SEPTEMBER 29, 2014

**LECTURE 2**
**LINEAR REGRESSION MODEL AND OLS**

# Definitions

A common question in econometrics is to study the effect of one group of variables $X_i$, usually called the *regressors*, on another $Y_i$, the *dependent variables*. An econometrician observes the random data:

$$(Y_1, X_1), (Y_2, X_2), \ldots (Y_n, X_n), \tag{1}$$

where for $i = 1, \ldots, n$, $Y_i$ is a random variable and $X_i$ is a random $k$-vector:

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik} \end{pmatrix}.$$

A pair $(Y_i, X_i)$ is called the *observation*, and the collection of observations in (1) is called the *sample*. The vector $X_i$ collects the values of $k$ variables for observation $i$.

The joint distribution of (1) is called the *population*. The population does not correspond to any physical population, but to a probability space. In a *cross-sectional* framework (each observation is a different individual or a firm etc.), it is often natural to assume that all observations are independently drawn from the same distribution. In this case, the population is described by the distribution of a single observation $(Y_1, X_1)$, which can be stated as well as $(Y_i, X_i)$ are iid for $i = 1, \ldots, n$. Note that the iid assumption does not imply that $Y_i$ and $X_i$ are independent, but rather that the random vector $(Y_i, X_i)$ is independent from $(Y_j, X_j)$ for $i \neq j$. At the same time, $Y_i$ and $X_i$ are still can be related.

In cross-sections, the relationship between the regressors and the dependent variable is modelled through the conditional expectation $E(Y_i|X_i)$. The deviation of $Y_i$ from its conditional expectation is called the *error* or *residual*:

$$U_i = Y_i - E(Y_i|X_i). \tag{2}$$

Contrary to $X_i$ and $Y_i$, the residual $U_i$ is not observable, since the conditional expectation function is unknown to the econometrician.

In the *parametric* framework, it is assumed that the conditional expectation function depends on a number of unknown constants or *parameters*, and that the functional form of $E(Y_i|X_i)$ is known. In the linear regression model, it is assumed that $E(Y_i|X_i)$ is *linear in the parameters*:

$$\begin{aligned} E(Y_i|X_i) &= \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \\ &= X_i'\beta, \end{aligned} \tag{3}$$

where

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

is a $k$-vector of unknown constants. The linearity of $E(Y_i|X_i)$ can be justified, for example, by saying that $(Y_i, X_i)$ jointly has a multivariate normal distribution. Since $\beta_j = \frac{\partial E(Y_i|X_i)}{\partial X_{ij}}$, the vector $\beta$ is a vector of *marginal effects* of $X_i$, i.e. $\beta_j$ gives the change in the conditional mean of $Y_i$ per unit change in $X_{ij}$, while holding the values of other variables ($X_{il}$ for $l \neq j$) fixed. One of the objectives is *estimation* of unknown $\beta$ from the sample (1).

Note that combining together equations (2) and (3), one can write:

$$Y_i = X_i'\beta + U_i. \tag{4}$$

By definition (2),

$$E\left(U_i|X_i\right) = 0.$$

This implies, that the regressors contain no information on the deviation of $Y_i$ from its conditional expectation. Further, the Law of Iterated Expectation (LIE) implies that the residuals have zero mean: $EU_i = 0$. If $(Y_i, X_i)$ are iid, then the residuals $\{U_i : i = 1, \ldots, n\}$ are iid as well.

In the *classical regression model*, it is assumed that the variance of the errors $U_i$ is independent of the regressors and the same for all observations:

$$Var\left(U_i|X_i\right) = \sigma^2,$$

for some constant $\sigma^2 > 0$. This property is called *homoskedasticity*.

## Assumptions

In this section, we formally define the linear regression model. Let's define

$$\begin{aligned} X &= \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} \\ &= \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & \ldots & X_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{pmatrix}, \\ Y &= \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \text{ and} \\ U &= \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{pmatrix}. \end{aligned}$$

The following are the four classical regression assumptions:

**(A1)** $Y = X\beta + U$.

**(A2)** $E\left(U|X\right) = 0$ a.s.

**(A3)** $Var\left(U|X\right) = \sigma^2 I_n$ a.s.

**(A4)** rank$(X) = k$ a.s.[1]

[1] The column (row) rank of a matrix is the maximum number of linearly independent columns (rows). One can show that, for any matrix, the column and row ranks are equal. If $A$ is an $n \times k$ matrix, then rank$(A) \leq \min(n, k)$. If rank$(A) = n$ (or rank$(A) = k$), we say that $A$ has full row (column) rank. Properties:

$$\begin{aligned} \text{rank}\left(A\right) &= \text{rank}\left(A'\right) = \text{rank}\left(A'A\right) = \text{rank}\left(AA'\right), \\ \text{rank}\left(AB\right) &\leq \min\left(\text{rank}\left(A\right), \text{rank}\left(B\right)\right), \\ \text{rank}\left(AB\right) &= \text{rank}\left(A\right) \text{ if } B \text{ is square and of full rank.} \end{aligned}$$

Instead of conditioning on the observed values of the regressors, in the classical regression model, one can assume that $X$ is not random, i.e. the value of $X$ are fixed in repeated samples. In this case, (A2) is replaced by $E(U) = 0$ and (A3) is replaced by $Var(U) = \sigma^2 I_n$. Since conditioning on $X$ is basically equivalent to treating them as fixed, two sets of assumptions lead to the same results.

For inference purposes, sometimes it is assumed that:

**(A5)** $U|X \sim N\left(0, \sigma^2 I_n\right)$.

In the case of fixed regressors, it is assumed instead that the unconditional distribution of the errors is normal. Assumptions (A1)-(A5) define the *classical normal regression model.* In this case,

$$Y|X \sim N\left(X\beta, \sigma^2 I_n\right).$$

Note that since all covariance elements in (A5) are zeros, (A5) implies independence of the residuals. Assumptions (A1)-(A4) alone do not imply independence across observations. Actually, for many important results, independence across observations is not required. Nevertheless, sometimes we would like to assume independence without normality:

**(A6)** $\{(Y_i, X_i) : i = 1, \ldots, n\}$ are iid.

In the case of fixed regressors, (A6) can be replaced with the assumption that $\{U_i : i = 1, \ldots, n\}$ are iid.

Assumption (A2) says that $U$ is mean independent of $X$. This is actually a very strong assumption. As we have mentioned above, it is equivalent to assuming that $X$ are not random. In many economic applications, it is hard to justify. However, many important results may be obtained with a weaker assumption of uncorrelatedness:

**(A2*)** For $i = 1, \ldots, n$, $E\left(U_i X_i\right) = 0$ and $E\left(U_i\right) = 0$.

Note that under (A2*), the expression $X_i'\beta$ does not have the interpretation of the conditional expectation. In this case, one can treat (4) as a *data generating process*.

Assumption (A3) implies that, the residuals $U_i$ have the same variance for all $i$, and uncorrelated with each other: $E\left(U_i U_j | X\right) = 0$ for $i \neq j$. If desired, independence of the residuals may be achieved through (A5) or Assumptions (A1) and (A6).

Assumption (A4) says that the columns of $X$ are not linearly dependent. If it is violated, then one or more regressors simply duplicate the information contained in other regressors, and thus should be omitted.

Often, the first column of the matrix $X$ is a column of ones. In this case, its coefficient $\beta_1$ is called the intercept. The intercept gives the average value of the dependent variable when all regressors are equal to zero.

## Estimation by the method of moments

One of the objectives of econometric analysis is *estimation* of unknown parameters $\beta$ and $\sigma^2$. An *estimator* is *any function* of the sample $\{(Y_i, X_i) : i = 1, \ldots, n\}$. An estimator can depend on the unknown residuals $U_i$ or unknown parameters like $\beta$ only through the observed variables $Y$ and $X$. An estimator usually is not unique, i.e. there exists a number of alternative estimators for the same parameter.

One of the oldest methods of finding estimators is called the *method of moments* (MM). The MM says to replace population moments (expectations) with the corresponding sample moments (averages). Assumptions (A2) or (A2*) imply that at the true value of $\beta$,

$$\begin{aligned} 0 &= E\left(U_i X_i\right) \\ &= E\left(\left(Y_i - X_i'\beta\right) X_i\right). \end{aligned} \tag{5}$$

Let $\widehat{\beta}$ be an estimator of $\beta$. According to the MM, we replace expectation in (5) with the sample average:

$$\begin{aligned} 0 &= n^{-1}\sum_{i=1}^{n}\left(Y_i - X_i'\widehat{\beta}\right)X_i \\ &= n^{-1}\sum_{i=1}^{n}X_iY_i - n^{-1}\sum_{i=1}^{n}X_iX_i'\widehat{\beta}. \end{aligned}$$

(Note that $Y_i - X_i'\widehat{\beta}$ is a scalar). Solving for $\widehat{\beta}$, one obtains:

$$\begin{aligned} \widehat{\beta} &= \left(n^{-1}\sum_{i=1}^{n}X_iX_i'\right)^{-1} n^{-1}\sum_{i=1}^{n}X_iY_i \\ &= \left(\sum_{i=1}^{n}X_iX_i'\right)^{-1}\sum_{i=1}^{n}X_iY_i \\ &= \left(X'X\right)^{-1}X'Y. \end{aligned} \tag{6}$$

The matrix $\sum_{i=1}^{n} X_iX_i' = X'X$ is invertible due to Assumption (A4). To show that $X'X = \sum_{i=1}^{n} X_iX_i'$, note that

$$\begin{aligned} X'X &= \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}' \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} \\ &= \begin{pmatrix} X_1 & X_2 & \dots & X_n \end{pmatrix} \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} \\ &= X_1X_1' + X_2X_2' + \ldots + X_nX_n' \\ &= \sum_{i=1}^{n}X_iX_i'. \end{aligned}$$

The expression $X_i'\widehat{\beta}$ gives the *estimated regression line*, with $\widehat{Y}_i = X_i'\widehat{\beta}$ being the *predicted* value of $Y_i$, and $\widehat{U}_i = Y_i - X_i'\widehat{\beta}$ being the *sample residual,*

$$\widehat{U} = Y - X\widehat{\beta}.$$

The vector $\widehat{U}$ is a function of the estimator of $\beta$. In the case of the MM estimator, the sample residuals have to satisfy the sample *normal equation*:

$$\begin{aligned} 0 &= X'\widehat{U} \qquad (7) \\ &= \sum_{i=1}^{n}\widehat{U}_iX_i \\ &= \begin{pmatrix} \sum_{i=1}^{n}\widehat{U}_iX_{i1} \\ \sum_{i=1}^{n}\widehat{U}_iX_{i2} \\ \vdots \\ \sum_{i=1}^{n}\widehat{U}_iX_{ik} \end{pmatrix}. \end{aligned}$$

If the model contains an intercept, i.e. $X_{i1} = 1$ for all $i$, then the normal equation implies that $\sum_{i=1}^{n}\widehat{U}_i = 0$.

In order to estimate $\sigma^2$, write:

$$\begin{aligned}\sigma^2 &= EU_i^2 \\ &= E\left(Y_i - X_i'\beta\right)^2.\end{aligned}$$

Since $\beta$ is unknown, we must replace it by its MM estimator:

$$\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \left(Y_i - X_i'\widehat{\beta}\right)^2.$$

## Least Squares

Let $b$ be an estimator of $\beta$. The *the ordinary least squares* (OLS) estimator is an estimator of $\beta$ that minimizes the *sum-of-squared errors* function:

$$\begin{aligned}S(b) &= \sum_{i=1}^n \left(Y_i - X_i'b\right)^2 \\ &= \left(Y - Xb\right)'\left(Y - Xb\right).\end{aligned}$$

It turns out that $\widehat{\beta}$, the MM estimator presented in the previous section, is the OLS estimator as well. In order to show that, write

$$\begin{aligned}S(b) &= \left(Y - Xb\right)'\left(Y - Xb\right) \\ &= \left(Y - X\widehat{\beta} + X\widehat{\beta} - Xb\right)'\left(Y - X\widehat{\beta} + X\widehat{\beta} - Xb\right) \\ &= \left(Y - X\widehat{\beta}\right)'\left(Y - X\widehat{\beta}\right) \\ &\quad + \left(X\widehat{\beta} - Xb\right)'\left(X\widehat{\beta} - Xb\right) \\ &\quad + 2\left(Y - X\widehat{\beta}\right)'\left(X\widehat{\beta} - Xb\right) \\ &= \left(Y - X\widehat{\beta}\right)'\left(Y - X\widehat{\beta}\right) \\ &\quad + \left(\widehat{\beta} - b\right)'X'X\left(\widehat{\beta} - b\right) \\ &\quad + 2\widehat{U}'X\left(\widehat{\beta} - b\right) \text{ (equals zero because of the normal equations)} \\ &= \left(Y - X\widehat{\beta}\right)'\left(Y - X\widehat{\beta}\right) + \left(\widehat{\beta} - b\right)'X'X\left(\widehat{\beta} - b\right).\end{aligned}$$

Minimization of $S(b)$ is equivalent to minimization of $\left(\widehat{\beta} - b\right)'X'X\left(\widehat{\beta} - b\right)$, because $\left(Y - X\widehat{\beta}\right)'\left(Y - X\widehat{\beta}\right)$ is not a function of $b$. If $X$ is of full column rank, as assumed in (A4), $X'X$ is a positive definite matrix, and therefore

$$\left(\widehat{\beta} - b\right)'X'X\left(\widehat{\beta} - b\right) \geq 0,$$

where $\left(\widehat{\beta} - b\right)'X'X\left(\widehat{\beta} - b\right) = 0$ if and only if $\widehat{\beta} - b = 0$.

Alternatively, one can show that $\widehat{\beta}$ as defined in (6) is the OLS estimator, by taking the derivative of $S(b)$ with respect to $b$, and solving the first order condition $\frac{dS(\widehat{\beta})}{db} = 0$. Write

$$S(b) = Y'Y - 2b'X'Y + b'X'Xb.$$

Using the fact that for a symmetric matrix $A$ we have that

$$\frac{\partial x'Ax}{\partial x} = 2Ax,$$

the first order condition is

$$\frac{\partial S\left(\widehat{\beta}\right)}{\partial b} = -2X'Y + 2X'X\widehat{\beta} = 0. \tag{8}$$

Solving for $\widehat{\beta}$, one obtains:

$$\widehat{\beta} = \left(X'X\right)^{-1} X'Y. \tag{9}$$

Note also that the first order condition (8) can be written as $X'\left(Y - X\widehat{\beta}\right) = 0$, which gives us normal equation (7).

## Properties of $\widehat{\beta}$

1. $\widehat{\beta}$ is a *linear* estimator. An estimator $b$ is linear if it can be written as $b = AY$, where $A$ is some matrix, which depends $X$ alone, and does not depend on $Y$. In the case of the OLS, $A = \left(X'X\right)^{-1} X'$.

2. Under Assumptions (A1), (A2) and (A4), $\widehat{\beta}$ is an *unbiased* estimator, i.e.

   $$E\widehat{\beta} = \beta.$$

   In order to show unbiasedness, first, plug-in the expression for $Y$ in Assumption (A1) into equation (9):

   $$\begin{aligned} \widehat{\beta} &= \left(X'X\right)^{-1} X'\left(X\beta + U\right) \\ &= \beta + \left(X'X\right)^{-1} X'U. \end{aligned} \tag{10}$$

   Next, consider the conditional expectation of $\widehat{\beta}$ given $X$:

   $$\begin{aligned} E\left(\widehat{\beta}|X\right) &= E\left(\beta + \left(X'X\right)^{-1} X'U|X\right) \\ &= \beta + E\left(\left(X'X\right)^{-1} X'U|X\right). \end{aligned}$$

   Now,

   $$\begin{aligned} E\left(\left(X'X\right)^{-1} X'U|X\right) &= \left(X'X\right)^{-1} X'E\left(U|X\right) \\ &= 0, \end{aligned}$$

   since, by Assumption (A2), $E\left(U|X\right) = 0$. Therefore,

   $$E\left(\widehat{\beta}|X\right) = \beta. \tag{11}$$

   Finally, the LIE implies that

   $$\begin{aligned} E\widehat{\beta} &= EE\left(\widehat{\beta}|X\right) \\ &= \beta. \end{aligned}$$

   Equation (11) shows that $\widehat{\beta}$ is conditionally unbiased given $X$. Note that unbiasedness cannot be obtained under Assumption (A2*).

3. Under Assumptions (A1), (A2) and (A4),

$$Var\left(\widehat{\beta}|X\right) = \left(X'X\right)^{-1} X'E\left(UU'|X\right) X \left(X'X\right)^{-1}.$$

Under the homoskedastic errors, Assumption (A3), the expression for the conditional variance simplifies to

$$Var\left(\widehat{\beta}|X\right) = \sigma^2 \left(X'X\right)^{-1}.$$

To show that, first by the definition of the variance we have:

$$\begin{aligned} Var\left(\widehat{\beta}|X\right) &= E\left(\left(\widehat{\beta} - E\left(\widehat{\beta}|X\right)\right)\left(\widehat{\beta} - E\left(\widehat{\beta}|X\right)\right)'|X\right) \\ &= E\left(\left(\widehat{\beta} - \beta\right)\left(\widehat{\beta} - \beta\right)'|X\right) \\ &= E\left(\left(X'X\right)^{-1} X'UU'X\left(X'X\right)^{-1}|X\right) \text{ (follows from (10))} \\ &= \left(X'X\right)^{-1} X'E\left(UU'|X\right) X\left(X'X\right)^{-1}. \end{aligned}$$

Under homoskedasticity, $E\left(UU'|X\right) = \sigma^2 I_n$, and

$$\begin{aligned} \left(X'X\right)^{-1} X'E\left(UU'|X\right) X\left(X'X\right)^{-1} &= \left(X'X\right)^{-1} X'\left(\sigma^2 I_n\right) X\left(X'X\right)^{-1} \\ &= \sigma^2\left(X'X\right)^{-1} X'X\left(X'X\right)^{-1} \\ &= \sigma^2\left(X'X\right)^{-1}. \end{aligned}$$

In the case of fixed regressors, $Var\left(\widehat{\beta}\right) = \sigma^2\left(X'X\right)^{-1}$.

4. If we add normality of the errors, i.e. under Assumptions (A1)-(A5), we obtain the following result:

$$\widehat{\beta}|X \sim N\left(\beta, \sigma^2\left(X'X\right)^{-1}\right).$$

It is sufficient to show that, conditional on $X$, the distribution of $\widehat{\beta}$ is normal. Then, $\widehat{\beta}|X \sim N\left(E\left(\widehat{\beta}|X\right), Var\left(\widehat{\beta}|X\right)\right)$. However, normality of $\widehat{\beta}|X$ follows from the facts that $\widehat{\beta}$ is a linear function of $Y$, and that $Y|X$ is normal by Assumption (A5) (see Normal distribution Section in Lecture 1). Again, in the case of fixed regressors, we simply omit conditioning on $X$,

$$\widehat{\beta} \sim N\left(\beta, \sigma^2\left(X'X\right)^{-1}\right).$$

5. *Efficiency* or the Gauss-Markov Theorem: Under Assumptions (A1)-(A4), the OLS estimator is the Best Linear Unbiased Estimator of $\beta$ (BLUE), where "best" means having the smallest variance, i.e. for any linear and unbiased estimator $b$, we have that $Var\left(b|X\right) - Var\left(\widehat{\beta}|X\right)$ is a positive semi-definite matrix:

$$Var\left(b|X\right) - Var\left(\widehat{\beta}|X\right) \geq 0.$$

Furthermore, if $\tilde{\beta}$ is a linear unbiased estimator and $Var(\tilde{\beta}|X) = Var\left(\widehat{\beta}|X\right)$, then $\tilde{\beta} = \widehat{\beta}$ with probability one.

Note that since the theorem discusses the conditional variance of the OLS estimator, unbiasedness in the statement of the theorem actually refers to unbiasedness conditional on $X$, i.e. $E\left(b|X\right) = \beta$.

**Proof:** Let $b$ be a linear and unbiased estimator of $\beta$ :

$$\begin{aligned} b &= AY, \\ E\left(b|X\right) &= \beta. \end{aligned}$$

These conditions imply that $AX = I$ with probability 1. Indeed,

$$\begin{aligned} E(b|X) &= E\left(A\left(X\beta + U\right)|X\right) \\ &= AX\beta + AE(U|X). \end{aligned}$$

By Assumption (A2), $E(U|X) = 0$, and thus, in order to obtain unbiasedness we need that $AX = I_k$. Next, we show that $Cov\left(\widehat{\beta}, b|X\right) = Var\left(\widehat{\beta}|X\right)$:

$$\begin{aligned} Cov\left(\widehat{\beta}, b|X\right) &= E\left(\left(\widehat{\beta} - \beta\right)(b - \beta)'|X\right) \\ &= E\left(\left(X'X\right)^{-1} X'UU'A'|X\right) \\ &= \left(X'X\right)^{-1} X'E\left(UU'|X\right)A' \\ &= \sigma^2\left(X'X\right)^{-1} X'A' \text{ (since, by Assumption (A3), } E\left(UU'|X\right) = \sigma^2 I_n) \\ &= \sigma^2\left(X'X\right)^{-1} \text{ (since } X'A' = I_k) \\ &= Var\left(\widehat{\beta}|X\right). \end{aligned}$$

Finally,

$$\begin{aligned} Var\left(\widehat{\beta} - b|X\right) &= Var\left(\widehat{\beta}|X\right) - Cov\left(\widehat{\beta}, b|X\right) - Cov\left(b, \widehat{\beta}|X\right) + Var\left(b|X\right) \\ &= Var\left(b|X\right) - Var\left(\widehat{\beta}|X\right). \end{aligned} \tag{12}$$

Now, since any variance-covariance matrix is positive semi-definite, we have that

$$Var\left(b|X\right) - Var\left(\widehat{\beta}|X\right) \geq 0.$$

To show uniqueness, suppose that there is a linear unbiased estimator $\tilde{\beta}$ such that $Var(\tilde{\beta}|X) = Var\left(\widehat{\beta}|X\right)$. Then, by (12), $Var\left(\widehat{\beta} - \tilde{\beta}|X\right) = 0$, and therefore $\tilde{\beta} = \widehat{\beta} + c(X)$ for some $k$-vector-valued function $c(X)$ that can depend only on $X$. However, since both $\widehat{\beta}$ and $\tilde{\beta}$ are conditionally unbiased given $X$, it follows that $c(X) = 0$ with probability one. □

Note that Assumption (A3), $E\left(UU'|X\right) = \sigma^2 I_n$, plays a crucial rule in the proof of the result. We could not draw conclusion about efficiency of OLS without it.