

# Notes on the least squares estimator

Daniel Sorensen\*

August 29, 2023

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

$$y = 1\alpha + \tilde{X}b + e \quad (1)$$

where  $y$  is a vector of observations with  $n$  elements,  $1$  is a vector of  $1$ 's with  $n$  elements, the scalar  $\alpha$  is the intercept,  $\tilde{X}$  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  $\tilde{X}$ , with mean 0 and variance  $I\sigma^2$ . Write (1) as

$$y = X\beta + e \quad (2)$$

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

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

and its variance is

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

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

$$\begin{aligned} X'X &= \begin{bmatrix} n & \sum_i X_i \\ \sum_i X_i & \sum_i X_i^2 \end{bmatrix} \\ &= n \begin{bmatrix} 1 & \bar{X} \\ \bar{X} & \frac{1}{n} \sum_i X_i^2 \end{bmatrix} \end{aligned} \quad (5)$$

---

\*Center for Quantitative Genetics and Genomics, Aarhus University, C F Møllers Alle 3, building 1130, 8000 Aarhus, Denmark

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

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

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

$$(X'X)^{-1} = \frac{1}{\sum_i (X_i - \bar{X})^2} \begin{bmatrix} \frac{1}{n} \sum_i X_i^2 & -\bar{X} \\ -\bar{X} & 1 \end{bmatrix}. \quad (6)$$

The least squares estimator (3) can be expressed as

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

and

$$\hat{b} = \frac{\sum_i X_i y_i - n\bar{X}\bar{y}}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i (X_i - \bar{X})(y_i - \bar{y})}{\sum_i (X_i - \bar{X})^2}. \quad (7)$$

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

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

where  $\frac{1}{n} \sum_i (X_i - \bar{X})^2$  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

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

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

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

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  $x_i = (X_i - \bar{X})$ . The equation for the mean of  $y$  given  $x$  is

$$E(y|x) = xb. \quad (11)$$

The least squares estimator is now

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

as in (7) with sampling variance

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

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  $\alpha$

$$\hat{\alpha} = \bar{y} - \hat{b}\bar{X}. \quad (13)$$

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

$$y = 1\alpha + Xb + e \quad (14)$$

where  $y$  is the vector of records with  $n$  elements, as before,  $1$  is the column vector of  $1$ 's with  $n$  elements,  $\alpha$  is the scalar intercept,  $X = \{X_{ij}\}$ ,  $i = 1, \dots, n$ ;  $j = 1, \dots, p$ , 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

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

where the operator  $P$  is given by

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

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

$$x = X - \bar{X} \quad (17)$$

and the  $n \times p$  matrix  $\bar{X}$  has the  $i$ th generic row equal to  $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)$ ,  $\bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{ij}$ . Therefore the least squares estimator for  $b$  is

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

with sampling variance

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

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

$$\hat{\alpha} = \bar{y} - \sum_{i=1}^p \bar{X}_i \hat{b}_i. \quad (20)$$

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

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

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

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

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

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

These are sampling variances of the  $i$ th 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

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

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

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

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  $X_1, X_2, \dots, X_p$  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

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

This is minimised with

$$\alpha = E(y_0) - b'E(X), \tag{23a}$$

$$b = [Var(X)]^{-1}Cov(X, y_0). \tag{23b}$$

The best linear predictor is

$$\hat{y}_0 = E(y_0) + b'(X - E(X)). \tag{24}$$

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

$$y_i = \alpha + X_i' b + e_i \quad (25)$$

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

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

This yields

$$\hat{\alpha} = \bar{y} - \sum_{i=1}^p \bar{X}_i \hat{b}_i \quad (26a)$$

$$\begin{aligned} \hat{b} &= (x'x)^{-1} x'y \\ &= \left( \frac{1}{n} x'x \right)^{-1} \frac{1}{n} x'y. \end{aligned} \quad (26b)$$

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

## NOTE

- The  $i$ th row of  $x'$  is

$$(X_{1i} - \bar{X}_i) \quad (X_{2i} - \bar{X}_i) \quad \dots \quad (X_{ni} - \bar{X}_i)$$

and the  $j$ th row of  $x'y$  is

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

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

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

- A fitted value evaluated at  $X_i$  is

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

with variance

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

One can compute an average variance that takes the form

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

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

## References