Estimate the new model. In Stata, we can add `level(90)` as an option in the `reg` command to report the 90% confidence interval.]