Lecture 23: The Standard Linear Model: Estimation

Relevant textbook passages:

Larsen–Marx [2]: Sections 11.1, 11.2, 11.3.

Larsen and Marx [2, Chapter 11] treat the standard linear model (or the simple linear model) for the special case where there are two variates $X$ and $Y$ (and a constant term). This is unfortunate since it obscures the simplicity and symmetry of the general model with variates $X_1, \ldots, X_K$ and $Y$. Since you have had some linear algebra in Ma 1, I will use the matrix approach to the more general problem.

23.1 The standard model

The standard linear model is so called, not because standard statistical problems satisfy the assumptions, but because the standard assumptions make the model nice to deal with. The basic premise of the model is that the expected value of the left-hand side variate $Y$ conditional on the value of $K$ right-hand side variates $X_1, \ldots, X_K$ is a linear function of the $X_k$s.

The right-hand side variates may be random variables, or they may be chosen by the experimenter. For instance, if we are interested in the effects of irrigation ($X$) on the sugar content of grapes ($Y$), there are two ways to collect data on the amount of water the grapes receive. One is to set out a rain gauge and see how much it rains, the other is to control the amount of water by using an irrigation system. From the point of view of the experimenter, the rainfall is a random variable, while the amount of irrigation is not random, it is chosen.

A function that describes $E(Y \mid X)$ is called a regression function and the variates $X_1, \ldots, X_K$ are called regressors.

The values of the regressors $X_1, \ldots, X_K$ are viewed as determining the value of $Y$ up to some random error. One frequent interpretation of the random error is that it is the sum of many omitted variables that we cannot/will not observe or measure.

Examples:
• The first time I ever heard of regression analysis was a few decades back. When I was a teen-ager in the sticks of Riverside County, I regularly read a gearhead magazine called Road & Track. (Today I only read it at my dentist’s office.) They tested unaffordable exotic cars and dutifully reported on such things as quarter-mile times, 0–60 times, braking distances, lateral g-forces, etc. At some point, they got new testing gear for measuring performance that included a bicycle wheel strapped to the rear bumper connected to state-of-the-art transistorized electronic sensors. This new gear either weighed more or less than their old gear, so that any new tests they ran would not be comparable to their old test results. To maintain their journalistic integrity they took all their old results and performed a regression analysis of quarter-mile times on pounds-per-horsepower. I was impressed by the plot of their data and the regression line. They then used the results to “correct” the old results so they would be comparable to the new ones, or maybe it was vice-versa. Some day I may be able to go back and find this out because in 2012 the Road & Track archives were donated to the Stanford University Library. See this press release.

• Hedonic pricing. Cf. zillow.com

\[ \text{price} = \text{const} + \beta_1 \text{sq. ft.} + \beta_2 \text{no. rooms} + \ldots + \varepsilon \]

• Kepler’s 3rd Law.
The square of the orbital period of a planet is directly proportional to the cube of the semi-major axis of its orbit.

\[ P^2 = cA^3. \]

Or

\[ 2 \ln P = \ln c + 3 \ln A \]

• Hubble’s Law.

red shift = c · distance

• Newton’s Law of Gravity:

\[ F = G \frac{M_1 M_2}{d^2} \]

\[ \ln F = \ln G + \ln M_1 + \ln M_2 - 2 \ln d \]

That is, the random variate \( Y \) satisfies

\[ Y = X_1 \beta_1 + \cdots + X_K \beta_K + \varepsilon \tag{1} \]

where \( \varepsilon \) is the error term.
The assumption of linearity is less restrictive than it may seem. For instance, a polynomial in \( x \) is a linear function of \( x, x^2, x^3, \) etc. Economists are for a variety of reasons (see, e.g., [1]) fond of weighted geometric means:

\[ y = cx_1^{b_1} \cdots x_K^{b_K}, \]

which upon taking logarithms can be written as a linear relationship

\[ \log y = \log c + b_1 \log x_1 + \cdots + b_K \log x_K. \]

This is perhaps why Newcomb’s library’s table of logarithms were so noticeably worn.

We can allow for discrete categorical variates with \textit{dummy variables}, which are indicators that assume the value one if the observation fits the category and zero otherwise. This allows the different categories to have different intercepts.\(^a\)

\(^a\)If you use dummy variables for each category, you cannot have a constant term—as this will make the regressors linearly dependent.

The variates \( X_k \) may be fixed constants chosen by an experimenter or they may be random variables themselves. It is very common for one of the \( X_k \)s, typically \( X_1 \) (sometimes denoted \( X_0 \)), to be the constant 1. Larsen and Marx [2] include a constant term without mentioning it as a variate. The constant term reduces to the expected value of \( Y \) conditional on all the \( X_k \)s being zero.

It is quite often the case that the investigator has a special interest in the effect of \( X_1 \) on \( Y \), but it is known that the other \( X_2, \ldots, X_K \) have an effect on \( Y \), and these are included in order to “control for the effects of \( X_2, \ldots, X_K \).”

The standard linear model takes as its data a set of \( N \) \textit{observations} of the values \( x_1, \ldots, x_K \) and \( y \).

\[ y_t = x_{t,1} \beta_1 + \cdots + x_{t,K} \beta_K + \varepsilon_t \quad (t = 1, \ldots, N) \]

where the \( \varepsilon_t \)s are unobserved errors. The relationship among these data are usually summarized in a matrix equation

\[ y = X \beta + \varepsilon \quad (2) \]
where
\[ y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \text{ is a } N \times 1 \text{ column vector} \]
\[ X = \begin{bmatrix} x_{1,1} & \cdots & x_{1,K} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,K} \end{bmatrix} \text{ is a } N \times K \text{ matrix,} \]
\[ \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} \text{ is a } K \times 1 \text{ column vector,} \]
and
\[ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix} \text{ is a } N \times 1 \text{ column vector.} \]

The statistical assumptions are that the error vector \( \varepsilon \) satisfies
\[
E(\varepsilon | X) = 0, \\
\text{Var}(\varepsilon | X) = E(\varepsilon \varepsilon' | X) = \sigma^2 I_{N \times N}.
\] (3)

This last assumption is known as homoskedacity. It is possible to deal with more general error structures, but that takes us into the realm of the generalized linear model. When the observations correspond to different time-periods (days, weeks, months) it is unlikely that the \( \varepsilon_t \) and \( \varepsilon_{t+1} \) are uncorrelated. Special techniques have been developed to deal with time series that exhibit serial correlation.

23.2 Least Squares Estimation

Regression analysis, or simply regression, is concerned with estimating the components of \( \beta \) and testing hypotheses regarding them.

The method of least squares, also known as ordinary least squares (OLS), estimates \( \beta_1, \ldots, \beta_K \) by minimizing the sum over \( t \) of squared deviations or residuals of \( y_t \) from a linear combination of the \( x_{t,k} \)s. Computing such an estimate is usually called regressing \( Y \) on \( X_1, \ldots, X_K \), or running a regression of \( Y \) on \( X_1, \ldots, X_K \).

For instance, when there is one regressor of interest and a constant term, minimizing the sum of squared residuals fits a straight line through the set of points \((x_t, y_t), t = 1, \ldots, N\), so as to minimize the sum of the squares of the vertical distances from the line. See Figure 23.1.
Figure 23.1: The line $y = x$ minimizes the sum of the squares of the vertical distances from the three points $(-1, -2), (0, 2), (1, 0)$. 
Given a column $K$-vector $b$, $$y - Xb$$ is the vector of residuals, or differences of $y_t$ from $\sum_{k=1}^{K} x_{t,k}b_k$. The sum of squared residuals (SSR) is thus $y - Xb$ dotted with itself:

$$(y - Xb)'(y - Xb).$$

(4)

Expanding (4) yields

$$\text{SSR}(b) = y'y - 2y'Xb + b'X'Xb,$$

which is a convex quadratic function in the components of $b$.\footnote{To see that it is convex, note that its Hessian is $2X'X$, which is positive semidefinite as $x'(2X'X)x = 2(Xx) \cdot (Xx) \geq 0$.}

The gradient of this function is

$$\nabla \text{SSR}(b) = -2X'y + 2X'Xb.$$

By convexity, the minimum occurs whenever the gradient equals zero. Thus the minimizer $\hat{\beta}_{\text{OLS}}$ satisfies the first-order condition $\nabla \text{SSR}(\hat{\beta}_{\text{OLS}}) = 0$, or

$$X'y = X'X\hat{\beta}_{\text{OLS}}.$$  

(5)

This matrix equation is known as the normal equation for $\hat{\beta}_{\text{OLS}}$. The reason for the terminology will become clear in a bit.

---

On the hypothesis that $X'X$ (a $K \times K$ matrix) is nonsingular, we then have that $\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y$ minimizes the sum of squared residuals. This $\hat{\beta}_{\text{OLS}}$ is called the ordinary least squares (OLS) estimator of $\beta$.

---

23.2.1 Remark What if $X'X$ is singular? This happens only if the rank of $X$ is less than $K$, which means that there is a nonzero linear combination of the columns that sums to zero. Let $X^k$ denote the $k^{\text{th}}$ column of $X$, and let

$$a_1X^1 + \cdots + a_KX^K = 0,$$

where not all $a_k$ are zero. Then if

$$y = \beta_1X^1 + \cdots + \beta_KX^K + \varepsilon,$$

we also have

$$y = \beta_1X^1 + \cdots + \beta_KX^K + \varepsilon + c(a_1X^1 + \cdots + a_KX^K)$$

$$= (\beta_1 + ca_1)X^1 + \cdots + (\beta_K + ca_K)X^K + \varepsilon$$

for any value of $c$. Whenever $a_k$ is nonzero (and there is at least one), the coefficient on $X^k$ is not unique—in fact, by choosing the proper $c$, it can be whatever we want. That is, the data cannot tell us what the coefficient $\beta_k$ is, even if every error term is zero.
Note that $\hat{\beta}_{\text{OLS}}$ is a random vector. This is because by (2) and (6) we have

$$\hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon,$$

(7)

where $\varepsilon$ is a random vector.

An important property of the OLS estimator is that the vector $e$ of OLS residuals,

$$e = y - X\hat{\beta}_{\text{OLS}},$$

is orthogonal to each $k^{th}$ column vector of the values of the regressor $X_k$. In matrix terms, this can be written

$$X'e = 0.$$  (8)

To see this, observe that

$$X'e = X'(y - X\hat{\beta}_{\text{OLS}})$$

$$= X'y - X'X\hat{\beta}_{\text{OLS}}$$

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

$$= X'y - X'y = 0.$$

If the regressors include a constant term, then the fitted “plane” passes through the sample means. That is,

$$\bar{y} = \bar{x}_1\hat{\beta}_1 + \cdots + \bar{x}_K\hat{\beta}_K.$$

To see why, observe that

$$y = X\hat{\beta}_{\text{OLS}} + e,$$

so

$$1'y = 1'X\hat{\beta}_{\text{OLS}} + 1'e,$$  (9)

where $1$ is a $N$-vector of ones. Since it is one of the regressors, (8) implies $1'e = 0$. Dividing (9) by $N$ gives $\bar{y} = \bar{x}_1\hat{\beta}_1 + \cdots + \bar{x}_K\hat{\beta}_K$.

23.3 The geometry of OLS estimation

There is a simple geometric interpretation of the OLS estimator. In general, the vector $y$ of the $T$ observations on $Y$ is not an exact linear combination of the columns $X^k$ of the observations on the $X_k$s. The vectors all belong to the $T$-dimensional space Euclidean space, but the columns of $X$ span an at most $K$-dimensional subspace $M$. What OLS estimation does is project $y$ orthogonally onto the subspace $M$. That is, it decomposes $y$ into two parts

$$y = X\hat{\beta}_{\text{OLS}} + e,$$
where \(X\hat{\beta}_{\text{OLS}}\) belongs to \(M\) (it is a linear combination of the columns of \(X\)) and \(e\) is orthogonal to all the vectors in \(M\). This latter is what (8) says. If we rewrite (8), we obtain

\[
0 = X' e = X'(y - X\hat{\beta}_{\text{OLS}}) = X'y - X'X\hat{\beta}_{\text{OLS}},
\]

which is the normal equation (5). The normal equations just state the LS residuals are orthogonal the column space of \(X\). The term “normal” is sometimes a synonym for orthogonal.\(^2\)

If \(X\) has rank \(K\), then the columns of \(X\) constitute a basis for \(M\) so \(y\) can be written as unique linear combination of the columns. But even if \(X\) is not of full rank, any solution to the normal equations will minimize the sum of squared residuals.

\[\text{Figure 23.2. Geometry of OLS.}\]

### 23.4 OLS and MLE

When the error vector \(\varepsilon\) has a multivariate normal distribution \(N(0, \sigma^2 I)\) distribution, then the OLS estimator of \(\beta\) is also the Maximum Likelihood Estimator.

To see this, write \(\varepsilon = y - X\beta\). Then the density of \(y - X\beta\) is the multivariate normal density \(N(0, \sigma^2 I)\)

\[
\left(\frac{1}{\sqrt{2\pi}}\right)^N \frac{1}{\sqrt{\det \sigma^2 I}} e^{-\frac{1}{2}(y-X\beta)'(\sigma^2 I)^{-1}(y-X\beta)} = \left(\frac{1}{\sqrt{2\pi}}\right)^N \left(\frac{1}{(\sigma^2)^N}\right)^{\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)}
\]

\(^2\)This is a plausible etymology, but I haven’t tracked down whether it is correct.
Taking logs, we find the log likelihood function is

\[-\frac{N}{2} \log(2\pi) - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)'(y - X\beta).\]

Maximizing this with respect to \( \beta \) amounts to minimizing \((y - X\beta)'(y - X\beta)\), which is exactly what OLS does.

The first order condition for the maximum with respect to \( \sigma^2 \) is

\[-\frac{N}{2} \frac{1}{\sigma^2} + \frac{1}{2} (y - X\beta)'(y - X\beta) \frac{1}{(\sigma^2)^2} = 0.\]

Then as in Lecture 16, multiply by \( 2(\sigma^2)^2 \) to get

\[-N\sigma^2 + (y - X\beta)'(y - X\beta) = 0,\]

so

\[\hat{\sigma}_{\text{MLE}}^2 = \frac{e'e}{N},\]

where

\[e = y - X\hat{\beta}.\]

Bibliography


http://www.jstor.org/stable/1811556
