Stat 6021: Midterm 1

Pledge:
"On my honor, I pledge that I have neither given nor received help on this assignment."

Sign: HD McSpadden

Instructions:

1. For all questions, you may use a calculator, or you may use R as a calculator.
2. For all questions, you may use R to find p-values, critical values, and multipliers. Please indicate what functions you used to find these values, if needed.
3. You may refer to the textbook or materials from the class.
4. State all conclusions in context.
5. Show intermediate steps in calculations.
6. You are not to use the internet, other than using the textbook and class materials.
7. The number in parentheses at the end of each question indicates the point value of the question.

Name & Date: Helen Diana McSpadden 9/20/2020

1. (Use R) Load the "road" dataset from the MASS package in R. The data set contains the annual deaths in road accidents for a number of US states. The variables are:

   - deaths: number of deaths in road accidents
   - drivers: number of drivers (in 10,000s)
   - popden: population density in people per square mile
   - rural: length of rural roads, in 1000s of miles
   - temp: average daily maximum temperature in January
   - fuel: fuel consumption in 10,000,000 US gallons per year

   A client wishes to explore the relationship between number of deaths (Y) and the number of drivers (X). Fit an appropriate linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where ϵ_i are i.i.d. $N(0, \sigma^2)$. For the rest of the question, assume the regression assumptions are met.

   (a) Report the estimated regression equation. What is the interpretation of the estimated slope, in context? (5)

   (b) Is the number of deaths in road accidents linearly related to the number of drivers? Report the values of the test statistic and p-value from the corresponding hypothesis test from the output (do not need to show calculation), and state a conclusion in context at 0.05 significance level. (4)

   (c) Use R to derive the 95% confidence interval for the population slope of the regression line. What does the interval tell us regarding the linear association between the number of deaths in road accidents and the number of drivers? (4)

   (d) Based on your answers in the previous two parts, briefly explain whether or not your conclusions are consistent with each other. (3)

   (e) Be sure to upload your R script on Collab. (2)

2. (Do not use R) Data are collected regarding the eruption times and time between eruptions for the Old Faithful Geyser, for 272 eruptions. The variables of interest are *eruptions*, the duration of the current eruption in minutes, and *waiting*, the waiting time to the next eruption in minutes. The following simple linear regression model is fitted:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

   where ϵ_i are i.i.d. $N(0, \sigma^2)$. You may assume that the assumptions for a linear regression are met. The output for this question consists of some printout from R.

   (a) What is the estimated variance of the true errors, ϵ_i, for this model? Why is this value also the estimated variance of the distribution of Y for given values of X? (4)

(b) Based on the output, construct the corresponding ANOVA table for this model. Be sure to show all relevant calculations. (6)

Source of Variation	df	SS	MS	F
Regression				
Error				*********
Total			*********	*********

(c) A researcher believes that the expected waiting time for the next eruption increases by more than 10 minutes per minute increase in the current eruption. Carry out a corresponding hypothesis test to test this belief. Be sure to state your null and alternative hypotheses, calculate your test statistic, and state your conclusion in context. (8)

(d) The 95% confidence interval for the mean time to next eruption for current eruptions lasting 4.5 minutes is (80.8134, 82.7022) minutes. Compute the corresponding 95% prediction interval for the time to next eruption if the next eruption lasts 4.5 minutes. (6)

3. (Do not use R) An experiment is conducted to study the relationship between the speed of a car, in miles per hour (mph), and its stopping distance, in feet (ft). It is proposed to study this relationip using a simple linear regression model, $Y = \beta_0 + \beta_1 X + \epsilon$. Using least squares, the estimators of the regression coefficients are

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}. \end{aligned}$$

The (observed) residuals are defined as

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

The output for this question consists of some printout from R: scatterplot of Y against X, the residual plot, and a plot of the profile log-likelihoods for the parameter, λ, of the Box-Cox power transformation.

One property of the residuals is that

$$\sum_{i=1}^{n} e_i = 0. \tag{1}$$

(a) Based only on property (1), what should we expect to see in a residual plot (plot of residuals against fitted values)? (2)

(b) Based only on property (1), what should we expect to see in a scatterplot of Y against X? (2)

(c) Based on the output, comment on whether, and how (if necessary), you would transform the response variable or the predictor, or both. Be sure to address what the scatterplot, the residual plot, and the plot of the profile log-likelihoods of the Box-Cox transformation are informing you regarding the need to transform any of the variables. (4)