STAT 505/506 MS Comprehensive Exam 22 aug 2016

Instructions: This exam is closed notes and closed book though you may use a calculator. Please answer these questions on separate sheets of paper **making sure to include your name at the top of each page, number the pages, and only write on ONE SIDE of the page**. Due to the length of the exam complete sentences are not necessary. Please look over the entire exam before beginning. Good luck!

1. **Fun Facts of Linear Models**. We have spent a considerable amount of time in this course on the linear model expressed as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Let's first cover some basics about this type of model.

   (a) We typically assume that $\boldsymbol{X}$ is full column rank.

      i. Briefly explain (or define) what it means for a matrix to be full column rank. [5pts]

      ii. Explain why we make this assumption. [5pts]

   (b) To obtain parameter estimates, we often use ordinary least squares due to the desirable properties of the least squares estimator. These properties are summarized in the Gauss-Markov Theorem based on several assumptions of the error term $\boldsymbol{\epsilon}$.

      i. What assumptions does the Gauss-Markov Theorem make regarding $\boldsymbol{\epsilon}$? [5pts]

      ii. Under these assumptions what does the Gauss-Markov Theorem state regarding the OLS estimator? [5pts]

2. **Species Diversity on the British Isles**. Species diversity on islands and isles is often of interest to ecologists as these locations typically boast unique biodiversity. In this analysis, researchers are interested in predicting species diversity in the British Isles. The dependent variable used is the `species` (number of species on the island) with five potential independent variables: `area` (island area in square kilometers), `elevation` (maximum island elevation in meters), `soil` (number of soil types on the island), `nlat` (midpoint of latitude range in degrees north), and `distance` (distance from mainland Britain in kilometers) collected from 41 isles (note that the British mainland was excluded). Attached is some output that may or may not be helpful.

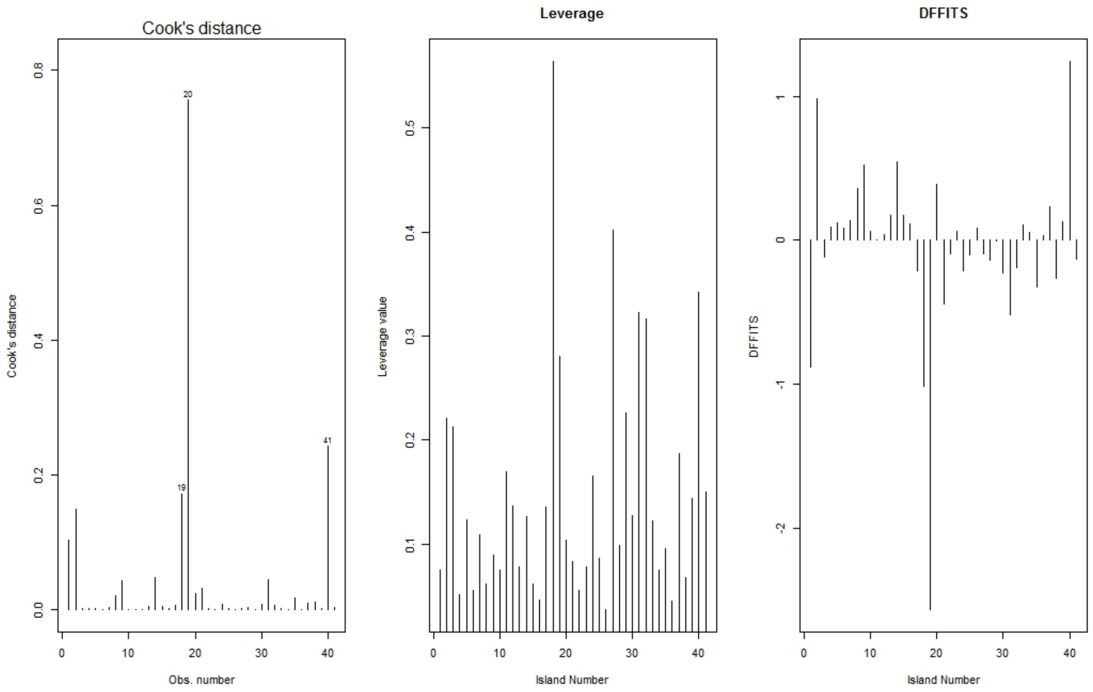
   (a) The researchers first considered a model with all five independent variables. Using the available output answer the following.

      i. Provide an interpretation of the slope coefficient associated with `area` *in context of the problem.* [3pts]

      ii. One concern of the researchers is that of multicollinearity. Explain what multicollinearity is and, referencing the appropriate diagnostics, evaluate if it is a major concern in this analysis. [7pts]

      iii. We frequently use a residuals versus plot as a model diagnostic. Explain what assumptions of the model we can evaluate using this plot and, using the plot provided, explain if there appears to be any concerns with violations of those assumptions. [8pts]

      iv. Below are plots of Cook's distance, h_{ii}, and DFFITS for each observation. Based on these plots explain if you have any concerns regarding the model. [6pts]



(b) Due to their knowledge of biological island diversity the researchers consider several potential models:

M1 : $Species_i = \beta_0 + \beta_1 Area_i + \beta_2 Elevation_i + \beta_3 Soil_i + \beta_4 nlat_i + \beta_5 Distance_i + \epsilon_i$

M2 : M1 + Quadratic terms are included for each independent variable

M3 : M1 + Interaction terms for each pair of independent variables

M4 : M1 + Quadratic terms for each independent variable and interaction terms for each pair of independent variables (this model is adding the additional terms added in M2 and M3 to those used in M1).

M5 : All independent variables are log transformed and species is not transformed

For all the models, assume the error terms are independent of one another and are normally distributed with constant variance. The residual (or error) sums of squares for each model are as follows:

Model	M1	M2	M3	M4	M5
RSS	472164	348766	252254	153003	253713

i. Is there evidence to support log transforming the independent variables? Support your answer with a test statistic or test statistics if possible. If it is not valid to conduct a hypothesis test explain why. [4pts]

ii. Is there evidence to support including interaction terms? Support your answer with a test statistic or test statistics if possible. If it is not valid to conduct a hypothesis test explain why. [4pts]

(c) In their analysis the researchers opted to exclude information from the British mainland (which is an island, though you probably knew that) as it is considerably larger and much more populated than the other 41 islands. Would it be reasonable for the researchers to use the results of their final model to predict species diversity on the British mainland? Why or why not? [5pts]

3. **Apple Juice Contamination**. *Alicyclobacillus acidoterrestris* is a bacterium whose spores are able to survive most pasteurization procedures and consequently can be found in fruit juices. A study was conducted by a company that manufactures apple juice to examine the effect of temperature (in degrees Celsius), pH, soluble solids concentration–referred to as brix concentration, and nisin concentration on the growth probability of *Alicyclobacillus acidoterrestris* CRA 7152 in apple juice. The ultimate goal is for the company to determine how to best inhibit growth of the bacterium. In this study 74 samples of apple juice were manufactured. Each sample was assigned to a given pH level, brix concentration (in IUI/mL), temperature (in Celsius), and nisin concentration (in IUI/mL) and the presence/absence of growth of the bacterium was noted for each sample. For your analysis, please consider pH, temperature, brix concentration, and nisin concentration to be quantitative. See attached output at the end of this exam to help you answer the following.

   (a) To start let's consider a model with all four of the potential explanatory variables.

      i. What assumptions are we making by using `family=binomial` in our code? [3pts]

      ii. Provide an interpretation of the coefficient associated with `temperature` *in context of the problem.* [3pts]

   (b) The company also wants to consider a more complicated model where the variable `pH` interacts with each of the other three variables.

      i. Explain what it means for two variables to interact. [4pts]

      ii. Conduct the appropriate test (if possible) to examine if there is evidence to include the interaction terms in the model. Make sure to clearly state your hypotheses, your test statistic (including its distribution) and how you would go about obtaining a p-value. [5pts]

   (c) **Regardless of your answer for the previous question, for this and the remaining questions assume the researchers are using the model where `pH` does NOT interact with the other three variables**. One common statistic that we have calculated is $\widehat{\phi} = \frac{\text{Residual Deviance}}{n-p}$.

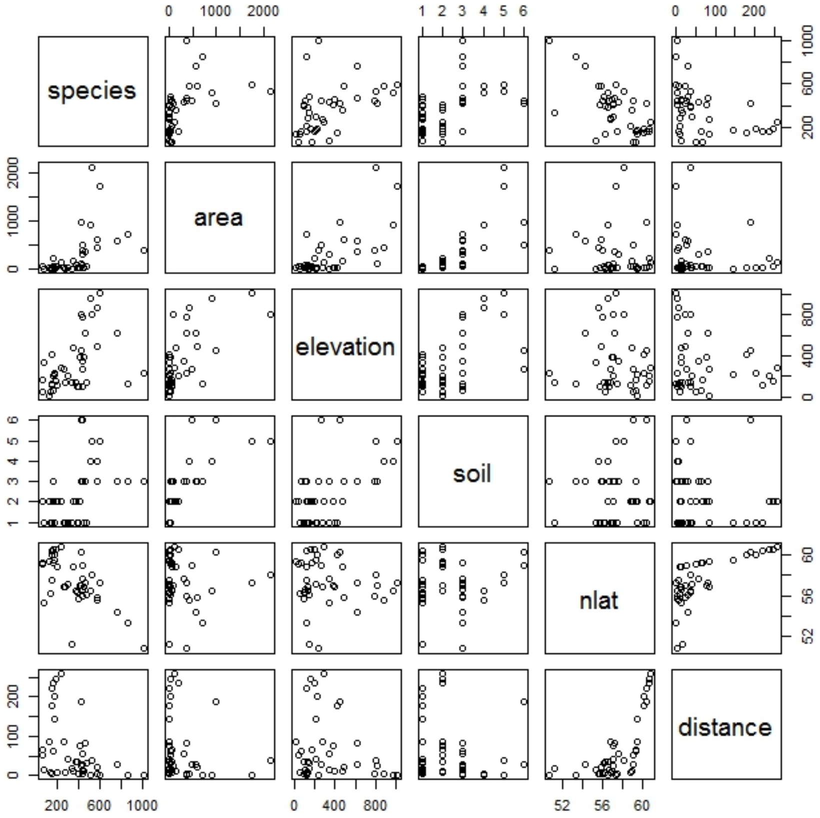
      i. Explain what this statistic is used for [3pts].

      ii. Calculate this statistic for this model and explain what the calculated value suggests [3pts].

      iii. Name one method that we can use if concerns are raised regarding our analysis based on this statistic [3pts].

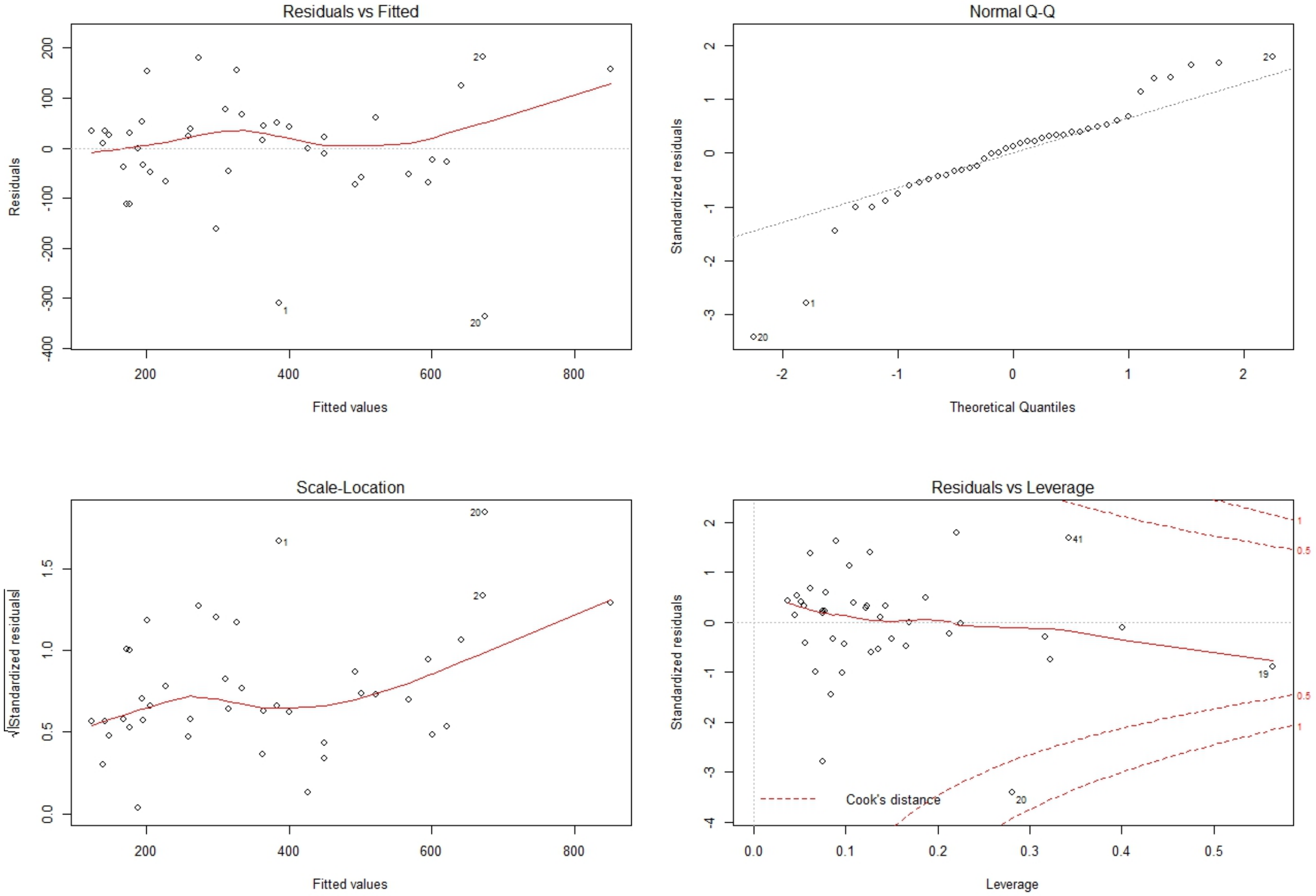
   (d) Of interest to the researchers is predicting the growth probability under typical room temperature conditions. One typical combination is a temperature of 30 degrees with brix and nisin concentrations of 15 and 35 IUI/mL respectively and a pH of 4.0. Would these conditions be sufficient to keep predicted growth probabilities close to 0 (which is considered to be a value of $< 10^{-5}$)? [4pts]

   (e) The company is also interested in calculating what they refer to as "critical values" which are the minimum value of a given factor that would inhibit growth. Obtain the critical value of pH that would inhibit growth (defined as a growth probability of 0.05) for a brix concentration of 19 IUI/mL, nisin concentration of 20 IUI/mL and a temperature of 35 degrees. [5pts]



```
> cor(islands2[,-1])
                 area  elevation        soil        nlat   distance    species
area       1.00000000  0.6661267  0.76832201 -0.07329581 -0.1604493  0.5158345
elevation  0.66612672  1.0000000  0.57889211 -0.10916042 -0.1971941  0.4471657
soil       0.76832201  0.5788921  1.00000000  0.02845539 -0.1267745  0.4956067
nlat      -0.07329581 -0.1091604  0.02845539  1.00000000  0.6990194 -0.6616449
distance  -0.16044930 -0.1971941 -0.12677447  0.69901940  1.0000000 -0.4340089
species    0.51583446  0.4471657  0.49560672 -0.66164490 -0.4340089  1.0000000

> islands.m1<-lm(species˜area+elevation+soil+nlat+distance,data=islands2)
> summary(islands.m1)

Call:
lm(formula = species ˜ area + elevation + soil + nlat + distance,
    data = islands2)

Residuals:
    Min      1Q  Median      3Q     Max
-336.15  -47.71   14.84   50.00  182.95

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4327.17527  616.65323   7.017 3.63e-08 ***
```

```
area           0.05966    0.06882   0.867  0.39196
elevation      0.06726    0.09167   0.734  0.46798
soil          60.32272   21.02917   2.869  0.00694 **
nlat         -72.93166   11.10905  -6.565 1.40e-07 ***
distance       0.59803    0.33888   1.765  0.08633 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 116.1 on 35 degrees of freedom
Multiple R-squared:  0.7409,    Adjusted R-squared:  0.7039
F-statistic: 20.02 on 5 and 35 DF,  p-value: 2.181e-09
> anova(islands.m1)
Analysis of Variance Table

Response: species
          Df Sum Sq Mean Sq F value    Pr(>F)
area       1 484868  484868 35.9417 7.836e-07 ***
elevation  1  35128   35128  2.6039   0.11558
soil       1  34179   34179  2.5336   0.12044
nlat       1 753876  753876 55.8825 9.400e-09 ***
distance   1  42013   42013  3.1143   0.08633 .
Residuals 35 472164   13490
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> vif(islands.m1)
     area elevation      soil      nlat  distance
 3.006675  1.862955  2.585686  2.036553  2.052533
```


Residuals vs Fitted
Residuals
-400
-300
-200
-100
0
100
200
200
400
600
800
Fitted values
1
20
20
Normal Q-Q
Standardized residuals
-3
-2
-1
0
1
2
-2
-1
0
1
2
Theoretical Quantiles
20
1
20
Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
200
400
600
800
Fitted values
20
1
20
Residuals vs Leverage
Standardized residuals
-4
-3
-2
-1
0
1
2
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
Cook's distance
1
0.5
0.5
1
41
19
20

```
> juice.log<-glm(growth~pH+nisin+temperature+brix,data=juice,
+ family=binomial(link=logit))
> summary(juice.log)

Call:
glm(formula = growth ~ pH + nisin + temperature + brix, family = binomial(link = logit),
    data = juice)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3614  -0.3990  -0.1585   0.6306   1.6200

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.24633    3.21864  -2.251 0.024362 *
pH           1.88595    0.54123   3.485 0.000493 ***
nisin       -0.06628    0.01905  -3.479 0.000503 ***
temperature  0.11042    0.04769   2.316 0.020585 *
brix        -0.31173    0.14317  -2.177 0.029458 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 95.945  on 73  degrees of freedom
Residual deviance: 52.331  on 69  degrees of freedom
AIC: 62.331

Number of Fisher Scoring iterations: 6

> juice.log2<-glm(growth~pH+nisin+temperature+brix
+ +pH*nisin+pH*temperature+pH*brix,data=juice,
+ family=binomial(link=logit))
> summary(juice.log2)

Call:
glm(formula = growth ~ pH + nisin + temperature + brix + pH *
    nisin + pH * temperature + pH * brix, family = binomial(link = logit),
    data = juice)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.08091  -0.33316  -0.01166   0.24397   1.62698

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -83.4684    35.4943  -2.352   0.0187 *
```

```
pH              16.5892     6.6523   2.494   0.0126 *
nisin           -0.6256     0.3520  -1.777   0.0755 .
temperature      1.4671     0.7012   2.092   0.0364 *
brix             2.1017     1.2181   1.725   0.0845 .
pH:nisin         0.1064     0.0651   1.634   0.1023
pH:temperature  -0.2636     0.1306  -2.018   0.0436 *
pH:brix         -0.4640     0.2398  -1.935   0.0530 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 95.945  on 73  degrees of freedom
Residual deviance: 40.400  on 66  degrees of freedom
AIC: 56.4

Number of Fisher Scoring iterations: 8
```