Info 271b
Fall 2014
Lab 3

Overview:
This lab has a series of 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 "lab report" file using a standard format (.pdf, .doc, etc). Please name this primary document "lab3_YourLastName.doc". Similarly, your script should be named "lab3_YourLastName.R"

This lab is due on Thursday, November 20th by 10am. You should email your two files (the main lab report and your R script) to Toshiro (tnish@berkeley.edu). Please plan ahead, we do not accept late papers.

Data Preparation, Analysis and Interpretation (100 Points)

Dataset (GSS):
Every other year, the General Social Survey collects responses to thousands of questions, covering a wide variety of topics. You will be using a subset of data from 1993, including a small number of variables.

While some variables may be self-explanatory, others may not make sense until you look at the GSS codebook. An easy way to investigate a variable is to look it up in the GSS mnemonic index, located at:

http://www3.norc.org/GSS+Website/Browse+GSS+Variables/Mnemonic+Index/

Before you run a statistical test on a variable, you should always read its description in the codebook, in order to understand what different values mean. For example, the codebook may explain if certain values stand for missing data. If this occurs, you should make sure those values are recorded as NA in R before proceeding.

Write a well-commented R script to perform each of the following tasks, following the best practices described in class. For each new variable, look for obvious errors and make sure that appropriate values are coded as NA.

Include all important output and answers to each question in your main lab report. You can also copy any graphics into the main lab report to make it easier for you to provide context for your answers. We should be able to understand what you did and what your answer is for each item in your main lab report without hunting for things in your R script.

1. Task 1: Conduct a chi-square test to determine if there is an association between marital status (marital) and political orientation (politics).

   A. **What are the null and alternative hypothesis for your test?**

   B. **What test statistic and p-value do you get?**

   C. **Conduct an appropriate effect size calculation for your relationship.**

   D. **Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results?**

2. Task 2: Conduct a correlation analysis to examine if there is an association between age when married (agewed) and hours of tv watched (tvhours).

   A. **What are the null and alternative hypotheses for your test?**

   B. **What test statistic and p-value do you get?**

   C. **Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results?**

3. Task 3: Create a new binary/dummy variable, “married”, that denotes whether an individual is currently married or not currently married. Conduct an independent sample t-test to evaluate the hypothesis that number of children (childs) is greater for those who are married than those who are not married.

   A. **What is the null and alternative hypotheses for your test?**

   B. **What test statistic and p-value do you get?**

   C. **Conduct an appropriate effect size calculation for your relationship.**

   D. **Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results?**

4. Task 4: We want to consider just the subpopulation of 23-year olds in this sample. Conduct a Wilcox rank-sum test to determine whether your new “married” variable from Task 3 is associated with the number of children (childs) *for respondents who are 23 years old*.

   A. **What is the mean of your new “married” variable among 23-year-olds (e.g., the proportion of cases in the category coded “1”)?**

   B. **What is the null and alternative hypotheses for your test?**

   C. **What test statistic and p-value do you get?**

   D. **Conduct a cohen’s d effect size calculation for your relationship (you can use the same code we used in our class R examples).**

   E. **Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results?**