The Least Squares Linear Regression Model

Henrique Veras

PIMES/UFPE

Introduction

Model builders are oftern interested in understanding the *conditional variation* of one variable relative to others rather than their *joint probability*

Question: What feature of the conditional probability distribution are we interested in?

Usually, the expected value $E[y|x]$, but sometimes might be:

Conditional median or other quantiles of the distribution (20th percentile, 5th percentile, etc), variance

Linear regression deals with **conditional mean**

The Linear Regression Model

$\mathbf{y} = f(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k) + \varepsilon$, where ε is called the **disturbance** term.

Our **theory** will specify the population regression equation $f(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k)$, which encompasses its format and the variables that matter.

Assumptions of the Linear Regression Model

The linear regression model consists of a set of assumptions about how a data set will be produced by an underlying "data generating process."

Assumption A1: The model specifies a linear relationship between y and $\mathbf{x}_1, \cdots, \mathbf{x}_k$:

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{x}_2\beta_2 + \cdots + \mathbf{x}_k\beta_k + \varepsilon$$

Notice that the assumption is about the linearity in the parameters rather than in the $\mathbf{x}$'s.

Linearity of the Regression Model

Each observation of a given data set looks like

$$y_1 = \beta_1 x_{11} + \beta_2 x_{21} + \cdots \beta_k x_{k1} + \varepsilon_1$$

$$y_2 = \beta_1 x_{12} + \beta_2 x_{22} + \cdots \beta_k x_{k2} + \varepsilon_1$$

$$\vdots$$

$$y_n = \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots \beta_k x_{kn} + \varepsilon_1$$

Linearity of the Regression Model

In Matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix}_{n\times k} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_n \end{bmatrix}_{k\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

Ful Rank

Assumption A2: The columns of X are linearly independent and there are at least k observations.

Assumption A2 states that there are no linear relationships among the variables.

Here's an example of a model that cannot be estimated, although we might be interested in quantifying each of the coefficients: the determinants of Monet's prices:

$$\ln \text{Price} = \beta_1 \ln \text{Size} + \beta_2 \ln \text{Aspect Ratio} + \beta_3 \ln \text{Height} + \varepsilon$$

where Size = Width $\times$ Height and Aspect Ratio $= Width/Height$

Regression

Assumption A3: The disturbance is assumed to have conditional expected value zero at every observation: $E(\varepsilon|\mathbf{X}) = 0$

No value of $\mathbf{X}$ conveys any information about ε. We assume that ε_i's are purely random draws from a population.

Moreover, we assume $E[\varepsilon_i|[\varepsilon_1, \cdots, \varepsilon_{i-1}, [\varepsilon_{i+1}, \cdots, [\varepsilon_n] = 0$.

Notice that by the **Law of Iterated Expectations**:

$$E[\varepsilon_i] = E_X[E[\varepsilon_i|\mathbf{X}]] = E_X[0] = 0$$

Regression

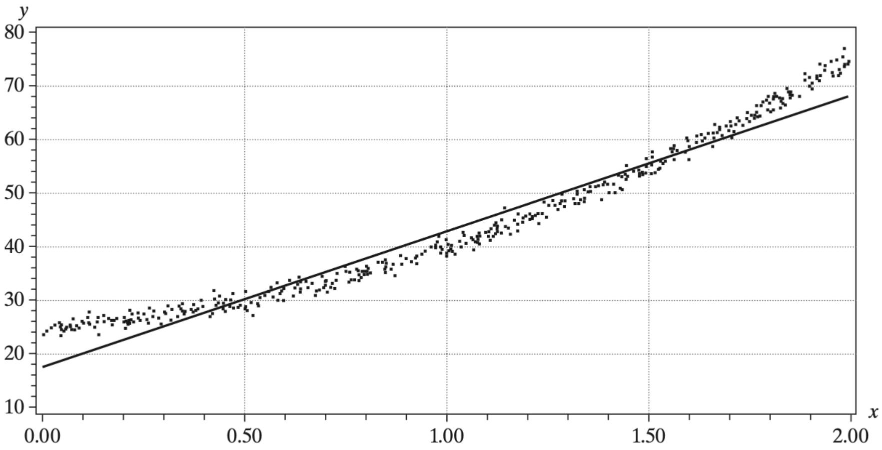
Point to note: $E[\varepsilon|\mathbf{X}] = 0 \Rightarrow Cov(\mathbf{X}, \varepsilon) = 0$. But the converse is not true: $E[\varepsilon] = 0$ **does not** imply that $E[\varepsilon|\mathbf{X}] = 0$.

Accordingly, $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$.

Assumptions **A1** and **A3** comprise the *linear regression model*.

What if $E[\varepsilon] \neq 0$?

FIGURE 2.2 Disturbances with Nonzero Conditional Mean and Zero Unconditional Mean.



Regression

Assumption **A3** is called the **exogeneity** assumption and it yields $E[\mathbf{y}] = \mathbf{X}\beta$.

Whenever $E(\varepsilon|x) \neq 0$, we say that x is **endogenous** to the model. One way that this can happen is when we leave out a variable that matters for the relationship.

Suppose the DGP of a given relationship is given by

$$Income = \gamma_1 + \gamma_2 educ + \gamma_3 age + u$$

but we estimate the model

$$Income = \gamma_1 + \gamma_2 educ + \varepsilon$$

How do we show that **A3** is not satisfied?

Homoskedasticity and Nonautocorrelated Disturbances

Assumption A4: $E[\varepsilon\varepsilon'|\mathbf{X}] = \sigma^2\mathbf{I}$

Also, notice that $Var[\varepsilon] = E[Var(\varepsilon|\mathbf{X})] + Var[E(\varepsilon|\mathbf{X})] = \sigma^2\mathbf{I}$

Data Generating Process for the Regressors

Assumption A5: $\mathbf{X}$ may be fixed or random.

Fixed $\mathbf{X}$: Experimental designs, whereby the researcher fixes the values of $\mathbf{X}$ to find $\mathbf{y}$.

Random $\mathbf{X}$: Observational studies. However, some columns of the $\mathbf{X}$ can be fixed, such as indicator variables for a given time period or time trends.

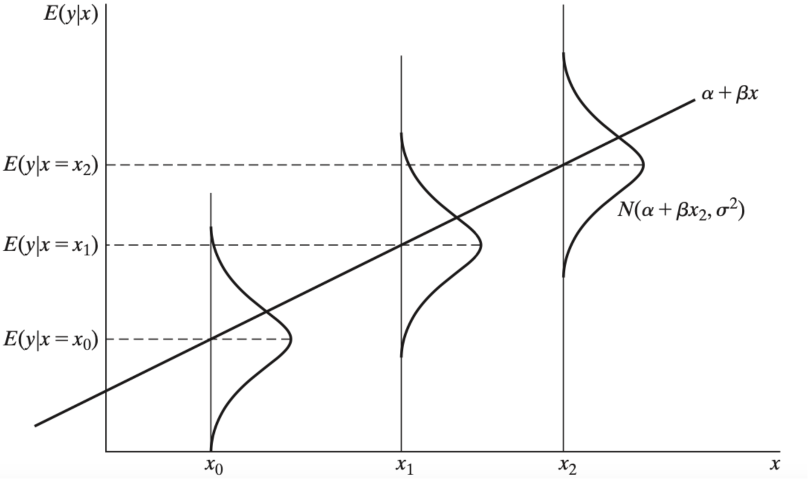
Normality

Assumption A6: $\varepsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

This assumption is useful for hypothesis testing and constructing confidence intervals but might not be needed as the Central Limit Theorem applies to sufficiently large data.

Visual Summary of the Assumptions

FIGURE 2.3 The Normal Linear Regression Model.



Computational Aspects of the Least Squares Regression

Let's now consider the algebraic problem of choosing a vector $\mathbf{b}$ so that the fitted line $\mathbf{x}_i'\mathbf{b}$ is *close* to the data.

We need to specify what do we mean by *close* to the data (the fitting criterion).

Usually, the fitting criterion is the *Least Squares* method: minimizing the sum of the squared deviations from the mean.

Crucial feature: LS regression provides us a device for "holding other things constant".

The LS Population and Sample Models

Recall the population regression model: $E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\beta$

We aim to find an estimate $\hat{y}_i = \mathbf{x}_i'\mathbf{b}$

Define the *residuals* from the estimated regression as

$$e_i = y_i - \mathbf{x}_i'b$$

Notice that $y_i = \mathbf{x}_i'\beta + \varepsilon_i = \mathbf{x}_i'b + e_i$

The LS Coefficient Vector

The Least Squares criterion requires us to minimize

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i' b)^2$$

In matrix terms, we minimize

$$S(\mathbf{b}) = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

Expanding, we have

$$S(\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

The LS Coefficient Vector

The necessary condition for a minimum is

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}$$

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

From **A2**, we know that $\mathbf{X}$ has full rank, which guarantees the existence of its inverse. Then, pre-multiplying both sides by $(\mathbf{X}'\mathbf{X})^{-1}$:

$$b_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

For the solution b_0 to minimize the sum of the squared residuals, the matrix $\frac{\partial^2 S(\mathbf{b})}{\partial \mathbf{b}^2} = 2\mathbf{X}'\mathbf{X}$ must be positive definite.

Example

Algebraic Aspects of the LS Solution

We have

$$\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{y} = -\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = -\mathbf{X}'\mathbf{e} = \mathbf{0}$$

Hence, for every column of $\mathbf{X}$, $\mathbf{x}_k'\mathbf{e} = 0$.

Denote the first row $\mathbf{X}$ as $\mathbf{x}_1 \equiv \mathbf{i}$, two implications follow:

1. The LS residuals sum to zero.
2. The regression hyperplane passes through the point of means of the data.

Table of Contents