Name: _____ Spring, 2019

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. (15 points) A businessman who flies frequently from New York to Philadelphia and also from New York to Boston believes that the in-air flight times are more variable for the Philadelphia flights (despite similar average in-air flight times). He carefully records the in-air flight times (in minutes) for his next ten flights to each location, and finds the following, where $\overline{y}$ and s represent the mean and standard deviation in the flight times to each location:

Destination	n	$\overline{y}$	s
Philadelphia	10	34.0	4.5
Boston	10	39.0	3.0

   (a) (10 points) Can the businessman conclude at the 0.05 significance level that the standard deviation in flight times is greater for flights to Philadelphia?

   (b) (5 points) What assumptions are needed in order for the the test in part (a) to be reasonable?

2. (30 points) A researcher was interested in comparing the yields (in bushels per acre) of four varieties (A, B, C, and D) of wheat. Each variety was planted in five randomly-assigned plots. The sum of squares for error was 3600, and the variety means were 95, 85, 60, and 80 (for A, B, C, and D, respectively).

   (a) (10 points) Provide a complete ANOVA table and do the appropriate test at level 0.05.

   (b) (10 points) Compute a 95% confidence interval for the difference between the mean for variety A and the average of the means for the other three varieties.

   (c) (10 points) Write down a complete set of orthogonal contrasts in the variety effects.

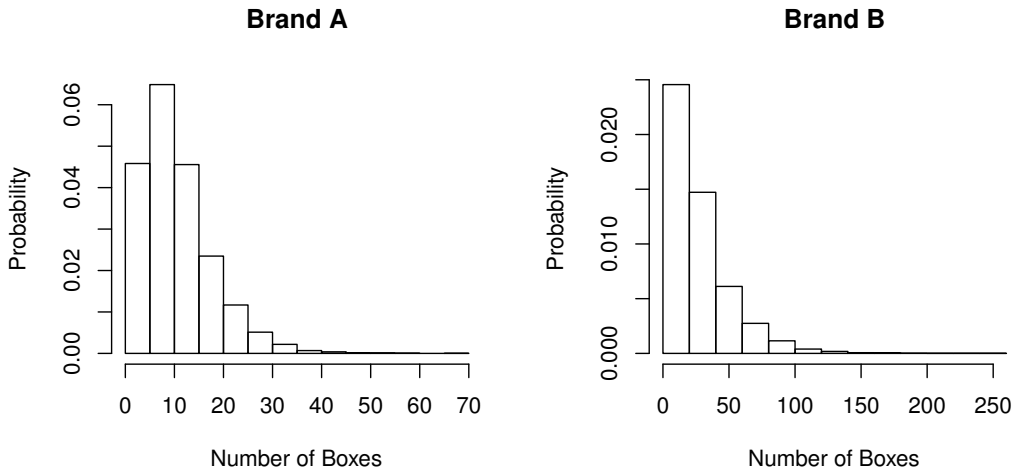
3. (18 points) A local grocery store owner is considering replacing the brand of frozen pizza they currently carry on the shelves, "Brand A," with a different brand of pizza, "Brand B." The owner decides she will make the switch to Brand B if there is sufficient evidence to indicate that there will be at least a 5 pizza box per day increase in sales, on average. During a 40 day trial period, she randomly chooses 20 days to sell Brand A and 20 days to sell brand B. She collects the following data, where $\overline{y}$ and s represent the mean and standard deviation in the number of pizza boxes sold per day:

Brand	n	$\overline{y}$	s
A	20	10.5	7.0
B	20	26.9	23.3

(a) (10 points) Create a 95% confidence interval for the difference in the mean number of Brand A and Brand B boxes sold each day. State clearly any assumptions that you make.

(b) (5 points) Use the confidence interval from part (a) to decide on a conclusion for the test of main interest at the 0.05 level. Briefly describe how you arrived at the conclusion and state the conclusion in terms of the problem.

(c) (3 points) Suppose that the histograms below show the true distributions of the number of pizza boxes sold per day for Brands A and B. Comment on the validity of the confidence interval and test in parts (a) and (b).



4. (30 points) An experiment was run to study the effect of two different levels of factor A and three different levels of factor B on a response variable. The data that were obtained are given in the table below. The researchers fit a fixed effects model without interaction.

   (a) (10 points) Write down the complete mathematical model, including all assumptions and explaining all terms.

   (b) (10 points) Find the design matrix X and the normal equations.

   (c) (10 points) Let μ be the overall mean and β_j the effect of the jth level of factor B. Determine whether each listed expression is estimable or not. Justify your answers. (i) $\mu + \beta_2$ (ii) $\beta_2 - \beta_3$

	Factor B		
Factor A	1	2	3
1	4	3	2
2	1	2, 6	10

5. (22 points) The average tax refund among all U.S. tax payers for the 2017 tax year was \$2,825. We are interested in determining whether this year's refunds are different from this value. Suppose that you are an accountant for a small tax preparation service in a major metropolitan area, and that it is currently very early in the tax season. We would like to use all currently submitted tax returns through your service to determine whether there is convincing evidence that the average refund for the 2018 tax year will be different from the 2017 tax year.

   (a) (5 points) Explain why the data collection procedure described above might be problematic with regard to answering this question.

   (b) (10 points) Suppose that the population standard deviation of all tax refunds is \$5,000, and that 250 tax returns have been submitted through your service thus far. Early estimates suggested that the average tax refund for the 2018 tax year would be approximately \$1,825. If this value is indeed the true average tax return for the 2018 tax year, calculate the power to detect a difference from \$2,825 with a sample of 250 tax refund at $\alpha = 0.05$.

   (c) (7 points) In principle, it would be possible to do a paired t-test to answer this question. Explain why the setup described above is NOT for a paired t-test, and also how you would do a paired t-test in this situation if you had the required data for it.

6. (25 points) An organization called Burtch Works Executive Recruiting has been interested in trends regarding the statistical/data science tools of SAS, R, and Python. In a survey conducted by them, a random sample of respondents were asked which of these three tools they preferred, and how many years of professional experience that they have.

   Among those with 0 through 5 years of experience, they found the following counts regarding which tool was the most preferred for each individual:

   - SAS: 37
   - R: 102
   - Python: 129

   Among those with 6 through 15 years of experience, they found the following counts regarding which tool was the most preferred for each individual:

   - SAS: 165
   - R: 179
   - Python: 154

   (a) (5 points) In order to determine whether there is evidence of an association between preferences and years of experience based on these data, what statistical test would you perform? Verify whether the required conditions for this test are satisfied.

   (b) (15 points) At $\alpha = 0.05$, perform the statistical test that you stated in the previous part.

   (c) (5 points) Can this survey be used to determine whether there is evidence that having more years of experience causes individuals to have a different preference? Why or why not?

7. (60 points) In a statistical analysis, we consider the 1982 data from the Panel Study on Income Dynamics (PSID). The information was collected from a nationally representative sample of 595 individuals. The purpose of this analysis is to build a model with `log(wage)` (natural log of monthly wage; wage in dollars) as the response and 7 potential explanatory variables: `education` (years of education), `experience` (years of full time work experience), `weeks` (weeks worked per year), `married` (married or not), `gender` (male or female), `union` (the individual's wage is set by a union contract or not), and `ethnicity` (African American or not).

$$\begin{aligned}\texttt{log(wage)} =& \beta_0 + \beta_1\texttt{education} + \beta_2\texttt{experience} + \beta_3\texttt{weeks} + \beta_4\texttt{married} \\ &+ \beta_5\texttt{gender} + \beta_6\texttt{union} + \beta_7\texttt{ethnicity} + \epsilon.\end{aligned}$$

Use the output to answer the following questions.

(a) (10 points) Give a 95% confidence interval for β_1 and interpret the interval in context of the problem.

(b) (10 points) Interpret the estimated coefficient associated with `union` on the original scale of wage (not log wage).

(c) (10 points) If possible, conduct the test of $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at $\alpha = 0.05$. If not possible, clearly explain how you would conduct the test and explain what distribution your test statistic follows.

(d) In the output for this problem, several diagnostic plots and VIF values are provided.

i. (10 points) Explain what assumptions of the model you can evaluate using plots **(a)**, **(b)** and **(c)**. Clearly identify which plot(s) can be used to check each assumption. Explain if there appear to be any concerns with violations of those assumptions.

ii. (5 points) Based on plot (d), do you have any concern about the fitted model? If so, discuss your concerns. If not, explain why not. For the same diagnostics purpose, provide an alternative method to plot (d).

iii. (5 points) Based on the VIF values, do you have any concern about the fitted model? If so, discuss your concerns. If not, explain why not. For the same diagnostics purpose, provide an alternative method to the VIF values.

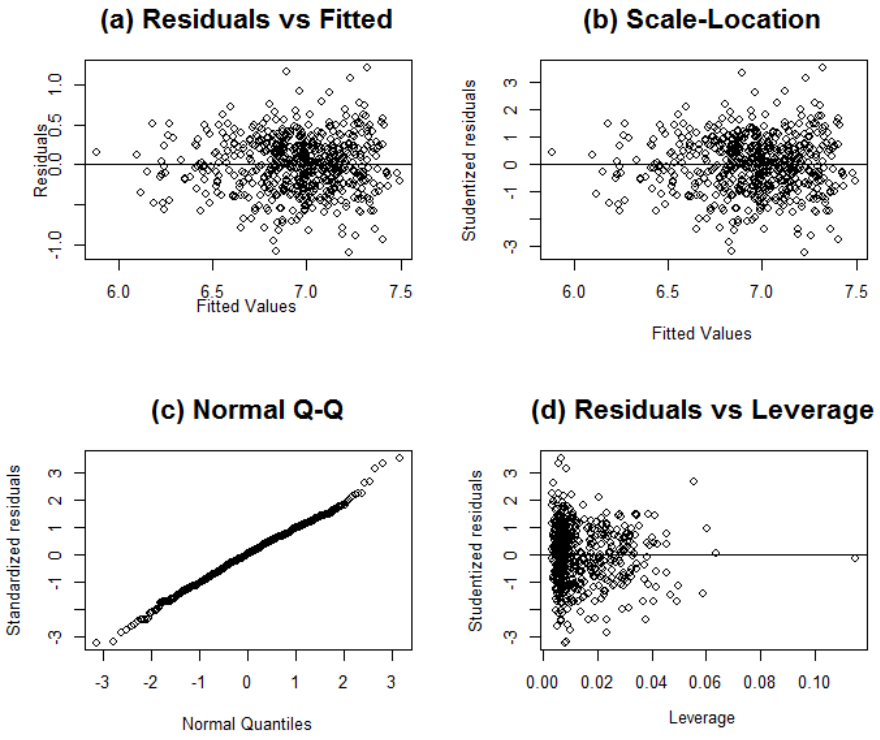
(e) (10 points) Given the model output, a fellow researcher decides to drop both variables, `weeks` and `marriedyes`, from the current model using $\alpha = 0.05$. Do you agree with her? If so, justify her decision. If not, explain your concerns.

```
> summary(model)
Call:
lm(formula = logwage ~ education + experience + weeks + married +
    gender + ethnicity + union, data = PSID1982)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09654 -0.23459  0.01556  0.23422  1.21751

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.513941   0.171859  32.084  < 2e-16 ***
education    0.080381   0.005496  14.624  < 2e-16 ***
experience   0.006795   0.001387   4.899 1.25e-06 ***
weeks        0.003870   0.002838   1.364 0.173212
marriedyes   0.096087   0.052055   1.846 0.065412 .
genderfemale -0.309227   0.064442  -4.799 2.03e-06 ***
ethnicityyes -0.175167   0.057116  -3.067 0.002263 **
unionyes     0.107387   0.031660   3.392 0.000741 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3474 on 587 degrees of freedom
Multiple R-squared:  0.3795,Adjusted R-squared:  0.3721
F-statistic: 51.29 on 7 and 587 DF,  p-value: < 2.2e-16
```




```
> vif(model)
   education   experience        weeks   marriedyes genderfemale ethnicityyes     unionyes
    1.157518     1.102745     1.065823     2.096836     2.045924     1.078382     1.147306
```