Math 150 - Methods in Biostatistics - Homework 6

your name here

Due: Wednesday, March 6, 2019, in class

```
knitr::opts_chunk$set(message=FALSE, warning=FALSE, fig.height=3, fig.width=5,
                      fig.align = "center")
library(tidyverse)
library(broom)
library(tidylog)
```

Note: there are two places to check for hints on R code. One is the class notes (http://st47s.com/Math150/Notes/, see R Examples) and the other is the R manual associated with the textbook which is on Sakai.

1. Chp 7, E9 no (d), Donner Party

(a) Create a logistic regression model using `Gender` and `Age` to estimate the probability of survival. Create a plot of the estimated probability of survival using `Age` as the explanatory variable and grouping the data by `Gender`. Use the plot and the model to interpret the coefficients in terms of the odds ratios.

```
donner <- read_csv("~/Dropbox/teaching/math150/PracStatCD/Data Sets/Chapter 07/CSV Files/C7 Donner.csv"
                   na="*")
names(donner) <- c("name", "gender", "age", "survived", "famsize", "X6", "X7", "X8", "X9", "X10",
                   "X11", "X12", "X13", "X14")
```

(b) Create and interpret a logistic regression model using `Gender`, `Age`, and `Gender*Age` to estimate the probability of survival. Create a plot of survival. Create a plot of the estimated probability of survival using `Age` as the explanatory variable and grouping the data by `Gender`.

(c) Explain any key differences between the plots created in parts (a) and (b). Discuss how adding the interaction term `Gender*Age` impacts the model.

2. Chp 7, E10 Variable Selection Techniques and Multicollinearity

(a) Crate a logistic regression model using `Radius`, `Concave`, and `Radius*Radius`, and `Radius*Concave` as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. [Note that you need to create the `Radius*Radius` variable before running the `glm`.]

```
cancer <- read_csv("~/Dropbox/teaching/math150/PracStatCD/Data Sets/Chapter 07/CSV Files/C7 Cancer2.csv
                   na="*")
cancer <- cancer %>%
  mutate(Radius2 = Radius*Radius)
```

```
## mutate: new variable 'Radius2' with 456 unique values and 0% NA
```

```
library(rms)  # rms to do drop in deviance tests, see the class notes
```

(b) Even though in part (a) Wald's test shows the highest p-value for `Radius`, it is typically best to attempt to keep the simplest terms in the model. Generally, keeping simpler terms in the model makes the model easier to interpret. Thus, we suggest as a first attempt keeping `Radius` in the model and eliminating the variable with the next highest p-value. Create a logistic regression model using `Radius`, `Concave`, and `Radius*Concave` as explanatory variables to estimate the probability that a mass is malignant.

Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance test to determine if `Radius*Radius` should be included in the model.

(c) Use a scatterplot to compare `Radius` to `Radius*Radius` and calculate the correlation between these two terms. Are the two variables highly correlated?

(d) Chapter 3 discusses **multicolinearity** (highly correlated explanatory variables). Explain whether you believe `Radius` is important in the logistic regression model. Why is the p-value for `Radius` so large in part (a) but very small in part (b)?

(e) Create a logistic regression model using `Radius` and `Concave` as explanatory variables to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance test to determine if `Radius*Concave` should be included in the model.

(f) Create a logistic regression model using only `Concave` as an explanatory variable to estimate the probability that a mass is malignant. Submit the logistic regression model and the likelihood ratio test results, including the log-likelihood (or deviance) values. Conduct the drop-in-deviance to test to determine if `Radius` should be included in the model.

(g) Submit a final model and provide a justification for choosing that model.