271B
Fall 2014
Final Exam

Overview:
This lab has a series of questions, as well as tasks using R. You will provide the answers to the questions below in a single document. You should include any key output and/or graphics in the primary report file as well. In addition, include your fully commented R script along with your submission (we should understand your answers without having to read the R script).

Submission:
Please organize all of your answers into one primary "final exam" file using a standard format (.pdf, .doc, etc). Please name this primary document "FinalExam_YourLastName.doc". Similarly, your script should be named "FinalExam_YourLastName.R"

This lab is due on Thursday, December 18th by 10am. You should email your two files (the main exam file and your R script) to Toshiro (tnish@berkeley.edu). Please plan ahead, we do not accept late papers.

Part 1: Multiple Choice (28 points)

For the following questions, please choose the best answer and provide the correct letter in your response.

Suppose you want to run an ordinary least squares regression to predict how long a contestant lasts on a reality-tv show called "Surviving in the Wild with Very Dangerous Animals" where contestants have to live on their own for a month on a deserted island with a selection of frightening wild animals such as mountain lions, bears, honey badgers, and ill-tempered ground squirrels. That is, your independent variable is Y = time spent on show. You hypothesize that math score (X1) and triceps strength (X2) are associated with how long a contestant will last on this reality-tv show. Your hypothesized model looks like this:

$$Y = b0 + b1 * X1 + b2 * X2 + \varepsilon$$

1. What statistical test should you use to test the significance of your overall model?
    a. t-test
    b. Ordinary Least Squares
    c. Chi-Square
    d. ANOVA
    e. Maximum Likelihood
    f. t-tests with correction for multiple comparisons

2. What is the null hypothesis for this test?
    a. At least one coefficient is equal to zero.
    b. All coefficients for each independent variable equal zero.
    c. b0 = 0
    d. b1 = 1
    e. b0 = b1 = b2

3. Suppose you rejected the null hypothesis for your test in Q1. What can you conclude?
    a. There are statistically significant relationships between each independent variable and the dependent variable.
    b. b0 is statistically significant
    c. b1 is not equal to zero
    d. The relationship in the underlying population is linear
    e. None of the above

4. A study of college students reveals an interesting relationship: People who use online dating sites (compared to those who do not use online dating sites) report higher levels of average relationship satisfaction. Assuming that we can measure the outcome (relationship satisfaction) with a questionnaire, which of the following is a *natural experiment* that you could use to examine the causality of this relationship?
    a. You randomly assign new college students to two conditions (use online dating, cannot use online dating) and provide access to online dating for the treatment group while restricting access to online dating for the other group.
    b. You distribute the outcome questionnaire (relationship satisfaction question) to a random sample of students at a university, where there are both online daters and non-online daters.
    c. You distribute the outcome questionnaire (relationship satisfaction question) to a non-random sample of students at a university, where there are both online daters and non-online daters.
    d. You examine a random sample of internet users in one Californian county, and compare them to a random sample of internet users in an adjacent California county where the local Internet Service Provider (ISP) has chosen to block access to online dating sites.
    e. You know from prior research that students on the west coast are more likely to use online dating than students on the east coast. So, you collect a random sample of students on the west coast and the east coast and compare them.

5. Suppose you have a random sample of Reddit.com users (N=200). What is the most appropriate way to tell if there is a difference in the average number of articles that that a person reads in their first three months versus their second three months?
    a. OLS Regression
    b. Paired samples t-test
    c. Effect size calculation
    d. Chi-square test of independence
    e. Independent samples t-test

6. You give a questionnaire about privacy concerns (all metric scales) to **every single Facebook user**—and *everyone* responds! What is the most appropriate way to tell if there is a practical difference in various privacy concerns among men and women?
    a. Chi-square test of independence
    b. OLS Regression
    c. Independent samples t-test
    d. Effect size calculation
    e. Paired samples t-test

7. Imagine that you select a random sample of hipsters (N=5) from the San Francisco bay area and find that the mean number of smooth jazz songs in their mp3 collection is 50 songs. The population mean of smooth jazz songs among hipsters is known to be 15 songs with a standard deviation of 4.3 songs. Would a sample mean this large be more likely or less likely if N=1000 hipsters?
    a. Equally likely.
    b. More likely.
    c. Less likely.
    d. There is nothing to base our guess upon; the sample means are random.

Part 2: Test Selection (24 points)

The Pew Internet and American Life Project collects survey data on a variety of topics related to online behavior. More information can be found at http://www.pewinternet.org. You will be working with a subset of data from a 2013 survey on online dating. The file name is Dating.csv. Note that you will have to properly import/read in the .csv in order to work with it in R or R Studio.

Recall that surveys are generally weighted in order to compensate for over- or under-representation of subgroups. These weights appear in the "weight" and "standwt" columns of the Pew dataset. For the sake of simplicity, however, you should ignore the weight values, and this will limit how well your findings generalize to the U.S. population.

In this section, there are several questions that apply to the Dating.csv dataset.

For each question, *select the most appropriate statistical procedure from the provided choices*, assuming that you do not meaningfully change or recode the original variables (except for dealing with missing values, if applicable). Please note that we are not asking you to conduct any tests in this section. Instead, you should examine the data and choose the most appropriate statistical procedure that you would use.

1. Is marital status (marital_status) related to using reddit (use_reddit)?
    a. chi-square
    b. t-test
    c. Pearson Correlation
    d. Wilcoxon Signed-Rank Test

2. Is the region of the country the respondent is from (region) related to his or her quality of life (life_quality)?
    a. Pearson Correlation
    b. Wilcoxon Signed-Rank Test
    c. Levene's Test
    d. ANOVA

3. Is flirting online (flirted_online) related to the number of years a respondent has spent in their relationship (years_in_relationship)?
    a. Fisher's exact test
    b. t-test
    c. Wilcoxon Signed-Rank test
    d. Multiple Regression

4. Is sexual orientation (lgbt) related to the number of adults in the respondent's household (adults_in_household)?
    a. chi-square
    b. ANOVA
    c. Pearson Correlation
    d. Wilcoxon Rank-Sum Test

5. Is the respondent's age related to the total number of children he or she has (children0_5 + children6_100 + children12_17)?
    a. Pearson Correlation
    b. Wilcoxon Rank-Sum Test
    c. Levene's Test
    d. ANOVA

6. Do 31-year-old men have more children than 31-year-old women?
    a. Pearson Correlation
    b. Wilcoxon Rank-Sum Test
    c. OLS Regression
    d. ANOVA
    e. t-test

Part 3: Short Answer (8 points)

1. What are the null and alternative hypotheses for the test in Question 2.1 (relating marital_status to use_reddit)?

2. Imagine that you have a dataset that has standardized GRE test scores for every US state. A colleague tells you that he conducted a single t-test and found that California has a statistically significantly higher test score than New York. What statistical test would you use to evaluate whether test scores differ by state? How would you evaluate all of the individual (pairwise) differences between states?

Part 4: Data Analysis (40 points)

1. OLS Regression

a. Data preparation: In this data, missing values are very obvious because they are strings such as "Refused". Properly deal with any missing values for *life_quality*, *years_in_relationship*, and *use_internet*. Note that R is treating these variables as factors (they have numbered responses, but the various missing codes are text strings). As we have seen in the textbook and R workshops, remember that you can use the as.numeric expression to tell R to treat factors as numeric variables after you deal with the missing values.

b. The *life_quality* variable measures quality of life on a 5-point scale, where 1 = excellent and 5 = poor. We would prefer, however, for higher numbers to be better. Reverse the scale for the variable so that 5 = excellent and 1 = poor. After you reverse the coding, what is the mean quality of life in the sample?

c. The *years_in_relationship* variable measures how long a respondent has spent in their current relationship. As you recode this variable, you may find that R converts each text string to the wrong number. For example, the string "0" may be converted to 2 or some other number (this happens because R's as.numeric function returns factor levels if they're available). If this happens, convert the variable to a character string before converting it to a numeric vector, as in the following expression:

```
as.numeric(as.character(Dating$years_in_relationship))
```

Notice that *years_in_relationship* equals zero for respondents that are not currently in a relationship. You should leave these values in the dataset for the purposes of this lab because that is a valid response. What is the mean of *years_in_relationship* in the sample?

d. To run a nested regression in R, your first step will be to select just the rows in your dataset that have no missing values in your final OLS model (see class R example from lecture). In this case, you will want just the rows that have non-missing values for *life_quality*, *years_in_relationship*, and *use_internet*. How many cases does this leave you with?

e. Fit an OLS model to the data from the previous step that predicts *life_quality* (dependent variable) as a linear function of *years_in_relationship* (independent variable). What is the slope coefficient you get? Is it statistically significant? What about practically significant?

f. Evaluate the regression diagnostics from your model.

g. Now fit a second OLS model to the data. Keep *life_quality* as your dependent variable, but now use both *years_in_relationship* and *use_internet* as your explanatory variables. What is the slope coefficient for *use_internet*? Is it statistically significant? What about practically significant?

h. Evaluate the regression diagnostics from your second model.

i. Compute the F-ratio and associated p-value between your two regression models. Assess and describe any improvement from your first model to your second.