Journal of Probability and Statistical Science
20(1), 184-190, August 2022

On Deriving the Least Squares Estimates in Introductory Regression Courses

Mark Inlow
Indiana State University

ABSTRACT
Introductory regression books typically begin their derivation of the least squares matrix estimation formula by considering the simple linear regression model. We suggest beginning with the zero-intercept model which has advantages. We provide two examples of this approach, one of which is a new, non-calculus derivation using the Cauchy-Schwarz inequality.
Keywords: teaching, regression, least squares, Cauchy-Schwarz, Gauss-Markov theorem

1 Introduction

Linear regression is an essential modeling methodology and, typically, the first one students study in depth. In addition, the matrix formulation of linear regression is often students' first encounter with treating samples as vectors. Given these facts, reconsidering how the matrix formulation of linear regression is taught in upper-level statistics courses seems worthwhile. Here we consider motivating and deriving the least squares estimation formula $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. In contrast to the standard approach which begins with the simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$, we suggest beginning with the zero-intercept model $y = \beta_1 x + \epsilon$ instead. We present two different ways to do this, one of which uses a new non-calculus derivation employing the Cauchy-Schwarz inequality.

□ Received July 2021, revised November 2021, in final form December 2021.
□ Mark Inlow (corresponding author) is affiliated with the Department of Mathematics and Statistical Sciences, Indiana State University

2 Current approaches for deriving least squares estimates

Essentially all introductory regression textbooks begin with a discussion of the simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$ where β_0 and β_1 are constants and $\epsilon \sim \text{Normal}(\mu = 0, \sigma^2)$ (e.g., Abraham and Ledolter 2006; Kutner, Nachtsheim, and Neter 2004; Mendenhall and Sincich 2020; Montgomery, Peck, and Vining 2021; Ryan 2009). One exception is Freund, Williams, and Sa (2006) who begin with the constant-only model $y = \mu + \epsilon$. While discussing the simple linear regression model, the equations for computing the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are derived. The standard approach is to use multivariate calculus methods to show that

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \text{ and}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

minimize the residual sum of squares $RSS(\tilde{\beta}_0, \tilde{\beta}_1) = \sum[y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_i)]^2$. Details of the second derivative test confirming that the critical point $(\hat{\beta}_0, \hat{\beta}_1)$ minimizes $RSS(\tilde{\beta}_0, \tilde{\beta}_1)$ are typically omitted.

A downside of this approach is that it assumes students have had multivariate calculus and are familiar with minimizing functions of two variables. This is problematic since multivariate calculus is not universally required for STEM majors - mathematics education majors for example - and is only recommended for students studying statistics (American Statistical Association Undergraduate Guidelines Workgroup 2014). In addition, students can be confused by the fact that $RSS(\tilde{\beta}_0, \tilde{\beta}_1)$ is differientiated with respect to $\tilde{\beta}_0$ and $\tilde{\beta}_1$ rather than the usual x and y. To avoid these issues, algebraic approaches for deriving $\hat{\beta}_0$ and $\hat{\beta}_1$ have been devised (e.g. Ehrenberg 1983; Stanley and Glass 1969). Although interesting, these have not been adopted by introductory regression textbooks.

In contrast to their fairly uniform treatment of the simple linear regression model, introductory regression texts differ substantially in their derivation of the matrix formula for the least squares coefficient estimates. Some (e.g., Mendenhall and Sincich 2020; Kutner, Nachtscheim, and Neter 2004) simply state the normal equations $(\boldsymbol{X'X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X'y}$ and then solve them to get $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$. Others (e.g., Abraham and Ledolter 2006; Montgomery, Peck, and Vining 2021) derive the normal equations using calculus. No mention is usually made of verifying that $\widehat{\boldsymbol{\beta}}$ actually minimizes $RSS(\widetilde{\boldsymbol{\beta}})$. A notable exception is Abraham and Ledolter (2006) who demonstrate this algebraicly; see Theorem 1 below. A similar proof is provided by Darlington (1969).

3 New approaches for deriving least squares estimates

As an alternative to deriving least squares formulas beginning with the simple linear regression model, we suggest starting with the regression through the origin or zero-intercept model $y = \beta_1 x + \epsilon$. In addition we recommend beginning with the simplest, $n = 2$ observation, case. An example of using this approach is the following:

1. Derive the least squares estimate $\hat{\beta}_1$ using two observations, (x_1, y_1) and (x_2, y_2), as follows:

1.1. Define the residual sum of squares $RSS(\tilde{\beta})$:

$$RSS(\tilde{\beta}_1) = (y_1 - \tilde{\beta}_1 x_1)^2 + (y_2 - \tilde{\beta}_1 x_2)^2$$

1.2. Differentiate $RSS(\tilde{\beta}_1)$ with respect to $\tilde{\beta}_1$:

$$RSS(\tilde{\beta}_1)' = -2x_1(y_1 - \tilde{\beta}_1 x_1) - 2x_2(y_2 - \tilde{\beta}_1 x_2)$$

1.3. Solve the estimating equation $RSS(\hat{\beta}_1)' = 0$ to get the critical point

$$\hat{\beta}_1 = \frac{x_1 y_1 + x_2 y_2}{x_1^2 + x_2^2}$$

1.4. (Optional) Use the second derivative test to show that $\tilde{\beta}_1 = \hat{\beta}_1$ minimizes $RSS(\tilde{\beta}_1)$:

$$RSS(\tilde{\beta}_1)'' = 2x_1^2 + 2x_2^2 > 0$$

2. Convert the system of equations $y_i = \beta_1 x_i + \epsilon_i,\ i = 1,2$, to vector form:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \beta_1 \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

$$= \beta_1 \boldsymbol{x} + \epsilon$$

3. Observe that $\hat{\beta}_1 = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y}$.

4. Note that the regression model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$ with n observations can be written in the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where the design matrix $\boldsymbol{X}$ is defined as

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix},$$

the coefficient vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)'$, and the random error vector $= (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$.

5. Motivated by the fact that $\hat{\beta}_1 = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y}$, conjecture that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is the least squares estimate of $\boldsymbol{\beta}$ and verify this using Theorem 1.

Lemma 1 *Suppose that the design matrix* $\boldsymbol{X}$ *is full rank so that* $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ *and* $\widehat{\boldsymbol{\beta}}$ *exist. then* $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$.

Proof: Multiply both sides of $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ by $(\boldsymbol{X}'\boldsymbol{X})$ and re-arrange the result. □

Theorem 1 *Suppose* $\boldsymbol{X}$ *is full rank. Then* $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ *exists and minimizes*

$$RSS(\widetilde{\boldsymbol{\beta}}) = \sum(y_i - \hat{y}_i)^2$$

$$= (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}})$$

Proof: Because $\boldsymbol{X}$ is full rank, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$and $\widehat{\boldsymbol{\beta}}$ exist. Let $\widehat{\boldsymbol{\beta}}_*$ be an alternate estimator of $\boldsymbol{\beta}$. Then

$\widehat{\boldsymbol{\beta}}_* = \widehat{\boldsymbol{\beta}} + \boldsymbol{\Delta}$ and

$$\begin{aligned} RSS(\widehat{\boldsymbol{\beta}}_*) &= [\boldsymbol{y} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}} + \boldsymbol{\Delta})]'[\boldsymbol{y} - \boldsymbol{X}(\widehat{\boldsymbol{\beta}} + \boldsymbol{\Delta})] \\ &= [(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) - \boldsymbol{X}\boldsymbol{\Delta}]'[(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) - \boldsymbol{X}\boldsymbol{\Delta}] \\ &= (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + \boldsymbol{\Delta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Delta} \end{aligned}$$

since the cross products $\boldsymbol{\Delta}'\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ and $(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'\boldsymbol{X}\boldsymbol{\Delta}$ are zero by Lemma 1. Since $\boldsymbol{X}$ is full rank, $\boldsymbol{X}\boldsymbol{\Delta} \neq \boldsymbol{0}$ and so

$$\begin{aligned} RSS(\widehat{\boldsymbol{\beta}}_*) &= RSS(\widehat{\boldsymbol{\beta}}) + \boldsymbol{\Delta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Delta} \\ &> RSS(\widehat{\boldsymbol{\beta}}). \quad \square \end{aligned}$$

As a calculus-free alternative to the preceding example, we propose deriving the least squares estimate $\hat{\beta}_1$ using the Cauchy-Schwarz inequality. One possibility is to replace 1.1 – 1.4 above with the following:

1.1' Suppose we have one observation (x_1, y_1) from the zero-intercept model $y = \beta_1 x + \epsilon$. Then a reasonable estimate of the slope β_1 is the rise over the run, $\tilde{\beta}_1 = y_1/x_1$.

1.2' (Optional) Derive the following properties of $\tilde{\beta}_1$:

$$\begin{aligned} &\tilde{\beta}_1 = \beta_1 + \frac{\epsilon_1}{x_1} \\ &E[\tilde{\beta}_1] = \beta_1 \\ &\text{Var}[\tilde{\beta}_1] = \frac{\sigma^2}{x_1^2} \end{aligned} \tag{1}$$

where σ^2 is the random error variance.

1.3' If we have two observations, (x_1, y_1) and (x_2, y_2), we can estimate β_1 by $\tilde{\beta}_{1,w} = \bar{y}_w/\bar{x}_w$ where $\bar{y}_w$ and $\bar{x}_w$ are weighted averages of the y and x values, e.g., $\bar{y}_w = w_1 y_1 + w_2 y_2$:

$$\tilde{\beta}_{1,w} = \frac{w_1 y_1 + w_2 y_2}{w_1 x_1 + w_2 x_2}$$

Equation (1) suggests using weights which depend on the magnitudes of the x values.

1.4' Derive the following properties of $\tilde{\beta}_{1,w}$:

$$\begin{aligned} &\tilde{\beta}_{1,w} = \beta_1 + \frac{w_1 \epsilon_1 + w_2 \epsilon_2}{w_1 x_1 + w_2 x_2} \\ &E[\tilde{\beta}_{1,w}] = \beta_1 \\ &\text{Var}[\tilde{\beta}_{1,w}] = \sigma^2 \frac{w_1^2 + w_2^2}{(w_1 x_1 + w_2 x_2)^2} \end{aligned} \tag{2}$$

1.5' Determine weights which minimize $\mathrm{Var}[\tilde{\beta}_{1,w}]$ using the Cauchy-Schwarz inequality. Cauchy-Schwarz states that for any w_1, w_2, x_1, and x_2

$$(w_1x_1 + w_2x_2)^2 \leq (w_1^2 + w_2^2)(x_1^2 + x_2^2) \tag{3}$$

with equality if and only if $w_i = cx_i, i = 1,2,$ for any constant c. Combining (2) and (3) yields

$$\mathrm{Var}[\tilde{\beta}_{1,w}] \geq \sigma^2 \frac{1}{x_1^2+x_2^2}$$

with equality if and only if $w_i = cx_i, i = 1,2$ for any $c \neq 0$. Thus the version of $\tilde{\beta}_{1,w}$ with minimum variance is

$$\tilde{\beta}_{1,min} = \frac{x_1y_1+x_2y_2}{x_1^2+x_2^2} \tag{4}$$

1.6' Show that $\tilde{\beta}_{1,min}$ is the least squares estimate of β_1 using Theorem 2, the scalar version of Theorem 1.

Lemma 2 *Let* $\tilde{\beta}_{1,min}$ *be defined by (4) then*

$$x_1(y_1 - \tilde{\beta}_{1,min}x_1) + x_2(y_2 - \tilde{\beta}_{1,min}x_2) = 0.$$

Proof: Replace $\tilde{\beta}_{1,min}$ by the right side of (4) and simplify. □

Theorem 2 *Let* $\hat{\beta}_{1,*}$ *be any other estimator of* β_1. *Then*

$$RSS(\hat{\beta}_{1,*}) > RSS(\tilde{\beta}_{1,min}).$$

Proof: Let $\hat{\beta}_{1,*} = \tilde{\beta}_{1,min} + \Delta$. Then

$$\begin{aligned}
RSS(\hat{\beta}_{1,*}) &= (y_1 - \hat{\beta}_{1,*}x_1)^2 + (y_2 - \hat{\beta}_{1,*}x_2)^2 \\
&= [y_1 - (\tilde{\beta}_{1,min} + \Delta)x_1]^2 + [y_2 - (\tilde{\beta}_{1,min} + \Delta)x_2]^2 \\
&= [(y_1 - \tilde{\beta}_{1,min}x_1) - \Delta x_1]^2 + [(y_2 - \tilde{\beta}_{1,min}x_2) - \Delta x_2]^2 \\
&= (y_1 - \tilde{\beta}_{1,min}x_1)^2 + (y_2 - \tilde{\beta}_{1,min}x_2)^2 \\
&\quad -2\Delta[x_1(y_1 - \tilde{\beta}_{1,min}x_1) + x_2(y_2 - \tilde{\beta}_{1,min}x_2)] + \Delta^2(x_1^2 + x_2^2) \\
&= (y_1 - \tilde{\beta}_{1,min}x_1)^2 + (y_2 - \tilde{\beta}_{1,min}x_2)^2 + \Delta^2(x_1^2 + x_2^2)
\end{aligned}$$

where the last line follows since the cross product term is zero by Lemma 2. Thus $RSS(\tilde{\beta}_{1,*}) = RSS(\tilde{\beta}_{1,min}) + \Delta^2(x_1^2 + x_2^2) > RSS(\tilde{\beta}_{1,min})$. □

4 Discussion

Rigorously deriving least squares coefficient formulas beginning with the zero-intercept model has several benefits. The first is that it reduces or eliminates the need for calculus and emphasizes the linear algebra and matrix concepts central to regression. For example, formulas for $\hat{\beta}_1$ can easily be written in vector form and thereby be seen as special cases of the general formulas, e.g., $\hat{\beta}_1 = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y}$ versus $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $\sigma^2_{\hat{\beta}_1} = \sigma^2(\boldsymbol{x}'\boldsymbol{x})^{-1}$ versus $\Sigma_{\widehat{\boldsymbol{\beta}}} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Similarly, the zero-intercept prediction formula, easily written in vector form,

$$\begin{aligned}\widehat{\boldsymbol{y}} &= \hat{\beta}_1\boldsymbol{x}\\ &= \left(\frac{\boldsymbol{x}'\boldsymbol{y}}{\boldsymbol{x}'\boldsymbol{x}}\right)\boldsymbol{x}\\ &= \boldsymbol{x}(\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y}\end{aligned}$$

is seen to be a special case of the general formula $\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, motivating the fact that in the general case $\widehat{\boldsymbol{y}}$ is the projection of $\boldsymbol{y}$ onto on the vector space spanned by the columns of $\boldsymbol{X}$. Further, Lemma 2, i.e., $\boldsymbol{x}'(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = 0$, provides a simple illustration of least squares geometry, namely, that the residual vector $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$ and $\boldsymbol{x}$ are orthogonal. Finally, an additional benefit of starting with the zero-intercept model is that the least squares formula for β_1 can be derived using Cauchy-Schwarz, e.g. 1.1'–1.6'. Combining this result with Theorem 1 provides a new, non-calculus derivation of the least squares matrix formula which has advantages compared to other non-calculus approaches (e.g., Darlington 1969; Ehrenberg 1983). A major advantage of the Cauchy-Schwarz derivation is that it's thoroughly statistical and emphasizes optimal estimation and linear algebra rather than calculus. In the process of deriving the least squares estimate of β_1 we prove that it's unbiased and that, among all estimators of the form $\tilde{\beta}_{1,w}$, it has the least variance. In other words, we prove a zero-intercept analogue of the Gauss-Markov theorem and show that $\hat{\beta}_1$ is the *best linear unbiased estimator* (BLUE) of β_1 among estimators of the form $\tilde{\beta}_{1,w}$. Thus students see optimal estimation properties of least squares estimates in a simple univariate setting prior to seeing them in the general case.

References

[1] Abraham, B., and Ledolter, J. (2006), *Introduction to Regression Modeling*, *Belmont, CA: Duxbury Press*.

[2] American Statistical Association Undergraduate Guidelines Workgroup. (2014) *2014 curriculum guidelines for undergraduate programs in statistical science*. Alexandria, VA: *American Statistical Association*. https://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf

[3] Darlington, R. (1969) "Deriving Least-Squares Weights without Calculus," *The American Statistician*, **23**, 41-42.

[4] Ehrenberg, A.S.C. (1983) "Deriving the Least Squares Equation," *The American Statistican*, **37**, 232-232.

[5] Freund, R., Williams, W., and Sa, P. (2006), Introduction to Linear Regression Analysis (2nd ed.), San Diego, CA: Academic Press.

[6] Kutner, M., Nachtsheim, C., and Neter, J. (2004), Applied Linear Regression Models (4th ed.), New York: McGraw-Hill/Irwin.

[7] Mendenhall, W., and Sincich, T. (2020), A Second Course in Statistics: Regression Analysis (8th ed.), Upper Saddle River, NJ: Pearson/Prentice-Hall.
[8] Montgomery, D., Peck, E., and Vining, G. (2021), Introduction to Linear Regression Analysis (6th ed.), New York: Wiley.
[9] Ryan, T. (2009), *Modern Regression Methods* (2nd ed.), New York: Wiley.
[10] Stanley, J., and Glass, G. V. (1969) "An Algebraic Proof That the Sum of the Squared Errors in Estimating Y from X via b_1 and b_0 Is Minimal," *The American Statistician*, **23**, 25-26.