October 13, 2019

Regression Model and Least Squares

Regression model

A common question in econometrics is to study the effect of one group of variables $\boldsymbol{X}_i$, usually called the *regressors*, on another Y_i, the *dependent variable.* An econometrician observes the random data:

$$(Y_1, \boldsymbol{X}_1), (Y_2, \boldsymbol{X}_2), \ldots (Y_n, \boldsymbol{X}_n), \tag{1}$$

where for $i = 1, \ldots, n$, Y_i is a random variable and $\boldsymbol{X}_i$ is a random k-vector:

$$\boldsymbol{X}_i = \begin{pmatrix} X_{i,1} \\ X_{i,2} \\ \vdots \\ X_{i,k} \end{pmatrix}.$$

A pair $(Y_i, \boldsymbol{X}_i)$ is called the *observation*, and the collection of observations in (1) is called the *sample.* The vector $\boldsymbol{X}_i$ 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 $(Y_1, \boldsymbol{X}_1)$, which can be stated as well as $(Y_i, \boldsymbol{X}_i)$, since (1) are iid for $i = 1, \ldots, n$. Note that the iid assumption does not imply that Y_i and $\boldsymbol{X}_i$ are independent, but rather that the random vector $(Y_i, \boldsymbol{X}_i)$ is independent from $(Y_j, \boldsymbol{X}_j)$ for $i \neq j$. At the same time, Y_i and $\boldsymbol{X}_i$ are still can be related.

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

$$e_i = Y_i - \mathbb{E}(Y_i|\boldsymbol{X}_i). \tag{2}$$

Contrary to $\boldsymbol{X}_i$ and Y_i, the residual e_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 $\mathbb{E}(Y_i|X_i)$ is known. In the linear regression model, it is assumed that $\mathbb{E}(Y_i|\boldsymbol{X}_i)$ is *linear in the parameters*:

$$\begin{aligned} \mathbb{E}(Y_i|\boldsymbol{X}_i) &= \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_k X_{i,k} \\ &= \boldsymbol{X}_i' \boldsymbol{\beta}, \end{aligned} \tag{3}$$

where

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

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

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

$$Y_i = \boldsymbol{X}_i'\beta + e_i. \tag{4}$$

By definition (2),

$$\mathbb{E}\left(e_i|\boldsymbol{X}_i\right) = 0.$$

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: $\mathbb{E}e_i = 0$. If $(Y_i, \boldsymbol{X}_i)$ are iid, then the residuals $\{e_i : i = 1, \ldots, n\}$ are iid as well.

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

$$\mathrm{Var}\left(e_i|\boldsymbol{X}_i\right) = \sigma^2,$$

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

Estimation by the method of moments

One of the objectives of econometric analysis is *estimation* of unknown parameters $\boldsymbol{\beta}$ and σ^2. An *estimator* is *any function* of the sample $\{(Y_i, \boldsymbol{X}_i) : i = 1, \ldots, n\}$. An estimator can depend on the unknown residuals e_i or unknown parameters like $\boldsymbol{\beta}$ only through the observed variables Y and $\boldsymbol{X}$. 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). The condition $\mathbb{E}\left(e_i|\boldsymbol{X}_i\right) = 0$ imply that at the true value of $\boldsymbol{\beta}$,

$$\begin{aligned} 0 &= \mathbb{E}\left(e_i\boldsymbol{X}_i\right) \\ &= \mathbb{E}\left(\left(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}\right)\boldsymbol{X}_i\right). \end{aligned} \tag{5}$$

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

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

(Note that $Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ is a scalar). Denote

$$
\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,k} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,k} & \cdots & X_{n,k} \end{bmatrix}_{n\times k} \quad \text{and } \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1}.
$$

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

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1} n^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i Y_i \\
&= \left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i Y_i \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}. \qquad (6)
\end{aligned}
$$

To show that $\boldsymbol{X}'\boldsymbol{X} = \sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'$, note that

$$
\begin{aligned}
\boldsymbol{X}'\boldsymbol{X} &= \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix}' \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 & \dots & \boldsymbol{X}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} \\
&= \boldsymbol{X}_1\boldsymbol{X}_1' + \boldsymbol{X}_2\boldsymbol{X}_2' + \ldots + \boldsymbol{X}_n\boldsymbol{X}_n' \\
&= \sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'.
\end{aligned}
$$

The expression $\boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ gives the *estimated regression line*, with $\widehat{Y}_i = \boldsymbol{X}_i'\widehat{\beta}$ being the *fitted (or predicted)* value of Y_i, and $\widehat{e}_i = Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ being the *sample residual,*

$$\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{e}_1 \\ \widehat{e}_2 \\ \vdots \\ \widehat{e}_n \end{pmatrix}.$$

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

$$\begin{aligned} 0 &= \boldsymbol{X}'\widehat{\boldsymbol{e}} \qquad (7) \\ &= \sum_{i=1}^{n} \widehat{e}_i \boldsymbol{X}_i \\ &= \begin{pmatrix} \sum_{i=1}^{n} \widehat{e}_i X_{i,1} \\ \sum_{i=1}^{n} \widehat{e}_i X_{i,2} \\ \vdots \\ \sum_{i=1}^{n} \widehat{e}_i X_{i,k} \end{pmatrix}. \end{aligned}$$

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

In order to estimate σ^2, write:

$$\begin{aligned} \sigma^2 &= \mathbb{E}e_i^2 \\ &= \mathbb{E}\left(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}\right)^2. \end{aligned}$$

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

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

Least Squares

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

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

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

$$\begin{aligned} S(\boldsymbol{b}) &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})' (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) \\ &= \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{b}\right)' \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{b}\right) \end{aligned}$$

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

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

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

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

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

$$
S(\boldsymbol{b})=\boldsymbol{Y}'\boldsymbol{Y}-2\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}.
$$

Using the fact that for a symmetric matrix $\mathbf{A}$ we have that

$$
\frac{\partial \boldsymbol{x}'\mathbf{A}\boldsymbol{x}}{\partial \boldsymbol{x}}=2\mathbf{A}\boldsymbol{x},
$$

the first order condition is

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

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

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

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