

NOTES ON SIMPLE LINEAR REGRESSION

1. INTRODUCTION

The purpose of these notes is to supplement the mathematical development of linear regression in Devore (2008). This development also draws on the treatment in Johnston (1963) and Larsen and Marx (1986). We begin with the basic *least squares* estimation problem, and next develop the moments of the estimators. Finally the fundamental optimality property of these estimators is established in terms of the *Gauss-Markov Theorem*.

2. LINEAR LEAST SQUARES ESTIMATION

The basic linear model assumes the existence of a linear relationship between two variables, x and y , which is disturbed by some random error, ε . Hence for each value of x the corresponding y -value is a *random variable* of the form

$$(2.1) \quad Y = \beta_0 + \beta_1 x + \varepsilon$$

where β_0 and β_1 are designated, respectively, as the *intercept parameter* and the *slope parameter* of the linear function, $\beta_0 + \beta_1 x$. If n values ($x_i : i = 1, \dots, n$) of x are observed, with corresponding errors ($\varepsilon_i : i = 1, \dots, n$), then the resulting random variables, ($Y_i : i = 1, \dots, n$), are given by

$$(2.2) \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad i = 1, \dots, n$$

In this context it is assumed that the random errors, ($\varepsilon_i : i = 1, \dots, n$), are *independently and identically distributed* (*iid*) with mean zero and variance σ^2 , so that

$$(2.3) \quad E(\varepsilon_i) = 0 \quad , \quad i = 1, \dots, n$$

$$(2.4) \quad \text{var}(\varepsilon_i) = \sigma^2 \quad , \quad i = 1, \dots, n$$

If values of y corresponding to ($x_i : i = 1, \dots, n$) are also observed, and are denoted by ($y_i : i = 1, \dots, n$), then the *least squares estimation problem* is to find estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, of the unknown parameter values, β_0 and β_1 , which minimize the sum of squared residuals [designated as $f(b_0, b_1)$ in Devore, p. 455]:

$$(2.5) \quad S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

This function is easily seen to be convex and differentiable in β_0 and β_1 , so that the unique solution $(\hat{\beta}_0, \hat{\beta}_1)$ is given by the first-order conditions:

$$(2.6) \quad 0 = \frac{\partial}{\partial \beta_0} S(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1)$$

$$(2.7) \quad 0 = \frac{\partial}{\partial \beta_1} S(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i)$$

If we let $\bar{x} = \frac{1}{n} \sum_i x_i$ and $\bar{y} = \frac{1}{n} \sum_i y_i$, then by (2.6)

$$(2.8) \quad \sum_i y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_i x_i = 0 \Rightarrow \frac{1}{n} \sum_i y_i - \hat{\beta}_0 - \hat{\beta}_1 \left(\frac{1}{n} \sum_i x_i \right) = 0$$

$$\Rightarrow \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$$

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

and by (2.7)

$$(2.9) \quad \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

To simplify (2.9) let the *estimated y-value* corresponding to $(\hat{\beta}_0, \hat{\beta}_1)$ be defined by

$$(2.10) \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \dots, n$$

and rewrite (2.9) as

$$(2.11) \quad \sum_i (y_i - \hat{y}_i) x_i = 0$$

Note also from (2.8) that

$$(2.12) \quad \sum_i (\hat{y}_i - y_i) = \sum_i \hat{y}_i - \sum_i y_i = \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) - n\bar{y}$$

$$= (n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i) - n\bar{y} = n(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) - n\bar{y}$$

$$= 0$$

To solve for $\hat{\beta}_1$ we first observe by subtracting (2.8) from (2.10) that

$$(2.13) \quad \hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$$

$$\Rightarrow (\hat{y}_i - y_i) + (y_i - \bar{y}) = \hat{\beta}_1(x_i - \bar{x}) \quad , \quad i = 1, \dots, n$$

Hence, multiplying both sides by $(x_i - \bar{x})$ and summing over i , we obtain

$$(2.14) \quad \sum_i (\hat{y}_i - y_i)(x_i - \bar{x}) + \sum_i (y_i - \bar{y})(x_i - \bar{x}) = \hat{\beta}_1 \sum_i (x_i - \bar{x})^2$$

But since (2.11) and (2.12) imply

$$(2.15) \quad \sum_i (\hat{y}_i - y_i)(x_i - \bar{x}) = -\sum_i (y_i - \hat{y}_i)x_i - \bar{x} \sum_i (\hat{y}_i - y_i) = 0$$

we may conclude from (2.14) that

$$(2.16) \quad \hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

[See expression (12.2) in Devore, p. 456.] Finally, by employing (2.8), we may solve for $\hat{\beta}_0$ in terms of $\hat{\beta}_1$ as

$$(2.17) \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

[See expression (12.3) in Devore, p. 456.]

3. MOMENTS OF THE ESTIMATORS

The estimators in (2.16) and (2.17) depend on the values of the random variables, $(Y_i : i = 1, \dots, n)$, and hence are themselves random variables. In particular, if the *sample mean* of the Y_i 's is denoted by

$$(3.1) \quad \bar{Y} = \frac{1}{n} \sum_i Y_i = \frac{1}{n} \sum_i (\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 \bar{x} + \frac{1}{n} \sum_i \varepsilon_i \quad ,$$

then it follows at once from (2.16) that $\hat{\beta}_1$ is a random variable of the form

$$(3.2) \quad \hat{\beta}_1 = \frac{\sum_i (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

and, similarly, that $\hat{\beta}_0$ is a random variable of the form

$$(3.3) \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

To compute the moments of the slope estimator, $\hat{\beta}_1$, it is convenient to simplify expression (3.2) as follows. By breaking (3.2) into two terms

$$(3.4) \quad \hat{\beta}_1 = \frac{\sum_i Y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} - \frac{\bar{Y} \sum_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

and observing that

$$(3.5) \quad \sum_i (x_i - \bar{x}) = \sum_i x_i - n\bar{x} = n \left(\frac{1}{n} \sum_i x_i - \bar{x} \right) = 0$$

we see that the second term vanishes, and hence that the estimator $\hat{\beta}_1$ can be written as a *linear combination* of the Y_i 's

$$(3.6) \quad \hat{\beta}_1 = \sum_i w_i Y_i$$

where the coefficients w_i are of the form

$$(3.7) \quad w_i = \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2}, \quad i = 1, \dots, n$$

and hence are *non-random* (i.e., depend only on the given values of the x_i 's). To analyze (3.6) we begin with several observations about the coefficient values in (3.7). First observe from (3.5) that

$$(3.8) \quad \sum_i w_i = \frac{\sum_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = 0$$

and moreover that

$$(3.9) \quad \sum_i w_i (x_i - \bar{x}) = \frac{\sum_i (x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} = 1$$

which together with (3.8) also implies

$$(3.10) \quad \sum_i w_i x_i = \sum_i w_i x_i - \bar{x} \sum_i w_i = \sum_i w_i (x_i - \bar{x}) = 1$$

To compute the mean of $\hat{\beta}_1$, observe from (2.2) and (2.3) that

$$(3.11) \quad E(Y_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i) = \beta_0 + \beta_1 x_i$$

so that by (3.6), together with (3.8) and (3.10),

$$(3.12) \quad \begin{aligned} E(\hat{\beta}_1) &= \sum_i w_i E(Y_i) = \sum_i w_i (\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_i w_i + \beta_1 \sum_i w_i x_i = (0) + \beta_1 (1) \\ &= \beta_1 \end{aligned}$$

Thus $\hat{\beta}_1$ is an *unbiased estimator* of β_1 . Moreover, since (3.1) and (2.3) imply that

$$(3.13) \quad E(\bar{Y}) = \beta_0 + \beta_1 \bar{x} + \frac{1}{n} \sum_i E(\varepsilon_i) = \beta_0 + \beta_1 \bar{x}$$

it follows from (3.3) together with (3.13) that

$$(3.14) \quad E(\hat{\beta}_0) = E(\bar{Y}) - E(\hat{\beta}_1) \bar{x} = [\beta_0 + \beta_1 \bar{x}] - \beta_1 \bar{x} = \beta_0$$

and thus that $\hat{\beta}_0$ is also an *unbiased estimator* of β_0

To compute the variance of $\hat{\beta}_1$, we again observe from (3.6) that

$$(3.15) \quad \begin{aligned} \hat{\beta}_1 &= \sum_i w_i (\beta_0 + \beta_1 x_i + \varepsilon_i) = \sum_i w_i (\beta_0 + \beta_1 x_i) + \sum_i w_i \varepsilon_i \\ &= \text{const} + \sum_i w_i \varepsilon_i \end{aligned}$$

and hence (from the independence of the ε_i 's that

$$(3.16) \quad \text{var}(\hat{\beta}_1) = \text{var}\left(\sum_i w_i \varepsilon_i\right) = \sum_i w_i^2 \text{var}(\varepsilon_i)$$

Hence we may conclude from (2.4) and (3.7) that

$$(3.17) \quad \text{var}(\hat{\beta}_1) = \sigma^2 \sum_i w_i^2 = \sigma^2 \sum_i \left(\frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} \right)^2$$

$$= \sigma^2 \frac{\sum_i (x_i - \bar{x})^2}{[\sum_j (x_j - \bar{x})^2]^2} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

[See expression (12.4) in Devore, p. 470.] Similarly, to determine the variance of $\hat{\beta}_0$, we observe from the above relations that

$$\begin{aligned} (3.18) \quad \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x} = \frac{1}{n} \sum_i Y_i - \bar{x} \sum_i w_i Y_i = \sum_i \left(\frac{1}{n} - \bar{x} w_i \right) Y_i \\ &= \sum_i \left(\frac{1}{n} - \bar{x} w_i \right) (\beta_0 + \beta_1 x_i + \varepsilon_i) \\ &= \sum_i \left(\frac{1}{n} - \bar{x} w_i \right) (\beta_0 + \beta_1 x_i) + \sum_i \left(\frac{1}{n} - \bar{x} w_i \right) \varepsilon_i \\ &= \text{const} + \sum_i \left(\frac{1}{n} - \bar{x} w_i \right) \varepsilon_i \end{aligned}$$

and hence that

$$\begin{aligned} (3.19) \quad \text{var}(\hat{\beta}_0) &= \sum_i \left(\frac{1}{n} - \bar{x} w_i \right)^2 \text{var}(\varepsilon_i) \\ &= \sigma^2 \sum_i \left(\frac{1}{n} - \bar{x} w_i \right)^2 = \sigma^2 \sum_i \left(\frac{1}{n^2} - \frac{2}{n} \bar{x} w_i + \bar{x}^2 w_i^2 \right) \\ &= \sigma^2 \left(\frac{1}{n^2} - \frac{2}{n} \bar{x} \sum_i w_i + \bar{x}^2 \sum_i w_i^2 \right) \\ &= \sigma^2 \left\{ \frac{1}{n} - (0) + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right\} = \sigma^2 \left\{ \frac{\sum_i (x_i - \bar{x})^2 + n \bar{x}^2}{n \sum_i (x_i - \bar{x})^2} \right\} \\ &= \sigma^2 \left\{ \frac{\sum_i x_i^2 - 2 \bar{x} \sum_i x_i + 2n \bar{x}^2}{n \sum_i (x_i - \bar{x})^2} \right\} = \sigma^2 \left\{ \frac{\sum_i x_i^2 - 2n \bar{x}^2 + 2n \bar{x}^2}{n \sum_i (x_i - \bar{x})^2} \right\} \\ &= \sigma^2 \left\{ \frac{\sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2} \right\} \end{aligned}$$

4. GAUSS MARKOV THEOREM

Finally we establish the fundamental optimality property of the above estimators. To do so, recall that for an independent random sample (Y_1, \dots, Y_n) from a population with mean, $\mu = E(Y)$, the sample mean, \bar{Y}_n , was shown to be a *best linear unbiased* (BLU) estimator of μ .

This optimality property turns out to be shared by the least-squares estimators $(\hat{\beta}_0, \hat{\beta}_1)$ above.

This result, known as the *Gauss-Markov Theorem*, provides the single strongest justification for linear least-squares estimation, and can be stated as follows:

GAUSS MARKOV THEOREM. *For any linear function, $L = a_0\beta_0 + a_1\beta_1$, of $(\hat{\beta}_0, \hat{\beta}_1)$, the least squares estimator, $\hat{L} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1$, has minimum variance among all linear unbiased estimators of L .*

Proof: We shall prove this assertion only for the linear function with coefficients $(a_0 = 0, a_1 = 1)$, i.e., for the estimate, $\hat{\beta}_1$, of the slope parameter, β_1 , (which is by far the most important of the two individual parameters). The argument for any linear function of β_0 and β_1 is essentially the same. To begin with, observe from (3.6) that $\hat{\beta}_1$ is indeed a linear estimator, i.e., is a linear function of the random variables $(Y_i: i = 1, \dots, n)$. Moreover, it was shown in (3.12) that $\hat{\beta}_1$ is also an unbiased estimator of β_1 . Hence it remains only to show that the variance of $\hat{\beta}_1$ never exceeds that of any other linear unbiased estimator. To do so, consider any other linear estimator, say

$$(4.1) \quad \tilde{\beta}_1 = \sum_i c_i Y_i$$

and suppose that $\tilde{\beta}_1$ is also unbiased estimator. Then by (3.12) we must have

$$(4.2) \quad \begin{aligned} \beta_1 &= E(\tilde{\beta}_1) = \sum_i c_i E(Y_i) \\ &= \sum_i c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_i c_i + \beta_1 \sum_i c_i x_i \end{aligned}$$

But since unbiasedness requires that (4.2) hold for all values of the unknown parameters β_0 and β_1 , it follows by setting $\beta_0 = 1$ and $\beta_1 = 0$ that

$$(4.3) \quad \sum_i c_i = 0$$

and in turn, by setting $\beta_1 = 1$, that

$$(4.4) \quad \sum_i c_i x_i = 1$$

Hence, in a manner identical with (3.15), these two conditions are seen to imply that

$$(4.5) \quad \tilde{\beta}_1 - \beta_1 = \sum_i c_i Y_i - \beta_1 = \sum_i c_i \varepsilon_i$$

and thus that the variance of $\tilde{\beta}_1$ is given by

$$(4.6) \quad \text{var}(\tilde{\beta}_1) = \sum_i c_i^2 \text{var}(\varepsilon_i) = \sigma^2 \sum_i c_i^2$$

To compare this with $\text{var}(\hat{\beta}_1)$, observe first that if the differences between the coefficients of $\tilde{\beta}_1$ and $\hat{\beta}_1$ in (4.1) and (3.6) are denoted by $d_i = c_i - w_i$, $i = 1, \dots, n$, then (4.6) can be rewritten as

$$(4.7) \quad \text{var}(\tilde{\beta}_1) = \sigma^2 \sum_i (w_i + d_i)^2 = \sigma^2 (\sum_i w_i^2 + 2 \sum_i d_i w_i + \sum_i d_i^2)$$

But by (4.3) and (4.4) together with (3.8) and (3.10) we must have

$$(4.8) \quad 0 = \sum_i c_i = \sum_i w_i + \sum_i d_i = (0) + \sum_i d_i \Rightarrow \sum_i d_i = 0$$

$$(4.9) \quad 1 = \sum_i c_i x_i = \sum_i w_i x_i + \sum_i d_i x_i = 1 + \sum_i d_i x_i \Rightarrow \sum_i d_i x_i = 0$$

which together imply that

$$(4.10) \quad \sum_i d_i w_i = \sum_i d_i \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2} = \frac{\sum_i d_i x_i - \bar{x} \sum_i d_i}{\sum_j (x_j - \bar{x})^2} = 0$$

Hence, recalling (3.7), we see that (4.7) reduces to

$$(4.11) \quad \text{var}(\tilde{\beta}_1) = \sigma^2 \sum_i w_i^2 + \sigma^2 \sum_i d_i^2 = \text{var}(\hat{\beta}_1) + \sigma^2 \sum_i d_i^2$$

and may conclude from the nonnegativity of $\sigma^2 \sum_i d_i^2$ that

$$(4.12) \quad \text{var}(\tilde{\beta}_1) \geq \text{var}(\hat{\beta}_1)$$

Thus $\hat{\beta}_1$ has *minimum variance* among all linear unbiased estimators, and the result is established.

5. REFERENCES

Devore, J.L., (2008) *Probability and Statistics for Engineering and the Sciences*, Seventh Edition, Duxbury Press, Belmont, California.

Larsen, R.J. and M.L. Marx, (1986) *An Introduction to Mathematical Statistics and its Applications*, Second Edition, Prentice-Hall, Englewood Cliffs, N.J.

Johnston, J., (1963) *Econometric Methods*, McGraw-Hill, N.Y.