The best fitting line

In our previous lecture we considered $\widehat{y}_i = 110 + 10x_i$ to predict $y_i =$ calories from $x_i =$ fat. One measure of how well this line fits the data is given by:

$$(y_1 - \widehat{y}_1)^2 + (y_2 - \widehat{y}_2)^2 + (y_3 - \widehat{y}_3)^2 + (y_4 - \widehat{y}_4)^2 + (y_5 - \widehat{y}_5)^2 = \sum_{i=1}^{5}(y_i - \widehat{y}_i)^2,$$

which is called the sum of squared errors, or SSE. Note that the SSE is a function of the slope and intercept that we are using, so for the linear equation $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ we can write that

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2.$$

For our line above for the cereal data we get

$$\begin{aligned} SSE(\beta_0 = 110, \beta_1 = 10) &= (110 - 110)^2 + (110 - 120)^2 + (120 - 120)^2 \\ &\quad + (130 - 130)^2 + (150 - 114)^2 \\ &= 200. \end{aligned}$$

Is this the smallest SSE possible? (Note that other criteria exist besides SSE for choosing a best fitting line) We can use calculus to find which slope and intercept will yield the smallest SSE. The solutions, called the least squares estimators, are given by:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{and } \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

For the cereal data the results are:

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = (-1.4)(-14) + (-.4)(-14) + (-.4)(-4)$$
$$+(.6)(6) + (1.6)(26) = 72.00 \text{ and}$$

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = (-1.4)^2 + (-.4)^2 + (-.4)^2 + (.6)^2 + (1.6)^2 = 5.2, \text{ so then}$$

$$\widehat{\beta}_1 = 72/5.2 = 13.85 \text{ and } \widehat{\beta}_0 = 124 - (13.85)(1.4) = 104.6,$$

so the least-squares line is $\widehat{y}_i = 104.6 + 13.85x_i$. When we calculate the SSE for the least-squares line we find it is equal to 123.1, so it does have smaller SSE than the previous line. The least squares estimates can be calculated in SAS in PROC REG (other SAS procedures can also be used). Look at the SAS code example and output for this lecture. A best fitting line that had SSE = 0 would indicate that the line perfectly fit the data. A large value of SSE can result from either i) much variation in the errors ε_i , or ii) when the true regression model is incorrect.