STAT 308 – Take Home Exam 2

This take home exam is worth a total of 50 points, and counts for half of your Exam 1 grade, along with the in class portion of the exam. For the problems in which calculations are needed, please include your `R` code with your answers. Turn in this exam to Sakai by Thursday, November 17 at 2:30 pm.

1. [3 pts each] For each of the following problems with a sample size of n, calculate an appropriate test statistic and p-value for a linear model with the given categorical variable with p possible categories as the lone explanatory variable in the model. Assume there are no other explanatory variables in the model.

a. $n = 50$, $p = 4$, $SSM = 150.32$, $SSE = 230.44$.

b. $n = 15$, $p = 3$, $SSM = 77.3$, $SSE = 125.44$.

c. $n = 60$, $p = 6$, $SSM = 50.32$, $SSE = 250.78$.

d. $n = 75$, $p = 5$, $SSM = 3.22$, $SSE = 9.87$.

e. $n = 28$, $p = 2$, $SSM = 60.44$, $SSE = 514.32$.

2. For this problem, consider the `mtcars` dataset available in base `R`. Use the command "?mtcars" to obtain information about the overall dataset, as well as the variables contained within it.

Suppose we would like to be able to predict a car's fuel efficiency in miles per gallon based on the car's horsepower.

a. [3 pts] Perform a formal hypothesis test to determine if horsepower is linear related to the car's fuel efficiency. Be sure to include all pieces of information needed to perform a hypothesis test. Assume $\alpha = 0.05$.

b. [2 pts] Check if the assumptions of homoscedasticity and normally distributed are violated. If they are not met, suggest how we can adjust our model so that these assumptions are not violated.

c. Suppose that, after discussion with fellow researchers, we decide to add a quadratic term for horsepower to our model. Perform a formal hypothesis test by answering the following questions.

- i. [2 pts] Write the appropriate null and alternative hypotheses.
- ii. [1 pt] What are the sums of squares that are added to the model when we add the quadratic term?
- iii. [1 pt] What are the error sums of squares for the polynomial model that includes both a linear and quadratic term for horsepower?
- iv. [2 pts] What is the F statistic and p-value?
- v. [2 pts] State your decision and conclusion in the context of the problem.

3. Consider the `lifeexp` dataset on the course webpage which contains information on countries from the year 2015. The variables contained in the dataset are defined as follows:

- `Status`: Development status of the country
- `Life.Expectancy`: Country's average life expectancy in years
- `Adult.Mortality`: Number of deaths of people between 15 and 60 per 1000 population
- `infant.deaths`: Number of infant deaths per 1000 population
- `Measles`: Number of measles cases per 1000 population
- `BMI`: average BMI for entire population
- `under.five.deaths`: Number of Under five deaths per 1000 population
- `Polio`: Polio immunization coverage for 1 year olds
- `Diptheria`: Diphtheria tetanus toxoid and pertussis immunization coverage for 1 year olds
- `HIV.AIDS`: Deaths per 1000 live births from HIV/AIDS (0-4 year olds)
- `Income.composition.of.resources`: Human Development Index in terms of income composition of resources (from 0 to 1)
- `Schooling`: Number of Years of Schooling

We are interested in creating a regression model that predicts a country's life expectancy based on the other factors in the dataset.

a. Suppose we know that both adult mortality and income composition of resources are significant in a linear model that predicts life expectancy. We would like to know if adding development status signficantly improves the predictive ability of our model.

- i. [2 pts] Report the least squares regression lines for predicting life expectancy for both developed and developing countries with adult mortality and income composition of resources included.
- ii. [1 pt] What is the expected difference in life expectancy between developed and developing countries for fixed levels of adult mortality and income composition of resources?
- iii. [4 pts] Perform a formal hypothesis test to determine if adding development status to our linear model significantly improves our model's predictive ability for life expectancy. Be sure to include all pieces of information needed to perform a hypothesis test. Assume $\alpha = 0.05$.
- iv. [5 pts] Perform a formal hypothesis test to determine if EITHER the linear relationship between adult mortality and life expectancy OR the linear relationship between income composition of resources and life expectancy is different for the two different development statuses. Be sure to include all pieces of information needed to perform a hypothesis test. Assume $\alpha = 0.05$.

b. We now would like to determine the "best" overall prediction model for life expectancy using all of the other variables in our dataset as predictors. Assume no polynomial or interaction terms are significant in our model.

- i. [5 pts] Use backward selection with AIC as your decision criteria to eliminate variables from your model. Be sure to state which variables were eliminated at each step.
- ii. [3 pts] Write your final regression model for predicting life expectancy.
- iii. [2 pts] Determine if the assumptions of homoscedasticity and normally distributed residuals for the final regression model are violated.