19 Simple Regression: Descriptive Statistics

19.B Appendix: Characterization of the Neutral Line

Let $\{(x_j, y_j)\}_{j=1}^{N}$ be a bivariate data set with $\sigma_x^2 > 0$, $\sigma_y^2 > 0$, and $\sigma_{x,y} \neq 0$. In this appendix, we show that the neutral line minimizes the sum of squared standardized distances to points in the data set.

At the start we will consider lines ℓ that can be described by equations of the form $y = ax + b$. In doing so we ignore vertical lines, but in the end we will show that a vertical line cannot solve the minimization problem.

To begin, we obtain an explicit formula for $(d_s((x, y), \ell))^2$, the squared standardized distance from (x, y) to the closest point on line ℓ. This is the solution to

$$\min_{\hat{x},\hat{y}} \left(\left(\frac{x - \hat{x}}{\sigma_x} \right)^2 + \left(\frac{y - \hat{y}}{\sigma_y} \right)^2 \right) \quad \text{subject to } \hat{y} = a + b\hat{x}.$$

Substituting in the constraint, we can rewrite the problem as

$$\min_{\hat{x}} \left(\frac{(x - \hat{x})^2}{\sigma_x^2} + \frac{(y - (a + b\hat{x}))^2}{\sigma_y^2} \right).$$

Taking the derivative of the objective function with respect to $\hat{x}$ and setting it equal to 0 yields

$$-\frac{2(x - \hat{x})}{\sigma_x^2} - \frac{2b(y - (a + b\hat{x}))}{\sigma_y^2} = 0. \tag{19.B.1}$$

Since the second derivative of the objective function is $2(\frac{1}{\sigma_x^2} + \frac{b^2}{\sigma_y^2}) > 0$, condition (19.B.1) is sufficient for solving the minimization problem. Solving (19.B.1) for $\hat{x}$ and then substituting the result into $\hat{y} = a + b\hat{x}$, we find that the closest point to (x, y) on ℓ is

$$(\hat{x}^*, \hat{y}^*) = \left(\frac{\sigma_y^2 x + b\sigma_x^2 y - b\sigma_x^2 a}{b^2\sigma_x^2 + \sigma_y^2}, \frac{b\sigma_y^2 x + b^2\sigma_x^2 y + a\sigma_y^2}{b^2\sigma_x^2 + \sigma_y^2} \right).$$

Thus the squared distance between (x, y) and ℓ is

$$
\begin{aligned}
\big(d_s((x_j, y_j), \ell)\big)^2 &= \left(\frac{x}{\sigma_x} - \frac{\hat{x}^*}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y} - \frac{\hat{y}^*}{\sigma_y}\right)^2 \\
&= \left(\frac{(b^2\sigma_x^2 + \sigma_y^2)x - \big(\sigma_y^2 x + b\sigma_x^2 y - b\sigma_x^2 a\big)}{\sigma_x(b^2\sigma_x^2 + \sigma_y^2)}\right)^2 \\
&\quad + \left(\frac{(b^2\sigma_x^2 + \sigma_y^2)y - \big(b\sigma_y^2 x + b^2\sigma_x^2 y + a\sigma_y^2\big)}{\sigma_y(b^2\sigma_x^2 + \sigma_y^2)}\right)^2 \\
&= \frac{(b^2\sigma_x^2 + \sigma_y^2)(bx - y + a)^2}{(b^2\sigma_x^2 + \sigma_y^2)^2} \\
&= \frac{(bx - y + a)^2}{b^2\sigma_x^2 + \sigma_y^2}. \qquad (19.B.2)
\end{aligned}
$$

It follows that we can express the sum of squared standardized distances of the data set $\{(x_j, y_j)\}_{j=1}^N$ to the line $y = ax + b$ as

$$\sum_{j=1}^{N} \big(d_s((x_j, y_j), \ell)\big)^2 = \sum_{j=1}^{N} \frac{(bx_j - y_j + a)^2}{b^2\sigma_x^2 + \sigma_y^2}. \qquad (19.B.3)$$

Our aim is to choose the values of a and b that minimize this sum.

Taking the derivative of (19.B.3) with respect to a and setting it equal to 0 yields

$$\frac{2}{b^2\sigma_x^2 + \sigma_y^2} \sum_{j=1}^{N} (bx_j - y_j + a) = 0,$$

which simplifies to

$$a = \mu_y - b\mu_x. \qquad (19.B.4)$$

Substituting this equation into $y = ax + b$ and rearranging yields

$$y - \mu_y = b(x - \mu_x), \qquad (19.B.5)$$

which says that the line passes through the mean point (μ_x, μ_y). Since the second derivative of (19.B.3) with respect to a is positive, (19.B.4) gives us the optimal choice of a for a given value of b. Substituting (19.B.4) into (19.B.3) yields

$$\begin{aligned}\sum_{j=1}^{N} \frac{(b(x_j - \mu_x) - (y_j - \mu_y))^2}{b^2\sigma_x^2 + \sigma_y^2} &= \sum_{j=1}^{N} \frac{b^2(x_j - \mu_x)^2 - 2b(x_j - \mu_x)(y_j - \mu_y) + (y_j - \mu_y)^2}{b^2\sigma_x^2 + \sigma_y^2} \\ &= \frac{N(b^2\sigma_x^2 - 2b\sigma_{x,y} + \sigma_y^2)}{b^2\sigma_x^2 + \sigma_y^2}.\end{aligned} \tag{19.B.6}$$

To find the optimal choice of b, we take the derivative of (19.B.6) with respect to b and set it equal to 0. This yields

$$\frac{N\big((2b\sigma_x^2 - 2\sigma_{x,y})(b^2\sigma_x^2 + \sigma_y^2) - (b^2\sigma_x^2 - 2b\sigma_{x,y} + \sigma_y^2) \cdot 2b\sigma_x^2\big)}{(b^2\sigma_x^2 + \sigma_y^2)^2} = 0,$$

which simplifies to

$$2\sigma_{x,y}(b^2\sigma_x^2 - \sigma_y^2) = 0. \tag{19.B.7}$$

The two solutions to (19.B.7) are $b = \frac{\sigma_y}{\sigma_x}$ and $b = -\frac{\sigma_y}{\sigma_x}$. By taking the second derivative of (19.B.6) with respect to b, one can check that of the two solutions, the local maximizer is the one whose sign is the same as that of $\sigma_{x,y}$. Thus

$$b = \pm\frac{\sigma_y}{\sigma_x}, \tag{19.B.8}$$

where, as in the text, we use the plus sign if $\rho_{x,y} > 0$ and the minus sign if $\rho_{x,y} < 0$. Substituting (19.B.8) into (19.B.5) yields

$$y - \mu_y = \pm\frac{\sigma_y}{\sigma_x}(x - \mu_x),$$

which is the neutral line.

The previous paragraph showed that the choice of b from (19.B.8) is the only one that is locally optimal. To finish the proof, we must show that lower values of the objective function cannot be obtained by choosing lines through the mean point with b approaching plus or minus infinity—that is, lines that are vertical or nearly so. To do so, we first compute the sum of the squared standardized distances to the neutral line: substituting expression (19.B.8) for b into (19.B.6), we obtain

$$\begin{aligned}\sum_{j=1}^{N}\left(d_s((x_j,y_j),\ell)\right)^2 &= \frac{N\left(\left(\frac{\sigma_y}{\sigma_x}\right)^2\sigma_x^2 - 2\left(\pm\frac{\sigma_y}{\sigma_x}\sigma_{x,y}\right)+\sigma_y^2\right)}{\left(\frac{\sigma_y}{\sigma_x}\right)^2\sigma_x^2+\sigma_y^2}\\ &= \frac{N\left(2\sigma_y^2\left(1-\left(\pm\frac{\sigma_{x,y}}{\sigma_x\sigma_y}\right)\right)\right)}{2\sigma_y^2}\\ &= N\big(1-(\pm\rho_{x,y})\big)\\ &= N\big(1-|\rho_{x,y}|\big), \qquad (19.B.9)\end{aligned}$$

where the last equality follows from the plus or minus matching the sign of $\rho_{x,y}$.

Now, imagine a standardized scatterplot of the data with a steep, positively sloped line drawn through the mean point (μ_x, μ_y) (as in equation (19.B.5)). Then imagine rotating this line counterclockwise. Using the picture, it is not hard to see that the sum of squared standardized distances generated by these lines changes continuously as the line is rotated, and in particular as the line passes through the vertical line $x = \mu_x$ to steep, negatively sloped lines. Furthermore, since the closest points from each data point to the vertical line are obtained by moving horizontally, it is easy to check that the sum of squared standardized distances for the vertical line equals N (and also that no other vertical line leads to a smaller sum—this follows from the derivation of the best constant predictor in Section 19.2.4). Since $\rho_{x,y} \neq 0$ by assumption, it follows that $N(1 - |\rho_{x,y}|) < N$. We therefore conclude that lines that are vertical or close to vertical do not generate smaller sums of squared standardized distances than the neutral line.