STA 101 Spring 2017
Homework 4 - Due Friday, May 17^{th}

Note: You do not have to use R Markdown to turn in the homework, but the homework must be turned in in a reasonable format. The answers to the questions should be in the body of the homework, and the code used to obtain those answers should be in an appendix. There should be no code in the body of the homework. You can accomplish this in R, Word, LaTex, Google Docs, etc.

1. Online you will find the file `alcohol.csv`. This dataset has the following columns:

   Column 1: `BrAC`: alcohol measured by breath in mg/L/hour (X)

   Column 2: `BAC`: blood alcohol content in g/L/hour (Y)

   The goal was to assess if a breath test was closely related to the results of a blood test given at the same time.
   Data Source: M. Pavlic, P. Grubweiser, K. Libisiller, Walter Rabl (2007). "Elimination Rates of Breath Alcohol," Forensic Science International

   (a) Find the best transformation for Y using the Box-Cox method. What was your value of λ? Plot the scatter plot of the transformed Y vs. X.

   (b) Test for equal variance using the linear model with the transformed Y from (a). What would you conclude at $\alpha = 0.05$?

   (c) Transform both X and Y using the Box-Cox method. Plot the scatter plot of the transformed X and Y.

   (d) Conduct a Shapiro-Wilks test and Brown-Forsythe test on the linear model which uses the transformed variables from (c). What are your conclusions for each?

   (e) Which transformation (Box-Cox for only Y, or for both X and Y) would you suggest, if any, and why? Justify your answer.

2. Online you will find a dataset `flu.csv`. The columns we are interested in are `shot` (1 indicates flu shot, 0 indicates no flu shot), age (the age of the patient), aware (an awareness score of health for the patient, where higher is more aware), and `gender` (the age of the subject, M or F).

   (a) Fit the logistic regression model and write down the estimated logistic-regression function.

   (b) Using (a), estimate the probability that a 55 year old female with an awareness score of 70 will get the flu shot.

   (c) Based on the sign of β_1, as your age increases does the probability of a flu shot go up or down?

   (d) Interpret the value of $exp(\hat{\beta}_1)$ in terms of the problem.

   (e) Find the 90% confidence interval for $exp(\hat{\beta}_1)$, and interpret it in terms of the problem.

   (f) Interpret the value of $exp(\hat{\beta}_3)$ in terms of the problem.

   (g) Find the 90% confidence interval for $exp(\hat{\beta}_3)$, and interpret it in terms of the problem.

3. Continue with question 2.

   (a) Does $\hat{\beta}_0$ have a practical meaning in this case? Explain.

   (b) Would it make sense to predict the probability of a flu shot for someone aged 12 based on this data set? Explain.

   (c) Find the error matrix using $\pi_0 = 0.50$, and calculate

      i. The sensitivity.
      ii. The specificity.
      iii. The overall error rate.

   (d) Find the AUC and the corresponding 95% confidence interval. Does this suggest that our model predicts Y well? Explain.

4. Online you will find a dataset `CHD.csv`. The columns we are interested in are `CHD` (1 indicates coronary heart disease (CHD), 0 indicates no CHD), and `age` (the age of the subject). These data come from Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression: Third Edition.

   (a) Fit the logistic regression model and write down the estimated logistic-regression function.

   (b) Using (a), estimate the probability that a 69 year old will have CHD.

   (c) Based on the sign of β_1, as your age increases does the probability of CHD go up or down?

   (d) Interpret the value of $exp(\hat{\beta}_1)$.

   (e) Find the 99% confidence interval for $exp(\hat{\beta}_1)$, and interpret it.

5. Continue with question 4.

   (a) Plot the logistic curve and display it. Does there seem to be a large effect of age on the probability of CHD?

   (b) Does $\hat{\beta}_0$ have a practical meaning in this case? Explain.

   (c) Would it make sense to predict the probability of having CHD for someone aged 44 based on this data set? Explain.

   (d) Find the error matrix, and calculate

      i. The sensitivity.
      ii. The specificity.
      iii. The overall error rate.

   (e) Find the AUC and the corresponding 95% confidence interval. Does this suggest that our model predicts Y well? Explain.

6. Answer the following with TRUE or FALSE. It is good practice to explain your answer.

(a) If the confidence interval for $\exp(\beta_1)$ contains 1, it suggests no influence of X_1 on the odds of the trait $(Y = 1)$.

(b) $\exp(\beta_1)$ gives the estimate for the odds of the trait $(Y = 1)$.

(c) β_0 always has a practical interpretation.

(d) If we fail to reject $H_0 : \beta_1 = 0$, we conclude that X_1 does not effect the probability of the trait.