STA 101 Spring 2018
Homework 5 - Due Wednesday, May 23^{rd}

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 a dataset `flu.csv`, which has the

   Column 1: shot (Y): If the subject got a flu shot ($y = 1$), or not ($y = 0$)

   Column 2: `age` (X_1): The age of the subject in years.

   Column 3: aware (X_2): The health awareness score, where a higher score indicates a higher level of awareness.

   Column 4: `gender` (X_3): `M` or `F`

   (a) Use forward step-wise selection. What X's were included in the final model?

   (b) Use backward step-wise selection. What X's were included in the final model?

   (c) Using the model from (a), write down the logistic regression model, and find the log-likelihood for that model.

   (d) Add an interaction term to the model from (a), and find the log-likelihood for that model.

   (e) State the null and alternative, and calculate the test-statistic and p-value for testing if the interaction term can be dropped from the model.

   (f) State your conclusion about (e) if $\alpha = 0.05$.

2. Continue with problem 1, using the model implied by 1(f).

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

   (b) Interpret the value of $exp(\hat{\beta}_2)$ in terms of the problem.

   (c) Predict if someone who was male, age 57, and had a health awareness score of 50 would get the flu or not.

   (d) Plot a histogram of Pearson's residuals. Are there any values above 4? If so, show the corresponding x and y values (the row/s).

   (e) Plot an index plot of $DFbeta$. Are there any values about 0.3? If so, show the corresponding x and y values (the row/s).

3. Online you will find a dataset `control.csv`, which has the

   Column 1: con (Y): The type of birth control, with `Short` (short term), `Long` (long term), or `None` (no birth control) the subject was currently on.

   Column 2: age (X_1): The age of the subject in months

   Column 3: `edu` (X_2): The level of education of the subject, with `A` (advanced), `G` (graduate or above), `M` (high school), `L` (below highschool).

   Column 4: `working` (X_3): `N` (they are not working) or `Y` (they are working).

   The purpose of the study was to examine contraceptive use in married women.

   (a) Use forward step-wise selection. What X's were included in the final model?

   (b) Use backward step-wise selection. What X's were included in the final model?

   (c) Using the model selected in (a), write down the two multinomial models.

   (d) Using the model selected in (a), and for the model that compares None to Long, interpret the value of $exp(\hat{\beta}_1)$.

   (e) Using the model selected in (a), and for the model that compares Short to Long, interpret the value of $\exp(\hat{\beta}_4)$.

4. Continue with problem 3, and using the model selected in 3(a).

   (a) For the model that compares None to Long, interpret the value of $exp(\hat{\beta}_2 - \hat{\beta}_4)$.

   (b) For the model that compares Short to Long, interpret the value of $\exp(\hat{\beta}_3 - \hat{\beta}_4)$.

   (c) State the null, alternative, calculate the likelihood ratio test-statistic, and state the p-value for testing if X_2 can be dropped from the model.

   (d) What is your conclusion for (c), assuming $\alpha = 0.10$?

   (e) Predict the probabilities for each category for a mother who is working, has a graduate education or above, and is 29 years old.

5. Continue with problem 3, and using **only the variable/s selected in 3(a).**

   (a) Create a dataset that combines only the `Long` and the `Short` categories, and find the logistic regression model. Write it down.

   (b) Plot a histogram of the standardized residuals. Are any observations larger than 3, or less than -3?

   (c) Plot an index plot of the change in the Pearson's test-statistic. Are any observations larger than 8? If so, show the corresponding x and y values (the row/s).

   (d) Based on the above, do you believe there are observations that should be removed? Explain.

6. Answer the following with TRUE or FALSE. To obtain partial credit on your answer should it be incorrect, explain your answer.

   (a) An influential point is always an outlier.

   (b) A multinomial model will have the same estimate β for all sub-models.

(c) In logistic regression, if we reject the null for a Likelihood ratio test, we conclude the smaller model fits better.

(d) A standardized residual of 0.50 would be unusual.

7. Online you will find a dataset `rat.csv`, which has the

   Column 1: `Weight`: The weight change in rats in grams (after 4 weeks).

   Column 2: `Amount`: The amount of food they were given (High, or Low).

   Column 3: `Type`: The type of food they were given (Beef, Pork, Cereal).

   This data describes the weight gain of rats (in grams), and the amount and type of diet they were fed. For this problem, we will be considering only Type as the explanatory variable (One Way ANOVA).

   (a) Find the mean, standard deviation, and group size for the weight of rats by the type of food they were given.

   (b) Plot the average weight by group, and a boxplot by group. Does there appear to be a difference in the means?

   (c) State the null and alternative for testing if the group means are equal.

   (d) Calculate the test-statistic and the p-value for the hypothesis in (c).

   (e) State your conclusion in terms of the problem, if $\alpha = 0.10$.

8. Continue with Problem 1.

   (a) Plot a qq plot of the residuals. Does the data appear to be approximately normal?

   (b) Find the p-value for the Shapiro-Wilks test, and state the conclusion of the hypothesis test, assuming $\alpha = 0.05$.

   (c) Plot the fitted values vs. the residuals. Does there appear to be constant variance by group?

   (d) Use the Modified-Levene test to test for constant variance. State the p-value and the conclusion of the hypothesis test, assuming $\alpha = 0.05$.

   (e) Does it appear that the assumptions of ANOVA are met in this problem? Explain.

9. Continue with Problem 1. Now, consider both Type and Amount as explanatory variables.

   (a) Create an interaction plot. Which group combination seems to have the highest average weight gain? Does there appear to be an interaction effect?

   (b) State the null and alternative for testing if there are interaction terms.

   (c) Find the test-statistic and p-value for the hypothesis in (a).

   (d) State the conclusion in terms of the problem, assuming $\alpha = 0.10$.

   (e) Find the average weight gain by groups.