Solutions to sheet 2

(2.1) In each case, the least squares line is $y = 3 + 0.5x$ and the squared correlation is 0.667. The predicted value at x=10 in each case is $3 + 0.5 \times 10 = 8$. To me, only the first pattern seems reasonably well summarized as "a linear part + residuals with no obvious pattern". For the second case, the quadratic curve in the scatterplot is reflected in the residuals. In the third case, it would be better to describe the pattern as a linear fit for most of the data, with one clear aberration. For the fourth plot, it is clear that one point is having a large effect on the fit. Anscombe's comment:

> If all observations are considered genuine and reliable, data set 4 is just as informative about the regression relation as data set 1; there is no reason to prefer either to the other. Yet in most circumstances we should feel that there was something unsatisfactory about data set 4. All the information about the slope of the regression line resides in one observation—if that observation were deleted the slope could not be estimated. In most circumstances we are not quite sure that every observation is reliable. If any one observation were discredited and therefore deleted from data set 1, the remainder would tell much the same story. That is not so for data set 4. Thus the standard regression calculation ought to be accompanied by a warning that one observation has played a critical value.

I would not be so generous with the interpretation even if the single point were totally reliable. The linear fit does not have much to do with most of the data. There is no real evidence to support the idea that the response is linear plus random noise. I would have no faith in the predicted value at $x = 10$.

The linear prediction in case 2 would also be clearly suspect. In case 3, I would have more confidence in a linear prediction based on the "linear part of the data", provided I could understand why the lone point was out by itself.

In these four cases the residual plots essentially just tell you what you could already see from the scatter plot. The general idea is that, if "linear fit + random noise" is a reasonable description of the data then the residual plot should have no obvious patterns—just a bunch of angry mosquitoes buzzing about. This sort of plot will be more useful when we have several explanatory variables going into the fit. In that case it is often not so easy to think of a plot that will reveal subtle departures from the "linear fit + random noise" description.

My main purpose in setting the exercise was to warn you that perusal of the r^2 is not enough to tell whether a linear fit is sensible. All four data sets have the same r^2, but their stories are completely different.

(2.2) Means and standard deviations for the final and classwork scores, after excluding the perfect case:

```
Variable         N     Mean    StDev
classwork       39    510.8     90.3
final           39    51.00    15.14
```

Define final1 = final - mean(final) and final2 = final1/stdev(final). Calculate classwork 1 and classwork2 similarly. The five regressions represented by the five lines in the Notes are given by:

```
final = 5.4 + 0.0893 classwork + residuals
final1 = - 45.6 + 0.0893 classwork + residuals
final2 = - 3.01 + 0.00590 classwork + residuals
final2 = 0.000 + 0.00590 classwork1 + residuals
final2 = 0.000 + 0.532 classwork2 + residuals
```

Write y_i for the final score, and x_i for the classwork score, of the ith student, $i = 1, \ldots, 39$. The five steps correspond to the five minimization problems: choose constants $(a_1, b_1), \ldots, (a_5, b_5)$ to minimize

$$\sum_i (y_i - a_1 - b_1 x_i)^2$$

$$\sum_i (y_i - \overline{y} - a_2 - b_2 x_i)^2 = \sum_i (y_i - (\overline{y} + a_2) - b_2 x_i)^2$$

$$\sum_i \left(\frac{y_i - \overline{y}}{s_y} - a_3 - b_3 x_i \right)^2 = \frac{1}{s_y^2} \sum_i \left(y_i - \overline{y} - s_y a_3 - s_y b_3 x_i \right)^2$$

$$\sum_i \left(\frac{y_i - \overline{y}}{s_y} - a_4 - b_4 (x_i - \overline{x}) \right)^2 = \sum_i \left(\frac{y_i - \overline{y}}{s_y} - (a_4 - b_4 \overline{x}) - b_4 x_i \right)^2$$

$$\sum_i \left(\frac{y_i - \overline{y}}{s_y} - a_5 - b_5 \frac{(x_i - \overline{x})}{s_x} \right)^2 = \sum_i \left(\frac{y_i - \overline{y}}{s_y} - a_5 - \frac{b_5}{s_x} (x_i - \overline{x}) \right)^2$$

Compare the right-hand side of each line with the left-hand side of the previous line to see why

$$\overline{y} + a_2 = a_1 = 5.4 \qquad \text{and} \qquad b_2 = b_1 = 0.0893$$

$$s_y a_3 = a_2 = -45.6 \qquad \text{and} \qquad s_y b_3 = b_2 = 0.0893$$

$$a_4 - b_4 \overline{x} = a_3 = -3.01 \qquad \text{and} \qquad b_4 = b_3 = 0.00590$$

$$a_5 = a_4 = 0 \qquad \text{and} \qquad \frac{b_5}{s_x} = b_4 \qquad \text{that is, } b_5 = 0.532$$

The equalities follow because we are minimizing the sum of squares at each step. You should check that the coefficients do satisfy these relationships.