INFX 573 Final Exam

Pamela Chakrabarty

December 13th, 2016

In this lab, you will need access to the following R packages:

```
#Loading all the required libraries
library(dplyr)
library(ggplot2)
library(car)
library(boot)
```

```
## Warning: package 'boot' was built under R version 3.3.2
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.3.2
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 3.3.2
```

```
library(tidyverse)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.3.2
```

```
library(MASS)
library(pROC)
library(arm)
```

```
## Warning: package 'arm' was built under R version 3.3.2
```

```
## Warning: package 'lme4' was built under R version 3.3.2
```

```
library(xgboost)
```

```
## Warning: package 'xgboost' was built under R version 3.3.2
```

```
library("ISLR")
```

```
## Warning: package 'ISLR' was built under R version 3.3.2
```

```
library(gbm)
```

```
## Warning: package 'gbm' was built under R version 3.3.2
```

```
## Warning: package 'survival' was built under R version 3.3.2
```

Problem 1:

In this problem we will use data about infidelitys, known as the Fair's Affairs dataset. The 'Affairs' dataset is available as part of the AER package in R. This data comes from a survey conducted by Psychology Today in 1969, see Greene (2003) and Fair (1978) for more information. The dataset contains various self-reported characteristics of 601 participants, including how often the respondent engaged in extramarital sexual intercourse during the past year, as well as their gender, age, year married, whether they had children, their religiousness (on a 5-point scale, from 1=anti to 5=very), education, occupation (Hillinghead 7-point classification with reverse numbering), and a numeric self-rating of their marriage (from 1=very unhappy to 5=very happy).

(a) Describe the participants. Use descriptive, summarization, and exploratory techniques to describe the participants in the study. For example, what proportion of respondents are female? What is the average age of respondents?

affairs numeric. How often engaged in extramarital sexual intercourse during the past year?

Description of data

gender factor indicating gender.

age numeric variable coding age in years: 17.5 = under 20, 22 = 20-24, 27 = 25-29, 32 = 30-34, 37 = 35-39, 42 = 40-44, 47 = 45-49, 52 = 50-54, 57 = 55 or over.

yearsmarried numeric variable coding number of years married: 0.125 = 3 months or less, 0.417 = 4-6 months, 0.75 = 6 months-1 year, 1.5 = 1-2 years, 4 = 3-5 years, 7 = 6-8 years, 10 = 9-11 years, 15 = 12 or more years.

children factor. Are there children in the marriage?

religiousness numeric variable coding religiousness: 1 = anti, 2 = not at all, 3 = slightly, 4 = somewhat, 5 = very.

education numeric variable coding level of education: 9 = grade school, 12 = high school graduate, 14 = some college, 16 = college graduate, 17 = some graduate work, 18 = master's degree, 20 = Ph.D., M.D., or other advanced degree.

occupation numeric variable coding occupation according to Hollingshead classification (reverse numbering).

rating numeric variable coding self rating of marriage: 1 = very unhappy, 2 = somewhat unhappy, 3 = average, 4 = happier than average, 5 = very happy

Using exploratory data analysis techniques on the dataset:

```
library(AER)
```

```
## Warning: package 'AER' was built under R version 3.3.2
```

```
## Warning: package 'lmtest' was built under R version 3.3.2
```

```
## Warning: package 'zoo' was built under R version 3.3.2
```

```
## Warning: package 'sandwich' was built under R version 3.3.2
```

```
# Load Affairs dataset of AER package and saving it into a local variable
data("Affairs")
Affairs_mydata <- Affairs

# Shows entire dataset
View(Affairs_mydata)
# Shows  structure of the dataset
str(Affairs_mydata)
```

```
## 'data.frame':    601 obs. of  9 variables:
##  $ affairs      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ gender       : Factor w/ 2 levels "female","male": 2 1 1 2 2 1 1 2 1 2 ...
##  $ age          : num  37 27 32 57 22 32 22 57 32 22 ...
##  $ yearsmarried : num  10 4 15 15 0.75 1.5 0.75 15 15 1.5 ...
##  $ children     : Factor w/ 2 levels "no","yes": 1 1 2 2 1 1 1 2 2 1 ...
##  $ religiousness: int  3 4 1 5 2 2 2 2 4 4 ...
##  $ education    : num  18 14 12 18 17 17 12 14 16 14 ...
##  $ occupation   : int  7 6 1 6 6 5 1 4 1 4 ...
##  $ rating       : int  4 4 4 5 3 5 3 4 2 5 ...
```

```
# Summary of dataset
summary(Affairs_mydata)
```

```
##     affairs          gender         age         yearsmarried    children
##  Min.   : 0.000   female:315   Min.   :17.50   Min.   : 0.125   no :171
##  1st Qu.: 0.000   male  :286   1st Qu.:27.00   1st Qu.: 4.000   yes:430
##  Median : 0.000                Median :32.00   Median : 7.000
##  Mean   : 1.456                Mean   :32.49   Mean   : 8.178
##  3rd Qu.: 0.000                3rd Qu.:37.00   3rd Qu.:15.000
##  Max.   :12.000                Max.   :57.00   Max.   :15.000
##  religiousness     education       occupation        rating
##  Min.   :1.000   Min.   : 9.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:14.00   1st Qu.:3.000   1st Qu.:3.000
##  Median :3.000   Median :16.00   Median :5.000   Median :4.000
##  Mean   :3.116   Mean   :16.17   Mean   :4.195   Mean   :3.932
##  3rd Qu.:4.000   3rd Qu.:18.00   3rd Qu.:6.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :20.00   Max.   :7.000   Max.   :5.000
```

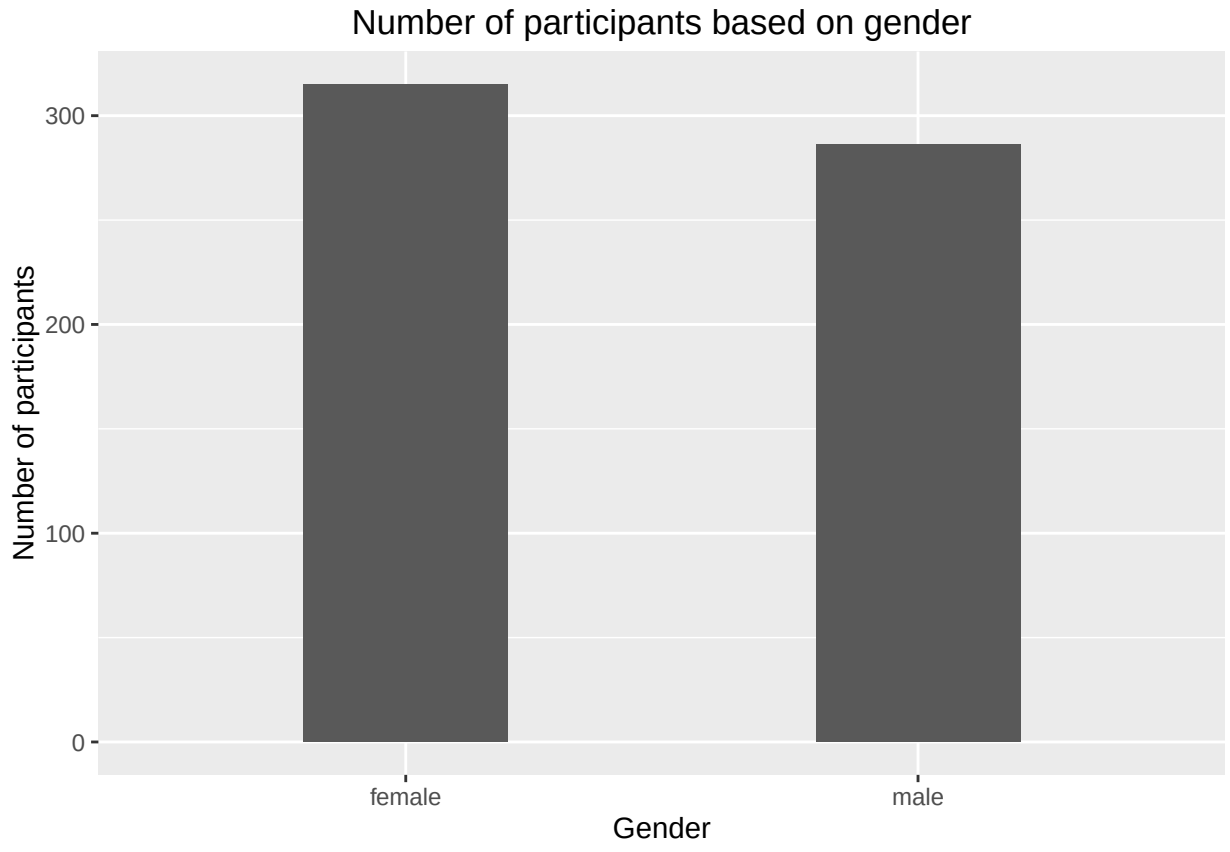
```
# Finding the proportion of male and female respondents
Affairs_mydata %>%
group_by(gender) %>%
summarise(total.participants = n()) %>%
ungroup() %>%
mutate(gender.prop = total.participants/sum(total.participants))
```

```
## # A tibble: 2 × 3
##   gender total.participants gender.prop
##   <fctr>              <int>       <dbl>
## 1 female                315   0.5241265
## 2   male                286   0.4758735
```

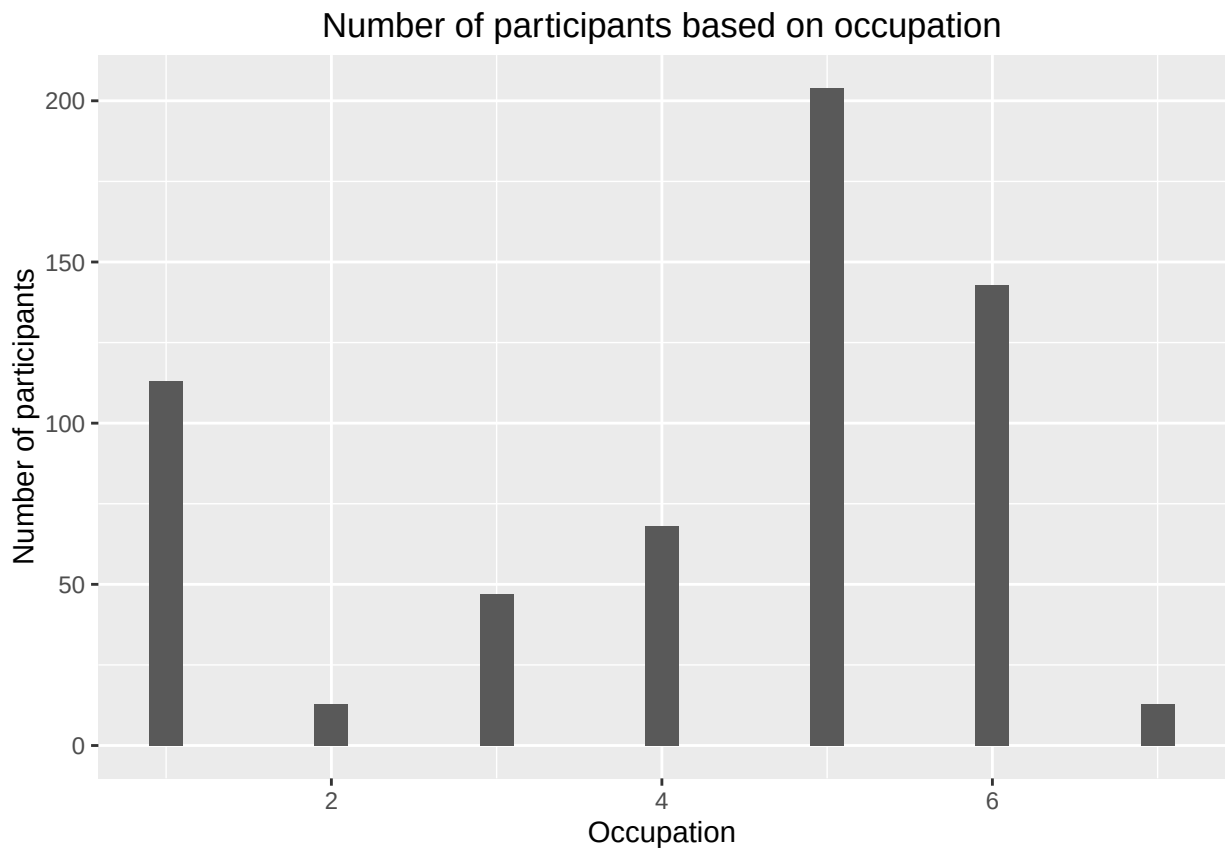
```
#Soln: Proportion of female participants: 0.524
#and, Proportion of male participants: 0.476

# Find average age of the participants
Affairs_mydata %>%
  summarise(avg.age = mean(age))
```

```
##    avg.age
## 1 32.48752
```

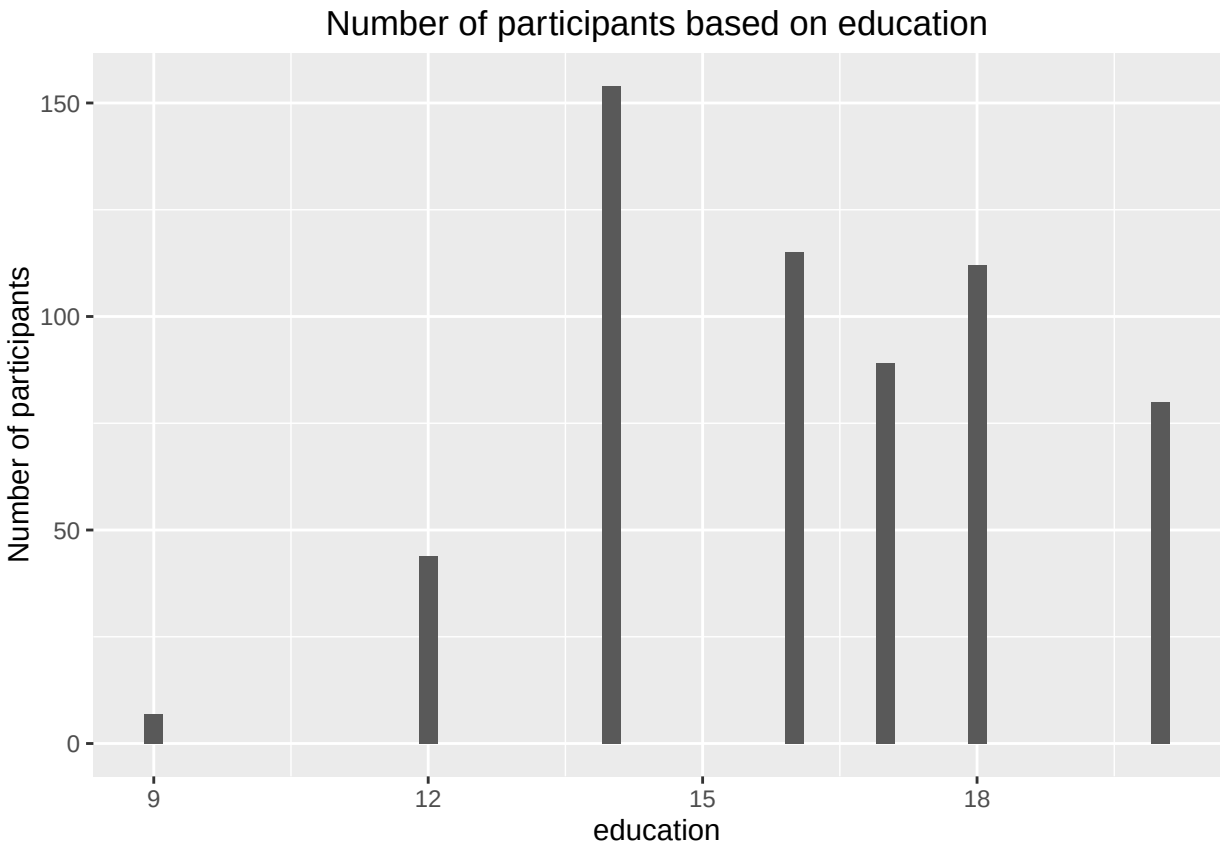
```
#Soln: Average age of  participants: 32.5


# Plotting number of participants based on gender
ggplot(Affairs_mydata, aes(gender)) + stat_count(width = 0.4) +
labs(x="Gender", y="Number of participants") +
  ggtitle("Number of participants based on gender")
```




```
#Based on gender, we see that there were more female participants and male participants.

# Plotting the frequency of participants based on occupation
ggplot(Affairs_mydata, aes(occupation)) + geom_bar(width = 0.2) +
labs(x="Occupation", y="Number of participants") +
ggtitle("Number of participants based on occupation")
```




```
#Based on occupation, highest number of participants were from Class 5 and least was from Class 2.

ggplot(Affairs_mydata, aes(education)) + geom_bar(width = 0.2) +
labs(x="education", y="Number of participants") +
ggtitle("Number of participants based on education")
```



Soln:Proportion of female participants: 0.524 and Proportion of male participants: 0.476 Average age of participants: 32.5

Based on gender as exploratory variable, we see from first plot that there were more female participants and male participants. Based on occupation as exploratory, we see from the second plot that the highest number of participants were from Class 5 and least was from Class 2.

(b) Suppose we want to explore the characteristics of participants who engage in extramarital sexual intercourse (i.e. affairs). Instead of modeling the number of affairs, we will consider the binary outcome - had an affair versus didn't have an affair. Create a new variable to capture this response variable of interest. Soln: To approach this problem, we can consider a new binary variable named "having.affair" with value 0 (meaning No affair) and value 1 (meaning having an affair). The count of affairs is used tocalculate the binary variable "having,afair". The binary variable will be labelled as Yes (with value 1) if number of affairs is more than 0 else labelled as No.

```
# Set a binary variable of someone having an affair(as 0) or not(as 1)
# If the number of affairs is equal to 0, binary variable is set to 0
Affairs_mydata$having.affair[Affairs_mydata$affairs == 0] <- 0
#if affair is more than 0, then we set binary variable as 1
Affairs_mydata$having.affair[Affairs_mydata$affairs  > 0] <- 1


# Using factor to label 0 as No and 1 as Yes
Affairs$having.affair <- factor(Affairs_mydata$having.affair, levels=c(0,1), labels=c("No","Yes"))

# Finding the count of new binary variables
data.frame(table(Affairs$having.affair))
```

```
##   Var1 Freq
## 1   No  451
## 2  Yes  150
```

Soln: We see that a total of 150 participants has an afair(Yes) while 451 participants did not have any affair (No).

(c) Use an appropriate regression model to explore the relationship between having an affair and other personal characteristics. Comment on which covariates seem to be predictive of having an affair and which do not.

```
# Fit a logistic regression model ( as we are dealing with a binary response varianle) with
#having.affair as the response variable and other personal characteristics as predictor variables
affair_predictors.fit <- glm(having.affair ~ gender + age + yearsmarried + children +
                 religiousness + education + occupation +rating,
               data=Affairs_mydata,family=binomial())

# Summary statistics of our fitted model.
summary(affair_predictors.fit)
```

```
##
## Call:
## glm(formula = having.affair ~ gender + age + yearsmarried + children +
##     religiousness + education + occupation + rating, family = binomial(),
##     data = Affairs_mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5713  -0.7499  -0.5690  -0.2539   2.5191
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.37726    0.88776   1.551 0.120807
## gendermale     0.28029    0.23909   1.172 0.241083
## age           -0.04426    0.01825  -2.425 0.015301 *
## yearsmarried   0.09477    0.03221   2.942 0.003262 **
## childrenyes    0.39767    0.29151   1.364 0.172508
## religiousness -0.32472    0.08975  -3.618 0.000297 ***
## education      0.02105    0.05051   0.417 0.676851
## occupation     0.03092    0.07178   0.431 0.666630
## rating        -0.46845    0.09091  -5.153 2.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 609.51  on 592  degrees of freedom
## AIC: 627.51
##
## Number of Fisher Scoring iterations: 4
```

Soln: Summary stats observations show:

i) The summary stats of the fitted model indicates age, yearsmarried, religiousness and rating as statistically significant variables with a p-value < 0.05. Religiousness and rating are statistically significant at 0.001 level while yearsmarried is at 0.01 level and age at 0.05 level. Hence, these can be used as predictor variables for determing if someone is having affair or not.

ii) Coefficient estimate of religiousness is negative (-0.3247). This means that for a increase in every 1 unit in religiousness, the odds of having an affair reduces by 0.3247

Coef estimate of rating is negative (-0.4685). This means that for a increase in every 1 unit in religiousness, the odds of having an affair reduces by 0.4685

Coef estimate of yearsmarried is positive(0.0948). This means that for a increase in every 1 unit in religiousness, the odds of having an affair rises by 0.0948

Coef estimate of age is negative(-0.0443). This means that for a increase in every 1 unit in religiousness, the odds of having an affair reduces by 0.4685

(d) Use an all subsets model selection procedure to obtain a "best" fit model. Is the model different from the full model you fit in part (c)? Which variables are included in the "best" fit model? You might find the bestglm() function available in the bestglm package helpful.

```
library(bestglm)
```

```
## Warning: package 'bestglm' was built under R version 3.3.2
```

```
## Loading required package: leaps
```

```
## Warning: package 'leaps' was built under R version 3.3.2
```

```
# Creating a new column caled bestglm
Affairs_mydata$bestglm <- Affairs_mydata$having.affair
# Replacing column "having.affair" of Affairs.data dataset to fit bestglm
Affairs.bestglm <- Affairs_mydata[,c("gender","age","yearsmarried","children",
                                     "religiousness", "education",
                                     "occupation", "rating", "bestglm")]

# Using bestglm to perform subset model selection
set.seed(1)
bestglm.fit<- bestglm(Affairs.bestglm, family = binomial,
                      method = "exhaustive")
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
# Displaying the summary statistic of the Best Model
bestglm.fit$BestModel
```

```
##
## Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Coefficients:
##   (Intercept)   yearsmarried  religiousness         rating
##       1.13820        0.05545       -0.33065       -0.45332
##
## Degrees of Freedom: 600 Total (i.e. Null);  597 Residual
## Null Deviance:       675.4
## Residual Deviance: 619.6     AIC: 627.6
```

Soln: We see that the bestglm best fit model differs from the previous logistic regression model in that bestb fit model using bestglm function shows 3 statistically relevant predictor variables namelely yearsmarried, religiousness and rating while the logistic regression model had shown 4 statitically significant predictor variables namely age along with yearsmarried, religiousness and rating.

(e) Interpret the model parameters using the model from part (d).

Soln:

```
summary(bestglm.fit$BestModel)
```

```
##
## Call:
## glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5554  -0.7481  -0.5775  -0.3506   2.3676
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.13820    0.45802   2.485 0.012953 *
## yearsmarried   0.05545    0.01901   2.917 0.003537 **
## religiousness -0.33065    0.08879  -3.724 0.000196 ***
## rating        -0.45332    0.08809  -5.146 2.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 619.64  on 597  degrees of freedom
## AIC: 627.64
##
## Number of Fisher Scoring iterations: 4
```

From summary stats of above best fit model, we see the intercept estimate as 1.1382. This indicates that there is a statistically significant association between response variable having.affair and the predictor variables. Additionally, we also see that

i) The coeffcient of yearsmarried(0.0555) is positive, meaning that for increase in every 1 unit in yearsmarried, the log odds of having an affair rises by 0.0555.

ii) the coeffcient of religiousness(-0.3306) is negative, meaning that for increase in every 1 unit in religiousness,the log odds of having an affair reduces by 0.3306.

iii) coeffcient of rating(-0.4533) is negative, meaning that for increase in every 1 unit in rating, the log odds of having an affair reduces by 0.4533.

(f) Create an artificial test dataset where martial rating varies from 1 to 5 and all other variables are set to their means. Use this test dataset and the predict function to obtain predicted probabilities of having an affair for case in the test data. Interpret your results and use a visualization to support your interpretation.

```
# Forming an artificial test dataset
arti_testdata <- data.frame(age=mean(Affairs_mydata$age), education=mean(Affairs_mydata$education),
                      yearsmarried=mean(Affairs_mydata$yearsmarried),
                      religiousness=mean(Affairs_mydata$religiousness),
                       rating=c(1:5))

# Using predict function to obtain predicted probabilities of having an affair in test data
arti_testdata$prob <- predict(bestglm.fit$BestModel, arti_testdata, type="response")

#Showing probability of affair corresponding from rating 1 to 5 respectively
arti_testdata$prob
```

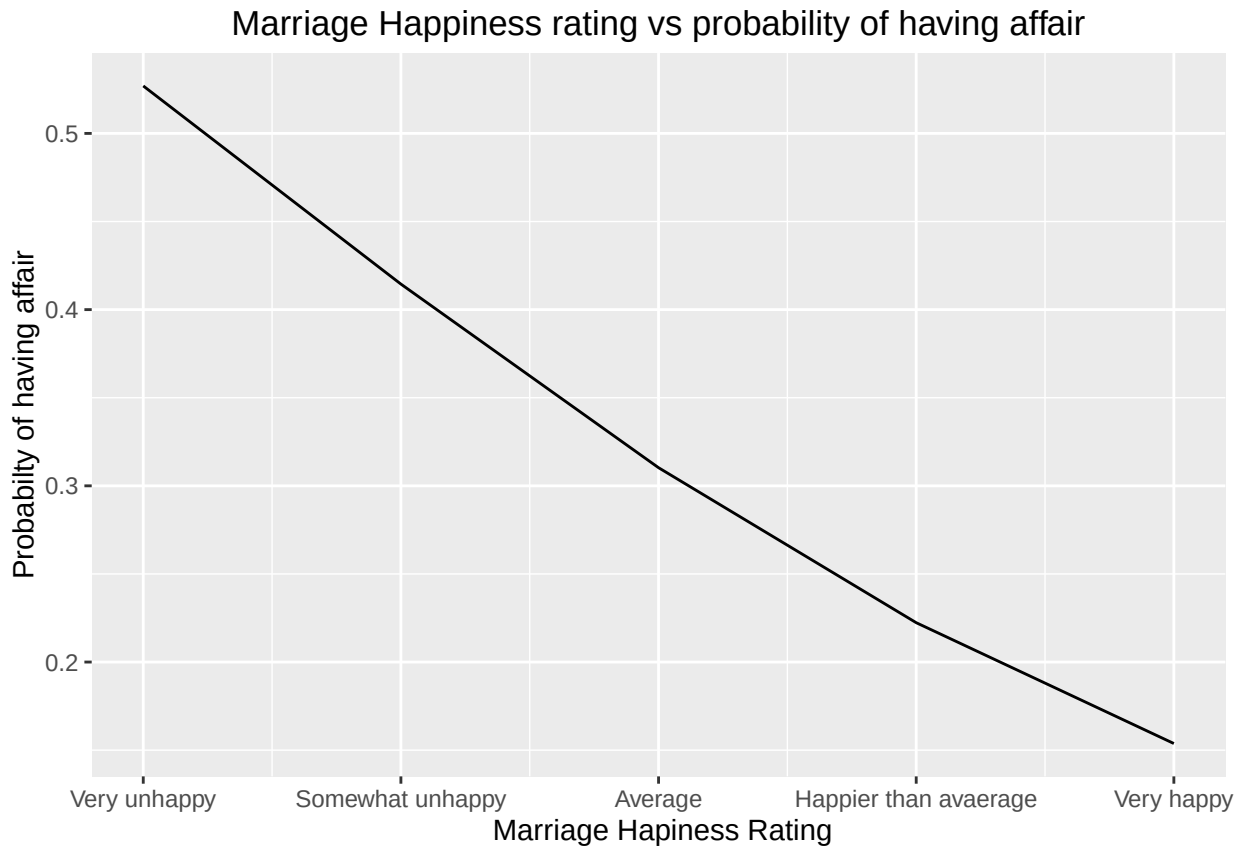
```
## [1] 0.5269478 0.4144913 0.3102921 0.2223405 0.1537609
```

```
# Showing our test data
arti_testdata
```

```
##        age education yearsmarried religiousness rating      prob
## 1 32.48752  16.16639     8.177696      3.116473      1 0.5269478
## 2 32.48752  16.16639     8.177696      3.116473      2 0.4144913
## 3 32.48752  16.16639     8.177696      3.116473      3 0.3102921
## 4 32.48752  16.16639     8.177696      3.116473      4 0.2223405
## 5 32.48752  16.16639     8.177696      3.116473      5 0.1537609
```

We see from above code chunk that probability of having an affair reduces from 0.527 to 0.154 with increase in ratings of hapiness in marriage (1 being very unhappy to 5 being very happy)

```
# Plotting rating of marriage vs probability of having an affair
ggplot(arti_testdata, aes(rating, prob)) +
labs(x=" Marriage Hapiness Rating",y="Probabilty of having affair") +
ggtitle("Marriage Happiness rating vs probability of having affair") + geom_line() +
scale_x_continuous(breaks=c(1:5), labels=c("Very unhappy", "Somewhat unhappy",
                           "Average", "Happier than avaerage",
                             "Very happy"))
```



As seen from the aboul visual plot, the plot supports the above observation that probability of having afair reduces as marriage happiness ratings increases from very unhappy to very happy

Problem 2:

In this problem we will revisit the state dataset. This data, available as part of the base R package, contains various data related to the 50 states of the United States of America. Suppose you want to explore the relationship between a state's Murder rate and other characteristics of the state, for example population, illiteracy rate, and more. Follow the questions below to perform this analysis. (a) Examine the bivariate relationships present in the data. Briefly discuss notable results. You might find the scatterplotMatrix() function available in the car package helpful.

```
#Loading R packages
library(Sleuth3)
```

```
## Warning: package 'Sleuth3' was built under R version 3.3.2
```

```
library(UsingR)
```

```
## Warning: package 'UsingR' was built under R version 3.3.2
```

```
## Warning: package 'HistData' was built under R version 3.3.2
```

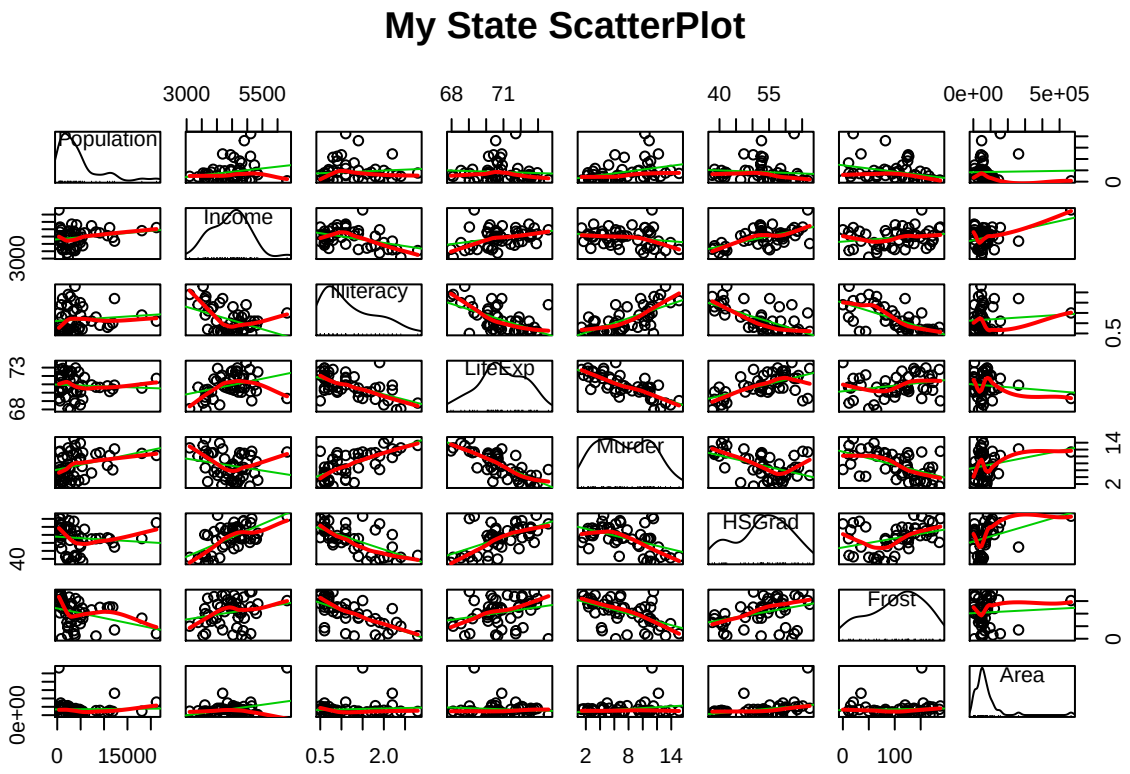
```
## Warning: package 'Hmisc' was built under R version 3.3.2
```

```
## Warning: package 'Formula' was built under R version 3.3.2
```

```
library(car)
library(MASS)


# Saving  state.x77 data into a local variable
state_mydata <- as.data.frame(state.x77)

# Renaming column names "HS Grad" and "Life Exp" to avoid spacing between words.
colnames(state_mydata)[colnames(state_mydata)=="HS Grad"] <- "HSGrad"
colnames(state_mydata)[colnames(state_mydata)=="Life Exp"] <- "LifeExp"

# Create correlation matrix
cor(state_mydata)
```

```
##             Population     Income  Illiteracy     LifeExp     Murder
## Population  1.00000000  0.2082276  0.10762237 -0.06805195  0.3436428
## Income      0.20822756  1.0000000 -0.43707519  0.34025534 -0.2300776
## Illiteracy  0.10762237 -0.4370752  1.00000000 -0.58847793  0.7029752
## LifeExp    -0.06805195  0.3402553 -0.58847793  1.00000000 -0.7808458
## Murder      0.34364275 -0.2300776  0.70297520 -0.78084575  1.0000000
## HSGrad     -0.09848975  0.6199323 -0.65718861  0.58221620 -0.4879710
## Frost      -0.33215245  0.2262822 -0.67194697  0.26206801 -0.5388834
## Area        0.02254384  0.3633154  0.07726113 -0.10733194  0.2283902
##                 HSGrad      Frost        Area
## Population -0.09848975 -0.3321525  0.02254384
## Income      0.61993232  0.2262822  0.36331544
## Illiteracy -0.65718861 -0.6719470  0.07726113
## LifeExp     0.58221620  0.2620680 -0.10733194
## Murder     -0.48797102 -0.5388834  0.22839021
## HSGrad      1.00000000  0.3667797  0.33354187
## Frost       0.36677970  1.0000000  0.05922910
## Area        0.33354187  0.0592291  1.00000000
```

```
# Plotting a scatterplot matrix to look at bivariate relationships
scatterplotMatrix(state_mydata, spread=FALSE, main="My State ScatterPlot")
```



Results observed from the scatter plot:

From the scatter plot, we can observe that Murder rate is bimodal and predictor variables show some sort of skewedness. We observe the fiollowing: Income increases with HSGrad while Illiteracy decreases with HSGrad Murder rate tends to increase with Population (r=0.344), Illiteracy(0.703) and Area(0.228) Murder rate tends to decrease with Income(r=-0.230), LifeExp(-0.781), HSGrad(-0.488) and Frost(-0.539)

(b) Fit a multiple linear regression model. How much variance in the murder rate across states do the predictor variables explain?

```
# Fit a multiple linear regression model with all 7 predictor variables
state_mydata.fit <- lm(Murder ~ Population + Income + Illiteracy + LifeExp +
                  HSGrad +  Frost + Area, data = state_mydata)

#summary statistics of above regression model
summary(state_mydata.fit)
```

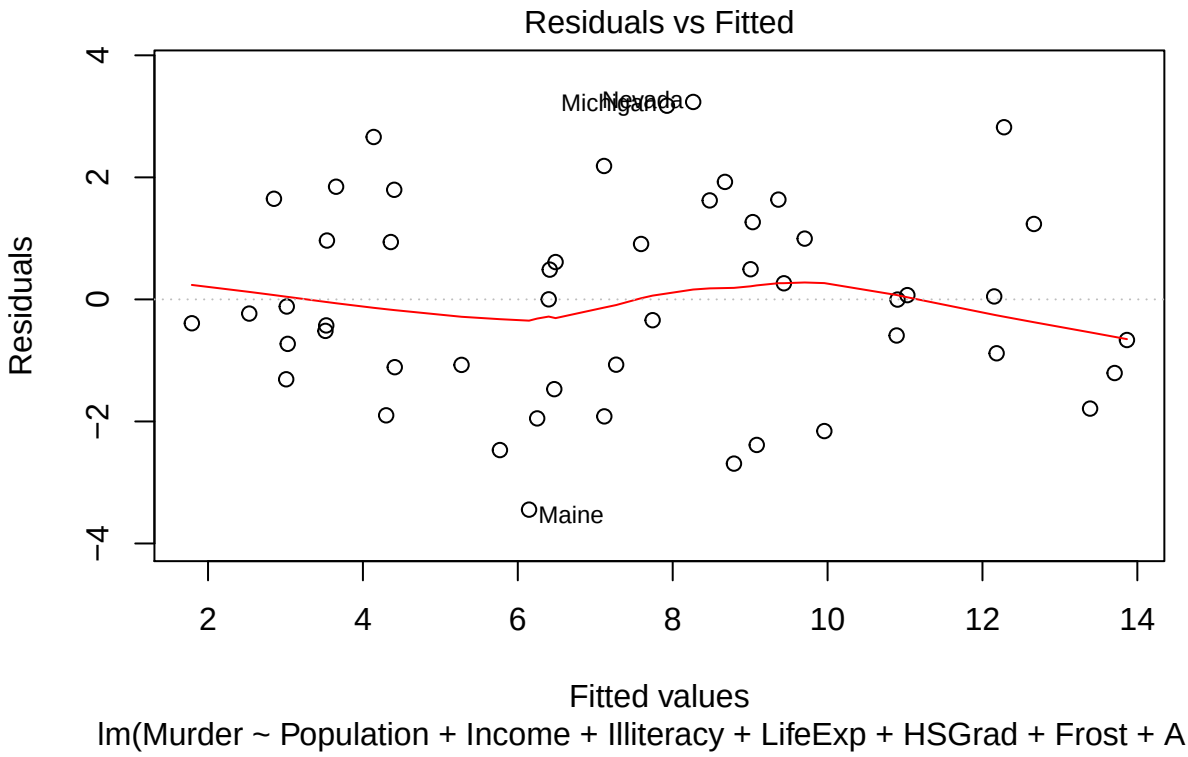
```
##
## Call:
## lm(formula = Murder ~ Population + Income + Illiteracy + LifeExp +
##     HSGrad + Frost + Area, data = state_mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4452 -1.1016 -0.0598  1.1758  3.2355
##
```

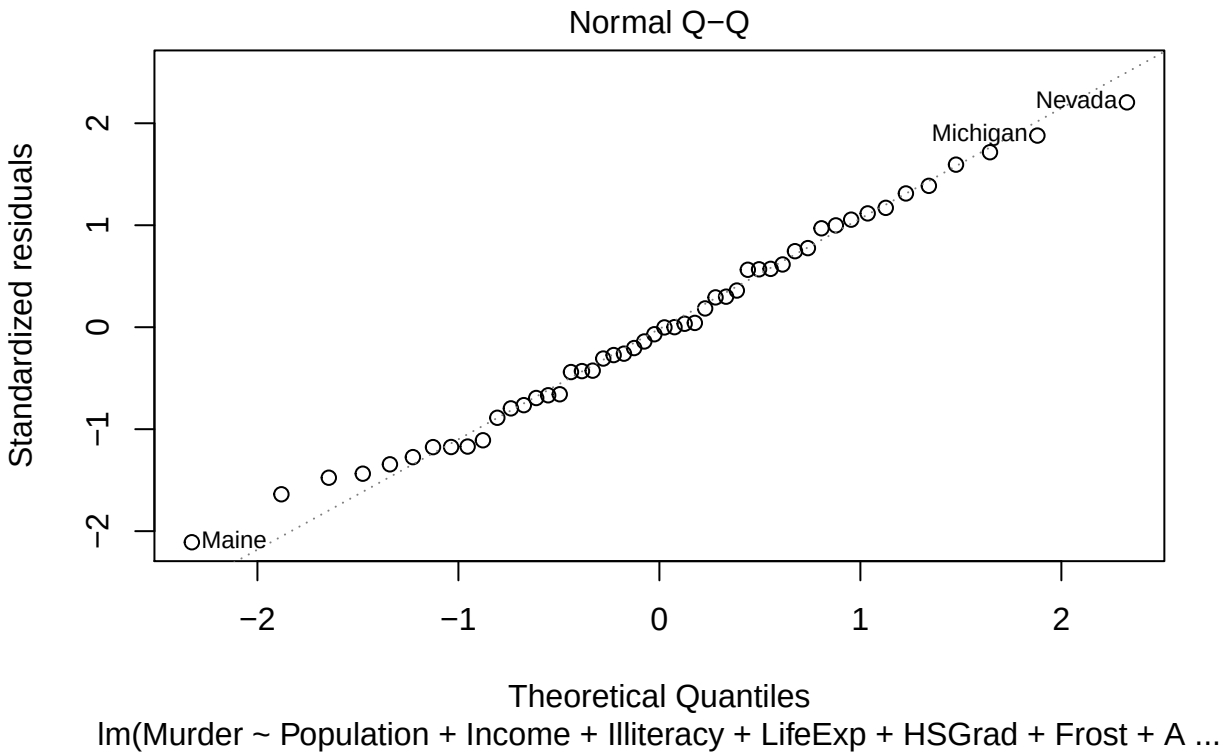
```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.222e+02  1.789e+01   6.831 2.54e-08 ***
## Population   1.880e-04  6.474e-05   2.905  0.00584 **
## Income      -1.592e-04  5.725e-04  -0.278  0.78232
## Illiteracy   1.373e+00  8.322e-01   1.650  0.10641
## LifeExp     -1.655e+00  2.562e-01  -6.459 8.68e-08 ***
## HSGrad       3.234e-02  5.725e-02   0.565  0.57519
## Frost       -1.288e-02  7.392e-03  -1.743  0.08867 .
## Area         5.967e-06  3.801e-06   1.570  0.12391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.746 on 42 degrees of freedom
## Multiple R-squared:  0.8083, Adjusted R-squared:  0.7763
## F-statistic: 25.29 on 7 and 42 DF,  p-value: 3.872e-13
```

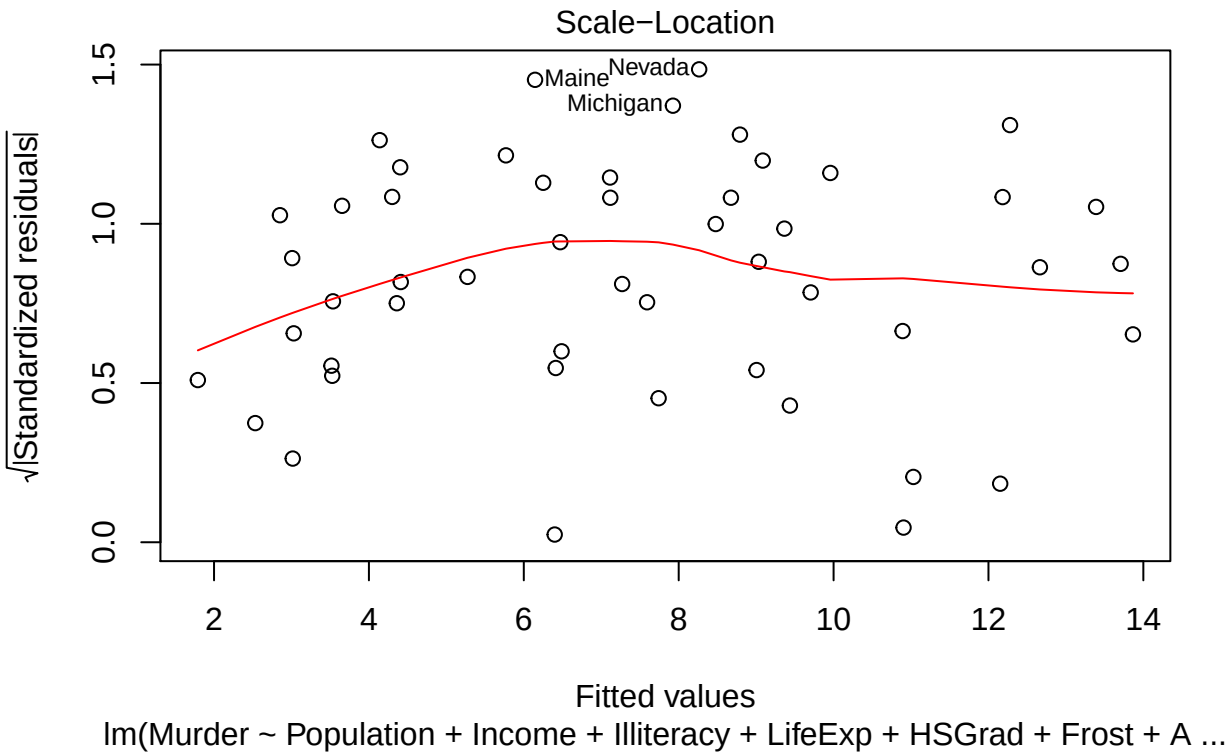
Answer: The summary results show Multiple R-suared value (indicating measure of variance as found in the model) as 0.8083. This indicates that about 80.83% of variance in murder rate across states can be predicted against the predictor variables. We also see from the summary stats that different predictor variables are statistically significant at different levels. For instance, Population is at 0.05 level and LifeExp at 0.001.

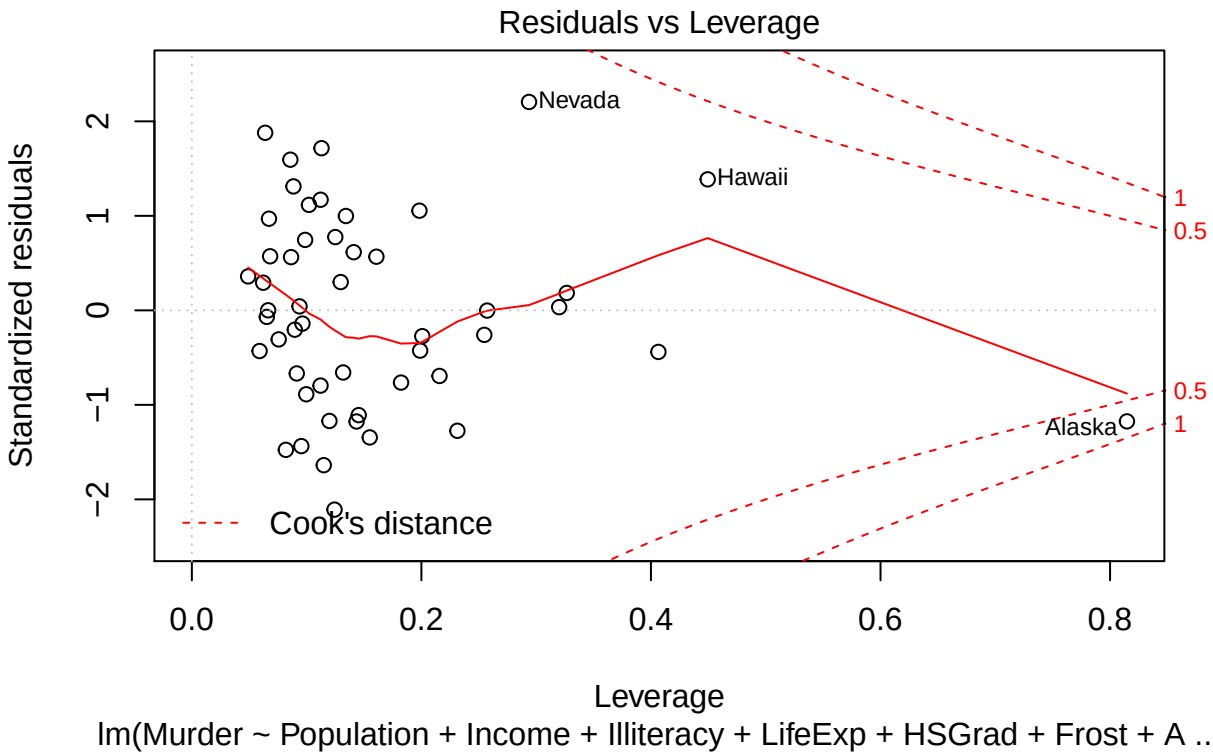
(c) Evaluate the statistical assumptions in your regression analysis from part (b) by performing a basic analysis of model residuals and any unusual observations. Discuss any concerns you have about your model.

```
# Ploting fitted model
plot(state_mydata.fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Murder ~ Population + Income + Illiteracy + LifeExp + HSGrad + Frost + A ...
Michigan
Nevada
Maine
4
2
0
−2
−4
2
4
6
8
10
12
14


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Murder ~ Population + Income + Illiteracy + LifeExp + HSGrad + Frost + A ...
Nevada
Michigan
Maine
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
Maine
Nevada
Michigan
2
4
6
8
10
12
14
Fitted values
lm(Murder ~ Population + Income + Illiteracy + LifeExp + HSGrad + Frost + A ...



Some statistical assumptions in the regression analysis are:

i) We assumed that each of the response variables are independent of each other. If indeed there was a dependency among them, then our model fit might not be said as accuarate
ii) We assumed that the dependent(response) variable has a normal distribution for a set or independent(predictor) values. Which is what we see in the Normal Q-Q plot where all the datapoints lie on the same straight line.
iii) We also assumed that the constant variance of the dependent variables is satisfied as can be seen from the plot.

Some concerns about the above model are:

i) We see from the plots some corelation between predictor variables like (HSGrad and Illiteracy) and (HSGrad and Income), but we do not know how corelation amongst each of these predictor variables effects the response variable(Murder Rate). This can reduces accuracy of the regression coefficients of the model.

ii) In the Residuals vs Fitted plot, we see that Nevada and Michigan (have high positive residuals) and Maine (has high negative residual values) which means that these are outliers and they need to be considered in the model analysis as they do not fit the model well.

(d) Use a stepwise model selection procedure of your choice to obtain a "best" fit model. Is the model different from the full model you fit in part (b)? If yes, how so?

```r
#Stepwise model selection to get a best fit model
state_mydata_bestfit <- step(state_mydata.fit, data = state_mydata)
```

```
## Start:  AIC=63.01
## Murder ~ Population + Income + Illiteracy + LifeExp + HSGrad +
##     Frost + Area
##
##              Df Sum of Sq    RSS    AIC
## - Income      1     0.236 128.27 61.105
## - HSGrad      1     0.973 129.01 61.392
## <none>                    128.03 63.013
## - Area        1     7.514 135.55 63.865
## - Illiteracy  1     8.299 136.33 64.154
## - Frost       1     9.260 137.29 64.505
## - Population  1    25.719 153.75 70.166
## - LifeExp     1   127.175 255.21 95.503
##
## Step:  AIC=61.11
## Murder ~ Population + Illiteracy + LifeExp + HSGrad + Frost +
##     Area
##
##              Df Sum of Sq    RSS    AIC
## - HSGrad      1     0.763 129.03 59.402
## <none>                    128.27 61.105
## - Area        1     7.310 135.58 61.877
## - Illiteracy  1     8.715 136.98 62.392
## - Frost       1     9.345 137.61 62.621
## - Population  1    27.142 155.41 68.702
## - LifeExp     1   127.500 255.77 93.613
##
## Step:  AIC=59.4
## Murder ~ Population + Illiteracy + LifeExp + Frost + Area
##
##              Df Sum of Sq    RSS    AIC
## <none>                    129.03 59.402
## - Illiteracy  1     8.723 137.75 60.672
## - Frost       1    11.030 140.06 61.503
## - Area        1    15.937 144.97 63.225
## - Population  1    26.415 155.45 66.714
## - LifeExp     1   140.391 269.42 94.213
```

```r
# Summary statistics of above model
summary(state_mydata_bestfit)
```

```
##
## Call:
## lm(formula = Murder ~ Population + Illiteracy + LifeExp + Frost +
##     Area, data = state_mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2976 -1.0711 -0.1123  1.1092  3.4671
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.202e+02  1.718e+01   6.994 1.17e-08 ***
## Population   1.780e-04  5.930e-05   3.001  0.00442 **
## Illiteracy   1.173e+00  6.801e-01   1.725  0.09161 .
## LifeExp     -1.608e+00  2.324e-01  -6.919 1.50e-08 ***
## Frost       -1.373e-02  7.080e-03  -1.939  0.05888 .
## Area         6.804e-06  2.919e-06   2.331  0.02439 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.712 on 44 degrees of freedom
## Multiple R-squared:  0.8068, Adjusted R-squared:  0.7848
## F-statistic: 36.74 on 5 and 44 DF,  p-value: 1.221e-14
```

Soln: The stepwise best fit model is different from the full model in the following ways:

i) Unlike the previous full model, the best fit model shows only 5 predictor variables ,three of them (Population, LifeExp and Area) statistically significant with p-value <0.05 and the other two (Illiteracy and Frost) being statistically significant at 0.1 level.

ii) We had seen that the previous full model indicated Population and LifeExp as statistically significant predictor variables with p-value <0.05. But the stepwise model additionally indicates Area as another statistically significant predictor variable along with Population and LifeExp.

iii)We can can also find that Residual standard error has went down from 1.75 (full model) to 1.71 (best fit model). Adjusted R-squared value rose from 0.776(full model) to 0.785(best fit model). Multiple R-squared value has not shown any change as such while F-statistic has increased from 25.3(full model) to 36.7(best fit model).

(e) Assess the model (from part (d)) generalizability. Perform a 10-fold cross validation to estimate model performance. Report the results.

```
# Using glm on model
state_mydata.fit <- glm(Murder ~ Population + Income + Illiteracy + LifeExp +
                  HSGrad +  Frost + Area, data = state_mydata)

# Fitting a stepwise model selection for best fit model
state_mydata.bestfit <- step(state_mydata.fit, data = state_mydata)
```

```
## Start:  AIC=206.91
## Murder ~ Population + Income + Illiteracy + LifeExp + HSGrad +
##     Frost + Area
##
##              Df Deviance    AIC
## - Income      1   128.27 205.00
## - HSGrad      1   129.01 205.29
## <none>            128.03 206.91
## - Area        1   135.55 207.76
## - Illiteracy  1   136.33 208.05
## - Frost       1   137.29 208.40
## - Population  1   153.75 214.06
## - LifeExp     1   255.21 239.40
```

```
##
## Step:  AIC=205
## Murder ~ Population + Illiteracy + LifeExp + HSGrad + Frost +
##     Area
##
##              Df Deviance    AIC
## - HSGrad      1   129.03 203.30
## <none>            128.27 205.00
## - Area        1   135.58 205.77
## - Illiteracy  1   136.98 206.29
## - Frost       1   137.61 206.51
## - Population  1   155.41 212.60
## - LifeExp     1   255.77 237.51
##
## Step:  AIC=203.3
## Murder ~ Population + Illiteracy + LifeExp + Frost + Area
##
##              Df Deviance    AIC
## <none>            129.03 203.30
## - Illiteracy  1   137.75 204.57
## - Frost       1   140.06 205.40
## - Area        1   144.97 207.12
## - Population  1   155.45 210.61
## - LifeExp     1   269.42 238.11
```

```
# Finfing mean squared error of  best fit model which results to be 2.58
mean((state_mydata$Murder-predict(state_mydata.bestfit, state_mydata))^2)
```

```
## [1] 2.580632
```

```
# Using 10 fold cross validation
set.seed(1)
#Using cv.glm function to find estimated K-fold cross-validation error for linear model
cv.state_data <- cv.glm(state_mydata, state_mydata.bestfit, K = 10)

# Geting cross validation results
cv.state_data$delta
```

```
## [1] 3.842053 3.755144
```

```
# the first is standard K-fold cross validation estimate result which is 3.84
# thr second is the biased modified result which is 3.76
```

Soln: The mean squared error estimates from linear model(2.58) is lower than the cross validation error estimate. "‘ (f) Fit a regression tree using the same covariates in your "best" fit model from part (d). Use cross validation to select the "best" tree.

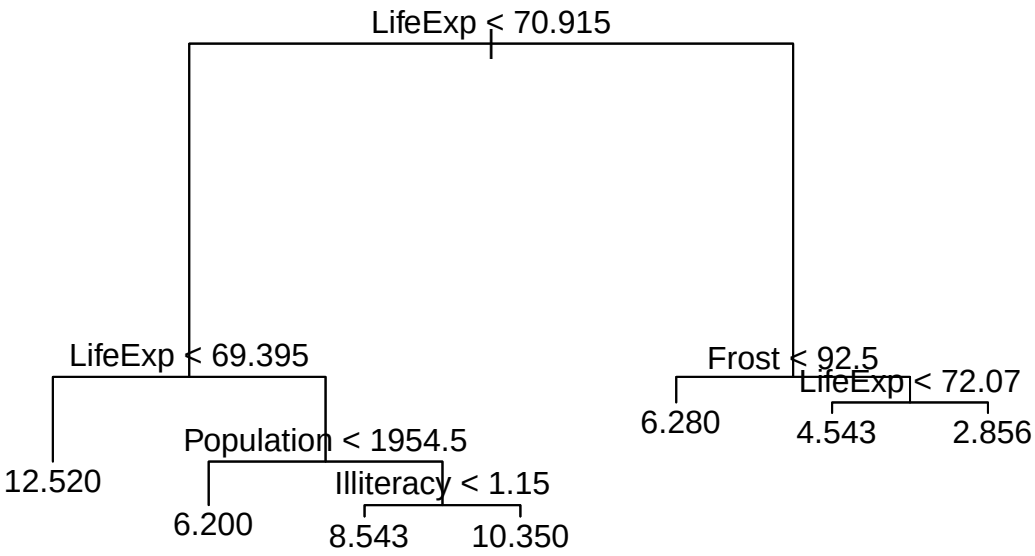
```
library(tree)
```

```
## Warning: package 'tree' was built under R version 3.3.2
```

```r
#We are fiting the tree with same covariates as the best fit stepwise model from question (d)
# Fit regression tree using same covariates in best fit model
tree.state_mydata <- tree(Murder ~ Population + Illiteracy + LifeExp + Frost + Area,
                  data = state_mydata)
# summary statistics
summary(tree.state_mydata)
```

```
##
## Regression tree:
## tree(formula = Murder ~ Population + Illiteracy + LifeExp + Frost +
##     Area, data = state_mydata)
## Variables actually used in tree construction:
## [1] "LifeExp"    "Population" "Illiteracy" "Frost"
## Number of terminal nodes:  7
## Residual mean deviance:  2.813 = 121 / 43
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.50000 -1.18900  0.02222  0.00000  0.74290  4.02000
```

```r
# Plotting tree
plot(tree.state_mydata)
#Labelling the tree nodes conditions
text(tree.state_mydata, pretty =0)
```



```r
# Doing cross validation
set.seed(1)
cv.tree_state_mydata<- cv.tree(tree.state_mydata)
# summary statistics of our decision tree
summary(cv.tree_state_mydata)
```

```
##        Length Class  Mode
## size   7      -none- numeric
## dev    7      -none- numeric
## k      7      -none- numeric
## method 1      -none- character
```

```r
#We pruned the decision tree to size 7 to get the best fit tree.
#Using prune.tree function as based on best=7 which was the tree size found from cv summary stats
prune.state_mydata <- prune.tree(tree.state_mydata, best = 7)
# Summary statistics of above pruned tree
summary(prune.state_mydata)
```

```
##
## Regression tree:
## tree(formula = Murder ~ Population + Illiteracy + LifeExp + Frost +
##     Area, data = state_mydata)
## Variables actually used in tree construction:
## [1] "LifeExp"    "Population" "Illiteracy" "Frost"
## Number of terminal nodes:  7
## Residual mean deviance:  2.813 = 121 / 43
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.50000 -1.18900  0.02222  0.00000  0.74290  4.02000
```

(g) Compare the models from part (d) and (f) based on their performance. Which do you prefer? Be sure to justify your preference.

Soln: We calculate mean squared error (MSE) for both models from part (d) and (f). The MSE of linesr best fit model is 2.58 which is more than the MSE of pruned tree model. So, we should prefer the prune tree model form part (f) as it shows relatively less mean squared error of the two.

```r
# Finding error rate of the linear best fit model from part (d)
mean((state_mydata$Murder-predict(state_mydata.bestfit, state_mydata))^2)
```

```
## [1] 2.580632
```

```r
# Finding  error rate of  pruned tree model from part (f)
mean((state_mydata$Murder-predict(prune.state_mydata, state_mydata))^2)
```

```
## [1] 2.41919
```

Problem 3:

Problem 3 (25 pts)
The Wisconsin Breast Cancer dataset is available as a comma-delimited text file on the UCI Machine Learning Repository http://archive.ics.uci.edu/ml. Our goal in this problem will be to predict whether observations (i.e. tumors) are malignant or benign. (a) Obtain the data, and load it into R by pulling it directly from the web. (Do not download it and import it from a CSV file.) Give a brief description of the data.

```
#loading data directly from the web
breastCancer_mydata <- read.csv("http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer
```

The Wisconsin Breast Cancer data shows 699 observations for 11 variables. The varaiables measure characteristics of tissue sample, storing it as a value between 1 to 10. The variables are as follows:

Sample code number referring to the Id number of the patient Clump Thickness: (1 - 10) Uniformity of Cell Size: (1 - 10) Uniformity of Cell Shape: (1 - 10) Marginal Adhesion: (1 - 10) Single Epithelial Cell Size: (1 - 10) Bare Nuclei: (1 - 10) Bland Chromatin: (1 - 10) Normal Nucleoli: (1 - 10) Mitoses: (1 - 10) Class: 2 for benign, 4 for malignant

(b) Tidy the data, ensuring that each variable is properly named and cast as the correct data type. Discuss any missing data.

```
head(breastCancer_mydata)
```

```
##        V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11
## 1 1000025  5  1  1  1  2  1  3  1   1   2
## 2 1002945  5  4  4  5  7 10  3  2   1   2
## 3 1015425  3  1  1  1  2  2  3  1   1   2
## 4 1016277  6  8  8  1  3  4  3  7   1   2
## 5 1017023  4  1  1  3  2  1  3  1   1   2
## 6 1017122  8 10 10  8  7 10  9  7   1   4
```

```
# Renaming the column names
colnames(breastCancer_mydata) <- c("Sample.Code.Number", "Clump.Thickness", "Uniformity.Of.Cell.Size",

# Looking for any missing data in the dataset
breastCancer_mydata[breastCancer_mydata == "?"] <- NA

# Deleting rows of dataset with missing values
breastCancer.data <- na.omit(breastCancer_mydata)

# To see structure of each variable of the dataset
str(breastCancer_mydata)
```

```
## 'data.frame':    699 obs. of  11 variables:
##  $ Sample.Code.Number         : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561
##  $ Clump.Thickness            : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ Uniformity.Of.Cell.Size    : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ Uniformity.Of.Cell.Shape   : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ Marginal.Adhesion          : int  1 5 1 1 3 8 1 1 1 1 ...
##  $ Single.Epithelial.Cell.Size: int  2 7 2 3 2 7 2 2 2 2 ...
##  $ Bare.Nuclei                : Factor w/ 11 levels "?","1","10","2",..: 2 3 4 6 2 3 3 2 2 2 ...
```

```
##  $ Bland.Chromatin           : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ Normal.Nucleoli           : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ Mitoses                   : int  1 1 1 1 1 1 1 1 5 1 ...
##  $ Class                     : int  2 2 2 2 2 4 2 2 2 2 ...
```

```
# Recasting "class"'s data type as from interger to factor type
breastCancer_mydata$Class <- as.factor(breastCancer_mydata$Class)

# Recasting datatype of "Bare Nuclei" from factor to integer type
breastCancer_mydata$BareN.uclei <- as.integer(breastCancer_mydata$Bare.Nuclei)
```

As seen above, some data tidy steps taken are as follows:

i) Renamed all the column variable names appropriately.
ii) Considered missing obervations in the dataset and removed rows with missing values
iii) Recasted datatype of "Class" from integer to factor type and recasted datatype of "Bare Nuclei" from factor to integer type

(c) Split the data into a training and validation set such that a random 70% of the observations are in the training set.

```
# Counting total rows in dataset
row.count <- nrow(breastCancer_mydata)
# Setting a random seed to produce results
set.seed(1)

# Assiging 70% of the observations to be in training dataset
#Using sample function to randomly calculate training dataset indices.
training <- sample(row.count, as.integer(0.7*row.count))
# Finding length of training dataset
length(training)
```

```
## [1] 489
```

```
# Defining the training datasset based on the indices obatined
training_Data <- breastCancer_mydata[training, ]
# Finding total number of observations in training dataset
nrow(training_Data)
```

```
## [1] 489
```

```
# Creating testing dataset
test_Data <- breastCancer_mydata[-training, ]
# Finding total number of observations in testing dataset
nrow(test_Data)
```

```
## [1] 210
```

Training data comprises 489 observations and the test data contains 210 observations.

(d) Fit a regression model to predict whether tissue samples are malignant or benign. Classify cases in the validation set. Compute and discuss the resulting confusion matrix.

```r
# Fit a logictic regression model to find if tissue samples are malignant or benign
breastCancer_reg.fit <- glm(Class ~ Clump.Thickness + Uniformity.Of.Cell.Size + Uniformity.Of.Cell.Shape
training_Data, family = binomial)

# Summary statistics of the fitted model
summary(breastCancer_reg.fit)
```

```
##
## Call:
## glm(formula = Class ~ Clump.Thickness + Uniformity.Of.Cell.Size +
##     Uniformity.Of.Cell.Shape + Marginal.Adhesion + Single.Epithelial.Cell.Size +
##     Bare.Nuclei + Bland.Chromatin + Normal.Nucleoli + Mitoses,
##     family = binomial, data = training_Data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.42292  -0.04390  -0.02409   0.01297   2.44277
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -12.1750     2.4575  -4.954 7.26e-07 ***
## Clump.Thickness                0.2993     0.1904   1.572 0.116037
## Uniformity.Of.Cell.Size       -0.2175     0.2487  -0.875 0.381838
## Uniformity.Of.Cell.Shape       0.6675     0.3156   2.115 0.034424 *
## Marginal.Adhesion              0.2457     0.1674   1.468 0.142142
## Single.Epithelial.Cell.Size    0.5367     0.2484   2.161 0.030708 *
## Bare.Nuclei10                  4.8206     1.4263   3.380 0.000726 ***
## Bare.Nuclei2                   2.3270     1.4243   1.634 0.102306
## Bare.Nuclei3                   3.8088     1.5917   2.393 0.016716 *
## Bare.Nuclei4                   3.6283     1.7346   2.092 0.036463 *
## Bare.Nuclei5                   1.9811     1.7124   1.157 0.247327
## Bare.Nuclei6                  16.5757  2966.9816   0.006 0.995542
## Bare.Nuclei7                   0.7333     1.5829   0.463 0.643197
## Bare.Nuclei8                   1.6095     1.2834   1.254 0.209816
## Bare.Nuclei9                  20.2657  1968.6561   0.010 0.991787
## Bland.Chromatin                0.5981     0.2409   2.483 0.013025 *
## Normal.Nucleoli                0.1725     0.1476   1.168 0.242756
## Mitoses                        0.5314     0.3406   1.560 0.118733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 617.213  on 474  degrees of freedom
## Residual deviance:  57.377  on 457  degrees of freedom
##   (14 observations deleted due to missingness)
## AIC: 93.377
##
## Number of Fisher Scoring iterations: 17
```

From the above fitted logistic regression model, we find that predictor variables Single.Epithelial.Cell.Size is statistically significant at level 0.05, Marginal.Adhesion and Bare.Nuclei are significant at level 0.01 level, while Clump.Thickness is statistically significant at level 0.001.

(e) Fit a random forest model to predict whether tissue samples are malignant or benign. Classify cases in the validation set. Compute and discuss the resulting confusion matrix.

From the confusion matrix, it can be observed that the model has a prediction accuracy of 0.974 or 97.4%. The test error rate is 0.0265, with only 5 obervations incorrectly classified as malignant.

(f) Compare the models from part (d) and (e) using ROC curves. Which do you prefer? Be sure to justify your preference.

Problem 4:

Problem 4 (25 pts) Please answer the questions below by writing a short response. (a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal in each application inference or predictions? Explain your answer.

Referenced : https://rpubs.com/ppaquay/65557

Soln: Three real-life applications of Classification could be

Example 1) Is this Netflix series campaign going to be successful or not.

Goal: Prediction Response: Success/Failure Predictors: Money spent, Running Time, Producer, TV Channel, Air time slot, Frequency of running the ad

Example 2) Should this applicant be admitted into University of Washington's MSIM program or not.

Goal: Prediction Response: Admit/Not admit Predictors: GRE Scores, GPA, SOP Essay, Relevant Work Experience, Letter of Recommendation, Career interest etc

Example 3) Does a wearable technology motivate students to exercise - Successful/Not Successful.

Goal: Prediction Response: Does a wearable technology motivate students to exercise Predictors: Time in hand, Number of times one workout in a week, Control/Test group, etc.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal in each application inference or predictions? Explain your answer. Answer: Three real-life applications of Regression could be

Example 1) GDP Growth in India

Goal: Inference Response: What is the GDP of India predicted to be by 2050 Predictors: Population, Per capita income, Education, Average life expectancy, Tax Revenue etc.

Example 2) What is the average apartment price in Bellevue over the next 3 years?

Goal: Inference Response: Average apartment in Bellevue will sell for $Y next year, $Z the year after, $T after that, etc. Predictors: Proximity to transit, Parks, Schools, Average size of family, Average Income of Family, Crime Rate, Price Flux in surrounding neighborhoods etc.

Example 3) which factors affect CEO salary

Goal: Inference Response: Bases on a dataset on top 500 US firms, these are factors that influence a CEO's salary Predictors: annual profit made, number of employees, industry, the CEO salary etc

(c) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Soln: The advantages of a very flexible approach are that it may give a better fit for non-linear models as it reduces the bias and it could help in making more accurate predictions.

The disadvantages are that it requires estimating a greater number of parameters, overfitting and it increases the variance and not good for linear models. Also, does not account for interpretability of results

A more flexible approach might be preferred in cases where we are interested in prediction and not the interpretability of the results. A less flexible approach would be preferred in cases where we are interested in inference and the interpretability of the results.

Citing reference: https://statlearning.wordpress.com/2013/12/22/chapter-2-conceptual-exercises/

Problem 5 (??? 5 pts) Suppose we have a dataset with five predictors, X1 = GPA, X2 = IQ, X3 = Gender (1 for Female, and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model and get Bˆ 0 = 50, Bˆ 1 = 20, Bˆ 2 = 0.07Bˆ 3 = 35, Bˆ 4 = 0.01, and Bˆ 5 = ???10. (a) Which answer is correct and why? i. For a fixed value of IQ and GPA, males earn more on average than females. ii. For a fixed value of IQ and GPA, females earn more on average than males. iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough. iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

Soln: We calculate the least square line which is given by Y=50+20GPA+0.07IQ+35Gender+0.01GPA×IQ ??? 10GPA×Gender

and which for the males will be YM=50+20GPA+0.07IQ+0.01GPA×IQ,

and which for females will be: YF=85+10GPA+0.07IQ+0.01GPA×IQ.

So we see that the starting salary for males is higher than for females on average if 50+20GPA>=85+10GPA which is equivalent to GPA>=3.5. So, iii. is the correct answer.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

Soln: Salary for this female= 50+20X4+0.07X110+35X1+4X110X0.01+4X1X(-10) =137.1 (in thousands of dollars) sO, the starting salary for a female with given condition is 137100 dollars

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is little evidence of an interaction effect. Justify your answer.

Soln: False. Although the coefficient for GPA/IQ interaction term is very small, yet we cannot infer that there is little evidence of interaction efect as the coefficient has no direct reltion with the evidence. We would need to test the hypothesus H0:B4$^{=0H0:B4}$=0 and consider the p-value associated with the tt to draw a conclusion. Also, it could be could be possible that the data has very low variance around the fit which would lead to a high interaction effect evn if it is a small value.

Problem 6 - Extra Credit (??? 5 pts) Apply boosting, bagging and random forests to a dataset of your choice that we have used in class. Be sure to fit the models on a training set and evaluate their performance on a test set. (a) How accurate are the results compared to simple methods like linear or logistic regression?

```
titanic_data <- read.csv('C:/Users/iGuest/Downloads/titanic.csv') #load data
str(titanic_data) # explore data structure
```

```
## 'data.frame':    1309 obs. of  14 variables:
##  $ pclass   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived : int  1 1 0 0 0 1 1 0 1 0 ...
##  $ name     : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 22 24 25 26 27 31 46 47 51 55 ...
##  $ sex      : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
```

```
##  $ age      : num  29 0.917 2 30 25 ...
##  $ sibsp    : int  0 1 1 1 1 0 1 0 2 0 ...
##  $ parch    : int  0 2 2 2 2 0 0 0 0 0 ...
##  $ ticket   : Factor w/ 929 levels "110152","110413",..: 188 50 50 50 50 125 93 16 77 826 ...
##  $ fare     : num  211 152 152 152 152 ...
##  $ cabin    : Factor w/ 187 levels "","A10","A11",..: 45 81 81 81 81 151 147 17 63 1 ...
##  $ embarked : Factor w/ 4 levels "","C","Q","S": 4 4 4 4 4 4 4 4 4 2 ...
##  $ boat     : Factor w/ 28 levels "","1","10","11",..: 13 4 1 1 1 14 3 1 28 1 ...
##  $ body     : int  NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest: Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 232 232 232 238 163 25 23 230 ...
```

```
titanic_data$pclass <- as.factor(titanic_data$pclass)
```

```
#Splitting data into a training (80% of observations) and test(20%)
#Finding out 80% of total number of rows
count80 <- floor(NROW(titanic_data)* 0.8)
set.seed(125)
#Creating a sample of size equal to 80% of the data set
sample_set <- sample(seq_len(nrow(titanic_data)), size = count80)
#Creating train and test data sets
train_data <- titanic_data[sample_set, ]
test_data <- titanic_data[-sample_set, ]
```

(b) Which of the approaches yields the best performance?

Problem 7 - Extra Credit (??? 5 pts) Suppose that X1, . . . Xn form a random sample from a Poisson distribution for which the mean ?? is unknown, (theta > 0). (a) Determine the MLE of theta, assuming that at least one of the observed values is different from 0. Show your work.

Soln: The likelihood function is fn(x|theta) = exp(ntheta)thetaˆy/n to i=1 xi! So, if y = Sum of n where i=1 xi > 0 and L(theta) = log fn(x|theta), the maximum of L(theta) is attained at ˆtheta = y/n = ¯xn

(b) Show that the MLE of theta does not exists if every observed value is 0.

Soln: If y = 0, since L(theta) is a decreasing function of theta and theta = 0 is not in the parameter space, the MLE does not exist.

Statement of Compliance: Please copy and sign the following statement.

I affirm that I have had no conversation regarding this exam with any persons other than the instructor (Dr. Emma Spiro). Further, I certify that the attached work represents my own thinking. Any information, concepts, or words that originate from other sources are cited in accordance with University of Washington guidelines as published in the Academic Code (available on the course website). I am aware of the serious consequences that result from improper discussions with others or from the improper citation of work that is not my own.

(signature): Pamela Chakrabarty (date): 12/12/2016