1 Review problem set 2 solutions

Here are solutions to review problem set 2:

11.30 and 11.31) See the SAS computer file on the website for the computer solution. For the data from problem 11.18, 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}=30, using α=.05. Then we have (using log base e of biological recovery) $\widehat{y}_{n+1} = 3.85 - (0.0366)(30) = 2.75$, $n = 13$, $\overline{x} = 30.$, $S_{xx} = 4550.$, $t_{.025,11} = 2.201$, and that $s_\varepsilon = .256$. Then we have

$$t_{\alpha/2}\, s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{S_{xx}}} = (2.201)(.256)\sqrt{\frac{1}{13} + \frac{(30-30)^2}{4550}} = .156,$$

so that the confidence interval is 2.75 $\pm$ .16 or (2.59, 2.91). Similarly, for the prediction interval

$$t_{\alpha/2}\, s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{S_{xx}}} = (2.201)(.256)\sqrt{1 + \frac{1}{13} + \frac{(30-30)^2}{4550}} = .584,$$

so the prediction interval is 2.75 $\pm$ .58 or (2.17, 3.33).

If you used log base 10 instead to transform biological recovery, your prediction is $\widehat{y}_{n+1} = 1.67 - (0.0159)(30) = 1.195$, your confidence interval is (1.127, 1.263) and your prediction interval is (0.941, 1.45).

11.38 a) The data can be plotted using a computer package (see files), and looks fairly linear.

b) The prediction equation is $\widehat{y}_i = 3.37 + 4.07x_i$.

c) The residual-by-predicted value plot does not follow a horizontal band about 0, so additional terms may be needed in the model.

11.43 a) $r^2_{yx} = 1168280/1172398 = .997$. This equals the value in the output.

b) The sign of the slope agrees with the correlation coefficient, so $r_{yx} = +\sqrt{r^2_{yx}} = .998$.

c) If the range of x or y is restricted then the value of r_{yx} will decrease, see the discussion in section 11.6.

11.76 a) If the taxpayers' group is correct, we expect an inverse relationship between expenditure and town population.

b) The output disagrees with the opinion of the group.

11.77) Yes, the regression line from 11.76 is misleading due to the influential point that is far different than the rest of the data.

11.78 a) The data point was forcing the least-squares solution to meet it, completely changing the pattern that exists for the rest of the data.

b) The rest of the data (after excluding the influential point) appear to agree with the taxpayers' group opinion.

11.79) a) The new prediction equation is $\widehat{y}_i = 180.38 - 1.35x_i$.

b) Without the influential point, $\widehat{y}_i = 180.38 - 1.35(37) = 130.4$. With all of the data, $\widehat{y}_i = 117.61 + .658(37) = 142.0$.

c) The one data point was able to completely change the fitted model due to its position in the covariate space.

For the final problem with the cereal data, we have:

$y_1 = \beta_0 + \varepsilon_1,\ y_2 = \beta_0 + \varepsilon_2,\ y_3 = \beta_0 + \varepsilon_3,\ y_4 = \beta_0 + \varepsilon_4,\ y_5 = \beta_0 + \varepsilon_5$, where x_{i1}, for example, is the ith observation's value of variable x_1. This can be rewritten as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [\beta_0] + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix},$$

so

$$\mathbf{X'X} = [1\ 1\ 1\ 1\ 1] \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = [5]\,,\ \mathbf{X'Y} = [1\ 1\ 1\ 1\ 1] \begin{bmatrix} 110 \\ 110 \\ 150 \\ 130 \\ 120 \end{bmatrix} = [620]\,, \text{ and}$$

$$(\mathbf{X'X})^{-1} = [5]^{-1} = [1/5] \text{ so } \widehat{\beta} = \widehat{\beta}_0 = (\mathbf{X'X})^{-1}\mathbf{X'Y} = (1/5)(620) = 620/5 = 124$$