

# Derivation of the Least Squares Regression Line Without Calculus

## Jonathan D. Cryer

Consider the data pairs  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  and with  $n > 1$ ,  $s_x > 0$ , and  $s_y > 0$ . Let  $x^*$  and  $y^*$  be standardized  $x$  and  $y$  values, that is,  $x_i^* = (x_i - \bar{x})/s_x$  and  $y_i^* = (y_i - \bar{y})/s_y$ . We wish to find the equation of the line,  $y^* = \beta_0 + \beta_1 x^*$ , that fits the standardized data in the least squares sense. We will show that the “best” (that is, least squares) slope is given by  $\hat{\beta}_1 = r$ , (the correlation between  $x$  and  $y$ ), and the best intercept is  $\hat{\beta}_0 = 0$ . First note that, after standardization, the variables have mean zero and standard deviation 1, so that  $\sum_{i=1}^n x_i^* = 0$  and  $\sum_{i=1}^n (x_i^*)^2 = n - 1$ . Similarly for

the  $y^*$ ’s. Also  $\sum_{i=1}^n x_i^* y_i^* = (n - 1)r$ . For arbitrary intercept  $\beta_0$  and slope  $\beta_1$  the sum of squared residuals (errors) may be written

$$\begin{aligned} SSE(\beta_0, \beta_1) &= g(\beta_0, \beta_1) = \sum_{i=1}^n [y_i^* - (\beta_0 + \beta_1 x_i^*)]^2 \\ &= \sum_{i=1}^n [(y_i^*)^2 - 2\beta_0 y_i^* - 2\beta_1 x_i^* y_i^* + 2\beta_0 \beta_1 x_i^* + \beta_0^2 + \beta_1^2 (x_i^*)^2] \\ &= (n - 1) - 0 - 2(n - 1)\beta_1 r + 0 + n\beta_0^2 + (n - 1)\beta_1^2 \\ &= (n - 1)(1 - 2r\beta_1 + \beta_1^2) + n\beta_0^2 \end{aligned}$$

Now  $n\beta_0^2 \geq 0$ , so  $g(\beta_0, \beta_1) \geq g(0, \beta_1) = (n - 1)(1 - 2r\beta_1 + \beta_1^2) = (n - 1)[1 - r^2 + (\beta_1 - r)^2]$ . By inspection, this is smallest when  $\beta_1 = r$ . Thus the least squares solutions are  $\hat{\beta}_1 = r$  and  $\hat{\beta}_0 = 0$ .

We also note that the minimum value of  $SSE$  may be written:

$$\min_{\beta_0, \beta_1} SSE(\beta_0, \beta_1) = SSE(\hat{\beta}_0, \hat{\beta}_1) = g(0, r) = (n - 1)(1 - r^2)$$

Since the  $SSE$  is nonnegative, this shows that  $r$  must always be between  $-1$  and  $+1$ .

The least squares regression line *in original terms* is found by solving for  $\hat{y}$  in  $\frac{\hat{y} - \bar{y}}{s_y} = r \left( \frac{x - \bar{x}}{s_x} \right)$ .

We obtain  $\hat{y} = b_0 + b_1 x$  where  $b_1 = r \frac{s_y}{s_x}$  and  $b_0 = \bar{y} - b_1 \bar{x}$ .