

# LectureQuestions

*Xiangyi Xu*

*May 2015*

## Lecture 8a

The `principal()` function will perform a principal components analysis in R, starting with a matrix. The format is as the following. Which of the following description is NOT correct regarding the parameters?

`principal(r, nfactors=m, rotate=n, scores=p)`

- `r` is a covariance matrix or a raw data matrix
- `nfactors` specifies the number of principal components to extract (1 by default)
- `rotate` indicates the rotation to be applied (varimax by default)
- `scores` specifies whether or not to calculate principal component scores (false by default)

## Lecture 8b

Which criteria is correct for deciding how many components to retain in a PCA?

- All of these choices.
- Basing the number of components on prior experience and theory.
- Selecting the number of components needed to account for some threshold cumulative amount of variance in the variables.
- Selecting the number of components to retain by examining the eigenvalues of the  $k \times k$  correlation matrix among the variables.

## Lecture 9a

Assume that `library(ggplot2)` has been loaded and `mtcars` is its built-in database. Which of the following code will NOT achieve the purpose as the other three?

- `plot(wt~mpg, data=mtcars)`
- `plot(mtcars$wt, mtcars$mpg)`
- `qplot(mtcars$wt, mtcars$mpg)`
- `ggplot(mtcars, aes(x=wt, y=mpg)) + geom_point()`

## Lecture 9b

Assume that `library(ggplot2)` has been loaded and database `pressure` is built-in. Which of the following 2 codes are equivalent?

`library(ggplot2)`

1. `qplot(temperature, pressure, data=pressure, geom="line")`
2. `ggplot(pressure, aes(x=temperature, y=pressure)) + geom_line()`
3. `qplot(temperature, pressure, data=pressure, geom=c("line", "point"))`
4. `ggplot(pressure, aes(x=temperature, y=pressure)) + geom_line() + geom_point()`

- 1 and 2
- 1 and 3
- 2 and 3
- 2 and 4

## Lecture 10a

The following codes are supposed to implement a version of Newton's method for calculating the square root of  $y$ . Which one is NOT correct?

```
y <- 12345
x <- y/2
```

- `while (abs(x*x-y) < 1e-10) x <- (x + y/x)/2`
- `while (abs(x*x-y) > 1e-10) x <- (x + y/x)/2`
- `repeat {x <- (x+y/x)/2; if (abs(x*x-y) < 1e-10) break}`
- `repeat {x <- (x+y/x)/2; if (all(abs(x*x - y) < 1e-10)) break}`

## Lecture 10b

Which of the following statement is FALSE?

- Shiny is a Python package that makes it easy to build interactive web applications (apps) straight from R.
- Shiny apps have two components: a user-interface script and a server script.
- The user-interface script controls the layout and appearance of your app. It is defined in a source script (ui.R).
- The server.R script contains the instructions that your computer needs to build your app.

## Lecture 11a

To standardize each variable in a dataset for analysis, we may use `scale()` function. The function equals to which code snippet of the following?

- `df1 <- apply(mydata, 2, function(x) {(x-mean(x))/sd(x)})`
- `df2 <- apply(mydata, 2, function(x) {x/max(x)})`
- `df3 <- apply(mydata, 2, function(x) {(x+mean(x))/sd(x)})`
- `df4 <- apply(mydata, 2, function(x) {(x-mean(x))/mad(x)})`

## Lecture 11b

In the partitioning approach, the most common method is the K-means cluster analysis. Which of the following statement is correct?

- All of the statements.
- Select  $k$  centroids; assign each data point to its closet centroid.
- Recalculate the centroids as the average of all data in a cluster; assign data points to their closet centroids.
- Continue the other steps until the observations are not reassigned or the maximum number of iterations is reached.

## Lecture 12a

Suppose we define the following function in R. What is the result of running `cube(3)` in R after defining the function?

```
cube <- function(x, n){  
  x^3  
}
```

- The number 27 is returned
- The users is prompted to specify the value of 'n'.
- An error is returned because 'n' is not specified in the call to 'cube'
- A warning is given with no value returned.

## Lecture 12b

What is an environment in R?

- A collection of symbol/object pairs
- A list whose elements are all functions
- A special type of function
- An R package that only contains data

## Lecture 13a

```
f1 <- function(x1, x2)  
  return (-5-3*x1+4*x2+x1^2-x1*x2+x2^2)  
f2 <- function(x)  
  return (-5-3*x[1]+4*x[2]+x[1]^2-x[1]*x[2]+x[2]^2)
```

What is the result of the following expression?

```
f1(0, 0) == f2(c(0, 0)) && f1(1, 2) == f2(c(1, 2))
```

- TRUE
- FALSE
- MAYBE
- 42

## Lecture 13b

In ANOVA model, to denote the complete crossing variables, the code `y ~ A*B*C` expands to which of the following formula?

- $y \sim A + B + C + A:B + A:C + B:C + A:B:C$
- $y \sim A + B + C + A:B + A:C + A:B$
- $y \sim A + B + C + A:B:C$
- $y \sim A + B + C$

## Lecture 14a

In MASS package, `stepAIC()` function performs stepwise model selection (forward, backward, and stepwise) using an exact AIC criterion. Which of the saying is NOT correct?

- In **backward stepwise regression**, we start with a model that includes all predictor variables, and then delete them two at a time until removing variables would degrade the quality of the model.
- In **forward stepwise regression**, we add predictor variables one at a time, stopping when the addition of variables would no longer improve the model.
- **Stepwise stepwise regression** combine the forward and backward stepwise approaches by evaluating an entered variable and deleting it if it doesn't contribute to the model.
- Another effective method of variable selection is using the `regsubsets()` function from **leaps** package.

## Lecture 14b

The most common approach for evaluating the statistical assumptions in a regression analysis, is to apply the `plot()` function to the object returned by the `lm()`. Doing so produces four graphs that are useful for evaluating the model fit. One of the graphs is “Residual versus Leverage graph” which identifies outliers, high-leverage points, and influential observations. Which of the following saying is correct?

- All of these.
- An outlier is an observation that is not predicted well by the fitted regression model.
- An observation with a high leverage value has an unusual combination of predictor values.
- An influential observation is an observation that has a disproportionate impact on the determination of the model parameters.