Math 171.02 Spring 2004
Calculating the least-squares linear regression line
March 15, 2004

| $x_i$ | $y_i$ | $\hat{y}_i$ | $e_i$ |
|---|---|---|---|
| 1 | -1 | -0.5 | -0.5 |
| 2 | 3 | 2 | 1 |
| 3 | 4 | 4.5 | -0.5 |

The mean of $x$ is

$$\overline{x} = \frac{x_1 + x_2 + x_3}{3} = \frac{1+2+3}{3} = 2,$$

and the mean of $y$ is

$$\overline{y} = \frac{y_1 + y_2 + y_3}{3} = \frac{-1+3+4}{3} = 2.$$

We can also compute the variance of $x$

$$\text{Var}(x) = s_x^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2}{3-1} = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3-1} = 1,$$

and the variance of $y$

$$\text{Var}(y) = s_y^2 = \frac{(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + (y_3 - \overline{y})^2}{3-1} = \frac{(-1-2)^2 + (3-2)^2 + (4-2)^2}{3-1} = 7.$$

Using the formula on page 88 of Moore gives

$$\begin{aligned} r &= \frac{\left(\frac{x_1-\overline{x}}{s_x}\right)\left(\frac{y_1-\overline{y}}{s_y}\right) + \left(\frac{x_2-\overline{x}}{s_x}\right)\left(\frac{y_2-\overline{y}}{s_y}\right) + \left(\frac{x_3-\overline{x}}{s_x}\right)\left(\frac{y_3-\overline{y}}{s_y}\right)}{3-1} \\ &= \frac{\left(\frac{1-2}{1}\right)\left(\frac{-1-2}{\sqrt{7}}\right) + \left(\frac{2-2}{1}\right)\left(\frac{3-2}{\sqrt{7}}\right) + \left(\frac{3-2}{1}\right)\left(\frac{4-2}{\sqrt{7}}\right)}{3-1} \\ &= \frac{5}{\sqrt{28}}. \end{aligned}$$

Recall that the equation of the regression line is

$$\hat{y} = a + bx$$

where

$$b = r\frac{s_y}{s_x} \quad \text{and} \quad a = \overline{y} - b\overline{x}.$$

Thus,

$$b = \frac{5}{\sqrt{28}}\frac{\sqrt{7}}{1} = 2.5 \quad \text{and} \quad a = 2 - (2.5)(2) = -3$$

so that

$$\hat{y} = 2.5x - 3.$$

Since the residuals = observed − predicted, we have

$$e_i = y_i - \hat{y}_i.$$

In other words, $\hat{y}_i = e_i + y_i$ or $\hat{y} = e + y$. We can therefore calculate $\text{Var}(\hat{y})$, $\text{Var}(y)$, and $\text{Var}(e)$. From above, we have $\text{Var}(y) = 7$. We can also calculate $\overline{e} = 0$ and $\overline{\hat{y}} = 2$. By construction, $\overline{e} = 0$ and $\overline{\hat{y}} = \overline{y}$ always. Thus,

$$\text{Var}(\hat{y}) = \frac{(-0.5-2)^2 + (2-2)^2 + (4.5-2)^2}{3-1} = 6.25,$$

and

$$\text{Var}(e) = \frac{(-0.5-0)^2 + (1-0)^2 + (-0.5-0)^2}{3-1} = 0.75.$$

Note that

$$\text{Var}(y) = \text{Var}(\hat{y}) + \text{Var}(e).$$

This is always true BECAUSE $\hat{y}$ and $e$ are independent random variables.

**Claim:** $r^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)}$

That is, $r^2$ is the fraction of the variability of $y$ that is explained by the linear regression.

Unless the data points are perfectly linear, there will be variability. In fact, most of our survey samples will have variability. The goal of statistics is to *explain the variability.* Linear regression is one way of assessing this. If $|r|$ is near 1, then a linear model is reasonable. If $r^2$ is near one (if and only if $|r|$ is near 1), then most of the variability is accounted for. Notice that this does not tell us the linear regression is reducing variability. Instead it tells us that variability is reduced because of the strong linear fit.

**Proof:** By definition, since $\overline{\hat{y}} = \overline{y}$ we have

$$\frac{\text{Var}(\hat{y})}{\text{Var}(y)} = \frac{\frac{\sum(\hat{y}_i - \overline{\hat{y}})^2}{n-1}}{\frac{\sum(y_i-\overline{y})^2}{n-1}} = \frac{\frac{\sum(\hat{y}_i - \overline{y})^2}{n-1}}{\frac{\sum(y_i-\overline{y})^2}{n-1}}.$$

But, $\hat{y}_i = a + bx_i$ and $a = \overline{y} - b\overline{x}$ so that

$$(\hat{y}_i - \overline{y})^2 = (a + bx_i - \overline{y})^2 = (\overline{y} - b\overline{x} + bx_i - \overline{y})^2 = b^2(x_i - \overline{x})^2.$$

Thus,

$$\frac{\text{Var}(\hat{y})}{\text{Var}(y)} = \frac{\frac{\sum b^2(x_i-\overline{x})^2}{n-1}}{\frac{\sum(y_i-\overline{y})^2}{n-1}} = b^2\frac{s_x^2}{s_y^2}.$$

But $b = r\frac{s_y}{s_x}$ so that $b^2\frac{s_x^2}{s_y^2} = r^2$, or

$$\frac{\text{Var}(\hat{y})}{\text{Var}(y)} = r^2$$

as required.