# Introductory Econometrics I – Spring 2022

## Problem Set 1 – Due date: Mar 17

Last updated: February 28, 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. (**Regression on a binary variable**) Consider the following simple regression model

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

$y$ is the outcome of interest, and $x_1$ is a **binary** explanatory variable that can only take two possible values 0 and 1. $\{(y_i, x_{i1}) : 1 \le i \le n\}$ is a random sample of size $n$. For example, $y_i$ could be the income of the $i$th individual, and $x_{i1}$ denotes the gender of the $i$th individual ($x_{i1} = 1$ if $i$ is female, and $x_{i1} = 0$ if $i$ is male). Let $n_1$ denote the number of observations with $x_{i1} = 1$ and $n_0$ denote the number of observations with $x_{i1} = 0$.

   (a) Show that $\sum_{i=1}^n y_i = \sum_{i=1}^n x_{i1} y_i + \sum_{i=1}^n (1 - x_{i1}) y_i$.

   (b) Run a regression of $y$ on 1 and $x_1$. Denote the estimator of $\beta_0$ and $\beta_1$ by $\hat\beta_0$ and $\hat\beta_1$ respectively. Show that

$$\hat\beta_1 = \bar y_1 - \bar y_0, \quad \hat\beta_0 = \bar y_0 \quad \text{where } \bar y_1 = \frac{1}{n_1}\sum_{i=1}^n x_{i1} y_i, \quad \text{and } \bar y_0 = \frac{1}{n_0}\sum_{i=1}^n (1 - x_{i1}) y_i.$$

   [Hint: use part (a) and the fact that $n_1 = \sum_{i=1}^n x_{i1}$, $n_0 = \sum_{i=1}^n (1 - x_{i1})$, $x_{i1}^2 = x_{i1}$.]

   (c) How do you interpret the result in part (b)?

   (d) Let Assumptions SLR.1-SLR.4 in the text hold. In particular, $\mathbb{E}[u|x_1] = 0$. Verify that

$$\mathbb{E}[\hat\beta_1] = \mathbb{E}[y|x_1 = 1] - \mathbb{E}[y|x_1 = 0].$$

   [Hint: $\hat\beta_1$ is unbiased for $\beta_1$ under these assumptions.]

2. (**Interpretation of Simple Linear Regression Model**) Suppose we collect data on 230 mothers' number of cigarettes they have per day during pregnancy and the birth weight of their child (in grams). Suppose now that you run a regression of the birth weight ($W_i$) on the number of cigarettes ($C_i$) and find that

$$W_i = 3396 - 15 C_i + \hat u_i.$$

   (a) Thinking about the model in a descriptive manner. Explain what it means that the slope coefficient is -15.

   (b) Thinking about the model in a causal manner. Explain what it means that the slope coefficient is -15.

   (c) Do you think the causal interpretation is valid?

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

   - `id`, individual index
   - `gender`, 1 = male, 2 = female
   - `birthyear`, birth year

- `marriage`, marriage status: 1=married, 2=all other status
- `wage`, annual income (in yuan)
- `schooling_yr`, years of education

Please answer the following questions using the dataset:

(a) Calculate summary statistics (number of observations, mean, standard deviation, minimum, maximum, median) of `birthyear`, `wage`, and `schooling_yr`.

(b) Generate a new variable, `female`, takes on the value 1 if the individual is female, 0 otherwise. What fraction of the sample is female?

(c) Calculate the average annual income for females and males in the sample. Then, estimate the regression model:

$$\texttt{wage} = \beta_0 + \beta_1 \texttt{female} + u.$$

What do you find? (Think about your answer to Q1).

(d) We believe that age and education years are factors affecting income. Estimate the following model:

$$log(\texttt{wage}) = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{schooling\_yr} + u.$$

Note: the data were collected in 2023. Think of the model in a casual manner. If `schooling_yr` increases by one year, what's the estimated effect on `wage`?

(e) Calculate the predicted value and the residual of the above model. Report their mean. What do you find?

(f) What is the goodness of fit of the above model?

4. (**Frisch–Waugh–Lovell theorem**) [Note: this question is optional and will not be graded. However, we encourage you to try this question.] Consider a multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u.$$

Suppose that the parameter of interest is $\beta_1$. We have two strategies for estimating $\beta_1$:

- Run a regression of $y$ on 1, $x_1$, $x_2$, $\cdots$, and $x_k$. The estimator of $\beta_1$ from this regression is denoted by $\hat{\beta}_1$.
- Run a regression of $x_1$ on 1, $x_2$, $\cdots$, $x_k$. Let $\hat{v}$ be the corresponding OLS residual. Then, run a regression $y$ on $\hat{v}$ (without the intercept). The estimated coefficient of $\hat{v}$ is denoted by $\hat{\hat{\beta}}_1$.

The Frisch-Waugh-Levell theorem in econometrics says $\hat{\beta}_1 = \hat{\hat{\beta}}_1$. Verify this conclusion for the **special case** with $k = 2$.