Practice for Learning R and Learning Latex

Jennifer Pan

August, 2011

Latex Environments

A) Try to create the following equations:

1. $\frac{5+6}{\alpha} = \beta^2$
2. $Pr(-1.96 \leq Z \leq 1.96) = 0.95$
3. $\hat{\beta}_x = r_{xy|z} \left(\frac{s_y}{s_x} \right) \left(\frac{\sqrt{1-r^2_{yz}}}{\sqrt{1-r^2_{xz}}}\right)$
4. $$\frac{1}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}$$

Solution

```
##1 $\frac{5 + 6}{\alpha} = \beta^2$

##2 $Pr(-1.96 \leq Z \leq 1.96) = 0.95$

##3 $\hat{\beta}_x = r_{xy|z} \left(\frac{s_y}{s_x} \right) \left(\frac{\sqrt{1-r^2_{yz}}}{\sqrt{1-r^2_

##4
\begin{eqnarray*}
\frac{1}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}
\begin{pmatrix}
\sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i\\
-\sum_{i=1}^{n} x_i & n\\
\end{pmatrix} \\
\end{eqnarray*}
```

B) Create a list (numbered) of your favorite foods.

C) Create a list (with bullets) of your preferred news sources

R Basics (Objects, Data Types, Object Classes)

A) Objects, Data Types, and Object Classes

1. Create a vector of integers from 1 to 20.
2. In one line of code, add 2, multiply by 5, take the square root, and then take the log of each element in the vector.
3. Create a vector of your 5 favorite cities.
4. Create a 3×3 matrix where each element of every column corresponds to the column number.
5. Convert this matrix into a dataframe.
6. Create a $3 \times 5 \times 2$ array of all 0s.
7. Create a list containing your array, your dataframe and your two vectors.

Solution

```
ans.1 <- 1:20
ans.2 <- log(sqrt((ans.1 + 2) * 5))
ans.3 <- c("New York", "Beijing", "Hong Kong", "Paris", "Shanghai")
ans.4 <- matrix(c(1, 2, 3), ncol = 3, nrow = 3, byrow = T)
ans.5 <- as.data.frame(ans.4)
ans.6 <- array(0, dim = c(3, 5, 2))
ans.7 <- list(ans.6, ans.5, ans.3, ans.2)
```

R (Combining/Indexing/Subsetting, Commonly used functions)

A) Combining, Indexing, and Subsetting

1. Create a 10×3 (10 rows, 3 columns) matrix where each element of every row corresponds to the row number.
2. Create another matrix, 50×3 (50 rows, 3 columns), where the first column contains 4's, the second 5's, and third 7's
3. Combine the two matrices
4. Name each column of your matrix `G1`, `G2`, and `G3`
5. Create a new dataset where all observations in column 3, `G3`, is less than or equal to 6

6. Write this new smaller dataset as a separate file into your working directory in any format (i.e. .csv, .dta, .txt)

7. Store the large dataset and the new smaller dataset in a list with appropriate names

Solution

```
##1
mat1 <- matrix(c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6,7,7,7,8,8,8,9,9,9,10,10,10),
   nrow = 10, ncol = 3, byrow=T)
%mat1

##2
mat2 <- matrix(c(rep(4, times=50),rep(5, times=50),rep(7, times=50)),
   nrow=50, ncol=3)
%mat2

##3
my.final <- rbind(mat1, mat2)
%dim(my.final)

%#as.data.frame(my.final)
%#class(my.final)

##4
colnames(my.final) <- c("G1", "G2", "G3")
%colnames(my.final)

##5
my.new <- my.final[my.final[,3] <= 6,]
%dim(my.new)

##6
write.csv(my.new, file = "mynew.csv")

##7
my.list <- list(large = my.final, small = my.new)
%my.list
```

R (Importing Data, Key Packages, Manipulating Data, Figures, Tables)

A) Data and Common Functions

1. Load the `cambridge` dataset again. Find the mean, median, standard deviation, and 20th and 80th percent quantiles of the income variable.
2. How many observations are there in this dataset?
3. What are the average loan amount, income, and loan rate?
4. Which income-sex observation had the highest interest rate in the dataset?
5. Do more males or females have rates greater than 5?

Solution

```
##1
mean(loans$income)
[1] 165.3897

median(loans$income)
[1] 120

sd(loans$income)
[1] 215.5218

quantile(loans$income, probs = c(0.2, 0.8))
  20%   80%
 81.6 186.0

##2
nrow(loans)
[1] 929

##3
Colmeans(loans)

##4
loans[loans$rate == max(loans$rate), c("income", "sex")]
    income  sex
658    200 Male

##5
tally <- table(loans$sex[loans$rate > 5])
names(tally)[tally == max(tally)]
```

R Figures and Tables (cont.) and LaTeX Tables and Figures

A) Figures

1. In R, load the `ccarddata` dataset again. Create a scatterplot where `age` is on the x axis, `credit card expend` is on the y axis.
2. Label the x and y axis of this plot, and add the title "Scatterplot".
3. Create a histogram of `credit card expend`, add color to the plot, label all axis, and include a title.
4. Add a dashed red line at the mean `credit card expend` to the histogram.
5. Add a legend to the histogram to explain this dashed red line.
6. Put the scatterplot and histogram side by side (no need to include the dashed mean line or legend).
7. Save this figure into your working directory as "myfigure.pdf".
8. Add "myfigure.pdf" to your LaTeX document.

Solutions

```
##1
library(foreign)
ccarddata <- read.dta("ccarddata.dta")

plot(x=ccarddata$age, y=ccarddata$credit_card_expend, xlab = "Age", ylab = "Credit Card Expenditure",
   main = "Scatterplot")

##2
hist(ccarddata$credit_card_expend, col = "gold", xlab = "CC Expenditure", ylab = "Frequency",
main = "Credit Card Expenditures")

##3
abline(v = mean(ccarddata$credit_card_expend), col = "red", lty="dashed", lwd=2)

##4
legend(x="topright", legend=c("mean"), col="red", lty="dashed", lwd=2)

##5
par(mfrow = c(1, 2))
plot(x=ccarddata$age, y=ccarddata$credit_card_expend, xlab = "Age", ylab = "Credit Card Expenditure",
   main = "Scatterplot")
hist(ccarddata$credit_card_expend, col = "gold", xlab = "CC Expenditure", ylab = "Frequency",
main = "Credit Card Expenditures")
```

```
##6
pdf(file= "myfigure.pdf")
plot(x=ccarddata$age, y=ccarddata$credit_card_expend, xlab = "Age", ylab = "Credit Card Expenditure",
   main = "Scatterplot")
hist(ccarddata$credit_card_expend, col = "gold", xlab = "CC Expenditure", ylab = "Frequency",
main = "Credit Card Expenditures")
dev.off()
```

B) Table In LaTeX, create a table with information on three Gov department professors. The first column contains their name, the second their office location, and the third column their phone number. Your table should have three rows.

R (Writing Functions, For Loops, Sampling) and LaTeXBibliography

1. Load the `PErisk` dataset from the `Zelig` package.
2. Which country does the 35th observation belongs to? Use code to identify the country.
3. Create a new dataset that omits Kenya from the dataset.
4. Using the new dataset, now extract the `barb2` and `gdpw2` variables. Find the mean, median, standard deviation, and correlation of these two variables. Present the results in a nicely formatted table in LaTeX.
5. Create, $\mathbf{X}$, a matrix with two columns: the first column is a column of 1s and the second column is the `gdpw2` variable. Then, create a vector, $\mathbf{y}$, which is the `barb2` variable. After creating $\mathbf{X}$ and $\mathbf{y}$, use R to calculate $\hat{\boldsymbol{\beta}}$, where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

   Type out this equation in LaTeX and also report the values in $\hat{\boldsymbol{\beta}}$. $\hat{\boldsymbol{\beta}}$ should be a vector of length 2.

6. Write a function that takes in any $n \times k$ matrix for $\mathbf{X}$ and any $n \times 1$ vector for $\mathbf{y}$ and calculates $\hat{\boldsymbol{\beta}}$.
7. Do the following 1000 times with a for loop: Take a sample of 61 observations from the dataset with replacement. For each sample, calculate $\hat{\boldsymbol{\beta}}$, where $\mathbf{X}$ and $\mathbf{y}$ are defined the same as in question 5. Store your results in a 1000×2 matrix.

Solution

```
##1
library(Zelig)
data(PErisk)
```

```
##2
PErisk$country[35]

##3
new.data <- PErisk[PErisk$country != "Kenya", ]

##4
variables <- cbind(new.data$barb2, new.data$gdpw2)
colMeans(variables)

apply(variables, MARGIN = 2, FUN = sd)

cor(new.data$barb2, new.data$gdpw2)

##5
X <- cbind(1, new.data$gdpw2)
y <- new.data$barb2
beta.hat <- solve(t(X) %*% X) %*% t(X) %*% y
beta.hat

##6
beta.func <- function(X, y) {
  beta.hat <- solve(t(X) %*% X) %*% t(X) %*% y
  return(beta.hat)
}

##7
results <- matrix(NA, nrow=1000, ncol=2)
for(i in 1:1000){
  row.numbers <- sample(1:nrow(new.data), size=61, replace=T)
  samp.data <- new.data[row.numbers,] ## sample the data
  X <- cbind(1, samp.data$gdpw2)
  y <- samp.data$barb2
  results[i,] <- beta.func(X=X, y=y)
}
```