Lab 5: Advanced R

Michael Surdek

Southern New Hampshire University

|  |
| --- |
| Alt tag: Workflow overview |

### LAB Instructions

|  |  |
| --- | --- |
| Step | Action |
| 1 | **Prepare working environment for the Lab and load data files**  **First need to get the files created to use in R environment. For this lab, you will be using RStudio to create your file to use for manipulation –**   1. **Load lab01.txt and lab02.txt into Big Sheets** 2. Set the working directory to whatever directory you have stored your data. For example, I have my data stored at: M:/Users/<your\_user\_name>/Desktop /DAT510/”   On the console window type: (note—I am using my directory information as an example- you will need to change the directory path to yours)  **setwd(“M:/Users/g.britton/Desktop/DAT510")**   1. Download the file from the Module 5 Lab area inside of your learning environment. (Module5RLab2.r) 2. In the script window, open the script called “Module5Lab2.R”. (Click on “File”, “Open File” and click on file “Module5Lab2.R”).   Start R and Read the Data Set Back Into Your Workspace:  *NOTE: you will need to change the “path” that is listed in the file to whatever the path you have saved your files to is. Example – in the lab it uses the path “~/LAB01” – I saved my files at this path: M:/Users/g.britton/Desktop/DAT510/”*  *Everytime I see “~/LAB01” I will change that reference to my path.*   1. Execute the following commands from the script window:   **options(digits=3) options(width=68)**  **ls() ## load(file=”Labs.Rdata”) ls()**  **rm(lab2)**  **ds <- lab1 colnames(ds) <- c("income", "rooms")** |
| 2 | **Obtain summary statistics for Household Income and visualize data:**   1. Execute the following commands from the script window:   **summary(ds$income) range(ds$income) sd(ds$income) var(ds$income)**  **plot(density(ds$income)) # left skewed**   1. What is the mean? 67152 2. What is the median? 50300 3. What is the standard deviation? 68178 |
| 3 | **Obtain summary statistics for Number of rooms and visualize data:**  Execute the following commands from the script window:  **summary(ds$rooms) range(ds$rooms) sd(ds$rooms) plot(as.factor(ds$rooms))**  What is the mean? 5.63  What is the median? 6.00  What is the standard deviation? 1.99 |
| 4 | **Remove Outliers**  In a previous lab, you recorded the range of income. You observed that the minimum household income is 4, and the maximum is 1,620,560.   1. Does this make sense to you? Why? \* The maximum of the household income variable being 1,620,560 makes sense because it is a large amount yet it is reasonable that a household income would be 7 figures. The minimum value of 4 makes less sense. This might be the result of invalid data or an abnormality in the way that household income was calculated. 2. What happens if you throw out the top and bottom 10%? Execute the following line from the script window     **(m <- mean(ds$income, trim=0.10) )**  If you throw out the top and bottom 10%, the mean should be less affected by the outlier points and should be closer to the median than the original mean.   1. How does this compare to the previous mean of this variable? The mean of the income variable reduced from 67152 to 55347.5, which means that the top 10% of the values were pulling the mean more than the bottom 10%. 2. Execute the following commands from the script window:   **ds <- subset(ds, ds$income >= 10000 & ds$income < 1000000) summary(ds) quantile(ds$income, seq(from=0, to=1, length=11))**   1. How do these values vary from the values in the original data set?   By removing values of the income variable that are less than 10,000 or greater than 1,000,000, the values of the 1st quartile, median, 3rd quartile, and mean all became higher. This means that more values that were less than 10,000 than were greater than 1,000,000.   1. Do they make more sense?   These values do not make much more sense in my opinion, since more values were removed from the lower end of the distribution.   1. Which data set would you prefer to use?   To get a true representation of the original data set while reducing the impact of outliers, it is best in this case to remove an equal percentage of values from each side of the distribution. I would prefer to use the data set that removes the top and bottom 10% from the income variable.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  \*We might consider the high and low value as outliers, and get rid of them. On the other hand, as we will discover, income is best described via a lognormal distribution, and hence these values are in the extreme ends +- 3 sds from the mean. |
| 5 | **Stratify Variable – Household Income and plot the results:**  Stratify breaks that occur close to U.S. Guidelines for Poverty, Median Income, Wealth, and Rich (> $250k @ year)   1. Execute the following code (listed under comment heading “step 5” in the script file):   **breaks <- c(0, 23000, 52000, 82000, 250000, 999999) labels <- c("Poverty", "LowerMid", "UpperMid", "Wealthy", "Rich")  wealth <- cut(ds$income, breaks, labels) # add wealth as a column to ds ds <- cbind(ds, wealth) # show the 1st few lines. head(ds)**   1. Continue to execute the remaining part of the code in Step 5   **wt <- table(wealth) percent <- wt/sum(wt)\*100 wt <- rbind(wt, percent) wt plot(wt)**   1. Take another look at the relationship between wealth and income. Execute the following lines:   **# take another look -- wealth by rooms  nt <- table(wealth, ds$rooms) print(nt) plot(nt) # nice mosaic plot**   1. Execute this code from the script file. These lines will remove the variables wealth, breaks and labels, and then save the variables data set and write into a file named “Census.Rdata”.   **rm(wealth,breaks,labels) save(ds, wt, nt, file="Census.Rdata")** |
| 6 | **Plot Histogram and Distributions:**  Problem: How do you represent income given the range of values?   1. Select and execute the code under Step 6 Histograms and distributions in the script file.   **library(MASS)**  **with(ds, {  hist(income, main="Distribution of Household Income", freq=FALSE)  lines(density(income), lty=2, lwd=2)** # line type (lty) 2 is dashed **xvals = seq(from=min(income), to=max(income), length=100)  param = fitdistr(income, "lognormal")  lines(xvals, dlnorm(xvals, meanlog=param$estimate[1], sdlog=param$estimate[2]), col=”blue”)**  **})**   1. Now try the same thing with log10(income)   **logincome = log10(ds$income)**  **hist(logincome, main="Distribution of Household Income", freq=FALSE)** # line type lty(2) is a dashed line **lines(density(logincome), lty=2, lwd=2)  xvals = seq(from=min(logincome), to=max(logincome), length=100) param = fitdistr(logincome, "normal") lines(xvals, dnorm(xvals, param$estimate[1], param$estimate[2]), lwd=2, col=”blue”)** |
| 7 | **Compute Correlation between income and number of rooms:**  1. You need to consider your hypothesis.   * Your hypothesis is that the number of rooms in a house is predicted by household income (the rich can buy bigger houses), e.g. *lm(rooms ~ income)* * Therefore, our null hypothesis: no correlation between income and number of rooms. * Alternate hypothesis: there is a correlation between income and the number of rooms.  1. Execute the following code (listed after the comment line “Step7 in the script file).   **with(ds, cor(income, rooms))**  **with(ds, cor(log(income), rooms))) # this will give a better correlation**   1. For comparison, correlate rooms with a completely unrelated variable.   **n = length(ds$income) with(ds, cor(runif(n), rooms))** |
| 8 | **Create a Boxplot - Distribution of income as a factor of number of rooms:**   1. Select and execute the code (Listed after the comment line “Step 8”) in the script window. 2. Plot the distribution of income as a factor of # of rooms. ‘log=”y”’ plots income on log scale. We will suppress the outlier points and let the whiskers cover the full range of the data.   **boxplot(income ~ as.factor(rooms), data=ds, range=0, outline=F, log=”y”,** **xlab="# rooms", ylab="Income")**   1. Plot the # of rooms as a function of wealth level.   **boxplot(rooms ~ wealth, data = ds,** **main="Room by Wealth", Xlab="Category", ylab="# rooms")**  **# we’ll keep the outlier points in this one** |
| 9 | **Exit R:**   1. Type the following command into the RStudio command window:   **q()**   1. R will ask you if you want to save your workspace. Answer “**no**.” |

*End of Lab Exercise*