Chapter 2: simple regression model

Goal: understand how to estimate and more importantly interpret the simple regression
Reading: chapter 2 of the textbook
Advice: this chapter is foundation of econometrics. You better have a solid understanding.

Discuss:

1. How to show the class size x and rating of eco201 instructors y are related? Comment on the limits of the following ideas.
2. Idea 1: estimate the covariance of x and y.
3. Idea 2: estimate the correlation (coefficient) of x and y.
4. Idea 3: let sample 1 includes the class with fewer than 40 students, and sample 2 includes the class with more than 40 students. Compare the average rating in the two samples (using two-sample t test).

Simple regression

1. Simple regression is used when we try to use only one independent variable (explanatory variable, regressor) x to explain the dependent variable y. We have two variables in mind, and we want to know their relationship.
2. We start with simple regression not because it is good, but because it is "simple". Actually simple regression has serious drawback: y can depend on many factors; but the simple regression just uses one of them. Ignoring other factors may induce some bias. That is, the estimated simple regression may capture the effect of other factors, instead of the true effect of x. We will revisit this bias issue repeatedly.
3. For instance, y is rating of eco201 instructor; x is the class size (the number of students in the class). We want to know how class size affects the rating. Other factors such as whether the class meets before 9 am matters, but the simple regression ignores them. Discuss: can you think of other factors that may affect rating?
4. Mathematically, the simple regression model is

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

where

(a) β_0 is the (unknown) intercept coefficient

(b) β_1 is the (unknown) slope coefficient.

(c) Graphically, the function $y = \beta_0 + \beta_1 x$ is a straight line. β_0 is the intercept, and β_1 is the slope. The line is upward sloping if $\beta_1 > 0$, and downward sloping if $\beta_1 < 0$.

(d) the error term (innovation, disturbance term, shock) u represents the factors other than x that affect y. The letter u is the first letter of "unobserved". The simple regression effectively treats all other factors as unobserved (as if their data are unavailable).

(e) y, x and u are all random variables. y and x are observed (for which data are available); while u is unobserved.

5. The key parameter is β_1, which measures the so called marginal effect

$$\beta_1 = \frac{dy}{dx} \approx \frac{\Delta y}{\Delta x} \tag{2}$$

(a) The model (1) assumes the marginal effect of x on y is constant. That means, y changes by a constant amount (or changes at a constant rate) when x changes by one unit. Later you will learn how to modify the regression to allow for non-constant marginal effect. Discuss: do you think the rating of instructor will change at constant rate when more and more students are added into a class? Hint: Think about a graph where y is on the vertical axis, and x is on the horizontal axis.

(b) The marginal effect is a mathematical concept. It assumes u is held constant. This reminds the ceteris paribus. That is why simple regression can be used for proving causality, but it must be used with caution. You need to verify certain assumption. See below.

6. We call (1) a "model" because it is based on a key assumption

$$E(u|x) = E(u) = 0 \tag{3}$$

(a) The $E(u) = 0$ part in (3) is not restrictive. Provided that the intercept term is

included, we can always redefine a new error term with zero mean if the original error term has nonzero mean.

(b) The $E(u|x) = E(u)$ part is crucial, for purpose of proving causality. For forecasting purpose, $E(u|x) = E(u)$ is not needed.

(c) $E(u|x) = E(u)$ implies that as x varies, the mean of u conditional on x remains constant (and the constant is $E(u)$). Put differently, it is as if u and x are independent so that $E(u|x) = E(u)$. The independence also reminds the ceteris paribus.

(d) Assumption (3) is equivalent to the assumption that

$$E(y|x) = \beta_0 + \beta_1 x \tag{4}$$

Proof: under (3),

$$E(y|x) = E(\beta_0 + \beta_1 x + u|x) = \beta_0 + \beta_1 x.$$

Assumption (4) implies the conditional mean $E(y|x)$ is linear and is equal to $\beta_0 + \beta_1 x$. Formally, $E(y|x) = \beta_0 + \beta_1 x$ is called population regression function (PRF).

(e) Assumptions (3) and (4) both fail when $E(u|x)$ depends on x, i.e., when

$$E(u|x) = f(x) \neq \texttt{constant}. \tag{5}$$

In that case the true conditional mean is $E(y|x) = \beta_0 + \beta_1 x + f(x)$, but $f(x)$ is missing on the right hand side of (4). So $\beta_0 + \beta_1 x$ mis-specifies $E(y|x)$.

(f) For instance, Assumptions (3) and (4) hold if the class size x is independent of all other factors u. Assumptions (3) and (4) fail when, say, small class is more likely to meet before 9 am than big class.

(g) Assumption (3) is difficult to check because it involves the conditional mean. However, Assumption (3) implies that the regressor x is uncorrelated with u, and this is easier to verify

$$E(u|x) = E(u) = 0 \Rightarrow \texttt{cov}(x, u) = 0 \tag{6}$$

Put differently, Assumptions (3) implies the regressor x is exogenous.

(h) Assumption (3) must be violated when x is correlated with any other factor, or $\texttt{cov}(x, u) \neq 0$. In that case the regressor x is endogenous.

Regressor is exogenous when it is uncorrelated with the error term

Regressor is endogenous when it is correlated with the error term

Assumption (3) is violated when regressor is endogenous

(i) For example, the class size is endogenous if it is correlated with the meeting time of the class. Assumption (3) is violated in that case. The class size is exogenous if it is uncorrelated with all other factors. This is a very restrictive requirement.

Most likely, the regressor in the simple regression is endogenous.

Endogeneity means difficulty in proving causality.

(j) The simple regression is very likely to suffer from the endogeneity issue. That is why simple regression is not the best choice to prove causality. If simple regression has to be used, it must be used with caution. We will revisit this issue later.

7. We can show the statistical interpretation of β_1 is

$$\beta_1 = \frac{\texttt{cov}(y, x)}{\texttt{var}(x)} \tag{7}$$

Therefore the method of moments (MM) estimator for β_1 is given as

$$\hat{\beta}_1 = \frac{s_{y,x}}{s_x^2} \tag{8}$$

$$s_{y,x} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1} \tag{9}$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{10}$$

where $s_{y,x}$ is the sample covariance of y and x; s_x^2 is the sample variance of x. Basically the MM estimator replaces population variance and covariance with sample variance and covariance. Remarks

(a) Formula (8) is used by the stata command `reg`

(b) Formula (8) shows the estimated slope coefficient $\hat{\beta}_1$ has the same sign as the sample covariance. If y and x move in the same direction, then they have positive sample covariance and positive $\hat{\beta}_1$.

(c) For instance we may believe the class size is negatively correlated with the rating. The instructor may perform better and get higher rating in small class than in big class. Then we guess $\hat{\beta}_1$ would be negative. You may want to figure out why if the sign of actual $\hat{\beta}_1$ contradicts with your guess.

(d) Formula (8) indicates the magnitude of $\hat{\beta}_1$ can be manipulated, simply because the covariance can be manipulated. Find out what will happen to $\hat{\beta}_1$ if we multiply x by a constant c.

(e) $\hat{\beta}_1$ cannot be computed if x does not vary. $s_x^2 = 0$ if that happens, but we cannot put zero at the denominator of (8). This result implies that

$$\boxed{\text{Big variance (variation) in } x \text{ is desirable.}}$$

Intuitively big variation in x means there is a lot of information about x which can be used for estimation. If x does not change or changes a little, the estimation can be imprecise.

(f) For instance, we hope our sample contains both big class and small class. We can not estimate the effect of class size on rating if all class in our sample have the same size.

(g) (7) indicates how to interpret β_1 and its estimate $\hat{\beta}_1$.

$$\boxed{\text{In general the slope coefficient } \beta_1 \text{ just measures the association (correlation).}}$$

A common mistake is trying to over-interpret β_1. Whether or not β_1 measures causality depends on Assumption (3). In short

$$\boxed{\beta_1 \text{ measures causality only when the regressor is exogenous!}}$$

Intuitively, ceteris paribus is satisfied by exogenous regressor because x and u are (as if) independent. So we can change x while holding all other factors u constant. If x and u are dependent (correlated), then β_1 only shows association, not causality.

8. Critical thinking: how to interpret and estimate the intercept coefficient β_0?

   (a) Using Assumption (3) and show

   $$\beta_0 = \mu_y - \beta_1 \mu_x \tag{11}$$

   (b) The MM estimator for β_0 is given by

   $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{12}$$

   where $\bar{y}$ and $\bar{x}$ are sample means of y and x. Why does (12) makes sense?

   (c) The stata command `reg y x` automatically reports $\hat{\beta}_0$. It is the coefficient of variable `_cons`

   (d) It can be meaningless to try to interpret β_0 because

   > The intercept coefficient β_0 measures the mean value of y when x equals *zero*.

   In our data, all classes have at least one student. So the class size x can not equal zero. Then β_0 becomes uninteresting.

   (e) How to modify the regression so that the new β_0 measures the mean value of y when x equals its *mean value*?

9. One advantage of the MM estimator is that calculus is not used. There is an alternative estimator, called ordinary least squares (OLS) estimator. The OLS estimator is intuitive, but requires calculus. Many old textbooks only discuss the OLS estimator. But modern econometrics emphasizes the MM estimator more and more. The good news is that the OLS estimators for β_0 and β_1 are the same as the MM estimators.

   (a) Denote the OLS estimator by $\hat{\beta}_0$ and $\hat{\beta}_1$. The hat emphasizes they are estimates. For each observation we can compute the fitted (predicted) value:

   $$\hat{y}_i \equiv \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad (i = 1, \ldots, n) \tag{13}$$

   where the subscript i is used to index the observation. x_i is the i-th observation of x. In total we have n fitted values: $(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$.

   (b) Because we use sample instead of population, most likely $\hat{\beta}_0$ and $\hat{\beta}_1$ are different from the true value β_0 and β_1. That means in general $y_i \neq \hat{y}_i$. Their difference,

called residual, captures the sampling error:

$$\hat{u}_i \equiv y_i - \hat{y}_i \tag{14}$$

In total we have n residuals.

(c) Now we have a very important decomposition:

$$y_i \equiv \hat{y}_i + \hat{u}_i \tag{15}$$

The fitted value $\hat{y}$ represents the part of y explained by x; the residual $\hat{u}$ represents the unexplained part. Put differently,

Residual is the part of y remained after the effect of x has been netted out.

(d) Because residual represents error, we want to minimize it. More explicitly, the OLS estimator is obtained by minimizing the sum of squared residuals:

$$\hat{\beta}_0, \hat{\beta}_0 = \texttt{argmin} \sum_{i=1}^{n} \hat{u}_i^2 \tag{16}$$

The objective function $\sum_{i=1}^{n} \hat{u}_i^2$ is called residual sum squares (RSS). Squaring is intended to penalize big error and avoid cancelation of positive and negative errors. Also a squared (quadratic) function has some nice math properties.

(e) From calculus, a (differentiable or smooth) function is minimized when the derivative is zero. So Appendix 2A of the textbook (on page 66) obtains the first order condition (FOC) (or necessary condition) by setting the (partial) derivatives of RSS with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ to zero. The FOC are

$$\sum_{i=1}^{n} \hat{u}_i = 0 \tag{17}$$

$$\sum_{i=1}^{n} \hat{u}_i x_i = 0 \tag{18}$$

The OLS estimator is the solution of the system of equations (17) and (18). The OLS estimators are the same as (8) and (12).

(f) Another unknown parameter is $\sigma = \mathtt{se}(u) = \sqrt{\mathtt{var}(u)}$. Its OLS estimator is

$$\hat{\sigma} = \sqrt{\frac{\mathtt{RSS}}{n-k-1}} \tag{19}$$

where n is the sample size, and k is the number of regressors. For simple regression $k = 1$. The $\hat{\sigma}$ is called standard error of regression (SER).

SER meaures the other factors contained in the error term.

To put SER into perspective, you may compare SER to μ_y. The regression has good fit if SER is small relative to μ_y.

(g) There is a formal measure of goodness of fit, called R squared (coefficient of determination). The formula is

$$R^2 = 1 - \frac{\mathtt{RSS}}{\mathtt{TSS}} = 1 - \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{20}$$

where TSS is shorthand for total sum squares $\mathtt{TSS} \equiv \sum_{i=1}^n (y_i - \bar{y})^2$. R^2 measures the fraction of variation in y that can be explained by the regressor x.

A model fits data well if R^2 is close to one.

For example, $R^2 = 0.4$ means 40% variation of the dependent variable can be explained by the model. Because R^2 measures the degree of linear association, not casuality, we have

Causality is the focus of econometrics; R^2 is not.