W203: Exploring and Analyzing Data
Summer 2016
Final Exam

Instructions

Please construct an R markdown report that addresses Part 3 of the Final Exam.

You will receive a web-based link from bCourses. This link will include all questions from Part 1 as well as questions (where applicable) from Parts 2 and 3.

As a suggestion, complete your report first, and include your answers for all parts in the report. Then access the web-based link from bCourses and answer all of the questions found there. Finally, **be sure to upload your report to bCourses** (you will be able to do so at the end of the bCourses quiz).

You reduce the chance of grading mistakes by writing the answer to each question in the space provided under each question on bCourses.

Your report must include *your code,* and *R's output.*

Part 1: Multiple Choice (32 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 dependent 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 regression
   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

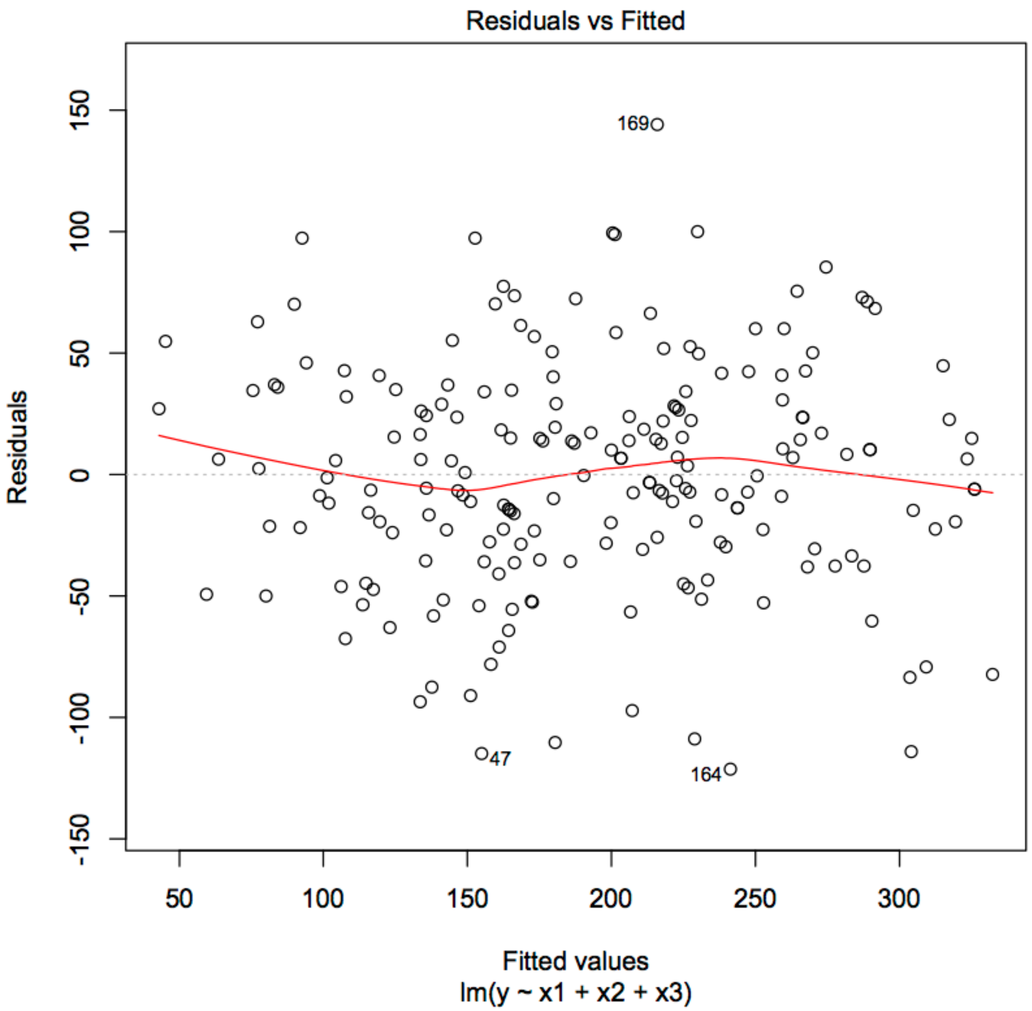
3. Suppose you rejected the null hypothesis for your test in Q1. What can you conclude?
    a. There are statistically significant relationships between each independent variable and the dependent variable.
    b. b0 is statistically significant
    c. b1 is not equal to zero
    d. The relationship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

4. Suppose you were to add in a third variable (X3) to the original model described above. Which of the following potential variables is *least likely* to be statistically significant if you add it to the original model?
    a. Fear of wild animals.
    b. Maximum number of pushups in 3 minutes.
    c. Outdoor camping experience.
    d. Knowledge of plants and animals.
    e. Previous game show experience.

5. 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.

6. The following graphic shows:



a. Non-linearity.
b. Heteroscedasticity and non-linearity.
c. Regression assumptions that have been met.
d. Heteroscedasticity.
e. Non-independence of errors.

7. When conducting a statistical test on a sample, which of the following is NOT a random variable?
   a. The test statistic
   b. The power of the test
   c. The p-value

d. The 95% confidence interval
e. The range of the variable in the sample

8. Researchers in several fields have noted the large number of published results that are not reproducible. Which of the following factors does NOT help explain why more than 5% of published results appear to be type-1 errors?
    a. Random variation in each sample drawn from a population will lead to larger or smaller p-values.
    b. Researchers may exploit their degrees of freedom to their advantage.
    c. Insignificant results are commonly relegated to the file drawer.
    d. Researchers may perform multiple comparisons without correcting the alpha level.

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

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 questions 9 through 14, *select the most appropriate statistical procedure from the provided choices*, assuming that you do not recode the variables in any way (except for dealing with missing values, if applicable).

Note: You do **NOT** need to execute the test(s) for questions 9 through 14.

9. 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

10. 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. Binary logistic Regression

    d. ANOVA

11. 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. ANOVA

12. 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

13. Is the respondent's age related to the total number of children he or she has (children0_5 + children6_11 + children12_17)?

    a. Pearson Correlation

    b. Wilcoxon Rank-Sum Test

c. Binary logistic Regression

d. ANOVA

14. 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. Dependent t-test

Part 3: Data Analysis (44 points)

Below are several additional questions that apply to the Dating.csv data set.

15. OLS Regression

a. 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 so that 5 = excellent and 1 = poor. What is the mean quality of life in the sample?

b. 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(D$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. What is the mean of years_in_relationship in the sample?

c. 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. 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?

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

e. 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?

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

16. Logistic Regression

a. What are the odds that a respondent in the sample has flirted online at some point (flirted_online)?

b. Conduct a logistic regression to predict flirted_online as a function of where a respondent lives (usr). What Akaike Information Criterion (AIC) does your model have?

c. According to your model, how much bigger are the odds that an urban respondent has flirted online than the odds that a rural respondent has flirted online? Is this effect practically significant?