# Regression Analysis

## Assignment 3A

**Problem 4.1 on page 184:** This exercise uses the NYJUICE dataset and covers polynomial transformations (see pages 144-147).

Problem 4.1: This problem is adapted from one in McClave et al. (2005). The **NYJUICE** data file contains data on demand for cases of 96-ounce containers of chilled orange juice over 40 sale days at a warehouse in New York City that has been experiencing a large number of out-of-stock situations. To better understand demand for this product, the company wishes to model demand, *Cases*, as a function of sale day, *Day*.

a. Construct a scatterplot for these data, and add a quadratic regression line to your scatterplot [computer help #15 and #32].
b. Fit a simple linear regression model to these data, that is, use statistical software to fit the model with *Day* as the single predictor [computer help #25] and write out the resulting regression equation.
c. Fit a quadratic model to these data, that is, use statistical software to fit the model with both *Day* and $Day^2$ as predictors and write out the resulting regression equation. You will first need to create the $Day^2$ term in the dataset [computer help #6 and #31].

   *Hint: To square a variable, X, in statistical software you may need to write it as "X**2" or "X^2".*

d. Based on the scatterplot in part (a), it appears that a quadratic model would better explain variation in orange juice demand than a simple linear regression model. To show this formally, do a hypothesis test to assess whether the $Day^2$ term is statistically significant in the quadratic model (at a 5% significance level). If it is, the more complicated quadratic model is justified. If not, the simpler simple linear regression model would be preferred.

   *Hint: Understanding the example on page 145 will help you solve this part.*

[8 points]

**Problem 4.3 on pages 184-185:** This exercise uses the INTERNET dataset and covers transformations for the response and predictors (pages 155-158). Compare the models for this problem by considering a hypothesis test for the slope and by visually assessing a scatterplot of the response variable versus the predictor variable with the least squares line added, particularly in terms of the regression assumptions of Section 3.4. The example in Section 4.1.4 on pages 151-154 uses this approach. You can also informally compare the models using the coefficient of determination (R-squared). However, you should not compare them using the regression standard error (s), which is based on the same units of measurement as the response variable, since the response variable is different in each model.

Problem 4.3: Recall Problem 2.1 from page 78 in which you fit a simple linear regression model to data for 212 countries with response *Int* (pecentage of the population that are Internet users) and predictor *Gdp* (GDP per capita in US$ thousands). This model is a reasonable one, but it is possible to improve it by transforming both the response and predictor variables. Investigate the use of transformations for this application using the data in the **INTERNET** data file. In particular, investigate natural logarithm and square root transformations for both the response and predictor variables. Which transformations seem to provide the most useful model for

understanding any association between *Int* and *Gdp*? Write up your results in a short report (no more than two pages) that compares and contrasts the different models that you fit. Include a few paragraphs describing your conclusions with respect to the various ways of comparing models and perhaps some scatterplots.

*Hint: Compare the following three models: (1) response Int and predictor Gdp; (2) response* $\log_e(Int)$ *and predictor* $\log_e(Gdp)$*; (3) response* $\sqrt{Int}$ *and predictor* $\sqrt{Gdp}$*. Use the methods from the example in Section 4.1.4 to guide your comparison.*

[18 points]