# 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.*

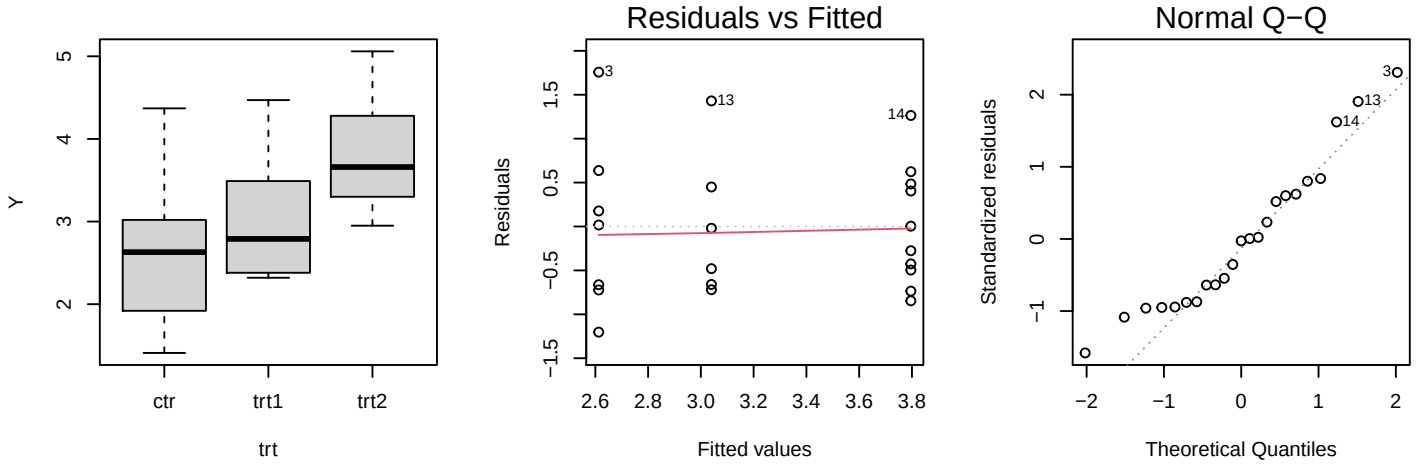
2. Suppose an experiment involving crop yields under three conditions (control, treatment 1, and treatment 2) resulted in the following table of observations:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 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.

The commands `lm.out <- lm(Y~trt)` and `anova(lm.out)` gives the ANOVA table, which is

```
        Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value  Pr(>F)
trt        2  6.0845 3.04224  4.5031 0.02429 *
Residuals 20 13.5118 0.67559
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From this we see that the $F$-statistic is equal to 4.5031.

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

> The $F$-statistic is measuring evidence against the null hypothesis in the pair of hypotheses
>
> $$H_0\text{: } \mu_1 = \mu_2 = \mu_3 \text{ versus } H_1\text{: Not all means are equal,}$$
>
> where $\mu_1$, $\mu_2$, and $\mu_3$ are the means of the Control, Treatment 1, and Treatment 2 groups.

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

> The assumption of equal variances in the treatment groups appears to be satisfied, because the spreads of the columns of points in the residuals versus fitted values plot are roughly equal.
>
> The assumption that the response values are Normally distributed around the treatment means appears to be satisfied because the Normal Q-Q plot of the residuals shows the points falling close to a straight line.
>
> The assumption that the response values are independent cannot be checked by looking at the data; we must trust to the experimental design.

(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?

> The $p$-value corresponding to the $F$-statistic, which is printed by calling `anova(lm.out)`, is equal to 0.02429, which is greater than 0.01, so we fail to reject the null hypothesis of equal treatment means at the 0.01 significance level.

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

> The $p$-value of 0.02429 is less than 0.05, so we *would* reject the null hypothesis of equal treatment means at the 0.01 significance level.

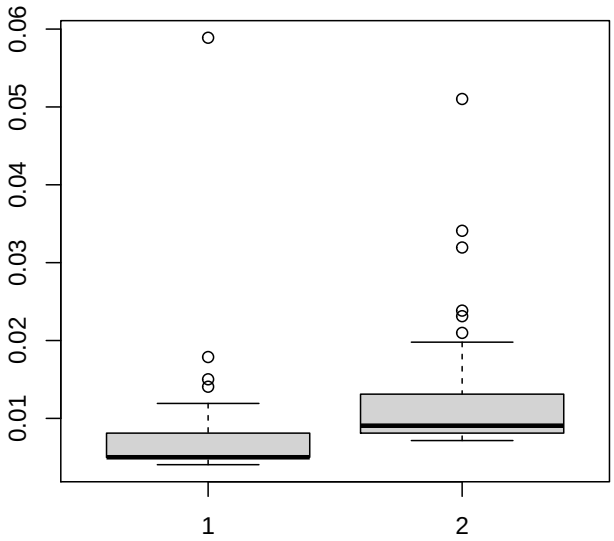
(e) In the cell means model

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

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

> This is the number given in the ANOVA table as the $\text{MS}_{\text{Error}}$. It is the value 0.67559.

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?

We should test

$$H_0\colon\ \mu_1 - \mu_2 = 0 \text{ versus } H_1\colon\ \mu_1 - \mu_2 \neq 0,$$

where $\mu_1$ and $\mu_2$ represent the mean computation times for the two functions.

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

We have

$$S_{\text{pooled}} = \sqrt{\frac{(50-1)(0.008034)^2 + (50-1)(0.00824)^2}{50+50-2}} = 0.008137747$$

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

We have

$$T_{\text{test}} = \frac{0.007482 - 0.012412}{\sqrt{(0.008137747)^2(1/50 + 1/50)}} = -3.029129$$

(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.

We have, assuming non-Normality, the interval

$$0.007482 - 0.012412 \pm \underbrace{z_{0.05/2}}_{2.575829} \cdot 0.008137747\sqrt{1/50 + 1/50} = (-0.00912224, -0.0007377596).$$

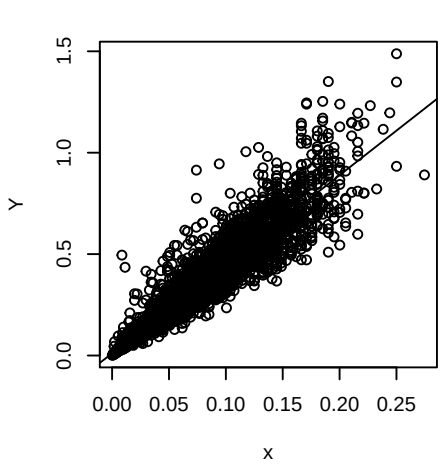
We could also use $t_{98,.01/2} = \texttt{qt(.995,98)} = 2.626931$, which gives the interval

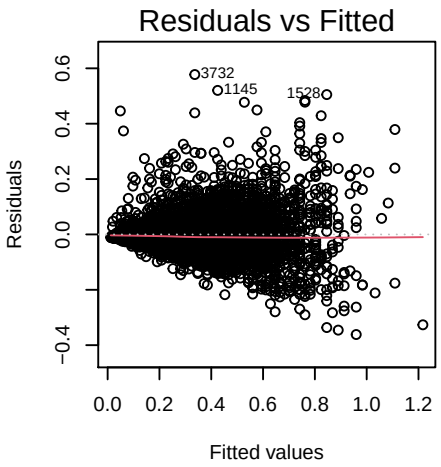
$$(-0.00920541, -0.0006545898).$$

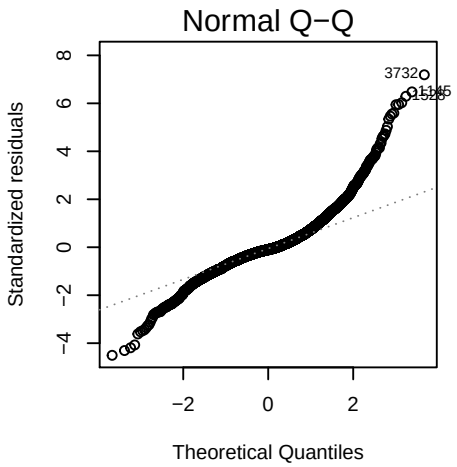
(e) State your conclusion about the hypotheses in part (a) at the $\alpha = 0.01$ significance level.

Since the 99% confidence interval for $\mu_1 - \mu_2$ does not contain 0, we reject the null hypothesis that $\mu_1 - \mu_2 = 0$. We therefore conclude that one of the functions is faster than the other.

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.

The Normal Q-Q plot of the residuals indicates that the responses are not Normally distributed around the regression line.

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

The residuals vs fitted values plot shows that the variability of the responses around the regression function is not constant.

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

The relationship between the shucked weight and the cubed diameter does appear to be linear.

(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.

This is the slope coefficient, which is estimated to be 4.390750.

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

This would be the value $0.011132 + 4.390750(0.15) = 0.6697445$.

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.

It is of interest to test

$H_0$: There is no association between the fasting blood sugar level and the type of chest pain

versus the alternative that there *is* an association.

(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.

The expected table is given by

| | typ. angina | atyp. angina | non-anginal pain | asymp. |
|---|---|---|---|---|
| fasting blood sugar $\leq$ 120 mg/dl | 17.04 | 35.78 | 67.30 | 109.89 |
| fasting blood sugar $>$ 120 mg/dl | 2.96 | 6.22 | 11.70 | 19.11 |

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

The value can be obtained with the `chisq.test()` function in R. It is 7.0334.

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

The degrees of freedom is 3.

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

The assumptions are *not* satisfied; we require, as a rule of thumb, that all of the expected counts must be greater than or equal to 5. We have one expected count of 2.96, so this requirement is not satisfied.

(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?

If the test is valid, then the $p$-value is 0.07084, so we would fail to reject the null hypothesis. There is *not* sufficient evidence in these data to claim that there is a relationship between whether the fasting blood pressure of a patient exceeds 120 and the type of chest pain the patient experiences.

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.

Regarding the number of people out of 10000 who click on the ad as a random variable, say $X$, we have $X \sim \text{Binomial}(10000, 0.005)$, so that

$$\mathbb{E}X = (0.005)10000 = 50.$$

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

This is

$$\begin{aligned} P(X \geq 60) &= 1 - P(X \leq 59) \\ &= 1 - \sum_{x=0}^{59} \binom{10000}{x} (0.005)^x (1-0.005)^{10000-x} \\ &= \texttt{1 - pbinom(59,10000,0.005)} \\ &= 0.09172118. \end{aligned}$$

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

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

We have that $\hat{p}_n \geq .006$ is equivalent to $X \geq 60$. The approximate probability of this is given by

$$\begin{aligned} P(\hat{p}_n > 0.006) &= P\left(Z > \frac{0.006 - 0.005}{\sqrt{0.005(1-0.005)/10000}}\right) \\ &= P(Z > 1.417762) \\ &= \texttt{1 - pnorm(1.417762)} \\ &= 0.07813013. \end{aligned}$$

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

We now have $X \sim \text{Binomial}(100, 0.005)$ and

$$P(X \geq 1) = 1 - P(X = 0) = 1 - (1 - 0.005)^{100} = 0.3942296.$$

(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.

We have $\hat{p}_n = 0.019$, and the 95% confidence interval is given by

$$0.019 \pm 1.96\sqrt{0.019(1-0.019)/1000} = (0.01053811, 0.02746189).$$