MTMG37 Example Solution to Class Exercise 7

1. The correlation is 0.82 (to two decimal places) for all four pairs, which suggests a strong, positive association in each case. The plots in Figure 1 show that the relationships are very different. The correlation is only meaningful for the first dataset, which has a roughly linear association. The second dataset has a strong, non-linear relationship that is not reflected in the correlation because it is a measure of only *linear* association. The third dataset appears to have an outlier; apart from that the relationship seems perfectly linear, but again this is not revealed by the correlation. The fourth dataset has only two distinct x-values, which makes it difficult to conclude anything about the relationship. Always plot the data if you are interested in the relationship between variables!

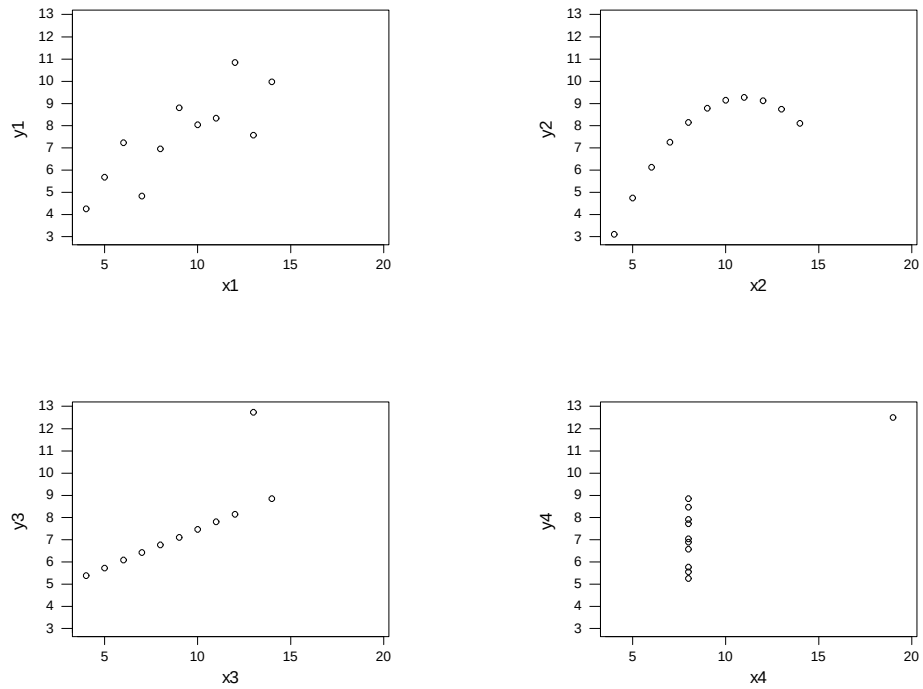

Figure 1. Scatter plots for four datasets.

2. The simple linear regression model is $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where Y_i is the response variable, x_i is the explanatory variable, and ε_i are independent $N(0,\sigma^2)$ random variables. For all four datasets the least-squares parameter estimates are $\hat{\beta}_0 = 3.0$ and $\hat{\beta}_1 = 0.5$.

3. The coefficients of determination are all $R^2 = 67\%$, and the p-values for testing $H_0 : \beta_1 = 0$ against a general alternative are all $p = 0.002$. These statistics indicate that the value of the x-variable is informative about the value of the y-variable in each case. In particular, the x-variables explain about two-thirds of the variation in the y-variables.

4. Diagnostic plots reveal that the fitted model is acceptable for the first dataset only. The plots of residuals against fitted values in Figure 2 reveals the non-linear structure that is missed by the model for the second dataset, and highlights the influential outliers in the third and fourth datasets. Histograms and probability plots (not shown) reveal that the assumption of Normal residuals is also untenable for datasets 2 and 3.

   Only the first fitted linear model is useful; for the other datasets, the models will give misleading results. The second dataset needs a non-linear model, the linear model for the third dataset should be fitted in such a way that the outlier is accounted for, and the fourth dataset is not appropriate for modelling the bivariate relationship.

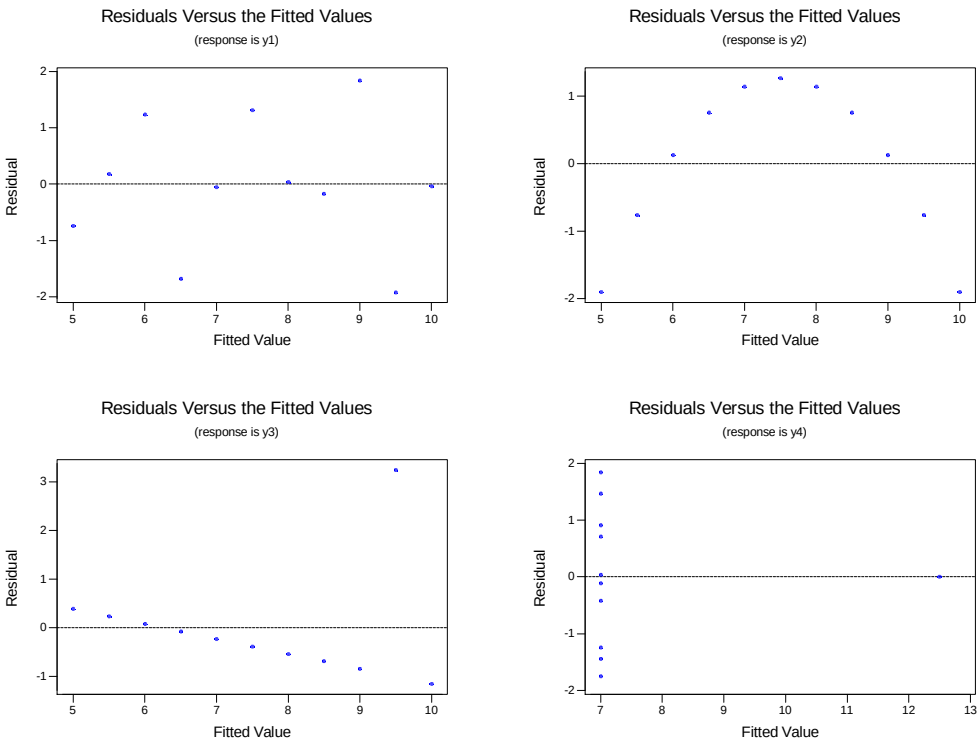

Figure 2. Diagnostic plots for the four linear regressions.