Exploratory Data Analysis on Affairs dataset

Maria George

December 14, 2015

```r
# Loading all the required libraries
library("dplyr")
library("ggplot2")
library("car") # Contains the scatterplotMatrix function
#install.packages("boot")
library("boot") # Perform crossvalidation
#install.packages("tree")
library("tree")
library("randomForest")
library(pROC) # Useful for computing and plotting classifer metrics
```

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).

Using descriptive, summarization, and exploratory techniques to describe the participants in the study.

```r
# install.packages("AER")
library(AER) # Contains Affairs dataset

# Loading the Affairs dataset of AER package and saving it into a local variable
data("Affairs")
Affairs.data <- Affairs

# Displaying the first few rows of the dataset
head(Affairs.data)
```

```
##    affairs gender age yearsmarried children religiousness education
## 4        0   male  37        10.00       no             3        18
## 5        0 female  27         4.00       no             4        14
## 11       0 female  32        15.00      yes             1        12
## 16       0   male  57        15.00      yes             5        18
## 23       0   male  22         0.75       no             2        17
## 29       0 female  32         1.50       no             2        17
##    occupation rating
## 4           7      4
## 5           6      4
## 11          1      4
## 16          6      5
## 23          6      3
## 29          5      5
```

```r
# Displaying the summary of the Affairs.data dataset
summary(Affairs.data)
```

```
##     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
```

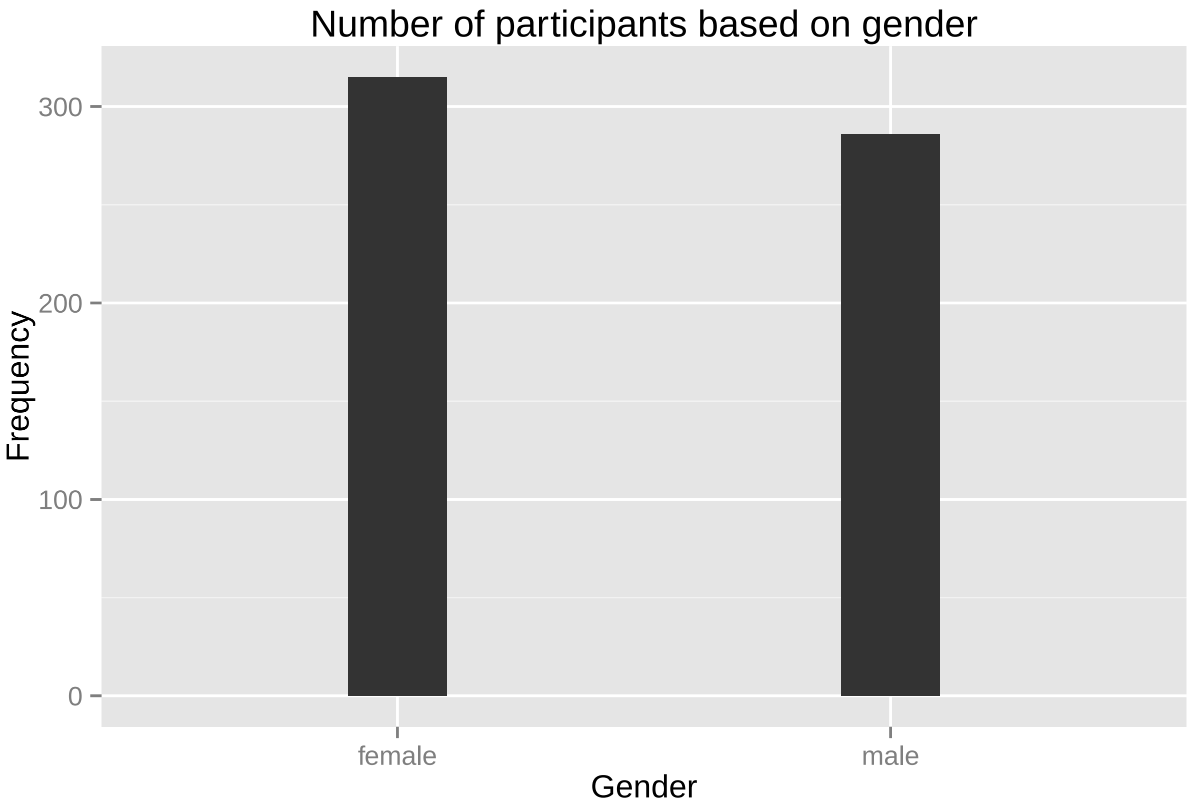
```r
# Displaying the structure of the Affairs.data dataset
str(Affairs.data)
```

```
## '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 ...
```

```
# Finding the proportion of male and female respondents
Affairs.data %>%
  group_by(gender) %>%
  summarise(total_participants = n()) %>%
  ungroup() %>%
  mutate(prop_gender = total_participants/sum(total_participants))
```

```
## Source: local data frame [2 x 3]
##
##   gender total_participants prop_gender
##   (fctr)              (int)       (dbl)
## 1 female                315   0.5241265
## 2   male                286   0.4758735
```

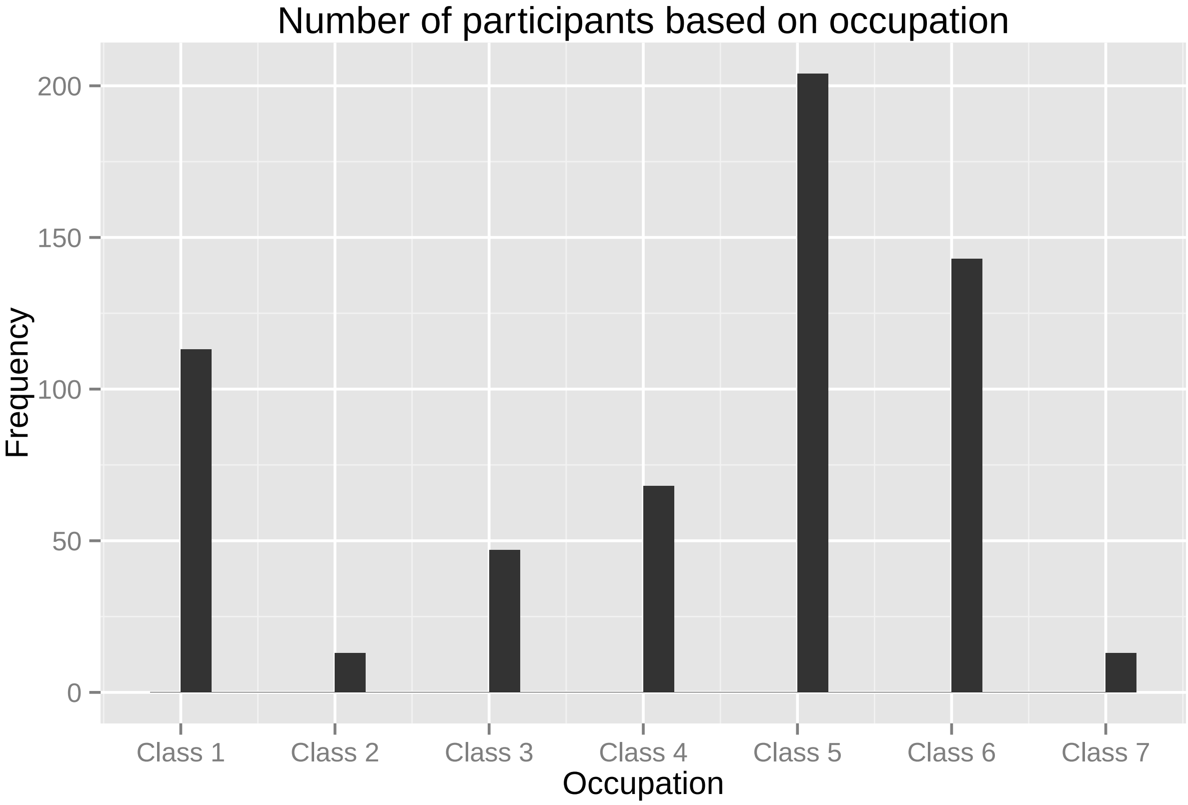
```
# Plotting the frequency of participants based on gender
ggplot(Affairs.data, aes(gender)) + geom_histogram(width = 0.2) +
  xlab("Gender") + ylab("Frequency") +
  ggtitle("Number of participants based on gender")
```



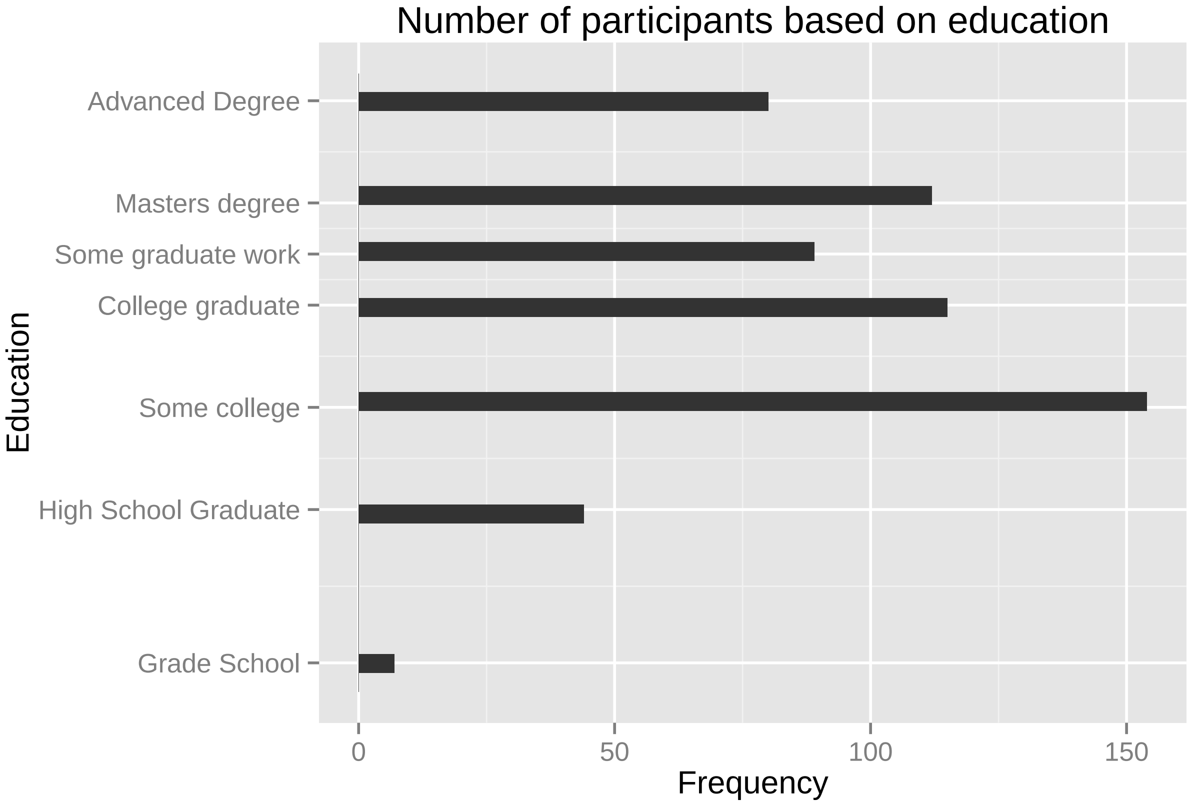
```
# Calculating the average age of the participants
Affairs.data %>%
  summarise(avg_age = mean(age, na.rm = TRUE))
```

```
##    avg_age
## 1 32.48752
```

```
# Plotting the frequency of participants based on occupation
ggplot(Affairs.data, aes(occupation)) + geom_histogram(width = 0.2) +
  xlab("Occupation") + ylab("Frequency") +
  ggtitle("Number of participants based on occupation") +
  scale_x_continuous(breaks=c(1, 2, 3, 4, 5, 6, 7),
                     labels=c("Class 1", "Class 2", "Class 3", "Class 4",
                              "Class 5", "Class 6", "Class 7"))
```



```
# Plotting the frequency of participants based on education
ggplot(Affairs.data, aes(education)) + geom_histogram(width = 0.2) +
  xlab("Education") + ylab("Frequency") +
  ggtitle("Number of participants based on education") +
  scale_x_continuous(breaks=c(9, 12, 14, 16, 17, 18, 20),
                     labels=c("Grade School", "High School Graduate",
                              "Some college", "College graduate",
                              "Some graduate work", "Masters degree",
                              "Advanced Degree")) + coord_flip()
```



The following observations were made regarding the participants of the Affairs.data dataset:

1. Proportion of female participants: 0.524, Proportion of male participants: 0.476
2. Average age of the participants: 32.5
3. Based on occupation, the maximun number of participants were from Class 5 (of Hollinghead 7-point classification with reverse numbering) and least from Class 2.
4. Based on level of education, the maximum number of participants were from the category "some college" and least number of participants had their level of education as grade school.

Exploring the characteristics of participants who engage in affairs. Instead of modeling the number of affairs, considering the binary outcome - had an affair versus didnt have an affair. Creating a new variable to capture this response variable of interest.

```r
# Creating a binary variable haveaffair denoting whether a participant have
# an affair or not
# If the number of affairs are greater than 0, binary variable is set to 1
Affairs.data$haveaffair[Affairs.data$affairs  > 0] <- 1
# If the number of affairs is equal to 0, binary variable is set to 0
Affairs.data$haveaffair[Affairs.data$affairs == 0] <- 0

# Converting the binary response variable to factor datatype with labels
# No and Yes for levels 0 and 1 respectively.
Affairs$haveaffair <- factor(Affairs.data$haveaffair,
                             levels=c(0,1),
                             labels=c("No","Yes"))

# Displaying the count of the new binary variable
table(Affairs$haveaffair)
```

```
##
##  No Yes
## 451 150
```

A new binary response variable haveaffair is created with level 0 indicating "No" and level "1" indicating "Yes". The binary variable is calculated based on the number of affairs. If the number of affairs is greater than 0, then haveaffair is set to "Yes"(level 1) or else "No"(level 0). It can be observed that the 451 participants did not have an affair whilde 150 of them had an affair.

Fitting a logistic regression model to explore the relationship between having an affair and other personal characteristics

```r
# Fitting a logistic regression model twith haveaffair as the response variable
# and all other personal characteristics as predictor variables
fit.allpredictors <- glm(haveaffair ~ gender + age + yearsmarried + children +
                  religiousness + education + occupation +rating,
                data=Affairs.data,family=binomial())
# Displaying the summary statistics of the fitted model.
summary(fit.allpredictors)
```

```
##
## Call:
## glm(formula = haveaffair ~ gender + age + yearsmarried + children +
##     religiousness + education + occupation + rating, family = binomial(),
##     data = Affairs.data)
##
## 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
```

From the summary statistics of the fitted model, it can be observed that age, yearsmarried, religiousness and rating are statistically significant with p-value less than 0.05. Thus these variables can be used for predicting having affairs or not. Religiousness and rating are significant at the 0.001 level, yearsmarried at the 0.01 level and age at the 0.05 significance level.

1. It can be observed that the coeffcient of age(-0.0443) is negative, indicating that for every one unit increase in age, the log odds of having an affair decreases by 0.0443.

2. It can be observed that the coeffcient of religiousness(-0.3247) is negative, indicating that for every one unit increase in religiousness,the log odds of having an affair decreases by 0.3247.

3. It can be observed that the coeffcient of yearsmarried(0.0948) is positive, indicating that for every one unit increase in yearsmarried, the log odds of having an affair increases by 0.0948.

4. It can be observed that the coeffcient of rating(-0.4685) is negative, indicating that for every one unit increase in rating (very unhappy to very h appy), log odds of having an affair decreases by 0.4685.

Using an all subsets model selection procedure to obtain a best fit model. Analyzing the best fit model and comparing it with the model fitted using all the predictor variables

```r
# Loading the bestglm package
#install.packages("bestglm")
library("bestglm")

# Creating a new column y, response variable to fit bestglm
Affairs.data$y <- Affairs.data$haveaffair
# Rearranging the columns of Affairs.data dataset to fit bestglm
Affairs.for.bestglm <- Affairs.data[,c("gender","age","yearsmarried","children",
                                        "religiousness", "education",
                                        "occupation", "rating", "y")]

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

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

```r
# Displaying the summary statistic of the Best Model
fit.reduced$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
```

Best fit model obtained by using bestglm function (default BIC is used to obatin the model) differs from the simple logistic regression model. The key difference between the two models are:

The best fit model only has predictor variables yearsmarried, religiousness and rating while the simple logistic regression had predictor variables age, yearsmarried, religiousness and rating to be statistically significant.

It can also be observed that the estimate of the model parameters of best fit model is slightly less that the models with all the predictors.

Interpreting the model parameters using the best fit model From the summary statistic of the best fit model, we can observe that the intercept estimate is 1.1382. This indicates that there is a significant association between response variable haveaffair with the predictor variables.

1. It can be observed that the coeffcient of religiousness(-0.3306) is negative, indicating that for every one unit increase in religiousness,the log odds of having an affair decreases by 0.3306.
2. It can be observed that the coeffcient of yearsmarried(0.0555) is positive, indicating that for every one unit increase in yearsmarried, the log odds of having an affair increases by 0.0555.
3. It can be observed that the coeffcient of rating(-0.4533) is negative, indicating that for every one unit increase in rating (very unhappy to very happy), the log odds of having an affair decreases by 0.4533.

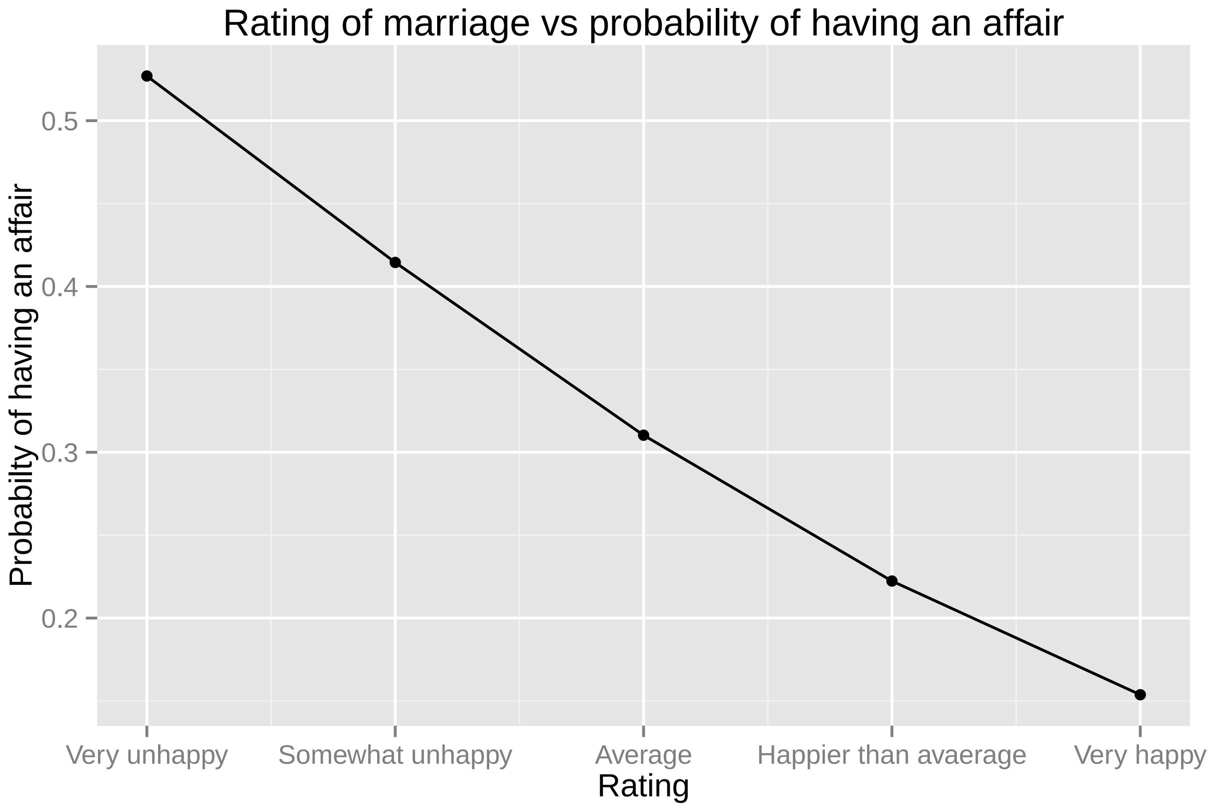
Creating an artificial test dataset where martial rating varies from 1 to 5 and all other variables are set to their means. Using this test dataset and the predict function to obtain predicted probabilities of having an affair for case in the test data.

```
# Creating an artificial test dataset
testdata <- data.frame(yearsmarried=mean(Affairs.data$yearsmarried),
                       religiousness=mean(Affairs.data$religiousness),
                       rating=c(1, 2, 3, 4, 5))

# Creating a new column prob to the test data containing the predicted
# probabilities
testdata$prob <- predict(fit.reduced$BestModel, testdata, type="response")
# Displaying the test data
testdata
```

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

```
# Plotting rating of marriage vs probability of having an affair
ggplot(testdata, aes(rating, prob)) + geom_point() + geom_line() +
  xlab("Rating") + ylab("Probabilty of having an affair") +
  ggtitle("Rating of marriage vs probability of having an affair") +
  scale_x_continuous(breaks=c(1, 2, 3, 4, 5),
                     labels=c("Very unhappy", "Somewhat unhappy",
                              "Average", "Happier than avaerage",
                              "Very happy"))
```



From the results, it can be seen that the probability of having an affair decreases from 0.527 to 0.154 when the rating of mariage increases from 1= very unhappy to 5= very happy, given yearmarried and religiousness are kept constant.

This can also be seen from the rating of marriage vs probability of having an affair graph, which clearly indicates that as the rating increases from very unhappy to very happy, the probability of having an affair decreases.