Homework 2
MSAN 601

due Monday, September 12, 2016 at 11.45pm

n.b. All deliverables are required to be typed and all graphs and statistical output generated in `R`. Deliverables with *any* handwritten elements will not be accepted and will receive a grade of zero. You are required to use either `R`Markdown or LaTeXto generate the `pdf` deliverable to be uploaded to Canvas. If using `R`Markdown, also upload the `Rmd` file; if using LaTeX, also upload the `R` file. Show **all** calculations unless otherwise instructed. Also, be sure to upload the two additional `.R` files for the final question.

Question 1

1. Using `SCENIC_data.csv`, regress Length of Stay (Y) on Average Daily Census (X). Report the R^2, b_0 and b_1 values.
2. Multiply the observations, both X and Y, by 192; we will refer to these as $X_{(2)}$ and $Y_{(2)}$. Regress $Y_{(2)}$ on $X_{(2)}$. Report the R^2, b_0 and b_1 values.
3. Multiply only Y by 47; we will refer to this as $Y_{(3)}$. Regress $Y_{(3)}$ on X. Report the R^2, b_0 and b_1 values.
4. Multiply only X by 12; we will refer to this as $X_{(3)}$. Regress Y on $X_{(3)}$. Report the R^2, b_0 and b_1 values.

Succinctly explain what you have gleaned from this exercise. Include a summarized tabular representation of the regression output and the associated R^2, b_0 and b_1 values.

Question 2

1. Using `platsicHardness.txt` data from Canvas. The first column of data represents plastic hardness in Brinell units (Y) and the second column data represents the number of hours elapsed since the plastic was molded (X).
2. Run a SLR model along with all relevant diagnostics—do not include numerical/graphical output, just a basic summary, written or tabular, of what you found. Succinctly discuss what you observe. Are there any issues with the model?
3. Run a Brown-Forsythe test to determine whether or not the error variances varies with the level of X. Divide the data into two groups, $X \leq 24$ and $X > 24$, and use $\alpha = 0.05$. State the decision rule and the conclusion. Discuss this result in context of what you discovered earlier.

Question 3

Read the articles entitled **False Psychology** and the **ASA Statement on Statistical Significance and** ***p*****-values**. You are not required to submit anything for these readings, but the material is testable in a quiz and final exam setting. Both articles can be found on Slack or on Canvas.

Question 4

`muscleMass.txt` contains data collected by a researcher that randomly selected 15 women from each ten-year agree groups to study the relationship between decreasing muscle mass and aging in women. The first column of data represents a measure of muscle mass (Y) and the second column of data represents age (X).

1. Explore the relationship by regressing muscle mass on age using a SLR model. Succinctly discuss what you observe. Are there any issues with the model?
2. Run a Breush-Pagan test to determine whether or not the error variances varies with the level of X, using $\alpha = 0.01$. State the alternatives, decision rule and the conclusion. Discuss this result in context of what you discovered earlier.

Question 5

In `SCENIC_data.csv`, LoS (Y) is regressed on Infection Rate (X). Both a confidence interval and a prediction interval were subsequently obtained for $X_h = 11.93$. R generated the following output for a 97% the confidence interval

```
       fit      lwr      upr
1 15.40861 13.47345 17.34377
```

and the following for a 97% prediction interval

```
       fit      lwr      upr
1 15.40861 11.34767 19.46955
```

Formulaically/mathematically explain the difference between the upper limit of the confidence interval, 17.34377, and the upper limit of the prediction interval, 19.46955. Stated alternatively, what term or terms (and their associated numerical values) specifically account for the difference?

$$19.46955 - 17.34377 = 2.12578$$

Question 6

In `R`, code two different functions which execute a Box-Cox (BC) procedure to find an optimal power transformation for Y.

1. Create a file called `linearBC.R` containing a function which executes a BC procedure that linearly searches for a minimal SSE value. Search through λ values from -3 to 3 in increments of 0.1. Generate **all**possible SSEs from -3 to 3 and *then* identify the minimal value.
2. Create a file called `bisectionBC.R` containing a function which executes a BC procedure using a bisection search to identify a minimal SSE value. Begin the bisection search with upper and lower bounds of -3 and 3 respectively. Be certain to explicitly state your stopping criteria (termination conditions), i.e., what is your tolerance to break out of your search.

Both functions should read in a `csv` from the current directory called `hw2_q5_data.csv`. The `csv` will contain two numeric columns: the first represents X and the second Y. Once the search in complete, the function should print the value of λ that minimizes SSE to the console. Use your own data sources to test your model. Both files will be `source()`d using test data.