STAT 200B 2019 Week11

Soyeon Ahn

1 Estimation

1.1 the method of least squares

minimizing the residual sum of squares

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

where $\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$

In the simple linear regression model fitting,

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right),$$

where $\hat{\beta}_0$ and $\hat{\beta}$ are least squares estimates, which minimize

$$f(\beta_0, \beta) = \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2.$$

The first-order partial derivatives of $f(\beta_0, \beta)$ are

$$\frac{\partial f(\beta_0, \beta)}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \beta_1 X_i)\right];$$
$$\frac{\partial f(\beta_0, \beta)}{\partial \beta} = -2 \sum_{i=1}^{n} X_i \left[Y_i - (\beta_0 + \beta_1 X_i)\right].$$

$\hat{\beta}_0$ and $\hat{\beta}$ should satisfy that

$$\frac{\partial f(\hat{\beta}_0, \hat{\beta})}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right] = 0,$$
$$\frac{\partial f(\hat{\beta}_0, \hat{\beta})}{\partial \beta} = -2 \sum_{i=1}^{n} X_i \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right] = 0$$

Thus,

$$\bar{Y} - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}\right) = 0$$

$$\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0$$

The least squares estimates are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

1.2 Expectation and Variance

Rewrite the equation as follows:

$$Y_i = \beta_0 + \beta_1 \bar{X} + \beta_1 (X_i - \bar{X}) + \epsilon_i \tag{1}$$

and the least squares estimates are

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} - \frac{\sum_{i=1}^n (X_i - \bar{X})\bar{Y}}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \sum_{i=1}^n w_i Y_i
\end{aligned}$$

where $w_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$ Note that $\sum_{i=1}^n w_i = 0$ and $\sum_{i=1}^n w_i (X_i - \bar{X}) = 1$.

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\
&= \sum_{i=1}^n \frac{1}{n} Y_i - (\sum_{i=1}^n w_i Y_i) \bar{X} \\
&= \sum_{i=1}^n \left(\frac{1}{n} - w_i \bar{X} \right) Y_i \\
&= \sum_{i=1}^n \left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) Y_i
\end{aligned}$$

$$\begin{aligned}
E[\hat{\beta}_1] &= E[\sum_{i=1}^n w_i Y_i] \\
&= \sum_{i=1}^n w_i E[Y_i] \\
&= \sum_{i=1}^n \frac{(X_i - \bar{X})(\beta_0 + \beta_1 \bar{X} + \beta_1 (X_i - \bar{X}))}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \beta_1 \\
V[\hat{\beta}_1] &= V[\sum_{i=1}^n w_i Y_i] \\
&= \sum_{i=1}^n w_i^2 \sigma^2 \\
&= \sum_{i=1}^n \left(\frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)^2 \sigma^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sigma^2}{n s_X^2}
\end{aligned}$$

$$\begin{aligned}
E[\hat{\beta}_0] &= E[\bar{Y} - \hat{\beta}_1 \bar{X}] \\
&= \frac{1}{n} E[\beta_0 + \beta_1 X_i] - \beta_1 \bar{X} \\
&= \beta_0 \\
V[\hat{\beta}_0] &= V[\sum_{i=1}^n \left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) Y_i] \\
&= \sum_{i=1}^n \left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)^2 \sigma^2 \\
&= \left(\sum_{i=1}^n \frac{1}{n^2} - \sum_{i=1}^n \frac{2}{n} \frac{(X_i - \bar{X})\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} + \sum_{i=1}^n \frac{(X_i - \bar{X})^2 \bar{X}^2}{(\sum_{i=1}^n (X_i - \bar{X})^2)^2} \right) \sigma^2 \\
&= \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \sigma^2 \\
&= \left(\frac{\frac{1}{n} \sum_{i=1}^n X_i^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right) \sigma^2 = \frac{\frac{1}{n} \sum_{i=1}^n X_i^2}{n s_X^2} \sigma^2
\end{aligned}$$

$$
\begin{aligned}
Cov[\hat{\beta}_0, \hat{\beta}_1] &= Cov[\sum_{i=1}^{n} w_i Y_i, \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right) Y_i] \\
&= \sum_{i=1}^{n} \left(\frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \times \left(\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \right) \sigma^2 \\
&= \sum_{i=1}^{n} \left(\frac{(X_i - \bar{X})}{s_X^2} \times \left(1 - \frac{(X_i - \bar{X})\bar{X}}{s_X^2}\right) \right) \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n s_X^2}(-\bar{X})
\end{aligned}
$$

2 Chapter 3. Linear Regression Analysis by George A.F. Seber and Alan J. Lee.

3

Linear Regression: Estimation and Distribution Theory

3.1 LEAST SQUARES ESTIMATION

Let Y be a random variable that fluctuates about an unknown parameter η; that is, $Y = \eta + \varepsilon$, where ε is the fluctuation or *error*. For example, ε may be a "natural" fluctuation inherent in the experiment which gives rise to η, or it may represent the error in measuring η, so that η is the true response and Y is the observed response. As noted in Chapter 1, our focus is on linear models, so we assume that η can be expressed in the form

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1},$$

where the *explanatory* variables $x_1, x_2, \ldots, x_{p-1}$ are known constants (e.g., experimental variables that are controlled by the experimenter and are measured with negligible error), and the β_j $(j = 0, 1, \ldots, p-1)$ are unknown parameters to be estimated. If the x_j are varied and n values, $Y_1, Y_2, \ldots, Y_n$, of Y are observed, then

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \qquad (i = 1, 2, \ldots, n), \tag{3.1}$$

where x_{ij} is the ith value of x_j. Writing these n equations in matrix form, we have

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ ]x_{n0} & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

or

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.2}$$

where $x_{10} = x_{20} = \cdots = x_{n0} = 1$. The $n \times p$ matrix $\mathbf{X}$ will be called the *regression matrix*, and the x_{ij}'s are generally chosen so that the columns of $\mathbf{X}$ are linearly independent; that is, $\mathbf{X}$ has rank p, and we say that $\mathbf{X}$ has *full rank*. However, in some experimental design situations, the elements of $\mathbf{X}$ are chosen to be 0 or 1, and the columns of $\mathbf{X}$ may be linearly dependent. In this case $\mathbf{X}$ is commonly called the *design matrix*, and we say that $\mathbf{X}$ has less than full rank.

It has been the custom in the past to call the x_j's the independent variables and Y the dependent variable. However, this terminology is confusing, so we follow the more contemporary usage as in Chapter 1 and refer to x_j as a *explanatory variable* or *regressor* and Y as the *response variable*.

As we mentioned in Chapter 1, (3.1) is a very general model. For example, setting $x_{ij} = x_i^j$ and $k = p - 1$, we have the polynomial model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i.$$

Again,

$$Y_i = \beta_0 + \beta_1 e^{w_{i1}} + \beta_2 w_{i1} w_{i2} + \beta_3 \sin w_{i3} + \varepsilon_i$$

is also a special case. The essential aspect of (3.1) is that it is linear in the unknown parameters β_j; for this reason it is called a *linear model*. In contrast,

$$Y_i = \beta_0 + \beta_1 e^{-\beta_2 x_i} + \varepsilon_i$$

is a nonlinear model, being nonlinear in β_2.

Before considering the problem of estimating $\boldsymbol{\beta}$, we note that all the theory in this and subsequent chapters is developed for the model (3.2), where x_{i0} is not necessarily constrained to be unity. In the case where $x_{i0} \neq 1$, the reader may question the use of a notation in which i runs from 0 to $p-1$ rather than 1 to p. However, since the major application of the theory is to the case $x_{i0} \equiv 1$, it is convenient to "separate" β_0 from the other β_j's right from the outset. We shall assume the latter case until stated otherwise.

One method of obtaining an estimate of $\boldsymbol{\beta}$ is the method of least squares. This method consists of minimizing $\sum_i \varepsilon_i^2$ with respect to $\boldsymbol{\beta}$; that is, setting $\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$, we minimize $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \|\mathbf{Y} - \boldsymbol{\theta}\|^2$ subject to $\boldsymbol{\theta} \in \mathcal{C}(\mathbf{X}) = \Omega$, where Ω is the column space of $\mathbf{X}$ ($= \{\mathbf{y} : \mathbf{y} = \mathbf{X}\mathbf{x} \text{ for any } \mathbf{x}\}$). If we let $\boldsymbol{\theta}$ vary in Ω, $\|\mathbf{Y} - \boldsymbol{\theta}\|^2$ (the square of the length of $\mathbf{Y} - \boldsymbol{\theta}$) will be a minimum for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ when $(\mathbf{Y} - \hat{\boldsymbol{\theta}}) \perp \Omega$ (cf. Figure 3.1). This is obvious geometrically, and it is readily proved algebraically as follows.

We first note that $\hat{\boldsymbol{\theta}}$ can be obtained via a symmetric idempotent (projection) matrix $\mathbf{P}$, namely $\hat{\boldsymbol{\theta}} = \mathbf{P}\mathbf{Y}$, where $\mathbf{P}$ represents the orthogonal projection onto Ω (see Appendix B). Then

$$\mathbf{Y} - \boldsymbol{\theta} = (\mathbf{Y} - \hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

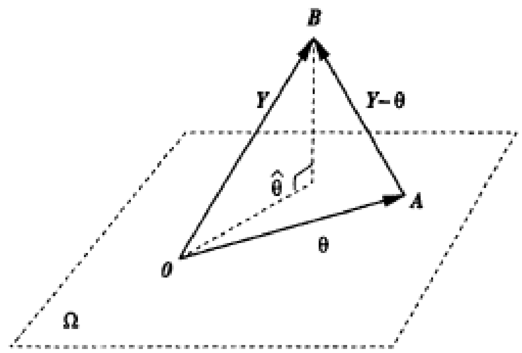

Fig. 3.1 The method of least squares consists of finding A such that AB is a minimum.

where from $\mathbf{P}\boldsymbol{\theta} = \boldsymbol{\theta}$, $\mathbf{P}' = \mathbf{P}$ and $\mathbf{P}^2 = \mathbf{P}$, we have

$$\begin{aligned}(\mathbf{Y} - \hat{\boldsymbol{\theta}})'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= (\mathbf{Y} - \mathbf{PY})'\mathbf{P}(\mathbf{Y} - \boldsymbol{\theta}) \\ &= \mathbf{Y}'(\mathbf{I}_n - \mathbf{P})\mathbf{P}(\mathbf{Y} - \boldsymbol{\theta}) \\ &= \mathbf{0}.\end{aligned}$$

Hence

$$\begin{aligned}||\mathbf{Y} - \boldsymbol{\theta}||^2 &= ||\mathbf{Y} - \hat{\boldsymbol{\theta}}||^2 + ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2 \\ &\geq ||\mathbf{Y} - \hat{\boldsymbol{\theta}}||^2,\end{aligned}$$

with equality if and only if $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Since $\mathbf{Y} - \hat{\boldsymbol{\theta}}$ is perpendicular to Ω,

$$\mathbf{X}'(\mathbf{Y} - \hat{\boldsymbol{\theta}}) = \mathbf{0}$$

or

$$\mathbf{X}'\hat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{Y}. \tag{3.3}$$

Here $\hat{\boldsymbol{\theta}}$ is uniquely determined, being the *unique* orthogonal projection of $\mathbf{Y}$ onto Ω (see Appendix B).

We now assume that the columns of $\mathbf{X}$ are linearly independent so that there exists a unique vector $\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Then substituting in (3.3), we have

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}, \tag{3.4}$$

the *normal equations*. As $\mathbf{X}$ has rank p, $\mathbf{X}'\mathbf{X}$ is positive-definite (A.4.6) and therefore nonsingular. Hence (3.4) has a unique solution, namely,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{3.5}$$

Here $\hat{\boldsymbol{\beta}}$ is called the (ordinary) *least squares estimate* of $\boldsymbol{\beta}$, and computational methods for actually calculating the estimate are given in Chapter 11.

We note that $\hat{\beta}$ can also be obtained by writing

$$\begin{aligned}\varepsilon'\varepsilon &= (\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)\\ &= \mathbf{Y}'\mathbf{Y}-2\beta'\mathbf{X}'\mathbf{Y}+\beta'\mathbf{X}'\mathbf{X}\beta\end{aligned}$$

[using the fact that $\beta'\mathbf{X}'\mathbf{Y}=(\beta'\mathbf{X}'\mathbf{Y})'=\mathbf{Y}'\mathbf{X}\beta$] and differentiating $\varepsilon'\varepsilon$ with respect to β. Thus from $\partial\varepsilon'\varepsilon/\partial\beta=\mathbf{0}$ we have (A.8)

$$-2\mathbf{X}'\mathbf{Y}+2\mathbf{X}'\mathbf{X}\beta=\mathbf{0} \tag{3.6}$$

or

$$\mathbf{X}'\mathbf{X}\beta=\mathbf{X}'\mathbf{Y}.$$

This solution for β gives us a stationary value of $\varepsilon'\varepsilon$, and a simple algebraic identity (see Exercises 3a, No. 1) confirms that $\hat{\beta}$ is a minimum.

In addition to the method of least squares, several other methods are used for estimating β. These are described in Section 3.13.

Suppose now that the columns of $\mathbf{X}$ are not linearly independent. For a particular $\hat{\theta}$ there is no longer a unique $\hat{\beta}$ such that $\hat{\theta}=\mathbf{X}\hat{\beta}$, and (3.4) does not have a unique solution. However, *a* solution is given by

$$\hat{\beta}=(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y},$$

where $(\mathbf{X}'\mathbf{X})^{-}$ is any generalized inverse of $(\mathbf{X}'\mathbf{X})$ (see A.10). Then

$$\hat{\theta}=\mathbf{X}\hat{\beta}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}=\mathbf{P}\mathbf{Y},$$

and since $\mathbf{P}$ is unique, it follows that $\mathbf{P}$ does not depend on which generalized inverse is used.

We denote the *fitted values* $\mathbf{X}\hat{\beta}$ by $\hat{\mathbf{Y}}=(\hat{Y}_1,\ldots,\hat{Y}_n)'$. The elements of the vector

$$\begin{aligned}\mathbf{Y}-\hat{\mathbf{Y}} &= \mathbf{Y}-\mathbf{X}\hat{\beta}\\ &= (\mathbf{I}_n-\mathbf{P})\mathbf{Y}, \quad \text{say,}\end{aligned} \tag{3.7}$$

are called the *residuals* and are denoted by $\mathbf{e}$. The minimum value of $\varepsilon'\varepsilon$, namely

$$\begin{aligned}\mathbf{e}'\mathbf{e} &= (\mathbf{Y}-\mathbf{X}\hat{\beta})'(\mathbf{Y}-\mathbf{X}\hat{\beta})\\ &= \mathbf{Y}'\mathbf{Y}-2\hat{\beta}'\mathbf{X}'\mathbf{Y}+\hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}\\ &= \mathbf{Y}'\mathbf{Y}-\hat{\beta}'\mathbf{X}'\mathbf{Y}+\hat{\beta}'[\mathbf{X}'\mathbf{X}\hat{\beta}-\mathbf{X}'\mathbf{Y}]\\ &= \mathbf{Y}'\mathbf{Y}-\hat{\beta}'\mathbf{X}'\mathbf{Y} \qquad \text{[by (3.4)],} && (3.8)\\ &= \mathbf{Y}'\mathbf{Y}-\hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}, && (3.9)\end{aligned}$$

is called the *residual sum of squares* (RSS). As $\hat{\theta}=\mathbf{X}\hat{\beta}$ is unique, we note that $\hat{\mathbf{Y}}$, $\mathbf{e}$, and RSS are unique, irrespective of the rank of $\mathbf{X}$.