APPENDIX 14.3: MORE ON σ_{est}

Chapter 14 provided the following definitions of σ_{est}:

$$\sigma_{est} = \sqrt{\frac{ss_{error}}{N}}$$

and

$$\sigma_{est} = \sqrt{(1-\rho^2)\sigma_y^2}.$$

We will now see why these are equivalent.

In Chapter 13, we stated that the total variability in the distribution of y-scores can be captured by ss_y, which can also be called ss_{total}. The following definitions were introduced in Chapter 13 using the statistics of samples and are rewritten here in terms of population parameters. In a population, we define ss_{total} (or ss_y) as

$$ss_{total} = \sum(y - \mu_y)^2.$$

ss_{total} can also be written as

$$ss_{total} = ss_{regression} + ss_{error}.$$

In the case of populations,

$$ss_{regression} = \sum(E(y \mid x) - \mu_y)^2$$

and

$$ss_{error} = \sum(y - E(y \mid x))^2.$$

The proportion of variability in y explained by regression is given by

$$\rho^2 = \frac{ss_{regression}}{ss_{total}}.$$

The proportion of variability in y not explained by regression is given by

$$1 - \rho^2 = \frac{ss_{error}}{ss_{total}}.$$

From the above equation, we can see that

$$ss_{error} = (1-\rho^2)ss_{total}^2.$$

Because variances are just average sums of squares, we can write

$$\frac{ss_{error}}{N} = (1-\rho^2)\frac{ss_{total}}{N}$$

as

$$\sigma_{error}^2 = (1-\rho^2)\sigma_{total}^2.$$

In regression, σ_{error}^2 is typically referred to as σ_{est}^2, and σ_y^2 is just another name for σ_{total}^2. Therefore

$$\sigma_{est} = \sqrt{\frac{ss_{Error}}{N}} = \sqrt{(1-\rho^2)\sigma_y^2}.$$

Table 14.A3.1 summarizes the quantities we've discussed and the relationships between them.

We can gain some insight into the relationship between the variance of the marginal distribution of y-scores (σ_y^2) and the variance of the conditional distributions (σ_{est}^2) by considering Figure 14.A3.1. Figures

TABLE 14.A3.1 ■ A Summary of the Sources of Variability in a Regression Analysis

Description	Raw-Score Formulas	Formal Symbols	Descriptive Symbols	Parametric Formulas
Variance of the y-scores.	$\sum(y-\mu_y)^2/N$	σ_y^2	σ_{total}^2	$\sigma_{regression}^2 + \sigma_{error}^2$
Variance of the expected values of y: $E(y\|x)$.	$\sum[E(y\|x)-\mu_y]^2/N$	$\sigma_{E(y\|x)}^2$	$\sigma_{regression}^2$	$\rho^2 * \sigma_{total}^2$
Variance of the deviations of the y scores from their expected values: $y - E(y\|x)$.	$\sum[y-E(y\|x)]^2/N$	$\sigma_{y-E(y\|x)}^2$	σ_{est}^2	$(1-\rho^2) * \sigma_{total}^2$

14.A3.1a, 14.A3.1c, and 14.A3.1e show three homoscedastic bivariate distributions. These figures show that all conditional distributions have the same standard deviations ($\sigma^2_{est} = 2.5$), and only the regression lines differ.

Figures 14.A3.1b, 14.A3.1d, and 14.A3.1f show probability density functions rotated 90° from their usual orientation with density on the *x*-axis and the dependent variable on the *y*-axis. This rotation makes it easier to see the connection between both sides of the figure. The conditional distributions are shown as blue lines in Figures 14.A3.1b, 14.A3.1d, and 14.A3.1f. Although all conditional distributions have $\sigma^2_{est} = 2.5$, their means [$E(y|x)$] depend on the slope of the regression line. The steeper the slope, the more widely separated the conditional distributions.

The gray lines in Figures 14.A3.1b, 14.A3.1d, and 14.A3.1f show the marginal distributions of *y*-scores. The standard deviation for each marginal distribution is denoted σ_y, and the variance is denoted σ^2_y. Figure 14.A3.1 shows that the steeper the slope of the regression line, the more variability there is in the marginal distribution of *y*-scores. Figures 14.A3.1b, 14.A3.1d, and 14.A3.1f show σ_y at the top and σ_{est} at the bottom, demonstrating that σ_y increases as the magnitude of β increases. When $\beta = 0$, the marginal distribution and the conditional distributions are identical (see Figure 14.A3.1f). In this case, regression explains none of the variability in the marginal distribution of *y* values because there is no variability in the expected value of *y*; i.e., $E(y|x) = \mu_y$ for all levels of *x*. For example, Figure 14.A3.1e shows that $E(y|x) = 24.5$, so $\mu_y = 24.5$.

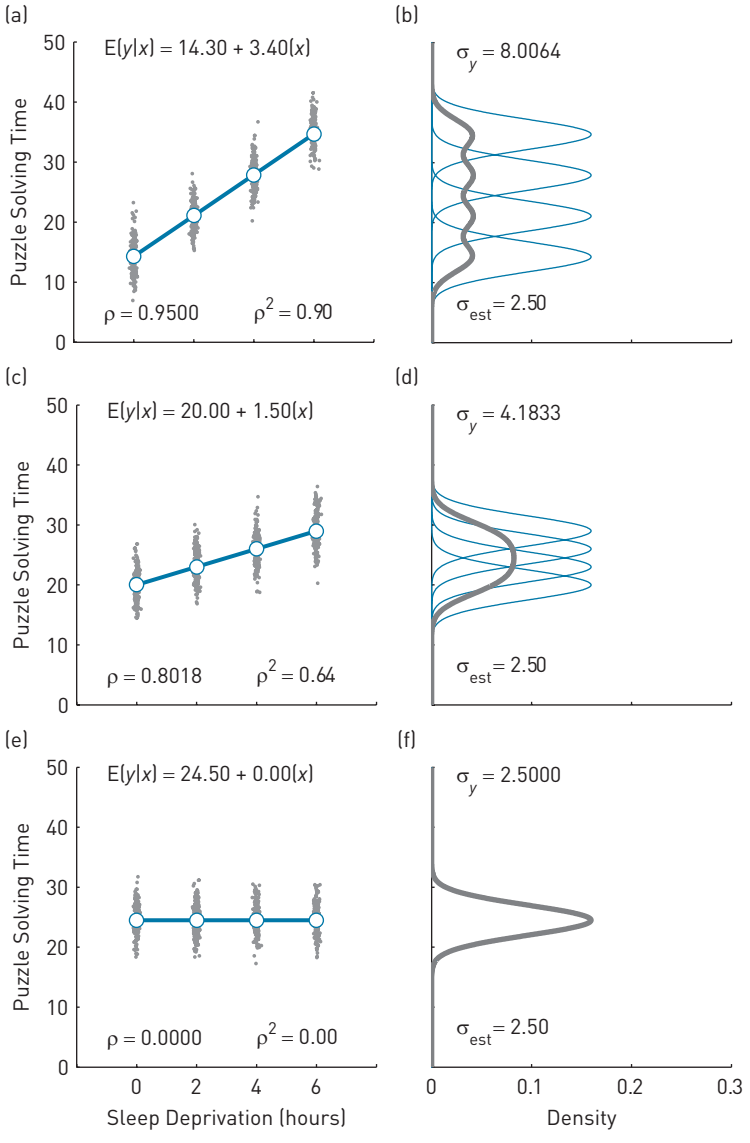
The marginal distribution becomes wider as β increases, while the standard deviations of the conditional distributions (σ_{est}) remain unchanged. In this case, regression explains some of the variability in the marginal distribution of *y* values because the increase in the standard deviation of the marginal distribution is caused by the increase in the variability in the expected values, $E(y|x)$. So, some of the variability in the marginal distribution is related to σ_{est} and some to the variability in the expected values, $E(y|x)$. As β increases, the proportion attributable to $E(y|x)$ increases. The exact proportion of the variability in the marginal distribution arising from variability in the $E(y|x)$ values is ρ^2.

Because β is defined in terms of ρ, the separation between the conditional distributions is also related to ρ^2. Figures 14.A3.1a, 14.A3.1c, and 14.A3.1e show ρ^2. Figures 14.A3.1b, 14.A3.1d, and 14.A3.1f show that as ρ^2 increases (from 0 to .56 to 0.9), the separation between the conditional distributions increases.

If we know ρ^2 and σ^2_y, then

$$\sigma^2_{est} = \left(1 - \rho^2\right)\sigma^2_y.$$

FIGURE 14.A3.1 ■ Conditional and Marginal Distributions



The relationship between the slope of the regression line and the total variability in *y*. (a, c, and e) Three homoscedastic bivariate distributions. In all cases, $\sigma_{est} = 2.5$. (b, d, f). The blue lines show the conditional distributions. The separation between the means of the conditional distributions depends on the slope of the regression line. The steeper the slope, the more widely separated the distributions. The gray lines show the marginal distributions of all *y*-scores. The variances of these distributions are $\sigma^2_y = \sigma^2_{est} / (1 - \rho^2)$.

In addition, if we know ρ^2 and σ^2_{est}, then

$$\sigma^2_y = \frac{\sigma^2_{est}}{1 - \rho^2}.$$