Stat 6021: Homework 2

1. The data set `mammals` from the `MASS` package contains the average brain and body weights for 62 species of land mammals. We wish to see how body weight (x) could explain the brain weight (y) of land mammals.

   (a) Create a scatter plot of brain weight against body weight of land mammals. Comment on the appearance of the plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

   (b) Fit a simple linear regression to the data, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

   (c) Based on your answers to parts 1a and 1b, do we need to transform at least one of the variables? Briefly explain.

   (d) For the simple linear regression in part 1b, create a Box Cox plot. What transformation, if any, would you apply to the response variable? Briefly explain.

   (e) Apply the transformation you specified in part 1d, and let y^* denote the transformed response variable. Create a scatterplot of y^* against x. Comment on the appearance of the plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

   (f) Fit a simple linear regression to y^* against x, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

   (g) Do we need to transform the x variable? If yes, what transformation(s) would you try? Briefly explain. Create a scatterplot of y^* against x^*. Do any assumptions for simple linear regression appear to be violated? If so, which ones?

   (h) Fit a simple linear regression to y^* against x^*, and create the corresponding residual plot. Do any assumptions for simple linear regression appear to be violated? If so, which ones? If the assumptions are not met, repeat with a different transformation on the predictor until you are satisfied.

   (i) Write out the regression equation, and if possible, interpret the slope of the regression.

2. For this question, we will use the `cornnit` data set from the `faraway` package. Be sure to install and load the `faraway` package first, and then load the data set. The data explore the relationship between corn yield (bushels per acre) and nitrogen (pounds per acre) fertilizer application in a study carried out in Wisconsin.

   (a) What is the response variable and predictor for this study? Create a scatterplot of the data, and interpret the scatterplot.

   (b) Fit a linear regression without any transformations. Create the corresponding residual plot. Based only on the residual plot, what transformation will you consider first? Be sure to explain your reason.

   (c) Create a Box Cox plot for the profile loglikelihoods. How does this plot aid in your data transformation?

   (d) Perform the necessary transformation to the data. Re fit the regression with the transformed variable(s) and assess the regression assumptions. You may have to apply transformations a number of times. Be sure to explain the reason behind each of your transformations. Perform the needed transformations until the regression assumptions are met. What is the regression equation that you will use?

   Note: in part 2d, there are a number of solutions that will work. You must clearly document your reasons for each of your transformations.

3. For this question, we will use the data set "nfl.txt", which contains data on NFL team performance from the 1976 season. The variables are:

   - y: Games won (14-game season)
   - x_1: Rushing yards (season)
   - x_2: Passing yards (season)
   - x_3: Punting average (yards/punt)
   - x_4: Field goal percentage (FGs made/FGs attempted)
   - x_5: Turnover differential (turnovers acquired - turnovers lost)
   - x_6: Penalty yards (season)
   - x_7: Percent rushing (rushing plays/total plays)
   - x_8: Opponents' rushing yards (season)
   - x_9: Opponents' passing yards (season)

   (a) Fit a multiple regression model for the number of games won against the following three predictors: the team's passing yardage, the percentage of rushing plays, and the opponents' yards rushing. Write the estimated regression equation.

   (b) Interpret the estimated coefficient for the predictor x_7 in context.

(c) A team with $x_2 = 2000$ yards, $x_7 = 48$ percent, and $x_8 = 2350$ yards would like to estimate the number of games it would win. Also provide a relevant interval for this estimate with 95% confidence.

(d) Using the output for the multiple linear regression model from part 3a, answer the following question from a client: "Is this regression model useful in predicting the number of wins during the 1976 season?" Be sure to write the null and alternative hypotheses, state the value of the test statistic, state the p-value, and state a relevant conclusion.

(e) Report the value of the t statistic for the predictor x_7. What is the relevant conclusion from this t statistic?

(f) Check the regression assumptions by creating the diagnostic plots. Comment on these plots.

(g) Consider adding another predictor, x_1, the team's rushing yards for the season, to the model. Interpret the results of the t test for the coefficient of this predictor. A classmate says: "Since the result of the t test is insignificant, the team's rushing yards for the season is not linearly related to the number of wins." Do you agree with your classmate's statement?

4. For this question, the data are from the `faraway` package in R. After installing the faraway package, load the `seatpos` dataset. Car drivers like to adjust the seat position for their own comfort. Car designers find it helpful to know where different drivers will position the seat. Researchers at the HuMoSim laboratory at the University of Michigan collected data on 38 drivers. The response variable is *hipcenter*, the horizontal distance of the midpoint of the hips from a fixed location in the car in mm. They measured the following eight predictors:

- x_1: *Age.* Age in years
- x_2: *Weight.* Weight in pounds
- x_3: *HtShoes.* Height with shoes in cm
- x_4: *Ht.* Height without shoes in cm
- x_5: *Seated.* Seated height in cm
- x_6: *Arm.* Arm length in cm
- x_7: *Thigh.* Thigh length in cm
- x_8: *Leg.* Lower leg length in cm

(a) Fit the full model with all the predictors. Using the `summary()` function, comment on the results of the t tests and ANOVA F test from the output.

(b) Briefly explain why, based on your output from part 4a, you suspect the model shows signs of multicollinearity.

(c) Provide the output for all the pairwise correlations among the predictors. Comment briefly on the pairwise correlations.

(d) Check the variance inflation factors (VIFs). What do these values indicate about multicollinearity?

(e) Looking at the data, we may want to look at the correlations for the variables that describe length of body parts: *HtShoes*, *Ht*, *Seated*, *Arm*, *Thigh*, and *Leg*. Comment on the correlations of these six predictors.

(f) Since all the six predictors from the previous part are highly correlated, you may decide to just use one of the predictors and remove the other five from the model. Decide which predictor out of the six you want to keep, and briefly explain your choice.

(g) Based on your choice in part 4f, fit a multiple regression with your choice of predictor to keep, along with the predictors $x_1 = Age$ and $x_2 = Weight$. Check the VIFs for this model. Comment on whether we still have an issue with multicollinearity.

(h) Conduct a partial F test to investigate if the predictors you dropped from the full model were jointly insignificant. Be sure to state a relevant conclusion.

(i) Produce a residual plot for your model from part 4g. Based on the residual plot, comment on the assumptions for the multiple regression model.

(j) Based on your results, write your estimated regression equation from part 4g. Also report the R^2 of this model, and compare with the R^2 you reported in part 4a, for the model with all predictors. Also comment on the adjusted R^2 for both models.

5. (You may only use R as a simple calculator or to find p-values or critical values) Data from $n = 113$ hospitals are used to evaluate factors related to the risk that patients get an infection while in the hospital. The response variable is *InfctRsk*, the percentage of patients who get an infection while hospitalized. The predictors are *Stay*, the average length of stay, *Cultures*, a ratio of the number of cultures performed per number of patients with no infection (times 100), *Age*, the average patient age, *Census*, the number of patients in the hospital, and *Beds*, the number of beds in the hospital. We consider the following multiple regression equation: $\text{E}(InfctRsk) = \beta_0 + \beta_1 Stay + \beta_2 Cultures + \beta_3 Age + \beta_4 Census + \beta_5 Beds$. Some R output is shown below. You may assume the regression assumptions are met.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2051282  1.2075929   0.170   0.8654
Stay        0.2055252  0.0660885   3.110   0.0024 **
Cultures    0.0590369  0.0103096   5.726  9.5e-08 ***
Age         0.0173637  0.0229966   -----   ------
Census      0.0010306  0.0034942   0.295   0.7686
Beds        0.0004476  0.0026781   0.167   0.8676
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.9926 on 107 degrees of freedom
Multiple R-squared:  ______,    Adjusted R-squared:  ______
F-statistic: 19.48 on 5 and 107 DF,  p-value: 9.424e-14
```

```
Analysis of Variance Table

Response: InfctRsk
           Df  Sum Sq Mean Sq F value    Pr(>F)
Stay        1  57.305  57.305 58.1676 1.044e-11 ***
Cultures    1  33.397  33.397 33.8995 6.154e-08 ***
Age         1   0.136   0.136  0.1376   0.71144
Census      1   5.101   5.101  5.1781   0.02487 *
Beds        1   0.028   0.028  0.0279   0.86759
Residuals 107 105.413   0.985
```

(a) What is the value of the estimated coefficient of the variable *Stay*? Write a sentence that interprets this value.

(b) Derive the test statistic, p-value, and critical value for the variable *Age*. What null and alternative hypotheses are being evaluated with this test statistic? What conclusion should we make about the variable *Age*?

(c) What is the R^2 for this model? Write a sentence that interprets this value in context.

(d) Suppose we want to decide between two potential models:

- Model 1: using x_1, x_2, x_3, x_4, x_5 as the predictors for *InfctRsk*
- Model 2: using x_1, x_2 as the predictors for *InfctRsk*

Carry out the appropriate hypothesis test to decide which of models 1 or 2 should be used. Be sure to show all steps in your hypothesis test.

(e) Suppose we want to decide between two potential models:

- Model 2: using x_1, x_2 as the predictors for *InfctRsk*
- Model 3: using x_1, x_2, x_3, x_4 as the predictors for *InfctRsk*

Carry out the appropriate hypothesis test to decide which of models 2 or 3 should be used. Be sure to show all steps in your hypothesis test.

6. We will revisit the data set `penguins` from the `palmerpenguins` package. The data set contains size measurements for adult foraging penguins near Palmer Station, Antarctica. In this set of questions, we focus on exploring the relationship between body mass (y) and bill depth (x_1) of three species of penguins.

(a) Create a scatterplot of the body mass against the bill depth of the penguins. How would you describe the relationship between these two variables?

(b) Create the same scatterplot but now with different colored plots for each species. Also be sure to overlay separate regression lines for each species. How would you now describe the relationship between the variables?

(c) Create a regression with interaction between bill depth and species, i.e.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \beta_4 x_1 I_1 + \beta_5 x_1 I_2 + \epsilon,$$

where I_1 and I_2 are indicator variables where $I_1 = 1$ for Chinstrap penguins and 0 otherwise, and $I_2 = 1$ for Gentoo penguins and 0 otherwise. Write down the estimated regression equation for this model.

(d) Carry out the relevant hypothesis test to see if the interaction terms can be dropped. What is the conclusion?

(e) Based on your answer in part 6d, write out the estimated regression equations relating body mass and bill depth, for each species of the penguins.

(f) Assess if the regression assumptions are met, for the model you will recommend to use (based on part 6d).

(g) Briefly explain if we can conduct pairwise comparisons for the difference in mean body mass among all pairs of species for given values bill depth, i.e.,

i. Adelie and Chinstrap,

ii. Adelie and Gentoo,

iii. Chinstrap and Gentoo.

If we are able to, conduct Tukey's multiple comparisons and contextually interpret the results of these hypothesis tests.

7. (You may only use R as a simple calculator or to find p-values or critical values) This question is based on data about teacher salaries from the 50 states plus DC (so $n = 51$) in the mid 1980s. The variables are:

- *PAY*, y: average annual public school teacher salary, in dollars.
- *SPEND*, x_1: Spending on public schools per student, in dollars.
- *AREA*: Region (North, South, West).

Table 1 below provides some summary statistics of the data:

Region	n	Mean PAY	Mean SPEND
North	21	\$24424	\$3901
South	17	\$22894	\$3274
West	13	\$26159	\$3919

Table 1: Summary Statistics of Teacher Pay

(a) Based only on Table 1, briefly comment on the relationship between geographic area and mean teacher pay.

(b) Based only on Table 1, briefly comment on the relationship between mean public school expenditure (per student) and mean teacher pay.

(c) Briefly explain why using a multiple linear regression model with teacher pay as the response variable with geographic area and public school expenditure (per student) can give further insight into the relationship(s) between these variables.

Use the following info to answer the rest of question 7.

We want to see if geographic region and spending on public schools affect the average public teacher pay. A regression with no interactions was fitted, i.e.,

$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 I_2 + \beta_3 I_3,$$

where I_2 and I_3 are the dummy codes for $AREA$. $I_2 = 1$ if $AREA =$ South, 0 otherwise, and $I_3 = 1$ if $AREA =$ West, 0 otherwise.

The following output from R is shown below

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.160e+04  1.334e+03   8.690 2.43e-11 ***
SPEND       3.289e+00  3.176e-01  10.354 1.03e-13 ***
AREASouth   5.294e+02  7.669e+02   0.690   0.4934
AREAWest    1.674e+03  8.012e+02   2.089   0.0422 *

############################################
##Variance-Covariance matrix for beta hats##
############################################

             (Intercept)        SPEND     AREASouth      AREAWest
(Intercept) 1780535.6980 -393.5597348 -491859.07243 -2.381145e+05
SPEND          -393.5597    0.1008967      63.18227 -1.870101e+00
AREASouth   -491859.0724   63.1822716  588126.71689  2.442380e+05
AREAWest    -238114.5499   -1.8701007  244238.02959  6.418738e+05
```

(d) What is the estimate of β_2? Give an interpretation of this value.

(e) Using the Bonferroni procedure, compute the 95% family confidence intervals for the difference in mean response for PAY between teachers in the

i. North region and the South region;

ii. North region and the West region;

iii. South region and the West region,

while controlling for expenditure.

(f) What do your intervals from part 7e indicate about the effect of geographic region on mean annual salary for teachers (while controlling for expenditure)?