1 Confidence interval, Prediction interval examples

For our small cereal data set, suppose we want to construct both a confidence interval for $\mu_{Y|X_0}$ and also a prediction interval for $Y_{X_0}$ when $X_0 = 1$, using $\alpha = .05$. Remember that $\hat{Y}_{X_0} = 104.62 + (13.85)(1) = 118.47$, $n = 5$, $\bar{X} = 1.4$, $S_x = 1.14$, $t_{3.975} = 3.182$, and that $S^2_{Y|X} = 41.03$. Then we have

$$t_{n-2,1-\alpha/2}S_{Y|X} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S^2_X}} = 3.182\sqrt{41.03} \sqrt{\frac{1}{5} + \frac{(1 - 1.4)^2}{(5-1)1.14^2}} = 9.79,$$

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

$$t_{n-2,1-\alpha/2}S_{Y|X} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S^2_X}} = 3.182\sqrt{41.03} \sqrt{1 + \frac{1}{5} + \frac{(1 - 1.4)^2}{(5-1)1.14^2}} = 22.61,$$

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:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{SS_{XY}}{\sqrt{SS_{X} SS_{Y}}},$$

Note also that $r = (S_X/S_Y)\hat{\beta}_1$. 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, $SS_{XY}/(n-1)$, is called the covariance of $X$ and $Y$:

The text has an interesting (optional) discussion of the bivariate normal distribution in this chapter.

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

$$r^2 = \frac{SS_Y - SSE}{SS_Y} = \frac{SSR}{SSY},$$

1
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. Note that the text discusses methods for testing hypotheses and constructing confidence intervals for $\rho$, many of which make use of

$$\text{Fisher’s Z transformation: } \frac{1}{2} \ln \frac{1 + r}{1 - r}.$$