Final Exam Study Guide

Robert Horton

May 16, 2015

Lecture 08a

catterbu If df is a dataframe, what does the following R code do?

```
princomp(df)
```

- It takes df and runs Principal Component Analysis (PCA) on its column features, creating components that explain the variance. It returns a princomp object.

- It takes df and runs Principal Component Analysis (PCA) on its row features, creating components that explain the variance. It returns a princomp object.

- It takes df and prepares a special kind of printing object.

- It takes df and analyzes its components to determine which ones explain the variance of df.

cpkaur What functions can you use to create a sequence over colors in R??

- All of these
- colorRampPalette
- colorRamp
- rainbow_hcl

vchaudhuri What is the difference between Principal component analysis (PCA) and exploratory Factor analysis (EFA)

- PCA is a data reduction technique that transforms a larger number of correlated variables into a much smaller set of uncorrelated variables and EFA is a technique to uncover the latent structure in a given set of variables
- PCA is a data expansion technique that transforms a smaller number of variables into a much larger set of correlated variables and EFA is a technique to uncover the explore the underlying factors in a data set
- PCA and EFA are essentially the same technique of data reduction technique to find correlated patterns among different variables.
- PCA is a technique to choose selective variables which matter in the data and compute the correlation in data only among those vraiables and EFA is a technique to extract linear relationship between all variables.

tdenatale One of the primary goals for using Principle Components Analysis is:

- To find the dominate components of the data and effectively reduce features needed to be analyzed to create an effective model
- The user should not normalize the data in order to see biased results
- The principle component values or eigen vectors have 0 length

- It is often useless as there often no components to analyze

nsh87 What should you be aware of when using the rainbow colormap with plots?

- All of these
- After the ROYGBIV colors are exhausted the colors will start repeating
- Certain colors, such as yellow, are hard to distinguish on some screens
- Our eyes tend to focus on colors that “pop”, potentially biasing our interpretation of the plot

lakarbatti What does the ggplot function **geom_point** do?

- The function is used to create scatterplots.
- This function is helpful to get the geometric points in a dataset.
- This function can be used to create a histogram.
- This function does not exist at all.

xxu26 The principal() function will perform a principal componets analysis in R, starting with a matrix. The format is as the following. Which of the following decription is NOT correct regarding the parmaters? `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)

sneha-krishna A principal component analysis is done on a data matrix, `data`, and a summary of results of this analysis is shown below. How many components are sufficient to summarize `data`?

```
> x <-princomp(data)
> summary(x)
Importance of components:
                        Comp.1     Comp.2     Comp.3      Comp.4
Standard deviation     2.0494032 0.49097143 0.27872586 0.153870700
Proportion of Variance 0.9246187 0.05306648 0.01710261 0.005212184
Cumulative Proportion  0.9246187 0.97768521 0.99478782 1.000000000
```

- 1
- 2
- 3
- 0

Lecture 08b

catterbu In the library `ggplot2`, one of the arguments to the function `ggplot()` is `aes`. For what does this stand and what does it do?

- It stands for “aesthetics.” In this argument, the columns of the data.frame being used are defined, as well as their use in the plot.

- It stands for “aesthetics.” In this argument, all aesthetic elements of the plot that do *not* involve the plot’s data.frame are defined.
- It stands for “Anti-Efficiency Selection.” In this argument, you optimize the R resource by telling R which elements should not be treated efficiently.

cpkaur What functions can you use to do Principal Component Analysis in R?

- `prcomp()`, `princomp()`, `PCA()`, `dudi.pca()` and `acp()`
- Only `prcomp()`, `princomp()` and `PCA()`
- `prcomp()` can be used but gives inaccurate results. `PCA()` is the only option
- Only `PCA()`

vchaudhuri What is the most common criteria to decide which components to be retained in PCA

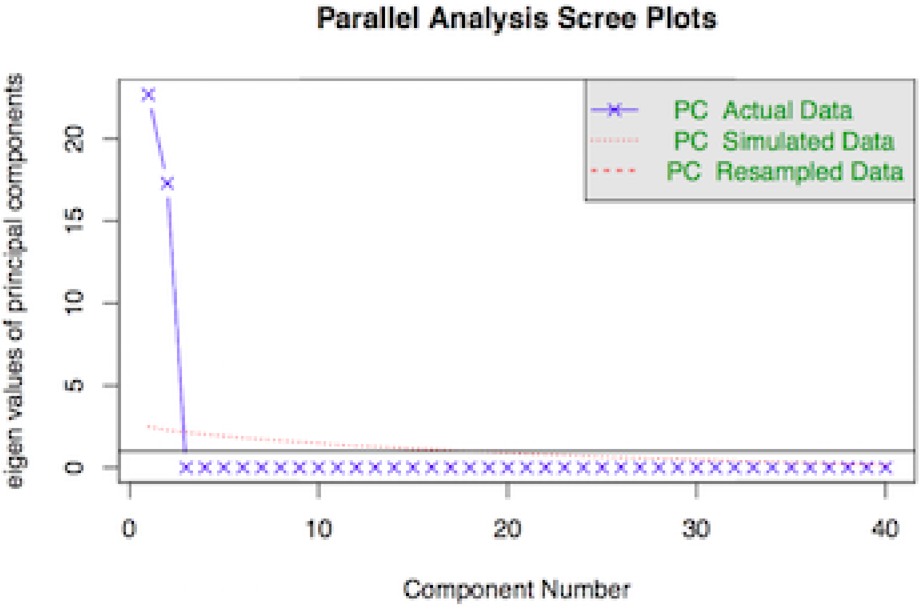
- Selecting the number of components to retain by examining the eigenvalues of the k correlation matrix among the variables
- Selecting the number of components needed to account for some threshold cumulative amount of variance in the variables (for example, 80%)
- Basing the number of components on existing knowledge that affects the data
- Scree plot: this is a graphical method in which you choose the factors until a break in the graph

tdenatale In the example code below in setting a parameter for a model, scientific notation is often used to set parameters. Why?

```
N<- 1e3
```

- It is often clearer and easier to read and modify if N is large such as ie9 or 10 to the ninth power versus 1000000000
- This is false in the code about setting N equal to 1000 is ALWAYS clearer
- This is false, counting 0’s is much easier
- Programmers prefer encoding the powers of ten, rather than being clearer as it increases job security

nsh87 A scree plot displays eigenvalues associated with components or factors and can help determine the number of factors that display most of the varaiability in a given data set. Using the scree plot below, what is the ideal number of factors for PCA from this data set?



- 2
- 1
- 3
- 40

lakarbatti In statistics, what is the meaning of **multicollinearity**?

- Its a phenomenon in which two or more predictor variables in a regression model are highly correlated.
- Its a model in which many values are linear.
- There is no such thing as multicollinearity in statistics.
- It is a model in which there is no relationship between multiple variables.

xxu26 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 examing the eigenvalues of the k*k correlation matrix among the variables.

sneha-krishna The `sweep` function returns an array obtained from an input array by sweeping out a summary statistic. Which of following lines of code will succesfully modify matrix `m`to generate matrix `x`?

```
> m
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
> x
     [,1] [,2]
[1,]   -1   -1
[2,]    0    0
[3,]    1    1
```

- `x <- sweep(m, 2, colMeans(m), "-")`
- `x <-sweep(m, colMeans(m), "-")`
- `x <- sweep(colMeans(m), "-")`
- `x <- sweep(m, 1, colMeans(m), "-")`

Lecture 09a

catterbu What type of statistics was used in the MUNI talk that was considered particularly interesting?

- Circular statistics

- Round statistics

- Linear statistics

- Machine Learning statistics

cpkaur What is the cut function used for?

- Convert numeric to factor
- Cut of the trailing floating point numbers
- Consider specific number of values from a list
- Round off numbers

vchaudhuri In the following chunk of code

```
library(ggplot2)
d=data.frame(beauty=c(1,2,6,4,4,6,7,8), intelligence=c(8,4,7,5,4,9,2,3), speed=c(7,6,9,5,7,6,7,8), gende
ggplot() +
scale_size_continuous(to=c(4,12)) +
geom_point(data=d, mapping=aes(x=intelligence, y=beauty, shape=gender, color=gender, size=speed)) +
opts(title="geom_point", plot.title=theme_text(size=40, vjust=1.5))
```

Which feature is defined by scale_size_continuous in ggplot?

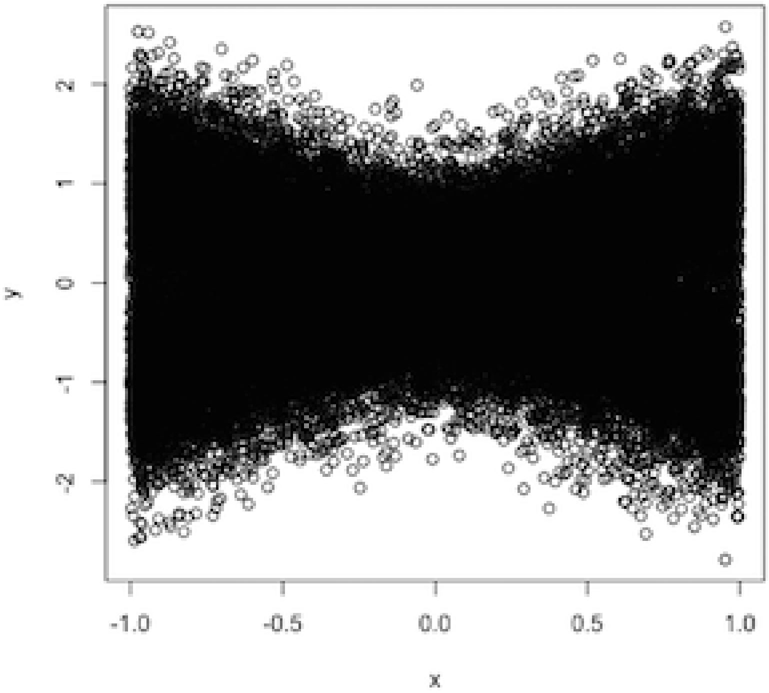
- Used to define the range of point sizes to use.
- Used for marking the scale of x and y axis at continuous intervals
- Used to define degrees of aesthetic scores.
- Used to size discrete variables

tdenatale Choose the correct statement regarding the following:

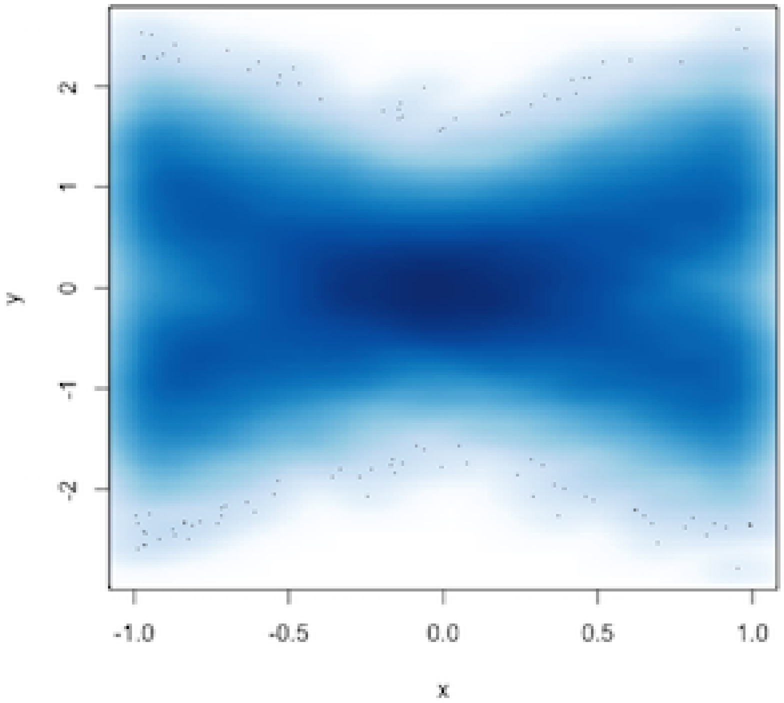
```
plot(num1 ~ num2, data=df)
lm(y ~ x)
```

- Changing “lm(y~x)” to “lm(y~x+Z)” makes the linear model function lm calculate results based on x and z
- The first plot statement plots a variable df in a dataframe named num2
- The use of tilde “~” is NOT valid in R
- Changing “lm(y~x)” to “lm(y~x+Z)” does nothing

nsh87 You create the plot below.



Realizing this display format does not tell you much about the data, you plot the same data, this time creating the new plot below.



What can be said about this new plot?

- It is a smooth scatter, or density plot, useful when a data set displays overplotting
- It is a blurred scatter, useful when a data set displays overplotting
- It not considered a scatter or density plot at all

- It requires a special package in R to produce

lakarbatti Given the piece of code below:

```r
N <- 100
df <- data.frame(
  var1 = runif(N, min=0, max=10),
  var2 = sample(letters[1:5], N, replace=T)
)
kable(head(df))
```

Which of the variables declared above are categorical?

- var2 is the categorical variable
- var1 is the categorical variable
- The sample does not have any categorical variable

xxu26 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, aex(x=wt, y=mpg)) + geom_point()

sneha-krishna Which of following lines of code will initialize a `ggplot` object for a data frame `df`?

- `ggplot() + geom_point(data = df, aes(x =x, y = y))`
- `ggplot(df) + geom_point(aes(x = x, y = y))`
- `ggplot(df, aes(x = x, y = y)) + geom_point()`
- All answers are correct

Lecture 09b

catterbu How does one add a violin plot in `ggplot`?

- `geom_violin()`

- `geom_viol()`

- `geomViolin()`

- `geomViol()`

cpkaur Which library out of the following is useful for manipulating atasets in R, focussing only on dataframes?

- `dplyr`
- `magrittr`

- `tidyr`
- `ggplot2`

vchaudhuri What does the corrgram package do?

- Calculates correlation of variables and displays the results graphically.
- Tabulates a correlation matrix and plots a kernel denisty map
- Draws a boxplot of the most correlated variables with independent variable
- Draws a heat map of the data

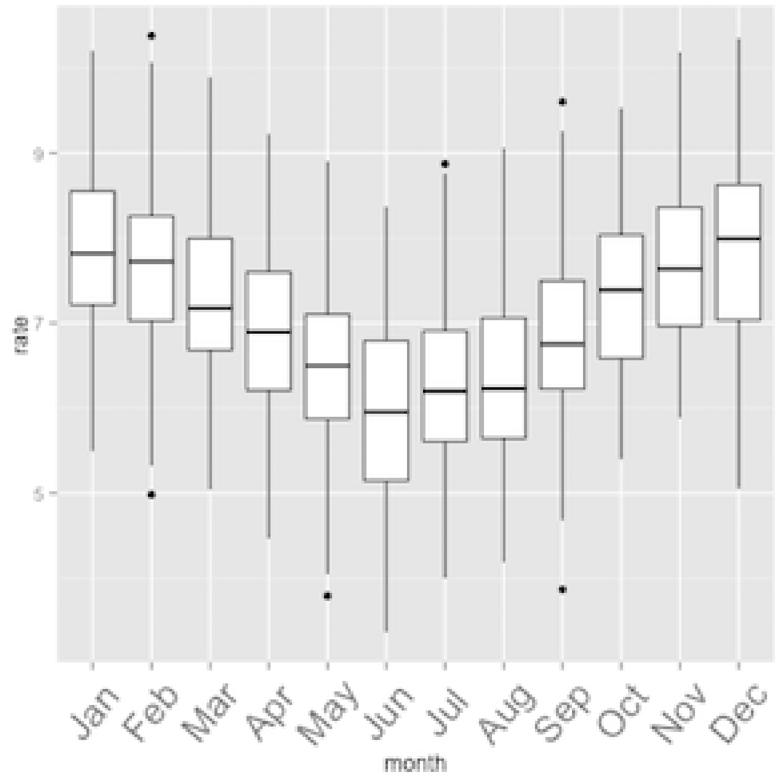
tdenatale The following are examples of R code to generate a correlation diagram and a scatter plot. Select the best answer below:

```
{r corrgram}
library(corrgram)
corrgram(df3, order=TRUE,
    lower.panel=panel.shade, upper.panel=panel.pts,
    text.panel=panel.txt, main="Corrgram")
{r ggplot_scattergram}
library(ggplot2)
g <- ggplot(data=flu, mapping=aes(x=date, y=rate)) + geom_point()
g
```

- All of these
- The corrgram package is used to show correlation between statistical data
- ggplot is a very powerful and easy to use graphical package for R programmers
- The code in the ggplot example is intended to make a scatter plot based on a data-frame named flu

nsh87 The following code generates the plot below:

```
g2 + geom_boxplot() + theme(axis.text.x=element_text(angle=50, size=20, vjust=0.5))
```



What do the box and whisker represent in the plot?

- The interquartile range (the middle half of the data set), and the median of the data, respectively
- The most common values in the data set, and the median of the data, respectively
- The interquartile range (the middle half of the data set), and the mean of the data, respectively
- The middle 25% of the data, and the median of the data, respectively

lakarbatti What does **floor(2.9)** return?

- Returns the number 2
- Throws an error since the function floor does not exist in R
- Rounds the number 2.9 to 3
- Returns the number 2.9

xxu26 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, geom="lines")
2. ggplot(pressure, aes(x=temperature, y=pressure)) + geom_line()
3. qplot(temperature, pressure, data=pressure, geom=c("line", "point"))
4. ggplot(pressure, aex(x=temperature, y=pressure)) + geom_line() + geom_point()
```

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

sneha-krishna If a scatterplot, g, is generated as shown below, which the following lines of code will succcesfully add a linear regression fit line to g?

```
> g <- ggplot(data, aes(x=x, y=y) ) +   geom_point()
```

- `g + geom_smooth(method=lm)`
- `g + geom_smooth()`
- `g$geom_smooth()`
- None are correct

Lecture 10a

catterbu In an R project using `shiny`, what function is used to define the parameters that can be adjusted?

- `sliderInput()`
- `slider_input()`
- `input()`
- `sliderOutput()`

cpkaur What would be the outcome of the following code?

```
x <- runif(1000, min=-10, max=10)
y <- x^2
z <- 2*y + 5*x + 6
plot(x,z)
library(rgl)
plot3d(x,y,z)
```

- An interactive 3D plot and a 2D plot
- A 2D plot
- A 3D plot
- There is an error in the code

vchaudhuri What function call is used to launch a shiny app in a console?

- runApp
- read.shinyApp
- read.App like read.csv to access a .csv file
- ui.r and server.r have to be called sperately

tdenatale Describe the brilliance of the Shiny app in R:

- Is that it can create an interactive environment available on the web useful by both nonprogrammers and programmers
- Is that anyone even non R users can learn how to GENERATE and use it easily
- Is that it cannot be used creatively for presentations

- Because it was created in a New York skyscraper, its was named (S)tay (h)igh (i)n (n)ew (y)ork or Shiny!

nsh87 Shiny apps for R are good for:

- All of these
- Exploratory data analysis
- Embedding interactions into R Presentations
- Sharing data sets with others on the web

lakarbatti What does **readRDS()** function do?

- Reads a binary file into a dataframe
- Reads a csv data set
- There is no such function in R

xxu26 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}

sneha-krishna Which of the following files are minimally necessary to create a Shiny app in R?

- `ui.R` and `server.R`
- Only `server.R` is necessary
- Only `ui.R` is necessary
- `image.R` and `server.R`

Lecture 10b

catterbu What are the names of the two R scripts used in a `shiny` R project?

- `ui.R` and `server.R`

- `gui.R` and `server.R`

- `ui.R` and `rserver.R`

- `rui.R` and `rserver.R`

cpkaur How many basic files are required to build a shiny application?

- 2

- 1
- 3
- 0

vchaudhuri The five Five basic verbs: “filter”, “select”, “arrange”, “mutate”, “summarise”,(plus group_by) are functionalities of the dplyr pckage or manipulate package or both

- dplyr
- manipulate
- both dplyr and manipulate
- neither

tdenatale The R package manipulate can be used to:

- To manipulate a plot interactively by a user
- To manipulate “big data” data-frames easily
- The app is restricted to only 2 variables which are controlled by pickers
- Works very quickly, regardless of underlying performance, such as size of data set or system speed

nsh87 Shiny apps have a number of widgets available to change parameters in functions and plots. When adjusting, for example, the Slider Input while running a Shiny app, the rest of the data will update…

- immediately
- only after you refresh the page
- only if the developer updates the app
- never

lakarbatti Which function returns the column names of a dataframe?

- names()
- getcols()
- readRDS()
- readdata()

xxu26 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.

sneha-krishna The `rgl1` package:

- is used to produce interactive 3-D plots
- contain the `plot3d` function that plots points within an rgl window
- can be used to construct geometric objects using functions such as `cube3d`
- All answers are correct.

Lecture 11a

catterbu What does the following code do?

```
hc1 <- hclust(dist(clusdat1[,c("x","y")]), method='average')
```

- Stores in `hc1` a hierarchical clustering object using a dissimilarity structure of `clusdata1`'s columns x and y and average linkage.

- Stores in `hc1` a hierarchical clustering object using the values of `clusdata1`'s columns x and y and a methodology that averages each column's importance.

- Stores in `hc1` a standard clustering object using a dissimilarity structure of `clusdata1`'s columns x and y and average linkage.

cpkaur Which algorithm is used for density based clustering?

- DBSCAN
- Hierarchical clustering
- K-means algorithm
- Expectation–maximization algorithm

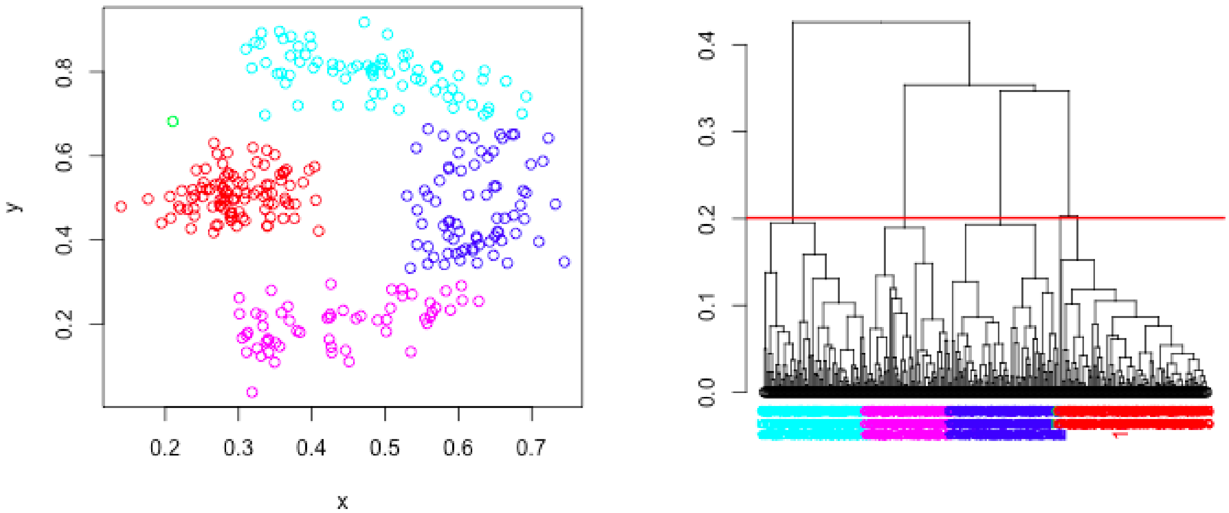
vchaudhuri What is the basic difference between hierarchial cluster vs fixed cluster analysis?

- In Hierarchial cluster plots are continually fused one-by-one in order of highest similarity or dissimilarity whereas in fixed cluster it is partition of the data, measured by the distance of the plot to the center of the cluster to which it belongs.
- In Hierarchial cluster plots are generated by partition of the data, measured by the distance of the plot to the center of the cluster to which it belongs, whereas in fixed cluster plots are continually fused one-by-one in order of highest similarity or dissimilarity.
- In Hierarchial cluster clusters are specified prior and the approach adopted is to cluster at a range of values in fixed cluster a phenomenon called chaining" is followed, where a single plot is continually added to the tail of the biggest cluster

tdenatale Executing the following code does what?

```
N <- 20
x <- 2 * runif(N) - 1
y1 <- -1 * x + rnorm(N/2, sd=1/2)
df2 <- data.frame(x, y=c(y1))
distances <-dist(df2, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
```

- The "dist" function using data from data-frame df2, creates a matrix called distance which represents the distances between the rows of df2
- The optional parameters diag, upper and p must be defined or unintended errors will always occur
- The runif or uniform distribution function creates data that is between 0 and 20
- The data.frame statement will result in an error because the variable c is NOT defined



nsh87
The type of clustering shown above is called…

- hierarchical clustering
- density clustering
- K-means clustering
- relationship clustering

lakarbatti What is **Centrality** in the igraph package?

- Degree of the graph
- The central point in a graph
- A point in the graph
- There is no such term in igraphics package

xxu26 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)})

sneha-krishna Which of the following statement regarding the `hclust` function is FLASE?

- The `hclust` function requires that the number of clusters to be extracted is specified.
- The `hclust` function performs hierarchical cluster analysis.
- The algorithm used in `hclust` proceeds iteratively. At each stage distances between clusters are recomputed.
- The `hclust` function in R uses defines the cluster distance between two clusters to be the maximum distance between their individual components.

Lecture 11b

catterbu What does the following code do?

```
getData <- function(file_path, sep=",", header=TRUE, quote="\"") {
    result <- tryCatch({
        data <- read.table(file_path, sep=sep, header=header, quote=quote,
                           stringsAsFactors=FALSE)
    }, warning = function(cond) {
        message <- "The file path and/or additional arguments could not be read"
        return(list(description=message, Rwarning=cond))
    }, error = function(cond) {
        message <- "The file path and/or additional arguments could not be read"
        return (list(description=message, Rwarning=cond))
    }, finally=NULL
    )
    return (result)
}
```

- Store a function in the variable `getData` that attempts to read a file into a *data.frame*, which is returned. The `tryCatch` function is used to catch an error in the event that the file cannot be read into `data` given its parameters.

- Store a function in the variable `getData` that attempts to read a file into a *list*, which is returned. The `tryCatch` function is used to catch an error in the event that the file cannot be read into `data` given its parameters.

- Store a function in the variable `getData` that attempts to read a file into a *data.frame*, which is returned. The `tryCatch` function is used to catch an error in the event that `R` becomes vengeful and and starts running away from you.

cpkaur What is the outcome of the following code?

```
g <- graph.formula( Alice-Bob-Cecil-Alice,Daniel-Cecil-Eugene, Cecil-Gordon )
plot(g)
```

- Undirected graph
- Directed graph
- Mixed graph
- Weighted graph

vchaudhuri Finish the sentence. In the layout function of igraph package which treats the edges as a set of springs with spring constant set by the weights parameter to the function. Vertices with high edge weight will in general, be ___________ to each other

- closer
- further

- parallel
- equidistant

tdenatale The following code was used in class to demonstrate a caffeine model, Which of the following are not characteristics of “networkD3”?

```
library("networkD3")
caffeineData <- data.frame(Source=c("N1", "C2", "N3", "C3a", "N4", "C5", "N6", "C7", "C7a", "C7a", "C3a
Target=c("C2", "N3", "C3a", "N4", "C5", "N6", "C7", "C7a", "N1", "C3a", "C7a", "CH3a", "CH3b", "CH3c",
simpleNetwork(caffeineData, height = 300, width = 700, fontSize=14)
```

- All of these answers are not characteristics of networkD3
- “networkD3” can only represent data in 2 dimensions
- “networkD3” is an old program that is going to be replaced by “HortonD4” an app that shows 4 dimensions
- “networkD3” cannot be used interactively by the user

nsh87 K-means clustering attempts to minimize…

- the within sum of squares value of each cluster
- the between sum of squares value
- the number of points in each cluster
- the randomness of cluster assignment

lakarbatti What is **Vertex and edge betweenness()** in the igraph package?

- The number of geodesics (shortest paths) going through a vertex or an edge
- The distance between two points in a graph
- Does not really mean anything

xxu26 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 obeservations are not reassigned or the maximum number of iterations is reached.

sneha-krishna The `closeness()` functions in the `igraph` packages measures:

- how many steps is required to access every other vertex from a given vertex
- the number of geodesics (shortest paths) going through a vertex or an edge
- the eigenvector centralities of positions v within it
- None of the choices are correct

Lecture 12a

catterbu What is the name of the general-purpose optimizer used figuring out optimal coefficients in linear regression?

- `optim()`

- `optimizer()`

- `optimal()`

- `optimization()`

cpkaur Which function(s) is/are used to fit linear models?

- `lm`, `glm`, `aov`
- `lm`, `glm`, `aov`, `ppr`
- None of these
- `lm`, `glm`, `ppr`

vchaudhuri In the following linear model

```
mod = lm(train_y ~ train_x).
```

if a list of X's(here train_x) is passed to get it's predicted Y . What would be the code in R?

- For train_x = 1, 2, and 3, use predict(mod, data.frame(train_x = c(1, 2, 3)))
- For train_x = 1, 2, and 3, use predict(mod, data.frame(train_x(1, 2, 3)))
- For train_x = 1, 2, and 3, use crossval(mod, data.frame(train_x = c(1, 2, 3)))
- For train_x = 1, 2, and 3, use coxph(mod, data.frame(train_x = c(1, 2, 3))).

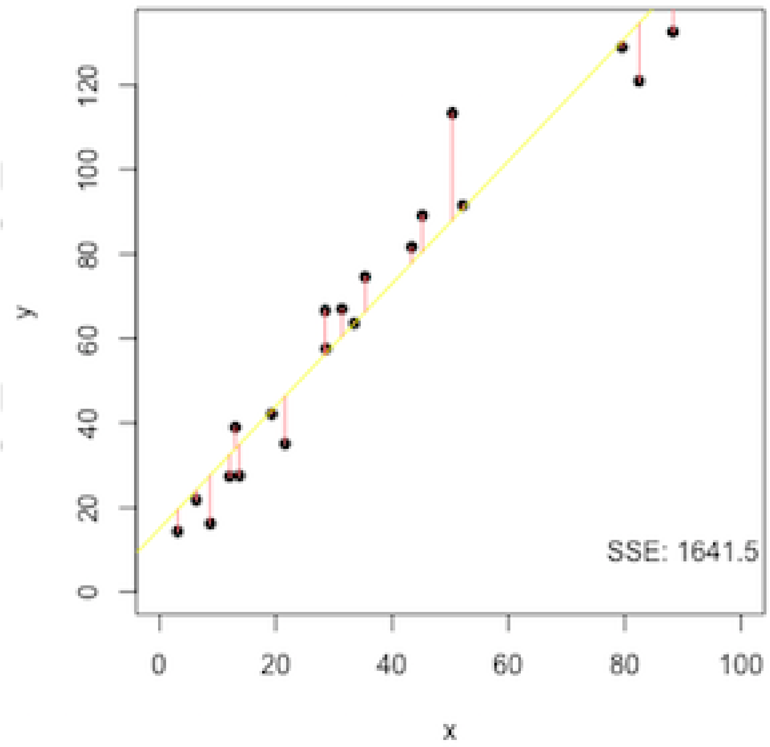
tdenatale Sample commands to implement Linear regression in R:

```
lmfit = lm( change ~ setting + effort )
summary(lmfit)
```

Which of the following statements are true?

- All of these
- Linear regression can be thought of as an optimization problem, where you are trying to find values for the coefficients m and b in y=mx+b that minimize the total squared error.
- The function "lm"" is used to fit linear models.
- The results of the "summary" function include coefficients and estimated errors

nsh87 Linear regression attempts to minimize the sum of squares error between the vertical distances of points and the fitted line. What is the plot below showing?



- How the error between the fitted line and points is measured
- A random quadratic equation plotting against some random points
- A poorly fit linear equation
- None of these answers is correct

lakarbatti What is the equation for a line?

- y = mx + b, where b is the y intercept, m is the slope
- y = mx + b, where y is the name of the line, m is the mean
- a + b + c = 0
- None of the choices

xxu26 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.

sneha-krishna A simple inear regression analysis is done on a data set containing vaiables `x` and `y`. Given the cofficent results below, which equation best fits this data set?

```
> fit <- lm(y ~ x)
> coef(fit)
(Intercept)           x
       36.0        1.94
```

- y=3.60+1.94*x
- x=3.60+1.94*y
- y=1.94+3.60*x
- x=1.94+3.60*y

Lecture 12b

catterbu What does the following code do?

```
factor(ifelse(df$prob > runif(N), "Yes", "No"))
```

- It takes a column of `df` called `prob` and checks if it is greater than a random uniform number of size `N`, which is likely the number of instances of `df`. A factor `"Yes"` is returned if true, and `"No"` if false.

- It takes a column of `df` called `prob` and checks if it is greater than a random uniform number of size `N`, which is likely the number of instances of `df`. A factor `"No"` is returned if true, and `"Yes"` if false.

- A single factor is returned, either `"Yes"` or `"No"` based on a general conclusion on if `df$prob` tends to be greater than `runif(N)`.

cpkaur Which function(s) is/are used to embed a shiny application within an R Markdown?

- `shinyAppDir`, `shinyApp`
- `shinyAppDir`
- `shinyApp`
- None of these

vchaudhuri Where is the method of Least squares most useful?

- Approximate solution in which there are more equations than unknowns
- Approximate solution in which there are more unknowns than equations
- Approximate solution in which there are equations and unknowns are equal
- Approximate solution in which equations is a multiples of the unknowns.

tdenatale Describe what the intent of the following R code inside a R markdown document?

```
shinyAppDir(
  system.file("examples/06_tabsets", package="shiny"),
  options=list(
    width="100%", height=550 , echo = FALSE
  )
)
```

- This example embeds a Shiny application located in another directory

- All of these
- Makes a data-frame with 6 examples separated by commas
- The height articulates how much of the horizontal screen the app can use

nsh87 What is the name of a common function in R that fits linear models?

- `lm()`
- `lin()`
- ’linfit()‘
- `fit()`

lakarbatti Which plotting function adds one or more straight lines through a current plot?

- abline()
- addline()
- moreline()
- None of the choices

xxu26 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

sneha-krishna What does the following code do?

```
library(manipulate)
manipulate(plot(1:x), x = slider(1, 100))
```

- Generates an interactive plot with “x” number of points where “x” can be manipulated to an interger value between 1 and 100.
- Generates a plot with 100 points.
- Generates an random number of points between 1 and 100.
- None of the choices are correct.

Lecture 13a

catterbu What does the following code do?

```
lm(response ~ stimulus:category, data=df)
```

- It fits data from the data.frame `df` using the column `response` versus `stimulus`, looking at `stimulus` by another column, called `category`.

- It fits data from the data.frame `df` using the column `response` versus the columns between `stimulus` and `category`.

- It fits data from the data.frame `df` using the column `response` compared to the columns `stimulus` and `category`.

- None of these are correct.

cpkaur Which of the following statements is FALSE?

- Linear regression attempts to model the relationship between independent variables.
- Logistic regression is useful when you are predicting a binary outcome from a set of continuous predictor variables.
- Poisson regression is useful when predicting an outcome variable representing counts from a set of continuous predictor variables.
- Survival analysis covers a set of techniques for modeling the time to an event.

vchaudhuri One cannot easily tell how one variable affects the prediction using RandomForest package. How can one still predict how response will change as one changes predictor?

- By partial dependence plot
- By partial independence plot
- By complete independence plot
- By complete dependence plot

tdenatale When would one use linear regression versus logistical regression (lm vs glm)?

- For example: if you wanted to see how body mass index predicts blood cholesterol (a continuous measure), you'd use linear regression. If you wanted to see how BMI predicts the odds of being a diabetic (a binary diagnosis), you'd use logistic regression.
- It doesn't matter either will result in what you are looking for
- It doesn't matter either will not result in what you are looking for
- None of these answers

nsh87 Linear regression treats the input variables as…

- numerical values
- boolean
- categories
- continuous densities

lakarbatti What does the generic **group_by** function do?

- The function groups a table by one or more variables
- The function is used in logistic regression to group similar labels
- The function groups different variables into a single variable
- There is no such function in R

xxu26 `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

sneha-krishna Which of following statement about Simpson's Paradox is true?

- Simpson's Paradox demonstates that a great deal of care has to be taken when combining small data sets into a large one.
- Simpson's Paradox can be caused by a lurking variable.
- Simpson's Paradox can be caused from unequal sized groups being combined into a single data set.
- All choices are correct.

Lecture 13b

catterbu What does the `manipulate` package allow you to do?

- Dynamically adjust different constants generating a plot so that you can see how the plot changes. This is useful in trying to get particular shapes to show up in plots.

- Statically adjust different variables for a plot. Then, code can be run to generate a new plot.

- Manipulate all of the variables of a given R script in one command, so that there is no need to search out every instance.

- Dynamically alter the type of the plot and the save it as a pdf file.

cpkaur What is the use of `dev.off()` function?

- Used to close open figure files and plotting windows
- Used to break the code run
- Used to assign value to dev attribute
- None of these

vchaudhuri Simpson's Paradox describes a situation in which variables X and Y are positively related overall, but suddenly become negatively related when conditioned on a third variable Z. What is statement is False:

- It is a forward stepwise regression, which aims to add the most important factors first.
- It is as simply being a problem arising from adding a minor factor to a model before including a major factor.
- It is similar to forward stepwise regression
- Simpson's is not a paradox in the first place, and instead is simply the effect of using variables in the wrong order.

tdenatale In the example below what is the intent of the `-1` in the lm statement?

```
> fit3 <- lm(response ~ category + dose:category - 1, data=df)
```

- Adjust the resultant intercept by one. This technique often allows better visualization and use of the result.
- There is always a minus 1 when using lm
- Identifies that the first item in the data-frame df should be ignored
- Is a keyword to invoke the Shiny app

nsh87 Which function included in R can be used to perform logistic regression?

- `glm()`
- `lm()`
- `fit()`
- `logit()`

lakarbatti In the following piece of code, what is the **cut** function used for?

```
N <- 10
age <- runif(N,7,10.5)
grade <- cut(age,breaks = 7:11,labels = 2:5,right =TRUE)
```

- To convert numeric values in the vector age to factors and store the values in the vector grade
- To cut and paste values from the vector age into vector grade
- To cut values from the vector age and store them in the enviroment variables
- **cut** throws an error

xxu26 In ANOVA model, to denote the complete crossing variables, the code `y ~ A*B*C` expands to which of the following formula?

- y ~ A + B + C + A:B + A:C + B:C + A:B:C
- y ~ A + B + C + A:B + A:C + A:B
- y ~ A + B + C + A:B:C
- y ~ A + B + C

sneha-krishna A logisitic function can be used to model:

- population growth
- the relationship between a scalar dependent variable and one or more explanatory variables
- the relationship between a Poisson distributed response variable and one or more explanatory variables
- Nones of these choices are correct.

Lecture 14a

catterbu How many plots are generated by the following code? What are they?

```
fit <- lm(weight ~ height, data=women)
par(mfrow=c(2,2))
plot(fit)
```

- Four: `Residuals vs. Fitted`, `Normal Q-Q`, `Scale-Location`, `Residuals vs. leverage`

- three: `Residuals vs.  Fitted`, `Normal Q-Q`, `Scale-Location`

- two: `Residuals vs.  Fitted`, `Normal Q-Q`

- three: `Residuals vs.  Fitted`, `Normal Q-Q`, `Residuals vs.  leverage`

cpkaur Which function out of the following can be used to calculate the False Positive Rate?

- `fall`
- `miss`
- `f`
- `pcfall`

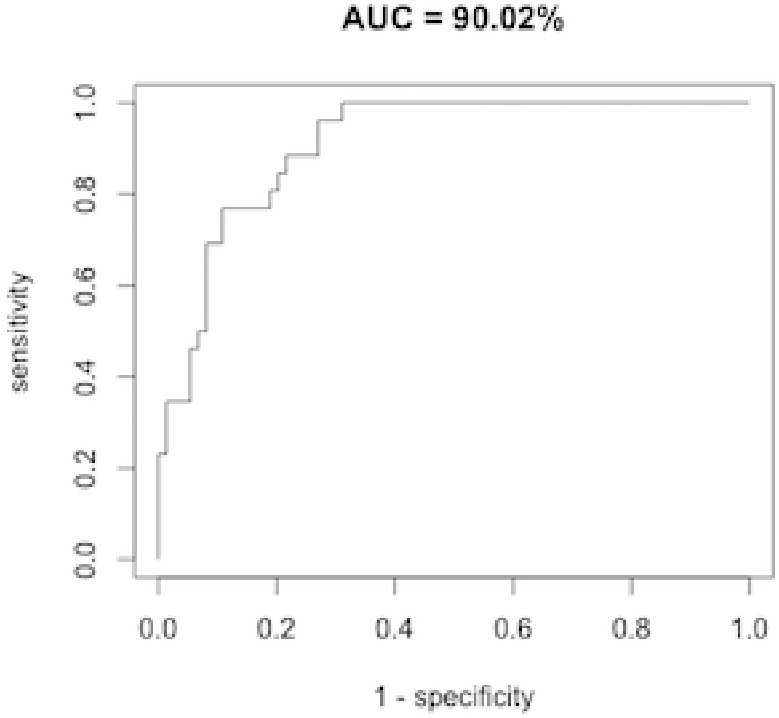
vchaudhuri What are ROCR's 3 commands to produce a simple ROC plot

- Prediction,performance, plot
- fit, prediction, plot
- fit, performance, plot
- lm,fit,plot

tdenatale When using model fitting, in the following example: glm or generalized linear model fits dataset d1. How best is it to use the data that you have at hand?

```
d1 <- sim_cancer(N)
fit1 <- glm(cancer ~ I(carcinogens - mean(carcinogens)) + I(age - mean(age)) + 1,
            data=d1, family="binomial")
```

- It is best to fit your data on one data set and evaluate the model on a second or test data set
- Always use all of your data to fit the model
- It is always OK to overfit your model
- The model is usually more accurate if the number of data points is small



nsh87
ROC curves, such as the one shown above, are a common tool used to…

- evaluate the performance of machine learning models
- evaluate the fit of a linear model
- plot points of a data set against predictions
- strike fear into the hearts of men

lakarbatti What is the **manipulate** function useful for?

- The **manipulate** function can be used to create interactive plots with slider, picker, checkbox or button
- The **manipulate** function can be used to manipulate a data frame
- The **manipulate** function can be used to change the data in a database table
- The **manipulate** function doesn't really do anything

xxu26 In `MASS` package, `stepAIC()`function peforms 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 stewsie approaches by evaluating an entered varialbe 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`.

sneha-krishna What does the `cut()` function do in the code below?

```
> x <- c(1:10)
> cut(x,breaks=2)
```

- It divides `x` into 2 intervals and codes the values in `x` according to which interval they fall.
- It returns two values from 'x'.
- It divides all values in `x` by 2.
- It return all values of x that are divisible by 2.

Lecture 14b

catterbu On a learning curve, the training and cross-validation sets will converge to what line when the model fits well?

- a line plotting the standard deviation of the random noise in the data set; the built-in error.

- a line plotting zero error; this is always the case.

- a line plotting one, the gaussian normal standard deviation.

- the two plots diverge.

cpkaur Where is AIC (Akaike information criterion) used?

- All of these
- In model selection
- Counting parameters
- Transforming data

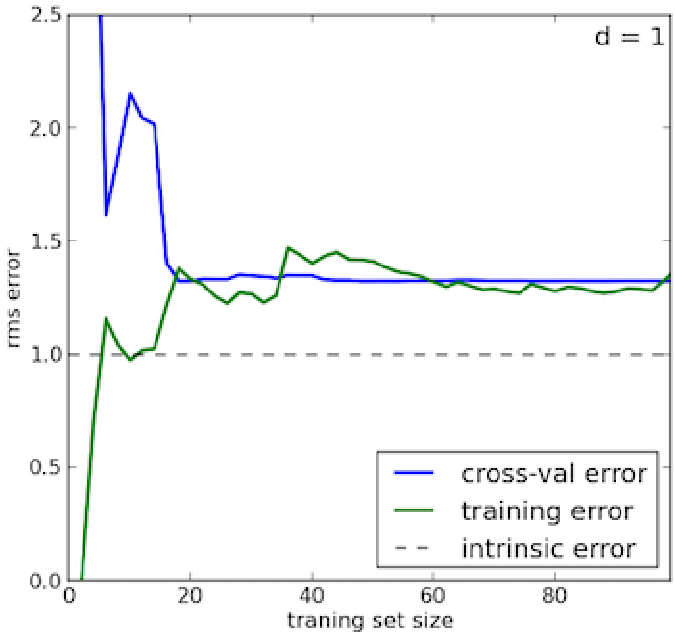
vchaudhuri What is the Bayesian Information criterion

- is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred.
- It is a measure of the relative quality of a statistical model for a given set of data
- It offers a relative estimate of the information lost when a given model is used to represent the process that generates the data
- is a method for selecting the most appropriate model among a set of competitors for a given data set

tdenatale Professor introduced heteroskedasticity, a technique to:

- Identify features through plotting residuals which may be a way to visualize nonlinearity better than plotting outcomes alone
- Reflect a consistent standard error that would be expected from a linear relationship
- Ability or method to identify unique bilinear features
- Show all model errors are consistent

nsh87 What can be said about the following learning curve?



- Adding additional points to the training data *will not* help improve the model
- Adding additional points to the training data *will* help improve the model
- The model is over-fitted
- The model is under-fitted

lakarbatti Which function can be used to fit **Generalized Linear Models**

- The **glm()** function
- The **lm()** function
- The **gen()** function

xxu26 The most common approach for evaluating the statistical assumptions in a gression 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 graphes is “Residual versus Leverage graph” which identifies outliers, high-leverage points, and inluential 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.

sneha-krishna What does the following code do?

```
a <- c(5,10,15,20)
b <- c(2,4,6,8)
df<- data.frame(a,b)
with (df, {df$a <- df$a+ 10; print(df)})
```

- modifies values of `a` in data frame `df` by adding 10 to each value in column `a`, then prints out `df`

- modifies column `a` in data frame `df` by adding the value 10 to the column, then prints out `df`
- modifies all values in data frame `df` by adding 10 to each value, then prints out `df`
- None of the above

Lecture 15a

vchaudhuri In Andrew Ng's lecture on Learning curves which of these statements is correct?

- A small training set results in a small training error but a large cross validation error
- Training error and cross validation error are directly proportaional to each other in a data set which fits a quadratic equation
- A small training data set leads to a large training error
- If a hypothesis has high bias increasing the training set size makes the line fit better

lakarbatti In statistics, what is **homoscedasticity**?

- A sequence or vector is **homoscedastic** if the variables in the sequence or vector have finite variance
- **homoscedasticity** is the science of measuring the coefficients in a dataset
- There is no such term as **homoscedasticity** in statistics
- A sequence is **homoscedastic** if the variables in the sequence have no or unequal variance

sneha-krishna True or false: one of the assumptions of the linear regression model is that there is no heteroscedasticity.

- True
- False
- Depend on the specific data set
- heteroscedasticity doesn't have anything to do with linear regression modelling.

Lecture 15b

lakarbatti Based on the video by Andrew Ng on learning curves, which of the following is a true statement for High Bias algorithms?

- If a learning algorithm is suffering from high bias, getting more data will not (by itself) help much in getting a lower cross validation or test set error
- High bias algorithms can easily be resolved with small data samples
- High bias algorithms are biased to sensitivity
- All of the statements are true