STA 302 / 1001 (A. Gibbs)
Sketch of Solutions to Exercises in Chapter 5 of Sheather

1. The scatterplot with points coded differently for residential and commercial accounts indicates that the relationship between number of days the payment is overdue and the amount of the overdue bill is different for residential accounts than it is for commercial accounts, but that both relationships appear to be linear.

   So we should fit the model:

   $$\text{LATE} = \beta_0 + \beta_1\text{BILL} + \beta_2 I_{[\text{RESIDENTIAL}]} + \beta_3 I_{[\text{RESIDENTIAL}]} * \text{BILL} + e$$

   where LATE and BILL are defined in the textbook, and $I_{[\text{RESIDENTIAL}]}$ is 1 if the account is residential and 0 if it is commercial.

   The fitted model is:

   $$\widehat{\text{LATE}} = 101.8 - 0.191\ \text{BILL} - 99.5 I_{[\text{RESIDENTIAL}]} + 0.357 I_{[\text{RESIDENTIAL}]} * \text{BILL}$$

   From the t-tests, we see that all of the coefficients are statistically significantly different from 0, given that the other terms are in the model.

   For commercial accounts, the fitted model is

   $$\widehat{\text{LATE}} = 101.8 - 0.191\,\text{BILL}.$$

   On average, the fitted model estimates that for each additional 10 dollars in the amount of the bill, the number of days it is overdue decreases by almost 2.

   For residential accounts, the fitted model is

   $$\widehat{\text{LATE}} = 2.21 + 0.166\,\text{BILL}.$$

   On average, the fitted model estimates that for each additional 10 dollars in the amount of the bill, the number of days it is overdue increases by about 1.7.

   Evaluating the model:
   - The plot of the standardized residuals versus the predicated values shows two points with unusually large, negative residuals. These belong to two residential accounts and should be checked for errors.
   - The plot shows no serious departures from constant variance and no indication of curvature.
   - The normal quantile shows no serious departures from normality, particularly if the two outliers noted above are ignored.
   - Looking at the influence statistics, we see that the outliers (observations 26 and 42) are also influential (Cook's distances of 0.161 and 0.146 as well as large values of DFFITS and DFBETAS), as is observation 82 (Cook's distance of 0.126). These points should be investigated further with the person who collected the data to understand why they are unusual before any further work is done.

2. New variables were created: `i2004` which is 1 if the year is 2004 and 0 if 1994, and `LowIncome_i2004` which is the product of `i2004` and `LowIncome`, the percentage of low-income students.

The analysis of covariance model

$$Y = \beta_0 + \beta_1\texttt{LowIncome} + \beta_1\texttt{i2004} + \beta_2\texttt{LowIncome\_i2004} + e$$

was first fit to the data. From this model, there is no evidence that the relationship between percentage of students repeating first grade and the percentage of low-income students differs between the two years (p-value for the interaction term is 0.6307).

A second model was fit without the interaction term. From this model there is no evidence of a difference in the percentage of students repeating first grade between the two years after accounting for the affect of the percentage of low-income students ($p = 0.5993$). There is strong evidence of a linear relationship between the percentage of students repeating first grade and the percentage of low-income students ($p = 0.0002$). The slope of this linear relationship is $0.07248 > 0$, so an increase in the percentage of low income students is associated with an increase in the percentage of students repeating first grade.

3. (a) From the SAS output, the test with null hypothesis that the coefficient of the interaction term is 0 (given the other terms are in the model) has p-value 0.0120, so we conclude that the rate of change of quality rating depends on whether or not there has been unwanted rain at harvest.

   (b) i. If there has been no unwanted rain at harvest, the estimated relationship between quality and the number of days since August 31 for the harvest is

   $$\widehat{\text{Quality}} = 5.16122 - 0.03145\,\text{EndofHarvest}$$

   which can be used to estimate that a delay to the end of harvest of about 32 days results in a decrease of 1 point in the quality rating when there has been no unwanted rain at harvest.

   ii. If there has been some unwanted rain at harvest, the estimated relationship between quality and the number of days since August 31 for the harvest is

   $$\widehat{\text{Quality}} = 6.94792 - 0.11459\,\text{EndofHarvest}$$

   which can be used to estimate that a delay to the end of harvest of about 9 days results in a decrease of 1 point in the quality rating when there has been some unwanted rain at harvest.