STA 138 Exam II Project, due
**Friday, March $8^{th}$ in lecture**

Read the following instructions carefully:

- **You may work in a group of two, or by yourself.**
- **You are not allowed to discuss the questions with anyone other than the instructor or TA and your group mate.**
- **Any outside help beyond that from the instructor or TA is considered plagiarism. This including asking a tutor, your classmates (for example, comparing answers), posting the questions to homework help sites, etc. Should we believe you have sought outside help, you will be reported to the Student Judicial Affairs office.**
- **You are allowed to use or modify your previous functions, or the instructors functions that are posted online.**
- **Do not share answers, or specific values for calculations, particularly on Piazza**.
- You may ask clarifying questions about code and general approach on Piazza, but do not give away any numerical answers. If you are concerned you may be giving something away, email me or the TA's directly.

## Multiple Logistic Regression

Choose one and only one of the following:

## Problem I

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:

- `y`: 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 where $Y$ = prostate cancer status.

## Problem II

We will be using the dataset online called `angina.csv`. This study was conducted in order to assess the relationship between various predictor variables, and if the subject had angina (a condition where heart muscles receive insufficient oxygen-rich blood). The data follows:

Column 1: `y`: 1 if angina present, 0 if absent

Column 2: `age`: The age of the subject in years

Column 3: `smoke`: With values current, ex, or never (indicating smoking history).

Column 4: `cig`: The average number of cigarettes smoked per day

Column 5: `hyper`: History of hypertension in the family, with values absent, mild, moderate.

Column 6: `myofam`: History of myocardial infarction in the family, with values yes, no.

Column 7: `strokefam`: History of stroke in the family, with values yes, no.

Column 8: `diabetes`: History of diabetes in the family, with values yes, no.

The goal of this problem is to build a model where $Y$ = angina status.

## The Report Format

This should be a 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. Consider outliers and influential points. Note: You may fit a model first, then consider outliers and influential points.

III: Model Selection/Analysis. Perform model selection. You may choose either correctness or prediction as your goal, but if you choose prediction you must use a model selection technique (you can not default to the largest model).
This section should include the results of your "best" model fit, 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 logistic regression function.
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: Interpretation: Interpret the coefficients and any confidence intervals or p-values that you calculated.

V: Prediction: Predict $\hat{\pi}$, and report back measures of prediction. Consider model goodness-of-fit measures, error matrices, etc.

- 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 the probability of angina for a 50 year old who has never smoked, with history of hypertension, angina, and stroke (and no other history of medical issues).

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

## 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. Double-sided pages.

iv. 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-VI
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.