1 Review exercises

Basic statistical concepts

1. Moments of random variables.

   Suppose there is a discrete random variable Y with distinct values y.

   (a) Define the expected value and the variance of Y, denoted by $\mathbb{E}(Y)$ and $\mathbb{V}(Y)$, respectively.

   (b) What do these two moments tell us about the probability distribution of Y?

2. The law of iterated expectations.

   Assume that we have a second discrete random variable X with distinct values x.

   (a) Show that it is true that:

   $$\mathbb{E}(Y) = \mathbb{E}_x[\mathbb{E}(Y|X = x)], \tag{1}$$

   where the outer expectation is over the support of X. Equation (1) is called the "law of iterated expectations", which is one of the most useful concepts in econometrics.

   (b) Why is this law useful? Give a substantive example of how this law can be used in practical data analysis.

3. Covariances.

   The covariance captures to co-movement between two variables (regression is another tool to do that, see below). The covariance is defined as:

   $$Cov(y, x) = \mathbb{E}[(y - \mathbb{E}(y))(x - \mathbb{E}(x))] \tag{2}$$

   In this course, we will often use some alternative formulations of the covariance.

   (a) Show that we can write the covariance alternatively as:

   $$Cov(y, x) = \mathbb{E}(y \cdot x) - \mathbb{E}(y)\mathbb{E}(x) \tag{3}$$

   (b) Further show that the covariance can also be written as:

   $$Cov(y, x) = \mathbb{E}[(y - \mathbb{E}(y))x] \tag{4}$$
   $$Cov(y, x) = \mathbb{E}[(x - \mathbb{E}(x))y] \tag{5}$$

4. Correlations

   The correlation between x and y is defined as:

   $$Corr(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}} \tag{6}$$

   Is it true that $Corr(x, y) = Corr(y, x)$?

The linear regression model

5. The simple linear regression model.

Assume we have the following population model relating the outcome variable y_i to the explanatory variable x_i:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{7}$$

(a) What assumptions are already built into the (structural) model given by equation (7)?

(b) Discuss the assumptions needed in order to identify β_0 and β_1.
Hint: Take expectations conditional on x.

(c) Derive an estimator for β_1 (there are several ways of doing this). What assumptions are needed in order to get an estimate of β_1?
Hint: The easiest way of deriving the estimator is by the method of moments, based on the two conditions derived in part (b).

(d) Derive the variance of $\hat{\beta}_1$ under the assumption of a constant conditional error variance, i.e. assume that $\mathbb{V}(\epsilon|x) = \mathbb{V}(\epsilon) = \sigma^2$. This assumption is called homoscedasticity.
Hint: Start with the result from the preceding exercise and then write the estimator as a function of the regression error term (by replacing y_i with the structural model). Then derive the variance conditional on x.

(e) How would you estimate the variance of $\hat{\beta}_1$ from the sample data?

(f) What assumptions are needed for consistency of this estimator?

(g) Asymptotically we have $\hat{\beta}_1 \overset{a}{\sim} N(\beta_1, \mathbb{V}(\hat{\beta}_1))$. How would you test the hypothesis that $\beta_1 = 0$?

6. Omitted variable bias.

Assume that the structural model is as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i \tag{8}$$

However, because z is unobserved to us (or can not be measured easily), we may be forced to estimate the following model instead:

$$y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i, \quad \varepsilon_i = \beta_2 z_i + u_i \tag{9}$$

That is, we just leave out z because we don't observe it (equation (8) is often labelled long regression, while equation (9) is called short regression).

(a) Show that $\hat{\alpha}_1$ is generally a biased estimator for the target parameter β_1.
Hint: Note that the estimator for α_1 is the estimator you already derived before for the simple linear regression model. However, also note that the 'true' model is now given by the long regression given by equation (8).

(b) Derive the bias of $\hat{\alpha}_1$. Also discuss the circumstances under which $\hat{\alpha}_1$ is an unbiased estimator for β_1.
Hint: You have to derive the expected value of $\hat{\alpha}_1$ conditional on x and z.

(c) Think of a substantive example: Is $\hat{\alpha}_1$ likely to be an unbiased estimator of β_1 in your example?
Hint: one example would be y_i is wage, x_i is education and z_i is IQ. Find another example.

(d) What happens to the bias term if $n \to \infty$? What do we make of this?