STA 101 Spring 2018

Project II - Due Friday, June 1^{st} **by the end of lecture**

Read the following instructions carefully:

- **You may work by yourself, or in a team of two people total.**
- **You are not allowed to discuss the questions with anyone other than the instructor or TA.**
- **Any outside help beyond that from the instructor or TA is considered plagiarism. This including asking a tutor, your classmates (no comparing answers), posting the questions to homework help sites, etc.**
- **You are allowed to use or modify your previous code, or the instructors code that is posted online.**
- **Formatting will be a significant portion of your grade for this take home portion. There should be an appendix of code, and no code in the body of the report.**
- **Your report should be in full paragraph form.** You are allowed to have tables, and/or use R Markdown, but it should have clearly labeled sections.

The Groups

From each group of problem, your group will choose **one** of the following datasets to explore:

Group I

Choose one and only one of the following:

Problem 1

We will be using the dataset online called `prostate.csv`. The rows contain information from patients who are being assessed for prostate cancer. The variables included are:

- `cancer`: Indicator of prostate cancer diagnosis (1) or no cancer diagnosis (0)
- `psa`: Serum prostate-specific antigen level (mg/ml)
- `c.vol` : Estimate of prostate cancer volume (cc)
- `weight`: Prostate weight (gm)
- `age`: Age of patient (years)
- `benign`: Amount of benign prostatic hyperplasia (cm^2)
- `inv`: Presence (”invasion”) or absence (”no-invasion”) of seminal vesicle invasion.
- `cap`: Degree of capsular penatration (cm)

The goal of this problem is to build a model to predict whether someone will be diagnosed with prostate cancer or not. In this problem, you should perform the following steps:

Problem 2

We will be using the dataset online called `student.csv`, which contains the following columns:

Column 1: `gender`: With values `female` and `male`.

Column 2: `schtype`: School type, with values `public`, `private`

Column 3: `read`: The reading score of a standardized test.

Column 4: `write`: The writing score of a standardized test.

Column 5: `math`: The math score of a standardized test.

Column 6: `science`: The science score of a standardized test.

Column 7: `honors`: Whether or not the student is enrolled in an honors class, with values `no`,`yes`.

Column 8: `ses`: Socioeconomic classs **Your response variable for this problem**, with values `low`, `medium`, `high`.

The Goal

The goal for either of these dataset follows:

- Specify the model framework you will be considering (ANOVA, linear regression, logistic regression, etc..).
- Look for outliers, and remove them if there are any.
- Perform model selection (i.e, decide which variables are most important for your prediction).
- Select a final and "best" model (this may be done via. hypothesis tests, or stepwise selection).
- Report some measures of how well your model is doing.
- Interpret the coefficients of your model.
- Report confidence intervals for parameters of interest (when possible).
- Report error matrices back, as well as overall error rate, and sensitivity and specificity when appropriate.
- If you choose Problem 1: **Based on your "best" model, i.e, you may not use all of these values**, predict the probability of prostate cancer diagnosis for someone with 10 `psa`, 5 `c.vol`, 40g for `weight`, `age` 67, with 2.5 `benign`, with no seminal vesicle invasion, and with 0.5 cm `cap`.
  If you choose Problem 2: **Based on your "best" model, i.e, you may not use all of these values**, predict what socioecnomic class you believe a female who went to private school, and who got 50 on all standardized tests, and who was not an honor student is.

Group II

Choose one and only one of the following:

Problem 1

We will be using the dataset online called `moon.csv`. The rows contain information on the average number of patients admitted per month, and what cycle the moon was in.

- `Admittance`: The average admittance of patients per day in one month
- `Moon`: Whether it was `Before` a full moon, `After` a full moon, or `During` a full moon.

Problem 2

We will be using the dataset online called `cows.csv`, which contains the following columns:

Column 1: `Weight`: The weight gain of a 1 year old cow after 1 year.

Column 2: `Grass`: The type of grass the cows were fed (A, B, C).

The Goal

The goal for either of these dataset follows:

- Specify the model framework you will be considering (ANOVA, linear regression, logistic regression, etc..).
- Look for outliers, and remove them if there are any.
- Interpret the coefficients of your model.
- Report confidence intervals for parameters of interest (when possible).
- Describe how the average changes by group (if at all).
- Predict the value of y for all groups of X.

3. The Report Format

Each question should be a short report. This means you write in full sentences, and have the following sections for each question, while being **as specific as you can** about your results:

I: Summary. This should include summary plots of describing the relationship between your explanatory and response variable, and any numerical summaries you find interesting.

II: Data Preparation. This section should include finding and removing any outliers in preparation for a model. How many outliers you found (if any) should be noted, and the rows with those outliers in them should be shown.

III: Model fitting. This section should include the results of your model fitting, including which model selection criteria you used, what your final model was, any confidence intervals or hypothesis tests you will interpret in a later section, and the estimated regression line. If you have a categorical variable, consider writing down the separate models it suggests. For this section, it is your choice whether or not to include interaction terms. Or, you may first determine which single terms are important, then see if any interactions to do with those terms are also important.

IV: Model Diagnostics: Perform diagnostics to see if the assumptions of linear regression hold. If you do not think they do, state this, but continue with the report.

V: Interpretation: Interpret the coefficients and any confidence intervals or p-values that you calculated.

VI: Prediction: Predict y, and report back measures of prediction.

VII: Conclusion: One or two sentences on what variables you found were most important to your model, and how they affected your outcome.

4. Details

Your report should be the following format:

i. Typed.

ii. A title page including your name/s, the name of the class, and the name of your instructor (me).

iii. Treat each question as a small, stand alone report. Then staple them together at the end.

iv. Double-sided pages.

v. An appendix of your R code used to produce the results. Do not include in R code in the body of your report.

For example, your project should be put together in the following order (stapled):

```
Cover Page
Parts I-V for first question
Parts I-V for second question
Code appendix
```

Feel free to make your cover page "unique" so that it is easy to find when I hand them back.

Notice: your project will be graded as a group effort (if you have two people). This means that you are responsible for your own work, and your partners work. I will not assign two different grades to one project.