STAT 512 Section 2 Assignment #5

See Canvas calendar for due date.
Reading: Ott & Longnecker section 12.8
44 points total, **4 points** per question unless otherwise noted.

1. For this problem use the data described in Ott and Longnecker Example 12.22 (p 664 in the 7th edition). The data are available from Canvas as "CKheart.csv". Read the description of the data in the book. You can use the output in the book to check your own R calculations.
   A. Use `glm()` to fit a logistic regression model that estimates the probability of a heart attack as a function of CK value. Include the Coefficients table in your assignment.
   B. Construct a plot of the data with the fitted logistic regression curve overlaid. Include the plot in your assignment.
   C. Give an estimate of the odds ratio corresponding to CK and an approximate 95% confidence interval.
   D. Give a one-sentence description of the odds of heart attack among those with a given level of CK, compared to the odds of a heart attack among those with a level of CK ten points higher. (**4 pts**)
   E. Calculate McFadden's pseudo R^2 for the model.
   F. Give an estimate of the CK level at which doctors would be 90% sure that a subject has had a heart attack.

2. An observational study was done to investigate risk factors associated with low infant birth weight. Data from 189 (singleton) pregnancies were collected at Baystate Medical Center, Springfield, MA during 1986. The response variable was **low** (1 if birth weight was less than 2.5 kg, 0 otherwise). The predictor variables included: **age** (mother's age in years), **mwt** (mother's weight in pounds prior to pregnancy), **race** (mother's race, 1= white, 2=black, 3=other) and **smoke** (1=mother smoked during pregnancy, 0 otherwise). The data is available from Canvas as "birthweight.csv". **Important note:** Be sure to define race and smoke as factors!
   A. To examine the relationship between **low vs race**: calculate the proportion of births resulting in low birthweight for each race category and present the p-value from a chi-square test. (**4 pts**)
   B. To examine the relationship between **low vs smoke**: calculate the proportion of births resulting in low birthweight for each smoke category and present the p-value from a chi-square test. (**4 pts**)
   C. Run a logistic regression with **smoke** as the only predictor variable. Calculate the **emmeans using type = "response"** for each smoke group (copy/paste the results to your assignment). Note: these should match your simple proportions from part B. (**4 pts**)
   D. Now consider all 4 predictors (age, mwt, race, smoke). Using best subsets selection with AIC criteria, which variables are included in the final model? Include the Coefficients table and Type3 Anova table in your assignment. (**4 pts**)

   NOTE: Use the selected model from the previous question for all further questions!
   E. Based on the model selected above, give the estimated odds ratio and corresponding 95% CI for Smokers vs Non-Smokers (smoke 1 vs 0).
   F. Calculate the **emmeans using type = "response"** for each **smoke group** (copy/paste the results to your assignment). Note that these values are different from what you found in part C because of the additional variables included in the model.
   G. Give the p-value corresponding to the Hosmer-Lemeshow test. Use `hoslem.test()` from the ResourceSelection package with g = 10 groups. Based on this test, is there evidence of lack of fit?