**HOMEWORK 7**

*Multiple Regression: Multicollinearity, Quadratic Relationships, Outliers, Partial F tests and Quantifying Predictive Accuracy*

Reading: This assignment focuses on content from your textbook, *STAT2: Building Models for a World of Data*, Sections 3.4, 3.5, 3.6, 4.3 and 4.4. Read these sections of your textbook.

Notes:

* Round all numbers to **3** decimal places unless otherwise specified.

1. 3.41 – Fluorescence experiment (quadratic). The study that generated these data is described in question 1.38 from Chapter 1. The data are in Fluorescence.jmp. Note that this data set is not identical to what is provided by the book.
   1. Fit a quadratic regression to predict Y = ProteinProp from the Calcium values. Make sure to turn off JMP’s Center Polynomials option. Report the

Intercept: 8.222

Linear coefficient: 1.498

Quadratic coefficient: 0.0700

* 1. Test for quadratic relationship, i.e. test whether the quadratic coefficient = 0. Report the p-value. If JMP reports < 0.0001, enter 0.0001.

p < .0001

* 1. Based on the test in the previous question, is it appropriate to use linear model (no squared term) to predict ProteinProb? (True = Yes, False = No)

No, False

1. 3.56 – Major League Baseball winning percentage.

The data in MLB2007standings.jmp is information about each major league baseball team in 2007. One team is omitted.

Fit a model to predict winning percentages (WinPct) using team batting average (BattingAvg), number of runs scored (Runs), number of triple hits (Triples), number of runs batted in (RBI), and number of games saved by the team’s pitchers (Saves).

* 1. For each of the variables below, indicate True if there are multicollinearity concerns for that variable and False if there are no concerns.

BattingAvg False

Runs True

Triples False

RBI True

Saves False

* 1. Using the fitted multiple regression from the previous question, test whether the regression coefficient for Runs = 0 AND the regression coefficient for RBI = 0 (i.e. simultaneously). Report the p-value.
     1. .04498
  2. Here some potential conclusions from results from the previous question and the fitted multiple regression. Indicate which are True (i.e. appropriate) and which are False (not appropriate). For simplicity, ignore the difference between no effect (coefficient = 0) and no evidence of an effect.

The regression coefficients for both Runs and RBI = 0 False

At least one of the two variables, Runs and RBI, has a non-zero regression coefficient True

Both regression coefficients (for Runs and RBI) are not 0. False

* 1. Calculate and report the PRESS RMSE statistic. This is the equivalent of rMSE, but based on the prediction error sum of squares.
     1. .03302

* 1. You are interested in how well the model predicts WinPct for the omitted team. Your choice is to report the rMSE or the Press rMSE.

Which is more appropriate?

Press rMSE. Bc Out of sample prediction

1. The National Center of Education Statistics conducts surveys and assessments with students.[[1]](#footnote-1) Results from a random sample of 200 high school seniors are provided in the hsb2.jmp data file. We will try to predict “math” which is a standardized math score. The explanatory variables are read (a standardized reading score) and prog (program student enrolled in after high school; vocational, general, or academic). Use general as the baseline/reference group.

Start by creating indicator variables for prog and then use Fit Model to fit a model using read and prog to predict math.

* 1. Report the estimated coefficients for:

Intercept 24.562

Read .5117

Vocational (this means the indicator variable you create with vocational = 1 and others =0) -1.783

Academic (this means the indicator variable you create with academic = 1 and others =0) 3.433

* 1. You want to test whether there is any difference between the three types of programs, after adjusting for effects of read (reading score). Conduct a partial F test (custom test in JMP) to test the null hypothesis of no differences among programs. Report:

The numerator df for this test 2

The p-value for this test .000145

1. 4.10 – Breakfast cereals. (page 190)

The data in cereal.jmp are nutrition information on 36 breakfast cereals. Fit a multiple linear regression to use Sugar and Fiber to predict Calories. You are interested in whether there are any outliers or influential points.

* 1. True/False: There are no observations that would be considered an unusually large positive or unusually large negative outlier False

* 1. Which row number is the most extreme positive outlier? 27
  2. True/False: There are no observations that would be considered unusually influential observations

True

* 1. Which row number is the most influential observation in this data set? 27

1. https://www.rdocumentation.org/packages/openintro/versions/1.7.1/topics/hsb2 [↑](#footnote-ref-1)