Name: _____ Fall, 2018

Applied Statistics Comprehensive Examination

- Calculators are permitted on this examination.
- When you compute a confidence interval, always give an interpretation of the interval in the context of the problem.
- When you perform a hypothesis test, always write down the null and alternative hypotheses, and write the conclusion in the context of the problem.
- There are 200 points on this examination.
- You must give complete explanations to receive full credit.
- Please put your answers and explanations on the separate sheets provided.

1. (20 points) A researcher is interested in whether the birth order within twins has an impact on their salary as adults. Although twins are typically born very close together, one does indeed come out before the other. The following data were collected on 5 sets of twins, with the salary (in thousands of dollars) at age 35 for each individual.

	Twin 1	Twin 2
Family 1	35	37
Family 2	80	76
Family 3	51	47
Family 4	52	44
Family 5	87	81

   (a) (2 points) Suppose that you want to test whether there is any difference in salary based on birth order. State the name of the test that you would perform here.

   (b) (3 points) For the test that you proposed in Part (a), does normality need to be assumed? If so, specifically what quantity needs to have a normal distribution? If not, explain why not.

   (c) (15 points) At $\alpha = 0.05$, perform the test that you proposed in Part (a).

2. (30 Points) Three diets were given to a group of adult volunteers to assess their effects on serum cholesterol levels. Twelve volunteers with similar starting cholesterol levels were randomly assigned to the three diets, and there were four volunteers for each diet. The data are as follows:

	Diet 1	Diet 2	Diet 3
	273	240	205
	302	221	211
	266	244	233
	247	238	224
Sum:	1088	943	873
Mean:	272.0	235.8	218.3

The partial ANOVA table is given here:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Diet				
Error				
Total	8,362			

a. (10 Points) Fill out the ANOVA table and test whether the mean cholesterol level is the same for the three diets.
b. (10 Points) Use Tukey's multiple comparisons procedure to determine which diets are significantly different. Make a recommendation as to which diet(s) minimize the cholesterol levels.
c. (10 Points) Construct a 95% confidence interval to estimate the difference in mean cholesterol between Diet 1 and Diets 2 and 3 combined.

3. (15 points) The 115th Congress of the U.S. House of Representatives contains 435 members. According to `pressgallery.house.gov`, the racial demographics of these 435 members are:

   - Caucasian: 330
   - African American: 47
   - Hispanic American: 42
   - Asian American: 14
   - Native American: 2

   Suppose that we want to know if there is sufficient evidence to suggest that the racial makeup of the 115th Congress is different from that of the U.S. population, which is approximately:

   - Caucasian: 61.4%
   - African American: 13.4%
   - Hispanic American: 18.1%
   - Asian American: 5.8%
   - Native American: 1.3%

   At $\alpha = 0.05$, perform a hypothesis test to answer this question.

4. (15 points) A woman who flies frequently from New York to Boston wishes to estimate the variability in the in-air flight times from New York to Boston. She carefully records the in-air flight times (in minutes) for her next five flights as 40, 38, 37, 39, and 46. The mean and standard deviation of these flight times are 40 and 3.54, respectively.

   (a) (10 points) Find a 95% confidence interval for the standard deviation in flight times from New York to Boston.

   (b) (5 points) What assumptions are needed in order for the confidence interval you found in Part (a) to be valid?

5. (30 Points) Tomato plants were grown in a greenhouse under treatments consisting of combinations of soil type and fertilizer type. There were two plants per soil type and fertilizer type combination. The following data on the yield (kg) of tomatoes were obtained for the 12 plants under study.

	Fertilizer Type		
Soil Type	**1**	**2**	**3**
I	5 7	5 5	3 5
II	5 9	2 3	4 6

An effects model with effects for soil type, fertilizer type, and the interaction was fit to the data. Let $\alpha_i,\ i = 1,2,$ be the effect of soil type 1 and 2, respectively. Let $\beta_j,$ $j = 1,2,3,$ be the effect of fertilizer types 1, 2, and 3, respectively. The Sum of Squares for Error was 14.5.

a. (10 Points) Test at the 0.05 level for the significance of the interaction between soil type and fertilizer types 2 and 3.
b. (5 Points) Construct a complete set of orthogonal interaction contrasts.
c. (10 Points) Using only one observation per soil type and fertilizer type combination, write the design matrix for the effects model above using sum-to-zero restrictions.
d. (5 Points) Is $(\alpha_1 - \alpha_2)$ estimable? Explain.

6. (20 points) According to a study done by Tufts University, 48.3% of U.S. college students voted in the 2016 U.S. presidential election. A Villanova professor wanted to know whether the percentage is different at Villanova, and so she conducted her own study in 2016 by personally interviewing 200 randomly chosen Villanova students from the U.S. She found that 84 of these students voted in the 2016 election.

   (a) (10 points) Can we conclude at the 0.05 significance level that the proportion of Villanova students from the U.S. who voted in the 2016 election was different than 0.483?

   (b) (10 points) If the true proportion of Villanova students from the U.S. who voted was 0.45, what was the power for the test in Part (a) for a random sample of 200 Villanova students from the U.S.?

7. (70 points) Species diversity on islands is of interest to ecologists as islands are isolated from the mainland and typically have unique biodiversity. In this analysis, researchers were interested in predicting species diversity in the British Isles. The dependent variable used was `species` (number of plant 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), `latitude` (midpoint of latitude range in degrees north), and `distance` (distance from mainland Britain in kilometers). Data were collected from 41 islands.

   (a) (40 points) The researchers first considered a model with all five independent variables. Using the available output, answer the following questions:

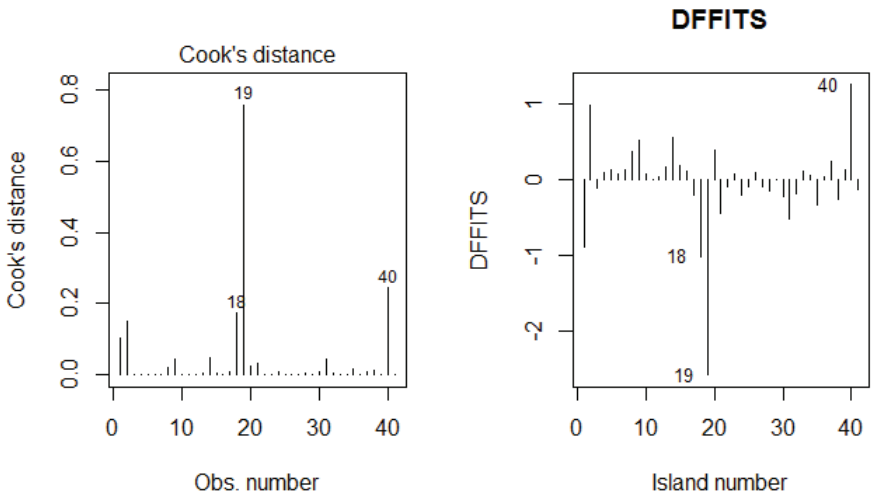
      i. (10 points) Conduct the global hypothesis test for the model at $\alpha = 0.05$.

      ii. (5 points) Interpret the slope coefficient associated with `soil` in the context of the problem.

      iii. (10 points) One concern of the researchers was multicollinearity. Explain the consequences of multicollinearity. Using appropriate diagnostics from the output, evaluate if it is a major concern in this analysis.

      iv. (10 points) Explain what assumptions of the model you can evaluate using the plots provided in the output, and identify specific plot(s) for each assumption. Explain if there appear to be any concerns with violations of those assumptions.

      v. (5 points) Below are plots of Cook's distance and DFFITS for each observation. Based on these plots, explain if you have any concerns regarding the model.



(b) (30 points) The researchers considered several potential models:
M1: $\texttt{species} = \beta_0 + \beta_1 \cdot \texttt{area} + \beta_2 \cdot \texttt{elevation} + \beta_3 \cdot \texttt{soil} + \beta_4 \cdot \texttt{latitude} + \beta_5 \cdot \texttt{distance} + \epsilon$
M2: M1 with quadratic terms included for each independent variable
M3: M1 with interaction terms for each pair of independent variables
M4: All variables from M1 are included with a log transformation, except `species`, which is in its original form.

The residual (or error) sums of squares (RSS) for each model are as follows:

Model	M1	M2	M3	M4
RSS	472,164	348,766	252,254	253,713

i. (10 points) Is there evidence to support log transforming the independent variables? Justify your answer by conducting an appropriate test ($\alpha = 0.05$), if possible. If it is not valid to conduct such a test, explain why.

ii. (10 points) Is there evidence to support including interaction terms? Justify your answer by conducting an appropriate test ($\alpha = 0.05$), if possible. If it is not valid to conduct such a test, explain why.

iii. (10 points) In addition to RSS (and related hypothesis tests), what other criteria can be used in model selection among these four models? Name three additional criteria and briefly explain how to use each of them in model selection.

Correlation Matrix

```
> cor(diversity[,-1])
                area  elevation        soil    latitude   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
latitude  -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
```

Model Summary

```
> diversity.m1<-lm(species~area+elevation+soil+latitude+distance, data=diversity)
> summary(diversity.m1)

Call:
lm(formula = species ~ area + elevation + soil + latitude + distance,
    data = diversity)

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 **
latitude     -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
```

VIF

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

ANOVA Table

```
> anova(diversity.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
latitude   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
```