Regression #1

Econ 671

Purdue University

Introduction

- The linear regression model is the workhorse of econometrics.
- In this series of lectures, we will cover the following topics related to the regression model:

The basic linear regression specification is given as follows:

$$y_i = x_i \beta + \epsilon_i.$$

- Here, i is a subscript that indexes the observations, $i = 1, 2, \ldots, n$.
- y_i is a scalar outcome variable. We seek to determine how changes in x affect y.
- x_i is a $1 \times k$ vector of explanatory variables. Specifically,

  $$x_i = [x_{i1} \;\; x_{i2} \;\; \cdots \;\; x_{ik}].$$

  There are k different variables employed in the regression model. Typically, $x_{i1} = 1 \;\; \forall i$ so that the model contains an intercept parameter.

$$y_i = x_i\beta + \epsilon_i.$$

- β is a $k \times 1$ vector of fixed, but unknown parameters. We seek to use the data $\{(x_i, y_i)\}_{i=1}^n$ to estimate β.
- ϵ_i is an error term, picking up factors that explain variation in y that are not captured by x. In some introductory texts, the following assumption is made:

$$E(\epsilon_i) = 0.$$

Complete on your own:

(Why is this not really much of an assumption? What if the mean of ϵ_i was, say, c instead?)

Complete on your own:

- We will also begin by assuming that the errors are independently distributed and *homoscedastic*, i.e.,

$$E(\epsilon_i^2|X) = \sigma^2 \quad \forall i.$$

Vector/Matrix Representation

$$y_i = x_i\beta + \epsilon_i.$$

This regression equation holds for each observation i. We can *stack* this information across observations as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

or, written out completely,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Vector/Matrix Representation

We write this compactly as

$$y = X\beta + \epsilon,$$

where y is $n \times 1$, X is $n \times k$, β is $k \times 1$ and ϵ is $n \times 1$. The columns of the X matrix list the different variables employed in the analysis while the rows list the values of all variables for each observation.

Example

To fix ideas, consider the following regression equation:

-

You record a bunch of information from a survey, stack the quantities into vectors and obtain

$$y = Wage = \begin{bmatrix} 15 \\ 25 \\ 18 \\ 35 \\ \vdots \\ 20 \end{bmatrix}, \quad Education = \begin{bmatrix} 12 \\ 12 \\ 16 \\ 11 \\ \vdots \\ 20 \end{bmatrix}.$$

Importantly,

-

Example, Continued

When writing the equation

$$y = X\beta + \epsilon,$$

note

$$X = \begin{bmatrix} & \\ & \\ & \\ & \\ \vdots & \vdots \\ & \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Example, Continued

When writing the equation

$$y = X\beta + \epsilon,$$

note

$$X = \begin{bmatrix} 1 & 12 \\ 1 & 12 \\ 1 & 16 \\ 1 & 11 \\ \vdots & \vdots \\ 1 & 20 \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The COLUMNS of X are formed of the various right-hand side explanatory variables. The first column, for example, is a column of ones (for the intercept parameter).

Example, Continued

When writing the equation

$$y = X\beta + \epsilon,$$

note

$$X = \begin{bmatrix} 1 & 12 \\ 1 & 12 \\ 1 & 16 \\ 1 & 11 \\ \vdots & \vdots \\ 1 & 20 \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Example, Continued

Other variables could be added in the obvious way ...

-

note

$$X = \begin{bmatrix} 1 & 12 & 0 \\ 1 & 12 & 1 \\ 1 & 16 & 3 \\ 1 & 11 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 20 & 1 \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

The new variable Siblings is added as the third column of X.

Vector/Matrix Representation

For later purposes, it is also useful to note an alternate way to represent the matrix $X'X$ as well as similar quantities. To this end, note:

$$\begin{aligned} X'X &= \begin{bmatrix} \overline{x}_1' \\ \overline{x}_2' \\ \vdots \\ \overline{x}_k' \end{bmatrix} [\overline{x}_1 \quad \overline{x}_2 \quad \cdots \overline{x}_k] \\ &= \begin{bmatrix} \overline{x}_1'\overline{x}_1 & \overline{x}_1'\overline{x}_2 & \cdots & \overline{x}_1'\overline{x}_k \\ \overline{x}_2'\overline{x}_1 & \overline{x}_2'\overline{x}_2 & \cdots & \overline{x}_2'\overline{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ \overline{x}_k'\overline{x}_1 & \overline{x}_k'\overline{x}_2 & \cdots & \overline{x}_k'\overline{x}_k \end{bmatrix}. \end{aligned}$$

where $\overline{x}_j \equiv [x_{1j} \quad x_{2j} \quad \cdots \quad x_{nj}]'$, the j^{th} *column* of X.

Vector/Matrix Representation

Likewise,

Vector/Matrix Representation

Similarly, other expressions like the $k \times 1$ vector $X'\epsilon$ can be written as

$$X'\epsilon = \sum_i x_i'\epsilon_i.$$

Writing the products in this way will prove to be convenient in terms of the asymptotic derivations that will come later.

Assumptions of the Regression Model

- Before discussing estimation, we must first set forth some assumptions regarding the regression model.
- Some of these assumptions are more critical than others, and some of these can be relaxed without too much difficulty.
- You should *not* think of these as being obvious or necessarily satisfied; the seeming validity of the assumptions that follow will vary with the application at hand.

Assumptions of the Regression Model

- The matrix X is full column rank, i.e., $Rank(X) = k$. (*What does this mean?*)
-
- When will this assumption most likely fail?
  Poor specification choice on the part of the researcher, e.g.:

$$y_i = \alpha_0 + \alpha_1 Male_i + \alpha_2 Female_i + \alpha_3 Education_i + \epsilon_i.$$

$$Wage_i = \beta_0 + \beta_1 age + \beta_2 Education + \beta_3 Experience + u_i.$$

Assumptions of the Regression Model

Another possible case where the assumption will fail is because of unexpected small sample problems:

$$y_i = \alpha_0 North_i + \alpha_1 South_i + Education_i + \epsilon_i$$

$$X = \begin{bmatrix} 1 & 0 & 12 \\ 1 & 0 & 12 \\ 0 & 1 & 16 \\ 0 & 1 & 16 \end{bmatrix}.$$

- In these cases, $(X'X)^{-1}$ does not exist. In many standard software packages, the program will often decide to drop a variable when this problem persists.
- In the context of a simple regression model, where only one x is included, you may have seen this assumption described as "There is some variation in x."

Assumptions of the Regression Model

With respect to the error terms, we will assume the following:

$$E(\epsilon|X) = 0,$$

an assumption commonly referred to as *mean-independence*.
In many cases, we could replace this with the weaker assumption:

$$E(X'\epsilon) = 0,$$

Complete on your own: Why is this a weaker assumption?

Assumptions of the Regression Model

Complete on your own:

To show that these two statements are not synonymous, can you find a case where $E(XY) = 0$ but $E(X|Y) \neq 0$?

Assumptions of the Regression Model

- The mean-independence assumption is critical, and you should not think it will be satisfied in all cases.
- Specifically, in models with *endogeneity problems*, models with *measurement error* in the right-hand side variables, and *simultaneous equations* models, this assumption will be violated.
- As a result, the properties of standard estimators are poor (as we will discuss later in the course), and other estimators can be used to restore these desirable properties.
- Consider:

$$Wage_i = \beta_0 + \beta_1 Education_i + \epsilon_i.$$

  Do you think $E(\epsilon | Education) = 0$?

Assumptions of the Regression Model

For now, we will also make an assumption regarding the second moment of ϵ:

$$E(\epsilon\epsilon'|X) = \mathrm{Var}(\epsilon|X) = \sigma^2 I_n.$$

As stated before, this is a *homoscedasticity* assumption.

This assumption is not required to derive many properties of the OLS estimator, but will be needed in order to characterize the asymptotics of this estimator.

Later, we will discuss how this assumption can be relaxed, and replaced by a more realistic assumption of *heteroscedasticity*.