STAT 308 – Take Home Final

This take home exam is worth a total of 50 points. For the problems in which calculations are needed, please include your `R` code with your answers. Turn in this exam to Sakai by Saturday, December 17 at 4:15 pm.

1. The World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications.

The dataset `heartdisease.csv` contains information on 3656 complete observations from a study of residents of Framingham, Massachusetts. Information on the variables can be found at https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression

a. Create a model predicting probability of 10 year risk of heart disease with only age and currentSmoker and cigsPerDay as explanatory variables.

(i). [3 pts] Interpret the parameter associated with age in the context of the problem.

(ii). [3 pts] Interpret the parameter associated with currentSmoker in the context of the problem.

b. Perform a formal hypothesis test to determine if the logistic regression model for predicting for 10 year risk of heart disease with age, currentSmoker, and cigsPerDay is statistically significant by doing the following tasks.

(i). [1 pt] State the appropriate null and alternative hypothesis.

(ii). [1 pt] State the appropriate test statistic and p-value.

(iii). [1 pt] What is the distribution of the test statistic under H_0?

(iv). [2 pts] Make a decision and conclusion in the context of the problem (assume $\alpha = 0.05$).

c. Perform a formal hypothesis test to determine if adding currentSmoker to a logistic regression model that already includes age and cigarettes per day is statistically significant.

(i). [1 pt] State the appropriate null and alternative hypothesis.

(ii). [1 pt] State the appropriate test statistic and p-value.

(iii). [1 pt] What is the distribution of the test statistic under H_0?

(iv). [2 pts] Make a decision and conclusion in the context of the problem (assume $\alpha = 0.05$).

(v). [1 pt] Intuitively, why does your decision from c(iv) make sense?

d. [5 pts] Determine what is the "best'' logistic regression model using backwards selection and AIC as your selection criterion. Write out the final regression model as a function of the logit of the probability of 10 year risk of heart disease.

e. [2 pts] What is the log-likelihood of the "best' ' model based on AIC backwards selection.

2. The following summary is from a linear regression on work-life balance as predicted from the average number of hours worked per week for MBA alumnus.

$$\sum_{i=1}^{n} x_i = 771, \sum_{i=1}^{n} y_i = 781.2, \sum_{i=1}^{n} x_i^2 = 40481, \sum_{i=1}^{n} y_i^2 = 44320.37, \sum_{i=1}^{n} x_i y_i = 39586.4, n = 15.$$

a. [3 pts] Write the least squares regression line.

b. [2 pts] What is the predicted work-life balance for an alumnus who works 50 hours per week?

c. [3 pts] What is the mean squared error?

d. [2 pts] What are the model sums of squares?

e. [4 pts] Calculate an appropriate F-statistic and p-value for testing if a simple linear model is statistically significant.

3. The Suggested Retail Price of an automobile in 2004 (in thousands of dollars) was regressed on Horsepower, Weight, Highway MPG and the number of Cylinders in the car. An incomplete ANOVA table is below.

	df	Sum Sq
Horsepower	1	42.79
Weight	1	0.95
HighwayMPG	1	1.06
Cylinders	1	0.034
Residuals	229	0.061

a. [2 pts] What is the mean squared error for the full model with all four explanatory variables included?

b. [2 pts] What are the sums of squares added to the model when we add weight, highway mpg, and number of cylinders to a model that already includes horsepower?

c. [3 pts] What are the mean squares for the model with Horsepower and Weight as explanatory variables?

d. [5 pts] Perform a formal hypothesis test to determine if adding weight to a linear model that already includes horsepower significantly improves the predictive ability of sales price.