# 651 Empirical Economics: Assignment 1

Per Hjertstrand* Andrew Proctor†

## Part A: Revisiting basics

1. Explain in your own words:

   **(a)** The meaning of a random variable.

   **Answer:** A random variable is a function that assigns numeric values to the outcomes of an experiment.

   **(b)** The difference between a continuous and discrete random variable.

   **Answer:** A random variable is called discrete if it can take only a finite or countable number of values. A random value is instead continuous if the range of values it can take on is uncountably infinite, each with a probability equal to zero (Wooldridge Appendix B-1a and B-1b).

   **(c)** The meaning of a probability distribution function (pdf).

   **Answer:** In this answer, we use probability distribution function to mean the non-cumulative functions known as the probability mass function (pmf) and probability density function (pdf), which are used to characterize the distributions of discrete and continuous random variables, respectively.

   A pmf, used to characterize the distribution of discrete random variables, assigns a probability between 0 and 1 to each possible value of the random variable, such that the sum of the probabilities is equal to 1.

   For continuous random variables, we instead characterize the probability distribution using the pdf. Since a continuous random variable is defined so that the probability of any single outcome is equal to zero, the same approach cannot be used for the pdf as in the pmf. Instead, the pdf is a function used to compute the probability that a random variable lies within any given interval (formally, by integrating over the specified interval) .

2. Let $X$ be the number of absent students from a class in 651 Empirical Economics on a Monday morning. The probability distribution for $X$ is

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $f(x)$ | 0.005 | 0.025 | 0.310 | 0.340 | 0.220 | 0.080 | 0.019 | 0.001 |

*Per.Hjertstrand@ifn.se

†Andrew.Proctor@phdstudent.hhs.se

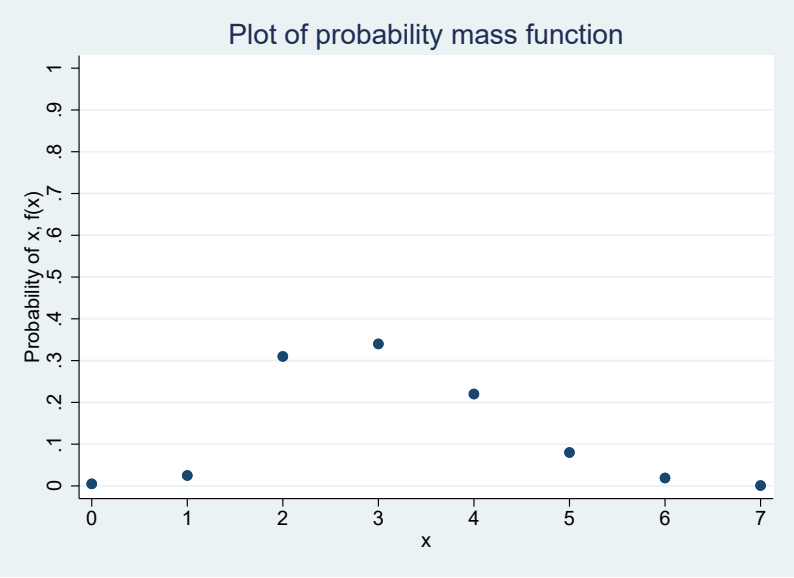


Figure 1: Probability distribution.

**(a)** Draw the probability distribution in a graph.

**Answer:** See Figure 1.

1. **(b)** Calculate the probability that $2, 3$ or $4$ students are absent.

   **Answer:** The probability is $0.310 + 0.340 + 0.220 = 0.87$. Thus, there is a 87% probability that $2, 3$ or $4$ students are absent.

   **(c)** Calculate the probability that *more than* 3 students are absent.

   **Answer:** The probability is $0.220 + 0.080 + 0.019 + 0.001 = 0.32$. Thus, there is a 32% probability that more than 3 students are absent.

   **(d)** Calculate the expected value of $X$ and interpret the result.

   **Answer:** The expected value $E(X)$ is equal to

$$\begin{aligned} E(x) &= 0.005 \times 0 + 0.025 \times 1 + 0.310 \times 2 + 0.340 \times 3 + 0.220 \times 4 + 0.080 \times 5 \\ &\quad + 0.019 \times 6 + 0.001 \times 7 \\ &= 3.066. \end{aligned}$$

   The expected value can be interpreted as the expected number of students absent from a class on Monday morning.

   **(e)** Calculate the variance of $X$ and interpret the result.

   **Answer:** The variance $Var(X)$ is equal to

$$Var(X) = E\left(X^2\right) - (E(X))^2,$$

   where

$$\begin{aligned} E\left(X^2\right) &= 0.005 \times 0^2 + 0.025 \times 1^2 + 0.310 \times 2^2 + 0.340 \times 3^2 \\ &\quad + 0.220 \times 4^2 + 0.080 \times 5^2 \\ &\quad + 0.019 \times 6^2 + 0.001 \times 7^2 \\ &= 10.578. \end{aligned}$$

Thus,

$$\begin{aligned} Var\left(X\right) &= 10.578 - 3.066^2 \\ &= 1.177644. \end{aligned}$$

We can interpret the variance as a measure of the spread of the observations from the expected value.

**(f)** Calculate the expected value and variance of $Y = 7X + 3$.

**Answer:** The expected value $E\left(Y\right)$ is equal to

$$\begin{aligned} E\left(Y\right) &= E\left(7X + 3\right) \\ &= 7 \times E\left(X\right) + 3 \\ &= 7 \times 3.066 + 3 \\ &= 24.462. \end{aligned}$$

The variance $Var\left(Y\right)$ is equal to

$$\begin{aligned} Var\left(Y\right) &= Var\left(7X + 3\right) \\ &= 7^2 \times Var\left(X\right) + Var\left(3\right) \\ &= 49 \times 1.177644 + 0 \\ &= 57.704556. \end{aligned}$$

# Part 2: Mechanics of OLS

Consider the following population model:

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

where $y_i$ is the consumption of food for household $i$ during one year and $x_i$ is the disposable income during that year for household $i$. Both variables are measured in thousands of Swedish Krona.

Suppose we collect data from 3 households where:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 15 \\ 12 \\ 5 \end{bmatrix} \text{ and } \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 10 \\ 15 \\ 7 \end{bmatrix}.$$

**(a)** Look at the data on $y$ and $x$. Is there something odd about this data?

**Answer:** The consumption of food for the first household is larger than the disposable income for that household which may seem odd. It is very important in empirical work to look carefully at the data to discover odd patterns like this.

**(b)** Calculate the OLS estimates of $\beta_0$ and $\beta_1$ by hand. Interpret your results.

**Answer:** We can use the MM assumptions 1 and 2 to calculate the OLS estimates. We rewrite the model as

$$u_i = y_i - \beta_0 - \beta_1 x_i.$$

MM assumption 1 says that the residuals should sum to zero, that is:

$$\sum_{i=1}^{3} u_i = 0 \longleftrightarrow$$

$$\sum_{i=1}^{3} (y_i - \beta_0 - \beta_1 x_i) = 0.$$

By plugging in our data, we get

$$(15 - \beta_0 - \beta_1 10) + (12 - \beta_0 - \beta_1 15) + (5 - \beta_0 - \beta_1 7) = 0 \longleftrightarrow$$

$$\beta_0 = \frac{32 - 32\beta_1}{3}.$$

MM assumption 2 says that the sum of the products between the residual and the explanatory variable should sum to zero, i.e.,

$$\sum_{i=1}^{3} u_i x_i = 0 \longleftrightarrow$$

$$\sum_{i=1}^{3} (y_i - \beta_0 - \beta_1 x_i)\, x_i = 0.$$

Plugging in our data, we get:

$$\begin{aligned}
\left(15 - \widehat{\beta}_0 - \widehat{\beta}_1 10\right) 10 + \left(12 - \widehat{\beta}_0 - \widehat{\beta}_1 15\right) 15 + \left(5 - \widehat{\beta}_0 - \widehat{\beta}_1 7\right) 7 &= 0 \longleftrightarrow \\
365 - 32\widehat{\beta}_0 - 374\widehat{\beta}_1 &= 0 \longleftrightarrow \\
374\widehat{\beta}_1 &= 365 - 32\widehat{\beta}_0 \\
&= 365 - 32\left(\frac{32 - 32\widehat{\beta}_1}{3}\right) \\
&= 365 - \frac{1024 + 1024\widehat{\beta}_1}{3} \\
&= \frac{1095}{3} - \frac{1024}{3} + \frac{1024\widehat{\beta}_1}{3} \\
&= \frac{71}{3} + \frac{1024\widehat{\beta}_1}{3} \longleftrightarrow \\
374\widehat{\beta}_1 - \frac{1024\widehat{\beta}_1}{3} &= \frac{71}{3} \longleftrightarrow \\
\frac{1122}{3}\widehat{\beta}_1 - \frac{1024\widehat{\beta}_1}{3} &= \frac{71}{3} \longleftrightarrow \\
\frac{98}{3}\widehat{\beta}_1 &= \frac{71}{3} \longleftrightarrow \\
\widehat{\beta}_1 &= \frac{71}{98}. \quad \sim.7244
\end{aligned}$$

The estimate for the intercept is:

$$\begin{aligned}\widehat{\beta}_0 &= \frac{32 - 32\beta_1}{3} \\ &= \frac{32}{3} - \frac{32}{3}\beta_1 \\ &= \frac{32}{3} - \frac{32}{3}\frac{71}{98} \\ &= \frac{3136}{294} - \frac{2272}{294} \\ &= \frac{864}{294} \\ &= \frac{288}{98}\end{aligned}$$

Summarizing our results: ~.293

$$\widehat{\beta}_0 = \frac{288}{98} \text{ and } \widehat{\beta}_1 = \frac{71}{98}$$

To simplify subsequent calculations, we also calculate the fitted values and the residuals. We have

$$\widehat{y}_1 = \frac{288}{98} + \frac{71}{98} \times 10 = \frac{998}{98}, \sim 10.18$$
$$\widehat{y}_2 = \frac{288}{98} + \frac{71}{98} \times 15 = \frac{1353}{98}, \quad 13.806$$
$$\widehat{y}_3 = \frac{288}{98} + \frac{71}{98} \times 7 = \frac{785}{98}. \quad 8.0102$$

The residuals are then yields:

$$\widehat{u}_1 = y_1 - \widehat{y}_1 = 15 - \frac{998}{98} = \frac{472}{98}, \sim 4.816$$
$$\widehat{u}_2 = y_2 - \widehat{y}_2 = 12 - \frac{1353}{98} = -\frac{177}{98}, \quad \sim 1.80$$
$$\widehat{u}_3 = y_3 - \widehat{y}_3 = 5 - \frac{785}{98} = \frac{295}{98}. \quad \sim 3.01$$

As a check, let us see if the MM assumptions are satisfied. MM assumption 1 holds since:

$$\widehat{u}_1 + \widehat{u}_2 + \widehat{u}_3 = \frac{472}{98} + \left(-\frac{177}{98}\right) + \left(-\frac{295}{98}\right) = 0.$$

Also, MM assumption 2 is satisfied because:

$$\begin{aligned}\widehat{u}_1 \times x_1 + \widehat{u}_2 \times x_2 + \widehat{u}_3 \times x_3 &= \frac{472}{98} \times 10 + \left(-\frac{177}{98}\right) \times 15 + \left(-\frac{295}{98}\right) \times 7 \\ &= \frac{4720}{98} - \frac{2655}{98} - \frac{2065}{98} \\ &= 0.\end{aligned}$$

We summarize our results in a table

| Obs | $y$ | $x$ | $\widehat{y}$ | $\widehat{u}$ |
|---|---|---|---|---|
| 1 | 15 | 10 | $\frac{998}{98}$ | $\frac{472}{98}$ |
| 2 | 12 | 15 | $\frac{1353}{98}$ | $\frac{177}{98}$ |
| 3 | 5 | 7 | $\frac{785}{98}$ | $\frac{295}{98}$ |

The OLS estimate of the intercept $\widehat{\beta}_0$ can be interpreted as the level of consumption of food at an disposable income level of zero. The OLS estimate of the slope $\widehat{\beta}_1$ can be interpreted as the additional unit of consumption of food for a unit increase in disposable income.

**(c)** Calculate the variances of the OLS estimates by hand. Interpret your results.

**Answer:** The variance for the slope coefficient is given by:

$$Var\left(\widehat{\beta}_1\right) = \frac{\widehat{\sigma}_u^2}{SST_x},$$

where the residual variance, the total sum of squares of $x$ and the mean of $x$ is:

$$\widehat{\sigma}_u^2 = \frac{1}{3-2}\sum_{i=1}^{3}(\widehat{u}_i)^2$$

$$SST_x = \sum_{i=1}^{3}(x_i - \overline{x})^2,$$

$$\overline{x} = \frac{1}{3}\sum_{i=1}^{3}x_i.$$

By plugging in our data, we get

$$\begin{aligned}\widehat{\sigma}_u^2 &= \frac{1}{3-2}\left[\left(\frac{472}{98}\right)^2 + \left(\frac{177}{98}\right)^2 + \left(\frac{295}{98}\right)^2\right] \\ &= \left(\frac{472}{98}\right)^2 + \left(\frac{177}{98}\right)^2 + \left(\frac{295}{98}\right)^2 \\ &= \frac{341138}{9604}. \quad \sim 35.52\end{aligned}$$

Moreover,

$$\begin{aligned}\overline{x} &= \frac{1}{3}(10+15+7) \\ &= \frac{32}{3},\end{aligned}$$

and,

$$\begin{aligned}SST_x &= \left(10 - \frac{32}{3}\right)^2 + \left(15 - \frac{32}{3}\right)^2 + \left(7 - \frac{32}{3}\right)^2 \\ &= \left(\frac{30}{3} - \frac{32}{3}\right)^2 + \left(\frac{45}{3} - \frac{32}{3}\right)^2 + \left(\frac{21}{3} - \frac{32}{3}\right)^2 \\ &= \left(-\frac{2}{3}\right)^2 + \left(\frac{13}{3}\right)^2 + \left(-\frac{11}{3}\right)^2 \\ &= \frac{2^2 + 13^2 + 11^2}{3^2} \\ &= \frac{4 + 169 + 121}{9} \\ &= \frac{294}{9}. \quad \sim 32.667\end{aligned}$$

Thus,

$$\begin{aligned}Var\left(\widehat{\beta}_1\right) &= \frac{\widehat{\sigma}_u^2}{SST_x} \\ &= \frac{\frac{341138}{9604}}{\frac{294}{9}} \\ &= \frac{341138}{9604} \times \frac{9}{294} \\ &= \frac{3070242}{2823576}. \quad \sim 1.087\end{aligned}$$

The OLS estimates are random variables. The variance of the OLS estimators gives a measure of the spread of the OLS estimates from their true values (since they are unbiased).

**(d)** Calculate the $R^2-$value by hand. Interpret your result.

**Answer:** The $R^2-$value is calculated as:

$$R^2 = \frac{SSE}{SST},$$

where the total sum oc squares and the explained sum of squares is:

$$SST = \sum_{i=1}^{3} (y_i - \overline{y})^2,$$
$$SSE = \sum_{i=1}^{3} (\widehat{y}_i - \overline{y})^2.$$

We have

$$\begin{aligned} \overline{y} &= \frac{1}{3}(15+12+5) \\ &= \frac{32}{3}, \\ &= \frac{158}{3}, \quad \sim 52.67 \end{aligned}$$

Moreover,

$$\begin{aligned} SST &= \left(15 - \frac{32}{3}\right)^2 + \left(12 - \frac{32}{3}\right)^2 + \left(5 - \frac{32}{3}\right)^2 \\ &= \left(\frac{13}{3}\right)^2 + \left(\frac{4}{3}\right)^2 + \left(-\frac{17}{3}\right)^2 \\ &= \left(\frac{13}{3}\right)^2 + \left(\frac{4}{3}\right)^2 + \left(-\frac{17}{3}\right)^2 \\ &= \frac{158}{3}, \end{aligned}$$

and

$$\begin{aligned} SSE &= \left(\frac{998}{98} - \frac{32}{3}\right)^2 + \left(\frac{1353}{98} - \frac{32}{3}\right)^2 + \left(\frac{785}{98} - \frac{32}{3}\right)^2 \\ &= \left(\frac{2994}{294} - \frac{3136}{294}\right)^2 + \left(\frac{4059}{294} - \frac{3136}{294}\right)^2 + \left(\frac{2355}{294} - \frac{3136}{294}\right)^2 \\ &= \left(-\frac{142}{294}\right)^2 + \left(\frac{923}{294}\right)^2 + \left(-\frac{781}{294}\right)^2 \\ &= \frac{1482054}{86436}. \quad \sim 17.146 \end{aligned}$$

Thus,

$$\begin{aligned} R^2 &= \frac{\frac{1482054}{86436}}{\frac{158}{3}} \\ &= \frac{1482054}{86436}\frac{3}{158} \\ &= \frac{741027}{2276148} \approx 0.326. \end{aligned}$$

$R^2$ is a statistical measure of how close the data are to the fitted regression line. In our case, the $R^2$ says that 32.6% of the variation in food consumption can be explained by disposable income.

**(e)** Calculate the OLS residuals by hand.

**Answer:** See table above.

**(h)** The true values are $\beta_0 = 0$ and $\beta_1 = 1$. Why are the OLS estimates different from the true values? *Hint:* Check whether MM assumptions 1 and 2 in the lecture notes are satisfied for the true values.

**Answer:** The OLS estimates are very different from the true values so we might suspect that the MM assumptions are violated. We begin by calculating the error terms for the population model

$$u_1 = y_1 - \beta_0 - \beta_1 x_1 = 15 - 0 - 1 \times 10 = 5,$$
$$u_2 = y_2 - \beta_0 - \beta_1 x_2 = 12 - 0 - 1 \times 15 = -3,$$
$$u_3 = y_3 - \beta_0 - \beta_1 x_3 = 5 - 0 - 1 \times 7 = -2.$$

MM assumption 1 holds since:

$$\sum_{i=1}^{3} u_i = 5 - 3 - 2 = 0.$$

However, MM assumption is violated because:

$$\sum_{i=1}^{3} u_i x_i = 5 \times 10 + (-3) \times 15 + (-2) \times 7 = -9.$$

Thus means that $Cov\,(u_i, x_i) \neq 0$, in which case the OLS estimators are biased and the reason why the OLS estimates are very different from the true values.

# Part 3: STATA exercises

Solutions for the Stata portion of the assignment are saved as Stata log output in assigment1.log. You can read this with any text editor and most internet browsers.