**STAT 21 SPRING 2019** **HOMEWORK 2** **DUE TUES FEB 19**

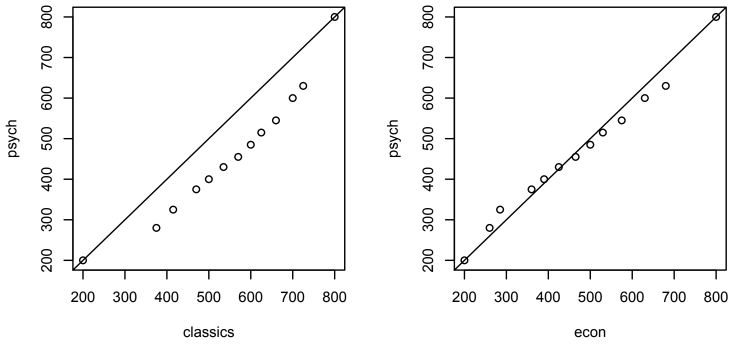
Homework policies reminder: Late homework is not accepted, with one exception: you may hand in an assignment late (by the beginning of the next class meeting) *once* during the semester. If you are planning to hand in a homework late, email me by the beginning of class on the original due date.

Your assignments should be typed, except for equations, symbols, and graphics, which may be handwritten. Please integrate graphics into the flow of your text; do not put all your graphics and tables in an appendix at the end of your written material.

I encourage you to discuss the homework with other students and with me. Your final answer should be written in your own words. On your paper, list the students with whom you had substantial collaborations.

### 1. Q-Q Plots (8 points)

**(a)** The following are quantile-quantile plots of GRE General Test Verbal scores for students intending graduate study in psychology, classics, and economics. Briefly describe the pattern in each of the Q-Q plots. (One sentence for each plot should suffice.)



**(b)** Suppose we have the SAT verbal and quantitative scores for all current Swarthmore students. How do the following two plots differ: (1) a Q-Q plot of the verbal and quantitative score distributions, and (2) a scatterplot plotting the verbal and quantitative scores for each student? What question is answered by each plot?

### 2. ANOVA Table (13 points)

The following is part of an ANOVA table from an ANOVA with 50 observations total and 2 groups.

| Source | degrees of freedom | Sum of Squares | Mean Square | F-ratio |
|---|---|---|---|---|
| Between | ______ | ______ | ______ | ______ |
| Within | ______ | ______ | 108.1 | |
| Total | ______ | 12463 | | |

**(a)** Fill in the seven blanks in the table above.

**(b)** Calculate $R^2$ for these data.

**(c)** Calculate $s$, the conditional standard deviation of Y.

**(d)** Calculate $s_Y$, the marginal standard deviation of Y.

### 3. Elephant Ivory (15 points)

Elephants have declined substantially in population, with losses estimated at 50–75% over the last half-century. A major reason for this decline is the killing of elephants for their ivory (another is habitat loss). Conservation officials would like to be able to trace the source of ivory in order to determine if it was legally or illegally obtained, and to determine where poaching might be taking place. However, this is difficult to do solely by visual inspection of ivory samples.

A dataset on stable isotope ratios of five elements was collected on ivory samples from Asia, West Africa, Central Africa, East Africa, and Southern Africa. The goal is to see if these isotope ratios differ in ivory from different regions; if so, then this information may be useful in locating the origin of the ivory. The dataset contains information on the country and region that 495 samples of ivory were obtained from; each sample's ratios of carbon, nitrogen, oxygen, hydrogen, and sulfur; and the latitude and longitude where the ivory was obtained.

Source: S. Ziegler, S. Merker, B. Streit, M. Boner, D.E. Jacob (2016): Towards understanding isotope variability in elephant ivory to establish isotopic profiling and source-area estimation. *Biological Conservation* 196, pp. 154-163.

**(a)** Download the dataset `ivory.csv` from Moodle. Then read the dataset into R using the following command:

```
data <- read.csv("~/Desktop/ivory.csv", header=T)
```

(You may have to change the path to the file, depending on where you saved the downloaded file.) Verify that your dataset has 495 rows and 9 columns. In this assignment, we will look only at the region of origin and carbon-13 isotope ratio. Define variables for region and origin as follows:

```
region <- data[,"Region"]
carbon <- data[,"delta13C"]
```

**(b)** We want to explore whether carbon ratios differ by region. Make boxplots of carbon by region using the following command:

```
boxplot(carbon ~ region)
```

Does ivory from different regions appear to vary in its carbon ratio? Do the assumptions of normality and equal variances for ANOVA appear to be violated?

**(c)** Fit an ANOVA model to the data using the following command:

```
fit1 <- lm(carbon ~ region)
anova(fit1)
```

Is there a significant difference in carbon ratios by region?

**(d)** If there is a significant different among regions, which pairs of regions are significantly different? Use the following commands to carry out a post-hoc test of means between pairs of regions:

```
posthoc <- TukeyHSD(aov(fit1))
posthoc
plot(posthoc)
```

You may have to re-size the plot window to see more of the y-axis labels. Which regions are significantly different?

**(e)** In part **(b)** we assessed the assumption of normality by eyeballing the symmetry of boxplots. Now let's use an NPP to accomplish the same goal. A strength of the NPP is that it can reveal subtle departures from normality, so perhaps we will be able to catch something that is not easily seen from the boxplots alone.

Rather than make separate NPPs for each region, it is typical to examine an NPP of the *residuals* of the entire dataset. Recall from your introductory stat course that a residual is defined as (*actual value – predicted value*) for an observation. Here the predicted value of an ivory sample is just the mean of the region corresponding to that sample. Subtracting off the mean does not change the distribution of the ivory samples; it merely shifts the distribution to the left or the right. Thus, if the data in each region are normally distributed, then the residuals will be as well. Furthermore, we can aggregate the regions together because if each region is normally distributed with the same variance, then combining the regions will result in a normal distribution with that variance.

To extract the residuals from an `lm` model object, use the `residuals()` command. Then make an NPP of the residuals and add the 45-degree line. Do the residuals appear normally distributed?

**(f)** Calculate R-squared. Would you say this is a relatively high value, or relatively low?

**(g)** Based on **(f)** and on your boxplots in **(b)**, would you say that an ivory sample's carbon ratio can be used to determine its region of origin? How is this question related to the difference between *statistical significance* and *effect size* that we explored in HW 1?

**(h)** Identify a simple way in which this analysis might be modified to improve its predictive power.

### 4. $F$ AND $R^2$ (6 POINTS)

In a simple linear regression, the $F$-statistic (the $F$-ratio) and $R^2$ both measure the strength of the relationship between Y and X. It should come as no surprise that the two quantities are related. For the case of $k = 2$ groups, demonstrate this relationship by showing that

$$R^2 = F \Big/ (F + (N - k))$$

(Hint: This is a straightforward proof and can be done in about five steps. Start with the right-hand side of the equation and substitute the definition of $F$, then simplify.)

### 5. IS SCIENCE BROKEN? (6 POINTS)

Read the following article: `https://fivethirtyeight.com/features/science-isnt-broken/`
and watch this John Oliver video: `https://youtu.be/0Rnq1NpHdmw` (warning: some parts NSFW)

**(a)** Briefly define *p-hacking* and *researcher degrees of freedom*.

**(b)** Why should you not believe a finding from any single scientific study? When *should* you consider a finding to be reliable?

**(c)**What are replication studies? Why are they rare?