STAT 515 fa 2020 Final Exam

Karl B. Gregory

This is a take-home test due to COVID-19. Do not communicate with classmates about the exam until after its due date/time. You may

- *Use your notes and the lecture notes.*

- *Use books.*

- *NOT work together with others.*

Write all answers on blank sheets of paper; then take pictures and merge to a PDF. Upload a single PDF to Blackboard.

1. Copy down this sentence on your answer sheet and put your signature underneath: *I have not collaborated with any other student on this exam. The work I have presented is my own.*

2. Suppose an experiment involving crop yields under three conditions (control, treatment 1, and treatment 2) resulted in the following table of observations:

Control	1.89	3.25	4.37	1.95	1.41	2.63	2.79			
Treatment 1	2.32	3.49	3.02	2.38	2.56	4.47				
Treatment 2	5.06	4.28	3.37	3.30	3.52	4.42	3.80	4.20	2.95	3.06

The data can be read into R using these commands:

```
Y <- c(1.89, 3.25, 4.37, 1.95, 1.41, 2.63, 2.79,
       2.32, 3.49, 3.02, 2.38, 2.56, 4.47,
       5.06, 4.28, 3.37, 3.30, 3.52, 4.42, 3.80, 4.20, 2.95, 3.06)
trt <- as.factor(c(rep("ctr",7),rep("trt1",6),rep("trt2",10)))
```

Some analyses of the data resulted in these plots:
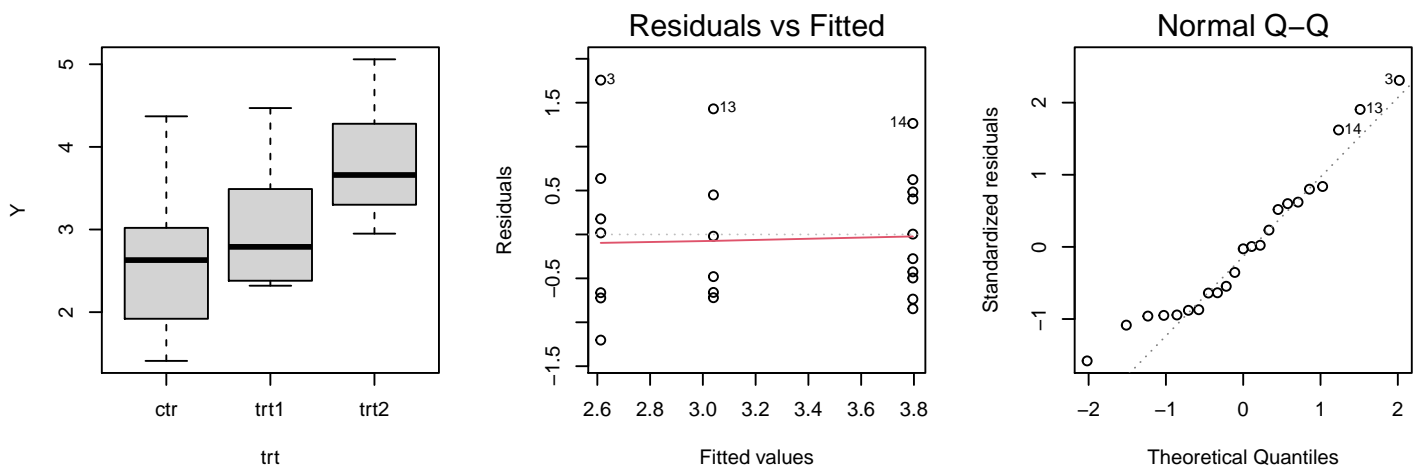
(a) Use the `lm()` and `anova()` functions in R (remember that you may go back and watch the recorded lectures) to get the value of the F-statistic.

(b) Write down the hypotheses with which the F-statistic is concerned.

(c) State the assumptions of the cell means model and say whether they appear to be satisfied.

(d) Assuming the assumptions are satisfied, what is your conclusion about the the control, treatment 1, and treatment 2

    i. at the $\alpha = 0.01$ significance level?

    ii. at the $\alpha = 0.05$ significance level?

(e) In the cell means model

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma_\varepsilon^2), \quad i = 1, \ldots, 3, \quad j = 1, \ldots, n_i,$$

what is the estimate of σ_ε^2 based on the data? Use the `lm()` and `anova()` functions.

3. To compare the speed of two functions for extracting the minimum value from a list of numbers, the two functions were executed 50 times on randomly generated sequences of 1,000 numbers, and the computation times were recorded in milliseconds. The computation times for the two functions are plotted in the side-by-side boxplots below:
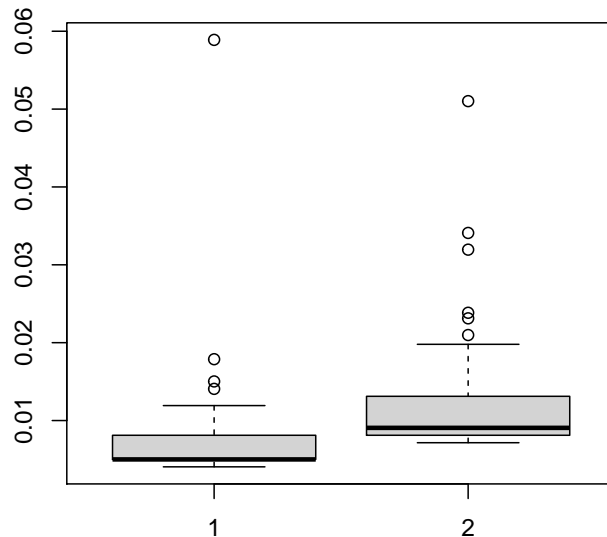
The means and standard deviations of the computation times were $\bar{X}_1 = 0.007482$, $\bar{X}_2 = 0.012412$, $S_1 = 0.008034$, and $S_2 = 0.00824$. Assume that the population variances are equal.

(a) To test whether either function is faster than the other, what are the hypotheses of interest?

(b) Give the value of S_{pooled}.

(c) Give the value of the test statistic for testing the hypotheses in part (a).

(d) Give a 99% confidence interval for the difference in means $\mu_1 - \mu_2$. State whether you have assumed Normality for the two populations.

(e) State your conclusion about the hypotheses in part (a) at the $\alpha = 0.01$ significance level.

4. The shucked weight Y and the cubed diameter x (diameter raised to the power 3) of 4,176 abalones were recorded. A linear regression of the shucked weights on the cubed diameters resulted in the following plots:
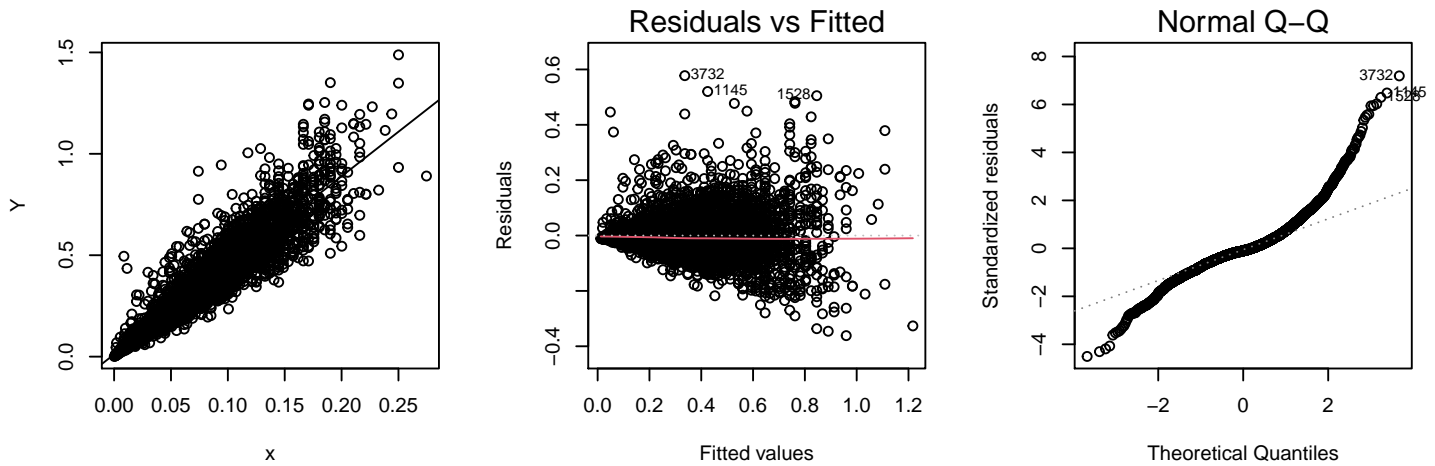
Consider the regression model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, 4176,$$

where $\varepsilon_1, \ldots, \varepsilon_{4176} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$. The output of the lm() function is

```
Call:
lm(formula = Y ~ x)

Residuals:
     Min       1Q    Median       3Q      Max
-0.36153 -0.03921 -0.00799  0.03051  0.57757

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.011132   0.002434   4.574 4.92e-06 ***
x           4.390750   0.026378 166.455  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08033 on 4174 degrees of freedom
Multiple R-squared:  0.8691,        Adjusted R-squared:  0.869
F-statistic: 2.771e+04 on 1 and 4174 DF,  p-value: < 2.2e-16
```

(a) State one way in which the assumptions of the model are not satisfied. Explain your answer.

(b) State another way in which the assumptions of the model are not satisfied. Explain your answer.

(c) State an assumption of the model that *does* appear to be satisfied. Explain your answer.

(d) Supposing that you could trust the model—that is, supposing the assumptions were satisfied—use the output (shown) of the `lm()` function to give:

    i. The change in the expected the shucked weight of abalones due to a one-unit increase in the cubed diameter.

    ii. The estimated expected shucked weight of abalones with cubed diameter equal to 0.15.

5. The table below divides 270 patients according to what type of chest pain they experienced and whether they have been diagnosed with heart disease:

	typ. angina	atyp. angina	non-anginal pain	asymp.
fasting blood sugar $\leq$ 120 mg/dl	15	37	62	116
fasting blood sugar $>$ 120 mg/dl	5	5	17	13

The data are taken from `https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29`.

(a) Write down the hypotheses of interest to a researcher studying possible relationships between fasting blood sugar levels and the type of chest pain experienced by heart patients.

(b) Give the table containing the expected numbers of patients in each cell if there were no relationship between the fasting blood sugar level and the type of chest pain.

(c) Give the value of Pearson's chi-squared test statistic.

(d) Give the degrees of freedom of the relevant chi-squared distribution for Pearson's test.

(e) State whether the assumptions are satisfied for conducting Pearson's chi-squared test for association. Explain your answer.

(f) Supposing that the test is valid, what is your decision at the $\alpha = 0.05$ significance level? Is there evidence of an association between the fasting blood sugar level and the type of chest pain?

6. Suppose you put an ad on the internet and that 0.50% of those who see it will click on it.

(a) Suppose 10,000 people see your ad. Find:

    i. The expected number of people who will click on it.

    ii. The exact probability that 60 or more people will click on it.

    iii. The approximate probability that 60 or more people will click on it based on the central limit result

$$\hat{p}_n \overset{\text{approx}}{\sim} \text{Normal}\left(p, \frac{p(1-p)}{n}\right) \quad \text{for large } n.$$

(b) Give the exact probability that at least one person out of 100 who see it will click on it.

(c) Suppose you devise a new ad and that after 1,000 people have viewed it, 19 have clicked on it. Based on this data, build a 95% confidence interval for the proportion of those who will click on your new ad if they see it.