1 Confidence interval, Prediction interval examples

For our small cereal data set, suppose we want to construct both a confidence interval for $E(y_{n+1})$ and also a prediction interval for $y_{n+1}$ when $x_{n+1}=1$, using $\alpha=.05$. Remember that $\hat{y}_{n+1} = 104.62 + (13.85)(1) = 118.47$, $n = 5$, $\bar{x} = 1.4$, $S_{xx} = 5.2$, $t_{0.025,3} = 3.182$, and that $s^2 = 41.03$. Then we have

$$t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}} = 3.182\sqrt{41.03} \sqrt{\frac{1}{5} + \frac{(1 - 1.4)^2}{5.2}} = 9.79,$$

so that the confidence interval is $118.47 \pm 9.79$ or $(108.68, 128.26)$. Similarly, for the prediction interval

$$t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}} = 3.182\sqrt{41.03} \sqrt{1 + \frac{1}{5} + \frac{(1 - 1.4)^2}{5.2}} = 22.61,$$

so the prediction interval is $118.47 \pm 22.61$ or $(95.86, 141.08)$. What is the interpretation of these intervals?

2 The correlation coefficient (r)

The correlation coefficient, $r$, is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

Note also that $r = (\sqrt{S_{xx}/S_{yy}})\hat{\beta}_1$. Some properties of $r$: 1) $-1 \leq r \leq 1$, 2) it is dimensionless, and 3) has the same sign as $\hat{\beta}_1$. $r$ is an index of linear association, and it estimates the population correlation coefficient $\rho$. A related quantity, $S_{xy}/(n-1)$, is called the covariance of X and Y:

If $\beta_1 = 0$ then our predicted value of $Y_i$ is $\hat{Y}_i = \bar{Y}$, and $S_{yy}/(n - 1)$ is an appropriate estimate of $\sigma^2$. On the other hand, if $\beta_1 \neq 0$ then $SS(\text{Residual})/(n - 2)$ is an appropriate estimate of $\sigma^2$. One way to quantify the reduction in variation in $Y$ due to $X$ is with:

$$r^2 = \frac{SS(\text{Total}) - SS(\text{Residual})}{SS(\text{Total})} = \frac{SS(\text{Regression})}{SS(\text{Total})},$$

which measures the strength of the relationship between $Y$ and $X$. Remember that $r^2$ does not measure 1) the magnitude of the slope of the regression line, or 2) the appropriateness of the simple linear regression model.