Stat 6021: Exam

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

Sign: _____

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. If you have any questions about this midterm, please post your questions on the discussion forum. I will not answer questions sent by email, since my answers have to be available to all students.
8. The number in parentheses at the end of each question indicates the point value of the question.

Name & Date: _____

1. 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, $Y = \beta_0 + \beta_1 X + \epsilon$. The output for this question consists of some output from R: scatterplot of Y against X (Figure 1), the residual plot, and a plot of the profile log-likelihoods for the parameter, λ, of the Box-Cox power transformation (Figure 2).

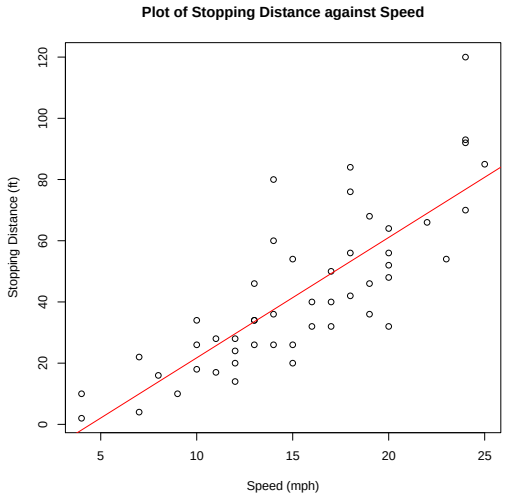

Figure 1: Scatterplot of Stopping Distance against Speed

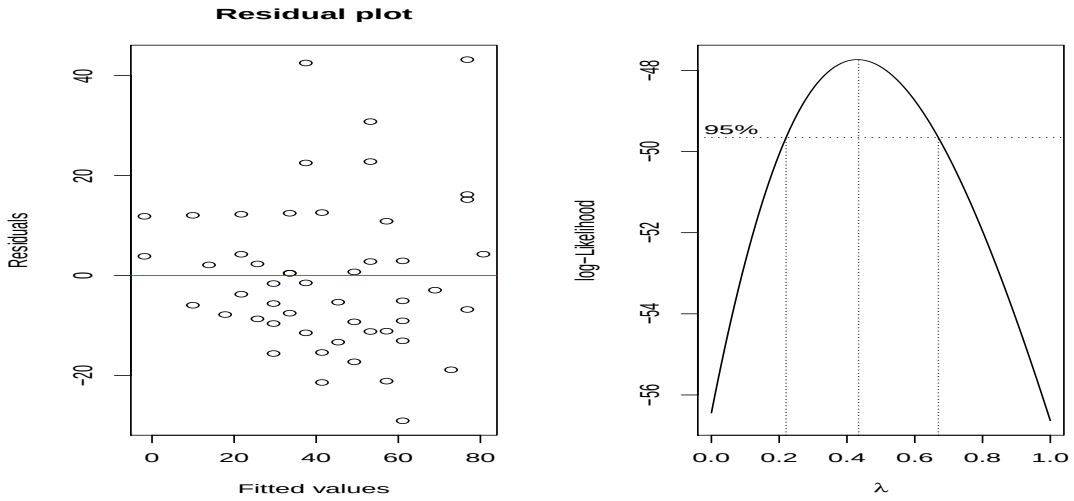

Figure 2: Left: Residual Plot, Right: Plot of the Profile Log-Likelihoods for λ of the Box-Cox Power Transformation

Based on Figure 1 and Figure 2, comment on whether, and how (if necessary), you would transform the response variable or the predictor, or both. Be sure to address each plot. **4 points**

2. A university medical center urology group was interested in the association between prostate-specific antigen, *PSA*, and a number of clinical measurements in men with advanced prostate cancer. Data were collected on 97 men. *PSA* was recorded in terms of antigen level, in milligrams per milliliters (mg/ml). The other variables are:

   - The Gleason score, *score*. The Gleason score is a grade of disease using total score of two patterns, and the total scores were either 6, 7, or 8, with higher scores indicating worse prognosis.
   - The prostate cancer volume, *volume*, in cc.
   - The presence of seminal vesicle invasion, *invasion*, which is coded as 1 if yes, and 0 if no.
   - The degree of capsular penetration, *capsular*, in cm.

   The simple linear regression model that relates *PSA* and *score* is written as:

$$PSA_i = \beta_0 + \beta_1 score_i + \epsilon_i. \tag{1}$$

   You may assume the regression assumptions are met. The output from R is shown below:

```
Call:
lm(formula = PSA ~ score, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -139.150     35.329  -3.939 0.000156 ***
score         23.687      5.109   4.637 1.13e-05 ***

Residual standard error: 37.02 on 95 degrees of freedom
Multiple R-squared:  0.1845,        Adjusted R-squared:  0.176
F-statistic:  ____ on 1 and 95 DF,  p-value: 1.129e-05
```

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

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

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

(c) One member of the urology group believes that prostate-specific antigen level increases, on average, by more than 15 mg/ml, per unit increase in Gleason score. Carry out a corresponding hypothesis test. Be sure to state the null and alternative hypotheses, calculate the test statistic, and state your conclusion in context. **7 points**

(d) Based on model (1), the 95% confidence interval for mean *PSA* when *score*= 7 is (19.09329, 34.22774). Compute the corresponding 95% prediction interval for a single observation of *PSA* for the same value for *score*. **6 points**

3. (This question is an extension of question 2). Suppose the urology group is dissatisfied with the model (1), and decides to consider additional predictors:

$$PSA_i = \beta_0 + \beta_1 score_i + \beta_2 volume_i + \beta_3 invasion_i + \beta_4 capsular_i + \epsilon_i. \qquad (2)$$

You may assume the regression assumptions are met. The output from R is shown below:

```
Call:
lm(formula = PSA ~ volume + invasion + capsular + score, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -40.769     33.244  -1.226 0.223199
score          6.393      5.025   1.272 0.206518
volume         2.028      0.584   3.473 0.000787 ***
invasionYes   17.857     10.751   1.661 0.100111
capsular       1.104      1.325   0.833 0.407104

Residual standard error: 30.99 on 92 degrees of freedom
Multiple R-squared:  0.4467,        Adjusted R-squared:  0.4226
F-statistic: 18.57 on 4 and 92 DF,  p-value: 3.246e-11

Analysis of Variance Table

Response: PSA
          Df Sum Sq Mean Sq F value    Pr(>F)
score      1  29466   29466 30.6815 2.866e-07 ***
volume     1  36200   36200 37.6940 2.072e-08 ***
invasion   1   4985    4985  5.1911   0.02502 *
capsular   1    666     666  0.6936   0.40710
Residuals 92  88354     960
```

(a) What is the p-value for testing $H_0 : \beta_4 = 0$ vs $H_a : \beta_4 \neq 0$ in model (2)? Your classmate says based on the result of the test, *capsular* is not linearly associated with *PSA*. Do you agree? If not, please briefly explain why. **3 points**

(b) Conduct an appropriate hypothesis test to decide between model (1) and model (2). Be sure to state the null and alternative hypotheses, calculate the test statistic, and state your conclusion in context. **6 points**

(c) How would you test $\beta_2 = \beta_3 = 0$ under the assumption that β_4 is 0? Be sure to state the null and alternative hypotheses, calculate the test statistic, and state your conclusion in context. **8 points**

4. The data for this question come from a sample of 200 high school students. The response variable is (y): *write*, the student's score on a standardized writing test, and the predictors are *race*, which is categorical with four levels (Hispanic/Asian/African American/Caucasian); *female*, which is an indicator that coded 1 if the student is female and 0 if the student is male; and (x_1): *read*, the student's score on a standardized reading test. The regression equation we are fitting is:

$$E(y) = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 F + \beta_5 x_1, \tag{3}$$

where $I_1 = 1$ if the student's race is Asian, $I_2 = 1$ if the student's race is African American, $I_3 = 1$ if the student's race is Caucasian, $F = 1$ if the student is female. Use the output for question 2 to answer the following. Assume that all the assumptions for fitting a regression model are met. The output from R is shown below:

```
Call:
lm(formula = write ~ race.f + female + read)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      19.29178    2.81823   6.845 9.77e-11 ***
race.fAsian       7.34591    2.57732   2.850  0.00484 **
race.fAfrican-Am  0.66265    2.12741   0.311  0.75577
race.fCaucasian   3.36449    1.58778   2.119  0.03536 *
female            5.26062    1.00307   5.245 4.08e-07 ***
read              0.53047    0.05061  10.482  < 2e-16 ***

##############################
##variance-covariance matrix##
##############################
> round(vcov(result.main),3)
                 (Intercept) race.fAsian race.fAfrican-Am race.fCaucasian
(Intercept)            7.942      -1.250           -1.912          -1.122
race.fAsian           -1.250       6.643            2.097           2.164
race.fAfrican-Am      -1.912       2.097            4.526           2.061
race.fCaucasian       -1.122       2.164            2.061           2.521
female                -0.584      -0.284           -0.193          -0.092
read                  -0.121      -0.014           -0.001          -0.019
                 female   read
```

```
(Intercept)      -0.584 -0.121
race.fAsian      -0.284 -0.014
race.fAfrican-Am -0.193 -0.001
race.fCaucasian  -0.092 -0.019
female            1.006  0.003
read              0.003  0.003
```

(a) For students who are Hispanic and male, what is the estimated regression equation for *write* and *read*? **2 points**

(b) What is the value of $\hat{\beta}_4$? Interpret this value in context. **4 points**

(c) Suppose that there are no significant interaction terms in this regression. Briefly explain what this means in terms of the relationship between *write* and *read*? **2 points**

(d) Compute a 95% confidence interval for the difference in mean score on the standardized writing test between students who are Caucasian and Hispanic, for given level of *gender*, and given value of *read*. Based on this interval, write an appropriate conclusion. **5 points**

5. For the following statements, state whether they are true or false. If false, briefly explain why.

(a) When one is comparing linear regression models with different number of predictors, multiple R^2 is a better criterion for model selection than adjusted R^2. **3 points**

(b) When using automated search procedures such as backward elimination, the model selected by the algorithm will be the most optimal model that we can get. **3 points**

(c) For a linear regression model, a plot of ordinary residuals, e_i, versus fitted values, $\hat{y}_i$, will help detect all outliers. **3 points**

(d) Based on Figure 3 below, the observation labeled "Case 1" is more likely to have a higher $DFFITS$ than the observation labeled "Case 3". **3 points**

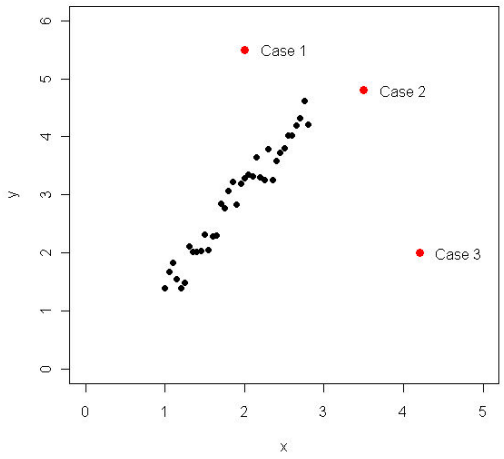

Figure 3: Scatterplot of Stopping Distance against Speed