Name: _____ Fall, 2020

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. Please use only the front side of these sheets.

1. (15 points) A 2015 study by two researchers named Cvetkovic and Vasiljevic collected the following data on "handedness" (right or left dominant) vs. eye color for girls aged 7-15 in Serbia:

	Right	Left
Green	67	13
Blue	101	11
Brown	417	40

   (a) (12 points) Can we conclude at the 0.05 level that there is an association between handedness and eye color for girls aged 7-15 in Serbia?

   (b) (3 points) Briefly explain whether the sample size is large enough for the inference in part (a) to be valid.

2. (30 points) Tesla Motors hopes to develop a compelling electric vehicle that can be sold for only $25,000. The key to achieving this goal is to reduce the cost of batteries. Researchers ran an experiment to study the effect of material type (A, B or C) and operating temperature (Low, Medium, or High) on a battery's lifetime as measured in discharge/charge cycles. Nine batteries were randomly selected from each material type and then randomly allocated to the three temperature levels. The data are given in the table below, where each number is a count of discharge/charge cycles in thousands for a single battery.

	Material		
Temperature	Type A ($j = 1$)	Type B ($j = 2$)	Type C ($j = 3$)
Low ($i = 1$)	3.9, 4.6, 5.2	4.5, 4.7, 3.7	4.1, 3.3, 4.8
Medium ($i = 2$)	1.2, 2.4, 2.2	4.0, 3.6, 3.1	5.1, 3.6, 4.1
High ($i = 3$)	2.1, 2.4, 1.7	2.0, 1.7, 1.3	3.0, 2.3, 2.4

The researchers fit the two-way main effects model $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$, where the $\epsilon_{ijk} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ are the error terms, the α_i are the Temperature main effects, and the β_j are the Material main effects. The sums of squares for the two-way main effects model are 22.016 for Temperature, 2.749 for Material, and 13.162 for Error.

(a) (10 points) Find a 95% confidence interval for the error variance σ^2 in the two-way main effects model.

(b) (10 points) Make interaction plots. Comment on whether there is interaction and, if so, the type of interaction.

(c) (10 points) Do a formal level–5% test to determine whether we should add interaction to the original main effects model. You may use the fact that the sum of squares for Interaction is 7.449.

3. (20 points) A Gallup poll in October 2020 sampled 1,500 likely voters and found that 30% of them identified Coronavirus as the most important problem facing the country. Assume that it was a simple random sample.

   (a) (10 points) Create a 95% confidence interval for the true proportion of likely voters who believe that Coronavirus is the most important problem facing the country.

   (b) (5 points) Consider conducting the same survey in November 2020. How large of a sample size must be chosen in order to create a 95% confidence interval with a margin of error of 1%?

   (c) (5 points) A pundit claims that, according to this poll, it is estimated that 30% of all citizens of the U.S. think Coronavirus is the most important problem facing the country. Explain to him/her using statistical language why inferring about the population of all citizens of the U.S. is not appropriate from this sample.

4. (25 Points) Suppose that we have data consisting of IQ scores for 27 pairs of identical twins, with one twin from each pair raised in a foster home and the other raised by the natural parents. The IQ for the twin raised in the foster home is denoted by Y, and the IQ for the twin raised by the natural parents is denoted by X_1. The social class of the natural parents (X_2) is also given :

$$X_2 = \begin{cases} 1 \text{ indicates the highest class} \\ 2 \text{ indicates the middle class} \\ 3 \text{ indicates the lowest class} \end{cases}$$

The goal is to predict Y using X_1 and X_2.

(a) (15 Points) Create indicator variables for social class and write the mathematical form of a regression model that will allow all three social classes to have their own y-intercepts and slopes. Be sure to interpret each term in your model.

(b) (10 Points) Describe how you would test the theory that the slope is the same for all three social classes. Be sure to state the hypothesis, general form of the test statistic, underlying probability distribution, and decision rule.

5. (30 points) The COVID-19 pandemic has resulted in the widespread implementation of extraordinary physical distancing interventions, with potential consequences for mental health. A study examined the relationship between age and anxiety levels, with the latter being assessed using an Anxiety score where higher scores indicate greater anxiety. The data are given below.

Age group	Anxiety score
18–39 years ($i = 1$)	64, 57, 38
40–64 years ($i = 2$)	55, 48, 52, 41
65 years and over ($i = 3$)	28, 32, 36

Using the one-way fixed effects model $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, where $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and τ_i is the effect for the ith age group, the ANOVA table is:

Source	DF	SS	MS	F	p
Age	2	762.9	381.5	5.298	0.040
Error	7	504.0	72.0		

(a) (10 points) Write down the normal equations for this model.

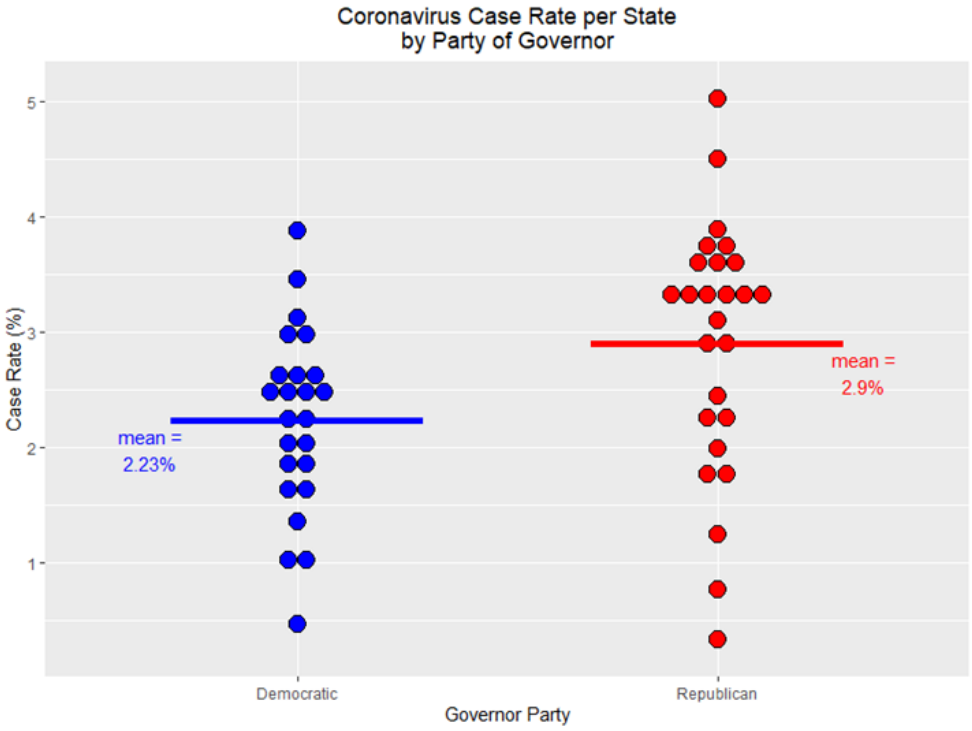
(b) (10 points) Using a level–5% test, determine whether there is a difference between the mean Anxiety score for the Age 65+ group and the average of the mean Anxiety scores for the Age 18–39 and Age 40–64 groups.

(c) (10 points) Compare the mean Anxiety scores for the three age groups by using the Bonferroni method of multiple comparisons to compute simultaneous confidence intervals for $\tau_1 - \tau_2$, $\tau_2 - \tau_3$ and $\tau_1 - \tau_3$. Use overall confidence level 85%, and provide an interpretation statement.

6. (10 points) Suppose we are interested in determining whether Villanova students get enough sleep each night. According to the Centers for Disease Control and Prevention, an adult needs at least 7 hours of sleep per night. If the true mean number of hours of sleep per night of Villanova students is 6, what is the power to determine that the true mean number of hours of sleep per night for Villanova students is less than 7 based on a sample size of 40? Assume that the standard deviation of hours of sleep per night is 1.4 hours and that the test will be conducted at the 5% level of significance.

7. (30 points) Recent data on cases of COVID-19 compared the case rate (percent of the population that has tested positive) of states with Democratic governors to states to Republican governors. A table with summary statistics and a graph of the data can be found below.

	n	$\overline{y}$	s
Democratic	24	2.23	0.80
Republican	26	2.90	1.10



(a) (15 points) Conduct the appropriate hypothesis test at the 5% level of significance to determine whether the mean case rate is different in states based on the party of the governor of the state.

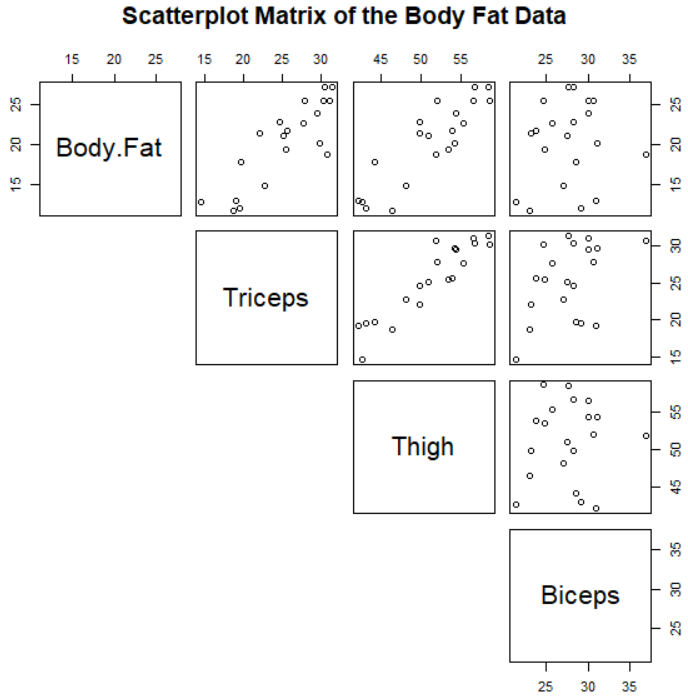
(b) (10 points) Conduct a test to determine if the variance in case rates is different in states with Republican versus Democratic governors.

(c) (5 points) Is it possible that the assumptions for the test in part (a) are valid but not in part (b)? Explain in context.

8. (40 Points) Hydrostatic weighing, or underwater weighing, is the most accurate way to determine body fat. However, this can be an expensive and time consuming (and wet!) way to determine body fat. In addition, places that perform hydrostat weighing can be difficult to find. The goal of this analysis is to determine if simple caliper measurements can be used to effectively predict body fat obtained from hydrostatic weighing. The following data were collected to determine if there is a relationship between body fat measurements taken at three locations on the body and the person's official body fat. The variables are:

Variable	Description
BODY FAT	Response Variable – Individual's official body fat percentage
TRICEPS	Body fat (mm) measured with a caliper at the triceps
THIGH	Body fat (mm) measured with a caliper at the thigh
BICEPS	Body fat (mm) measured with a caliper at the biceps

The following is a scatterplot matrix of the data in which each scatterplot presents the relationship between a pair of variables:



The following is a correlation matrix of the data:

```
          Body.Fat   Triceps     Thigh    Biceps
Body.Fat 1.0000000 0.8432654 0.8780896 0.1424440
Triceps  0.8432654 1.0000000 0.9238425 0.4577772
Thigh    0.8780896 0.9238425 1.0000000 0.0846675
Biceps   0.1424440 0.4577772 0.0846675 1.0000000
```

The following are the results of the full model along with the variance inflation factors:

```
Call:
lm(formula = Body.Fat ~ Triceps + Thigh + Biceps)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
Triceps        4.334      3.016   1.437    0.170
Thigh         -2.857      2.582  -1.106    0.285
Biceps        -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared:  0.8014,    Adjusted R-squared:  0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

Variance Inflation Factors:

```
Triceps    Thigh   Biceps
708.8429 564.3434 104.6060
```

After the model fitting process, it was determined that a simple linear regression using only the explanatory variable Triceps was the best model. The following are the results of the best model:

```
Call:
lm(formula = Body.Fat ~ Triceps)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1195 -2.1904  0.6735  1.9383  3.8523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4961     3.3192  -0.451    0.658
Triceps       0.8572     0.1288   6.656 3.02e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.82 on 18 degrees of freedom
Multiple R-squared:  0.7111,    Adjusted R-squared:  0.695
F-statistic:  44.3 on 1 and 18 DF,  p-value: 3.024e-06
```

The following are the raw residuals, hat values, studentized residuals, and PRESS residuals from the best model:

	resids	hat	student	press
1	-3.3190323	0.12028776	-1.27672766	-1.45130146
2	3.1235978	0.05076346	1.14690167	1.20823590
3	-6.1195212	0.11070971	-2.66219494	-2.99361746
4	-3.9480534	0.09214389	-1.52229542	-1.67680253
5	-1.9761577	0.13030800	-0.74205430	-0.85323804
6	1.2521300	0.05018152	0.44537195	0.46890217
7	1.6804482	0.12748590	0.62716326	0.71880014
8	2.9806010	0.06404591	1.09888594	1.17408103
9	3.8522828	0.07142552	1.46182821	1.57427136
10	-1.0621514	0.05007931	-0.37716115	-0.39704488
11	0.2376042	0.12004580	0.08731646	0.09922842
12	2.6376347	0.10414566	0.98760933	1.10242177
13	-2.8332831	0.14099570	-1.08975771	-1.26862894
14	2.4095304	0.11552793	0.90398282	1.02205921
15	1.7811816	0.28902790	0.73966476	1.04035694
16	0.1091026	0.08670618	0.03934815	0.04308378
17	0.3520383	0.06196427	0.12533014	0.13360913
18	1.0090720	0.09997820	0.36803891	0.40892222
19	-3.1620291	0.06415438	-1.17107693	-1.25135695
20	0.9950046	0.05002300	0.35312648	0.37172108

a. (12 Points) Write the mathematical form of the full model and state the assumptions. Be sure to explain how you would assess the validity of each assumption.
b. (10 Points) Comment on the presence of multicollinearity in the full model. Please identify four items in the output that point to the presence or lack of multicollinearity in the data.
c. (10 Points) Construct and interpret a 95% confidence interval for the slope in the best model.
d. (8 Points) Comment on the presence of outliers or influential observations in the best model. Briefly explain how you would determine if any such flagged observations have too much influence on the model.