Notes on the least squares estimator

Daniel Sorensen*

May 20, 2024

The argument leading to the asymptotic variance of the least squares estimator (6.20) requires clarification, as well as the subsection **Least Squares Prediction as an Approximation to Best Linear Prediction** on page 273. This file elaborates on the topics.

Linear regression with one predictor variable

Consider the simple linear regression model

$$y = 1\alpha + \widetilde{X}b + e \tag{1}$$

where y is a vector of observations with n elements, 1 is a vector of $1's$ with n elements, the scalar α is the intercept, $\widetilde{X}$ is the observed $n \times 1$ full rank matrix containing the values of the covariate across observations, b is the unknown regression parameter (a scalar) and e is the vector of n residuals, independent of $\widetilde{X}$, with mean 0 and variance $I\sigma^2$. Write (1) as

$$y = X\beta + e \tag{2}$$

where $X = \{X_i\}_{i=1}^n$ has the appended column of $1's$ in its first column and $\beta = (\alpha, b)$. The least squares estimator is

$$\hat{\beta} = (X'X)^{-1} X'y \tag{3}$$

and its variance is

$$Var\left(\hat{\beta}|X, \sigma^2\right) = (X'X)^{-1}\sigma^2. \tag{4}$$

It is easy to check that $X'X$ has the following form,

$$\begin{aligned} X'X &= \begin{bmatrix} n & \sum_i X_i \\ \sum_i X_i & \sum_i X_i^2 \end{bmatrix} \\ &= n\begin{bmatrix} 1 & \overline{X} \\ \overline{X} & \frac{1}{n}\sum_i X_i^2 \end{bmatrix} \end{aligned} \tag{5}$$

*Center for Quantitative Genetics and Genomics, Aarhus University, C F MÃ¸llers Alle 3, building 1130, 8000 Aarhus, Denmark

where $\overline{X}$ is $\sum_i X_i/n$, the average value of X. The determinant of (5) is

$$\begin{aligned} \det(X'X) &= n\left(\sum_i X_i^2 - n\overline{X}^2\right) \\ &= n\sum_i \left(X_i - \overline{X}\right)^2. \end{aligned}$$

The inverse matrix $(X'X)^{-1}$ is therefore

$$(X'X)^{-1} = \frac{1}{\sum_i \left(X_i - \overline{X}\right)^2} \begin{bmatrix} \frac{1}{n}\sum_i X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}. \tag{6}$$

The least squares estimator (3) can be expressed as

$$\begin{bmatrix} \hat{\alpha} \\ \hat{b} \end{bmatrix} = \frac{1}{\sum_i \left(X_i - \overline{X}\right)^2} \begin{bmatrix} \frac{1}{n}\sum_i X_i^2 & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix} \begin{bmatrix} \sum_i y_i \\ \sum_i X_i y_i \end{bmatrix},$$

and

$$\hat{b} = \frac{\sum_i X_i y_i - n\overline{X}\,\overline{y}}{\sum_i \left(X_i - \overline{X}\right)^2} = \frac{\sum_i \left(X_i - \overline{X}\right)(y_i - \overline{y})}{\sum_i \left(X_i - \overline{X}\right)^2}. \tag{7}$$

From (4) and (6), multiplying and dividing by n,

$$Var\left(\hat{b}|X,\sigma^2\right) = \frac{\sigma^2}{n}\left[\frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2\right]^{-1}, \tag{8}$$

where $\frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2$ is the sampling variance of X.

If X is allowed to vary, the unconditional (with respect to X) variance of the least squares estimator is

$$\begin{aligned} Var\left(\hat{b}|\sigma^2\right) &= E_X\left[Var\left(\hat{b}|X,\sigma^2\right)\right] + Var_X\left[E\left(\hat{b}|X,\sigma^2\right)\right] \\ &= E_X\left[Var\left(\hat{b}|X,\sigma^2\right)\right] \\ &= \frac{\sigma^2}{n} E\left\{\left[\frac{1}{n}\sum_i \left(X_i - \overline{X}\right)^2\right]^{-1}\right\}. \end{aligned} \tag{9}$$

As n increases, the sampling variance of X converges to the true variance of X, $Var(X)$, the inverse of the sampling variance converges to the inverse of the true variance, and

$$Var\left(\hat{b}|\sigma^2\right) \rightarrow \frac{\sigma^2}{n}[Var(X)]^{-1}. \tag{10}$$

Another line of argument followed in the book that leads to (10) is as follows. Instead of fitting model (1) consider fitting the model to the original data y excluding the intercept and using centred covariates $x_i = \left(X_i - \overline{X}\right)$. The equation for the mean of y given x is

$$E(y|x) = xb. \tag{11}$$

The least squares estimator is now

$$\begin{aligned} \hat{b} &= (x'x)^{-1}x'y \\ &= \left[\sum_i \left(X_i - \overline{X}\right)^2\right]^{-1} \sum_i \left(X_i - \overline{X}\right)y_i \\ &= \frac{\sum_i \left(X_i - \overline{X}\right)(y_i - \overline{y})}{\sum_i \left(X_i - \overline{X}\right)^2} \end{aligned}$$

as in (7) with sampling variance

$$\begin{aligned} Var\left(\hat{b}|x, \sigma^2\right) &= (x'x)^{-1}\sigma^2 \\ &= \frac{\sigma^2}{\sum_i \left(X_i - \overline{X}\right)^2} \end{aligned} \tag{12}$$

as in (8). Arguing as before, the same asymptotic unconditional variance (10) is obtained in the case of (12). Availability of $\hat{b}$ leads to the estimator of α

$$\hat{\alpha} = \overline{y} - \hat{b}\overline{X}. \tag{13}$$

These results obtained using a single covariate regression model as an example, extend to a model based on an arbitrary number p of (possibly) correlated covariates, provided $p < n$ and matrix X is of full rank.

Linear regression with multiple predictor variables

The multiple linear regression model takes the standard form

$$y = 1\alpha + Xb + e \tag{14}$$

where y is the vector of records with n elements, as before, 1 is the column vector of $1's$ with n elements, α is the scalar intercept, $X = \{X_{ij}\}$, $i = 1, \ldots, n$; $j = 1, \ldots, p$, is the full rank matrix of p covariates of order $n \times p$, b is the vector of p multiple regression coefficients and the random residuals are collected in the vector $e \sim (0, I\sigma^2)$. Given the model, the normal equations are

$$\begin{aligned} 1'1\hat{\alpha} + 1'X\hat{b} &= 1'y, \\ X'1\hat{\alpha} + X'X\hat{b} &= X'y. \end{aligned}$$

Absorbing $\hat{\alpha}$ in the second equation results in the system

$$X'(I-P)X\hat{b} = X'(I-P)y \tag{15}$$

where the operator P is given by

$$P = 1(1'1)^{-1}1' = \frac{1}{n}11' \tag{16}$$

and $I-P$ is symmetric and idempotent. It is easy to confirm that the effect of P on the system (15) is such that

$$\begin{aligned} X'(I-P)X &= x'x, \\ X'(I-P) &= x' \end{aligned}$$

where

$$x = X - \overline{X} \tag{17}$$

and the $n \times p$ matrix $\overline{X}$ has the ith generic row equal to $\left(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p\right)$, $\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$. Therefore the least squares estimator for b is

$$\hat{b} = (x'x)^{-1}x'y \tag{18}$$

with sampling variance

$$Var\left(\hat{b}|X,\sigma^2\right) = \sigma^2(x'x)^{-1}. \tag{19}$$

With $\hat{b}$ available, the estimator of α is

$$\hat{\alpha} = \overline{y} - \sum_{i=1}^{p} \overline{X}_i \hat{b}_i. \tag{20}$$

The marginal asymptotic variance of the least squares estimator is obtained arguing as before. Write (19) as

$$Var\left(\hat{b}|X,\sigma^2\right) = \frac{\sigma^2}{n}\left(\frac{1}{n}x'x\right)^{-1}. \tag{21}$$

The ith diagonal element of $\frac{1}{n}x'x$ is

$$\frac{1}{n}\left[\left(X_{1i}-\overline{X}_i\right)^2 + \left(X_{2i}-\overline{X}_i\right)^2 + \cdots + \left(X_{ni}-\overline{X}_i\right)^2\right]$$

and the element in row i and column j of $\frac{1}{n}x'x$ is

$$\frac{1}{n}\left[\left(X_{1i}-\overline{X}_i\right)\left(X_{1j}-\overline{X}_j\right) + \left(X_{2i}-\overline{X}_i\right)\left(X_{2j}-\overline{X}_j\right) + \cdots + \left(X_{ni}-\overline{X}_i\right)\left(X_{nj}-\overline{X}_j\right)\right].$$

These are sampling variances of the ith covariate and sampling covariances between covariates i and j, respectively. As n increases towards infinity, these sample moments converge to the true variances and covariances and

$$\frac{1}{n}x'x \to V$$

the true variance-covariance matrix of X. Therefore the marginal unconditional variance of $\hat{b}$ is

$$\begin{aligned}
Var\left(\hat{b}|\sigma^2\right) &= E_X\left[Var\left(\hat{b}|X,\sigma^2\right)\right] + Var_X\left[E\left(\hat{b}|X,\sigma^2\right)\right] \\
&= E_X\left[Var\left(\hat{b}|X,\sigma^2\right)\right] \\
&= \frac{\sigma^2}{n}E\left\{\left[\frac{1}{n}x'x\right]^{-1}\right\} \\
&\to \frac{\sigma^2}{n}V^{-1}. \qquad (22)
\end{aligned}$$

The same result (22) is arrived at if a model excluding the intercept and using centred covariates is fitted to the original data y. The model for the mean takes the form

$$E(y|x) = xb,$$

where x is defined in (17). This leads to the least squares estimator (18) and the remaining narrative leading to (22) is the same as before.

Least squares prediction as an approximation to best linear prediction

Consider the problem of predicting the scalar random variable y_0 from scalars $X_1, X_2, \ldots, X_p$ and assume that y_0 and the $X's$ have finite mean and variance. A linear function $\alpha + b'X$ predicts y_0 with mean squared error

$$E\left[(y_0 - \alpha - b'X)^2\right].$$

This is minimised with

$$\alpha = E(y_0) - b'E(X), \qquad (23a)$$
$$b = [Var(X)]^{-1}Cov(X, y_0). \qquad (23b)$$

The best linear predictor is

$$\widehat{y_0} = E(y_0) + b'(X - E(X)). \qquad (24)$$

Let $(y_1, X_1), (y_2, X_2), \ldots, (y_n, X_n)$ be an *iid* sequence of random vectors, where y_i are scalars and $X_i \in \mathbb{R}^p$, $i = 1, 2, \ldots, n$. If one postulates the linear model

$$y_i = \alpha + X_i' b + e_i \tag{25}$$

where α is a scalar intercept, X_i' is the ith row of the $n \times p$ full rank matrix X, then the least squares estimators of α and b are the solution to

$$\begin{bmatrix} n & 1'X \\ X'1 & X'X \end{bmatrix} \begin{bmatrix} \widehat{\alpha} \\ \widehat{b} \end{bmatrix} = \begin{bmatrix} 1'y \\ X'y \end{bmatrix}.$$

This yields

$$\begin{aligned} \widehat{\alpha} &= \overline{y} - \sum_{i=1}^{p} \overline{X_i}\widehat{b}_i && \text{(26a)} \\ \widehat{b} &= (x'x)^{-1} x'y \\ &= \left(\frac{1}{n} x'x\right)^{-1} \frac{1}{n} x'y. && \text{(26b)} \end{aligned}$$

In (26), $\overline{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$, $x = X - \overline{X}$, and $n \times p$ matrix $\overline{X}$ has the ith generic row equal to $\left(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p\right)$. As $n \to \infty$, $\overline{y}$ aproaches $E(y)$, $\overline{X}_i$ approaches $E(X_i)$, $(n^{-1}x'x)$ approaches $Var(X)$, $(n^{-1}x'y)$ approaches $Cov(y, X)$ and the least squares predictor $\widehat{y}_0 = \widehat{\alpha} + X_0'\widehat{b}$ approaches the best linear predictor (24) irrespective of the true relationship between y and X.

NOTE

- The ith row of x' is

$$\left(X_{1i} - \overline{X}_i\right) \quad \left(X_{2i} - \overline{X}_i\right) \quad \cdots \quad \left(X_{ni} - \overline{X}_i\right)$$

  and the jth row of $x'y$ is

$$\sum_{i=1}^{n} \left(X_{ij} - \overline{X}_j\right) y_i = \sum_{i=1}^{n} \left(X_{ij} - \overline{X}_j\right)(y_i - \overline{y}).$$

  Appealing to asymptotics, as $n \to \infty$

$$\frac{1}{n} \sum_{i=1}^{n} \left(X_{ij} - \overline{X}_j\right)(y_i - \overline{y}) \to Cov(y, X_j).$$

- A fitted value evaluated at X_i is

$$\widehat{y}_i = X_i'\widehat{b}$$

with variance

$$Var(\widehat{y}_i|X_i') = X_i'(X'X)^{-1}X_i\sigma^2.$$

One can compute an average variance that takes the form

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} Var(\widehat{y}_i|X_i') &= \frac{1}{n}tr\left[Var\left(X\widehat{b}|X\right)\right] \\
&= \frac{1}{n}tr\left[(X'X)^{-1}X'X\right]\sigma^2 \\
&= \frac{p+1}{n}\sigma^2,
\end{aligned}$$

which approaches $Var(y_i|X)$ as $p \to n$, indicating almost perfect fit. Such a model will do poorly in prediction of future data.

A more economical exposition of the consistency of the least squares estimator is as follows (see also page 273 in the book). Consider a random sample $y_i, x_i; i = 1, \ldots, n$, $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^{p+1}$. A linear function $x_i'b$ predicts y_i with mean squared error

$$E(y_i - x_i'b)^2.$$

This is minimised with

$$b = (E(x_i x_i'))^{-1}E(x_i y_i), \tag{27}$$

where b is a $(p+1) \times 1$ column vector, $E(x_i x_i')$ is a $(p+1) \times (p+1)$ matrix and $E(x_i y_i)$ is a $(p+1) \times 1$ column vector. The best linear predictor of y_i given x_i is

$$\widehat{y}_i = x_i'(E(x_i x_i'))^{-1}E(x_i y_i). \tag{28}$$

Define the least squares estimator of b as the minimiser of the residual sum of squares

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'b)^2.$$

The minimisation yields

$$\begin{aligned}
\widehat{b} &= \left(\sum\nolimits_{i=1}^{n} x_i x_i'\right)^{-1}\left(\sum\nolimits_{i=1}^{n} x_i y_i\right) \\
&= \left(\frac{1}{n}\sum\nolimits_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} x_i y_i\right). \qquad (29)
\end{aligned}$$

As $n \to \infty$, (29) converges in probability to (27) and the least squares predictor $x_i'\widehat{b}$ converges to the best linear predictor (28).

References