MULTIPLE REGRESSION

To describe multiple regression, we begin with simple linear regression. When dealing with a numeric explanatory variable, X, and a numeric response variable, Y, collected as (x,y) pairs, the first step is to produce a scatterplot and find their correlation. If the correlation is significant, then the shape on the scatterplot will have a linear trend and a simple linear regression line may be fit to the data. For example, if X is hours of study and Y is score on a test, then the regression model might be

$$score = 35 + 9hours$$

The equation is the old high school $y = mx + b$ slope-intercept form, where the slope is $m = 9\,\frac{points}{hour}$ and the intercept is 35 points. This would mean that those who studied zero hours for the test would score a 35, on average, and every hour of study would earn an average of 9 more points.

The only difference with multiple regression is that more than one explanatory variable is used. For example, if we also had the previous test score recorded, we could add it to the model. The result might look like:

$$score = -2 + 1.2hours + 0.9prev\_test$$

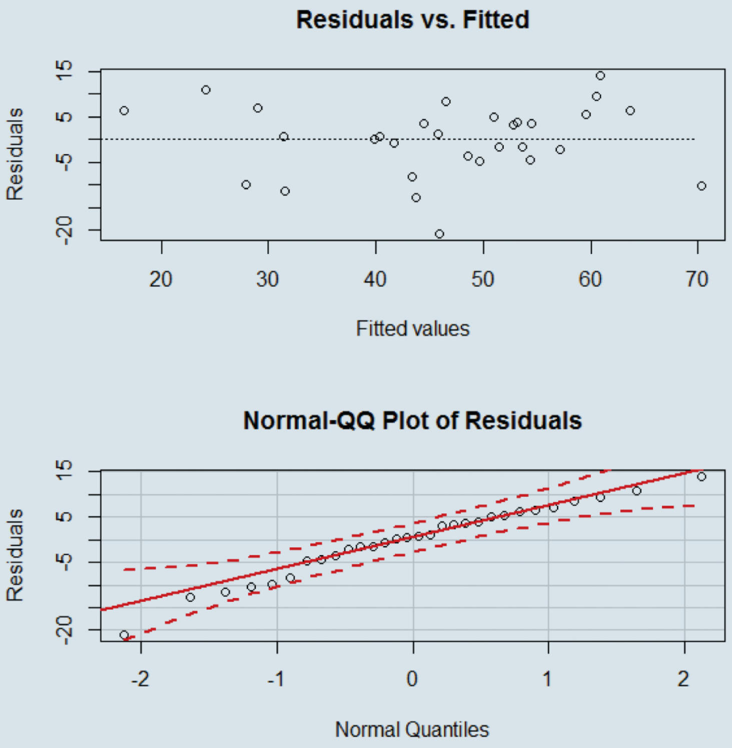

Figure 4. The residuals vs. fitted plot shows that the linearity and constant variance assumptions are reasonable since the points are random. The Normal Q-Q plot shows that distribution of the sample residuals is close to what we would expect with independent normal errors since the points fall within the 95% confidence bands.

In this case, students with a previous test score of 80 and five hours of study could be expected to score -2 + 1.2(5) + 0.9(80) = 76 on the test. If the intercept of -2 was not statistically significant, it could be removed, forming a "no intercept" model.

The three main assumptions for a multiple regression to be valid are that (i) the observations are independent, (ii) the differences between the response variable and the values predicted by the model are approximately normally distributed, and (iii) the differences have a common variance.

The model in Equation 1 is consistent with these assumptions. Assumption (i) is satisfied because the pitches were conducted under the same circumstances and there is no measurable effect of the order of the pitches. The correlation between pitch order (1,2,...,10) and Rating for the three pitchers, along with the *p*-value for the correlation test of the null hypothesis of zero correlation are $corr_A = -0.29$ ($p\text{-value}_A = 0.41$); $corr_B = 0.41$ ($p\text{-value}_B = 0.24$); and $corr_C = 0.53$ ($p\text{-value}_C = 0.11$). The differences between the response variable and the predicted values are called residuals.

Assumption (ii) is seen to be reasonable because the standardized residuals lie near the 45 degree line of the Normal Q-Q plot in Figure 4, indicating they "line up" with what is expected if they were normal (Shapiro-Wilks test of normality, $W = 0.97$; p-value = 0.55).

Assumption (iii) is reasonable because the plot of the residuals versus the fitted values in Figure 4 is random, indicating the variance does not depend on the level of the rating (non-constant error variance test $X^2 = 0.03$; p-value = 0.86).