Classification and Regression

Divya Krishnan

December 14, 2015

Multivariate Logistic Regression

```
# Standard libraries
library(RCurl)
library(leaps)
library(car)
library(randomForest)
library(pROC)
library(boot)
library(tree)
library(AER)
```

```
## Warning: package 'sandwich' was built under R version 3.2.3
```

```
library(bestglm)
# Setting seed
set.seed(1)
```

In this problem we will use the infidelity data, 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 (Hollingshead 7-point classiffication 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?

There are 601 participants in the study. About 430 participants have children and 171 participants have no children at the time of the study. The study includes newly married participants as well as participants who have been married for about 15 years. There were 52% females and 47% males in the study. The average age of participants in the study is 32. Histogram of number of extramartial affairs of the respondents shows a right-skewed distribution. Most of the respondents never had any extramartial affairs.

```
# Exploratory data analysis
data(Affairs)
str(Affairs)
```

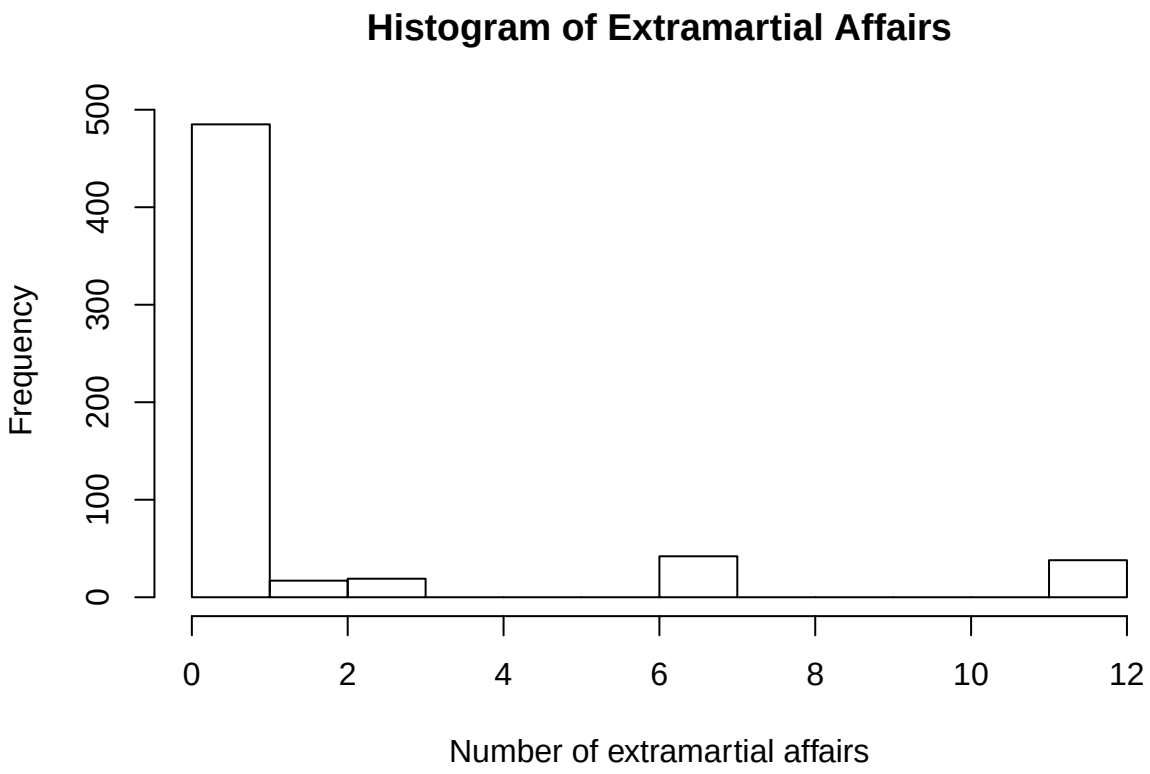
```
## 'data.frame':    601 obs. of  9 variables:
##  $ affairs      : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ 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(Affairs)
```

```
##     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
```

```
# Histogram of the number of extramartial affairs
hist(Affairs$affairs,xlab="Number of extramartial affairs",
     main="Histogram of Extramartial Affairs")
```




```
sex<-table(Affairs$gender)
# Proportion of females
sex[1]/sum(sex)
```

```
##    female
## 0.5241265
```

```
# Proportion of males
sex[2]/sum(sex)
```

```
##      male
## 0.4758735
```

```
# Average age of respondents
mean(Affairs$age)
```

```
## [1] 32.48752
```

(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.

```r
# Creating a binary variable 'hadAffair'
Affairs$hadAffair<-as.factor(ifelse(Affairs$affairs>0,1,0))
table(Affairs$hadAffair)
```

```
##
##   0   1
## 451 150
```

(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.

We may not want to consider the variable giving the number of extra martial afffairs(affairs variable) as the response variable, hadAffair has been computed from affairs variables. Hence, we will definitely see a relationship between the two variables and that will skew the model. Hence, to get a fair model we will use all predictor variables except number of affairs.

Based on the p-value, the following covariates seem to be predictive of having an affair - age, yearsmarried, religiousness, occupation and rating. Age, religiousness and rating seem to have a negative relationship with the response variable whereas yearsmarried and occupation seem to have a positive relationship. Self rating on their marriage and religiousness seem to have the strongest effect on the response variable.

```r
# Converting the categorical predictors to factors
Affairs$religiousness<-factor(Affairs$religiousness,levels=sort(unique(Affairs$religiousness)),
                              labels=c("Anti","Not at all","Slightly","Somewhat","Very"))
Affairs$education<-factor(Affairs$education,levels=sort(unique(Affairs$education)),
                              labels=c("Grade school","High school graduate",
                              "Some college","College graduate","Some graduate work",
                              "Masters degree","Advanced degree"))
Affairs$occupation<-factor(Affairs$occupation)
Affairs$rating<-factor(Affairs$rating,levels=sort(unique(Affairs$rating)),
                              labels=c("Very unhappy","Somewhat unhappy",
                              "Average","Happier than average","Very happy"))
# Logistic regression using all variables except affairs variable
glm.affair<-glm(hadAffair ~ .,data=Affairs[,2:10],family=binomial)
# Model summary
summary(glm.affair)
```

```
##
## Call:
## glm(formula = hadAffair ~ ., family = binomial, data = Affairs[,
##     2:10])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6423  -0.7464  -0.5177  -0.2266   2.8529
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   0.56598    1.19091   0.475  0.63461
## gendermale                    0.16984    0.26022   0.653  0.51397
## age                          -0.04851    0.01910  -2.540  0.01109 *
## yearsmarried                  0.10246    0.03372   3.039  0.00238 **
```

```
## childrenyes                     0.45771    0.30626   1.495  0.13503
## religiousnessNot at all        -1.01388    0.37779  -2.684  0.00728 **
## religiousnessSlightly          -0.61530    0.37862  -1.625  0.10413
## religiousnessSomewhat          -1.56309    0.38553  -4.054 5.03e-05 ***
## religiousnessVery              -1.46191    0.46699  -3.130  0.00175 **
## educationHigh school graduate   0.23783    1.04430   0.228  0.81985
## educationSome college           0.03267    1.01722   0.032  0.97438
## educationCollege graduate       0.01977    1.03839   0.019  0.98481
## educationSome graduate work     0.78593    1.03485   0.759  0.44757
## educationMasters degree         0.45191    1.04003   0.435  0.66391
## educationAdvanced degree        0.37611    1.07221   0.351  0.72576
## occupation2                     0.71437    0.76956   0.928  0.35326
## occupation3                     0.65274    0.44803   1.457  0.14514
## occupation4                     0.92041    0.41898   2.197  0.02803 *
## occupation5                     0.09588    0.35461   0.270  0.78686
## occupation6                     0.27977    0.41743   0.670  0.50271
## occupation7                     0.65393    0.70266   0.931  0.35203
## ratingSomewhat unhappy          0.09641    0.61452   0.157  0.87533
## ratingAverage                  -0.75876    0.61285  -1.238  0.21569
## ratingHappier than average     -1.03989    0.59042  -1.761  0.07819 .
## ratingVery happy               -1.53509    0.60323  -2.545  0.01093 *
## ---
## 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: 587.65  on 576  degrees of freedom
## AIC: 637.65
##
## Number of Fisher Scoring iterations: 4
```

References - https://cran.r-project.org/web/packages/AER/AER.pdf

(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.

The best model is different than the full model in part(c) as it only includes the predictor variables - gender, age, yearsmarried, religiousness and rating.

```
# All subsets model selection based on AIC
best.AIC<-bestglm(Affairs[2:10],family=binomial,IC="AIC")
```

```
## Morgan-Tatar search since family is non-gaussian.
## Note: factors present with more than 2 levels.
```

```
# Top 5 best models
best.AIC$BestModels
```

```
##   gender   age yearsmarried children religiousness education occupation
## 1   TRUE  TRUE         TRUE    FALSE          TRUE     FALSE      FALSE
```

```
## 2   TRUE  TRUE         TRUE     TRUE          TRUE     FALSE      FALSE
## 3  FALSE  TRUE         TRUE    FALSE          TRUE     FALSE      FALSE
## 4  FALSE  TRUE         TRUE     TRUE          TRUE     FALSE      FALSE
## 5  FALSE FALSE         TRUE    FALSE          TRUE     FALSE      FALSE
##   rating Criterion
## 1   TRUE  622.8157
## 2   TRUE  623.5260
## 3   TRUE  624.3088
## 4   TRUE  624.8531
## 5   TRUE  626.1983
```

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

The AIC for the best model is 624.8. The best model suggest that the predictor variables, age, yearsmarried, religiousness and rating have a statistically significant realtionship with the response variable of having/not having an affair.

```
# Coefficients for the best model
best.AIC$BestModel
```

```
##
## Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Coefficients:
##                 (Intercept)                   gendermale
##                     1.15421                      0.39216
##                         age                 yearsmarried
##                     -0.04284                      0.10931
##    religiousnessNot at all       religiousnessSlightly
##                     -0.95876                     -0.58073
##      religiousnessSomewhat           religiousnessVery
##                     -1.53041                     -1.40503
##     ratingSomewhat unhappy               ratingAverage
##                      0.09166                     -0.79456
## ratingHappier than average            ratingVery happy
##                     -1.06400                     -1.59818
##
## Degrees of Freedom: 600 Total (i.e. Null);  589 Residual
## Null Deviance:       675.4
## Residual Deviance: 600.8     AIC: 624.8
```

```
# Summary of the best model
summary(best.AIC$BestModel)
```

```
##
## Call:
## glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5744  -0.7527  -0.5523  -0.2980   2.4865
##
```

```
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   1.15421    0.75746   1.524 0.127564
## gendermale                    0.39216    0.21045   1.863 0.062403 .
## age                          -0.04284    0.01839  -2.329 0.019839 *
## yearsmarried                  0.10931    0.03047   3.587 0.000335 ***
## religiousnessNot at all      -0.95876    0.36582  -2.621 0.008771 **
## religiousnessSlightly        -0.58073    0.36953  -1.572 0.116062
## religiousnessSomewhat        -1.53041    0.37336  -4.099 4.15e-05 ***
## religiousnessVery            -1.40503    0.45263  -3.104 0.001909 **
## ratingSomewhat unhappy        0.09166    0.57984   0.158 0.874399
## ratingAverage                -0.79456    0.57488  -1.382 0.166935
## ratingHappier than average -1.06400    0.55077  -1.932 0.053378 .
## ratingVery happy             -1.59818    0.56485  -2.829 0.004664 **
## ---
## 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: 600.82  on 589  degrees of freedom
## AIC: 624.82
##
## Number of Fisher Scoring iterations: 4
```

(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.

The artificial test is created using the means of the variables for interval variables age and yearsmarried. For the ordinal variables (religiousness, education and occupation), median was used and for the nominal variables (gender,children), mode was used as a measure of central tendency. This test dataset would only predict 'not having an affair'(hadAffair variable=0) due to having all values set to means/median/mode and having zero variance. The predicted values are 0.15,0.22,0.27,0.45 and 0.47. The histogram shows that the number of records are somewhat distributed in a similar way within the 5 probabilitites and none of the probabilities are equal to or greater than 0.5. Hence all of the predicted classification would be not having an affair.

```
# Creating rating as a random sequence of 1 to 5
rating<-sample(c(1:5),nrow(Affairs),replace=TRUE)
rating<-factor(rating,levels=sort(unique(rating)),
               labels=c("Very unhappy","Somewhat unhappy",
                        "Average","Happier than average","Very happy"))
# Finding median values for ordinal variables
median(as.numeric(Affairs$religiousness))
```

```
## [1] 3
```

```
median(as.numeric(Affairs$education))
```

```
## [1] 4
```

```r
median(as.numeric(Affairs$occupation))
```

```
## [1] 5
```

```r
# Finding mode for nominal variables
table(Affairs$gender)
```

```
##
## female   male
##    315    286
```

```r
table(Affairs$children)
```

```
##
##  no yes
## 171 430
```

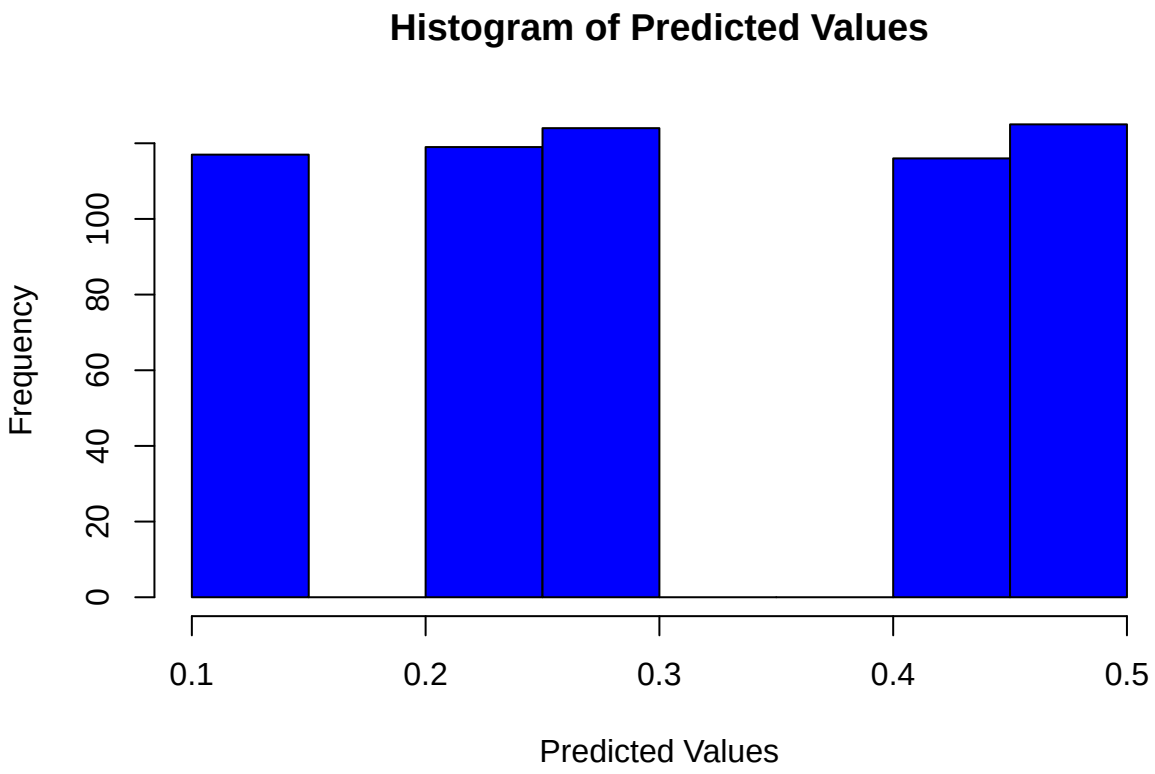
```r
# Creating the test dataset using measures of central tendency
affairsTest<-data.frame(gender=factor("female"),age=mean(Affairs$age),
                        yearsmarried=mean(Affairs$yearsmarried),
                        children=factor("yes"),religiousness=factor("Slightly"),
                        education=factor("College graduate"),
                        occupation=factor("5"),rating)
# Predicting for affairs
yhat.affair<-predict(glm.affair,affairsTest,type="response")
# Exploring predicted values
summary(yhat.affair)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1481  0.2220  0.2743  0.3135  0.4467  0.4706
```

```r
table(round(yhat.affair,4))
```

```
##
## 0.1481  0.222 0.2743 0.4467 0.4706
##    117    119    124    116    125
```

```r
glm.pred.affair<-rep(0,nrow(affairsTest))
# Predicting affair on threshold probability of 0.5
glm.pred.affair[yhat.affair>0.5]<-1
# Plotting the predicted probabilities
hist(yhat.affair,col="blue",xlab="Predicted Values",
     main="Histogram of Predicted Values")
```




```
# Predicted classification
table(glm.pred.affair)
```

```
## glm.pred.affair
##   0
## 601
```

Classification - Regression 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.

1. Garbage sorting - The response would be the garbage classification as compost, recycle or trash. The predictor variable could be properties of the waste such as its composition, biodegradable nature, lifespan of the waste etc. The goal in this application is prediction as we can sort waste automatically using this application. Based on the predictors, we would be able to decide whether the waste is compost, recycle or trash.

2. Gmail Classification - Gmail's mail classification classifies email as primary, social and promotions. The predictor variable could be presence of keywords (such as buy, discount, offer, login, membership), presence of more than 10 email ids in the receiver address. The goal of this application is prediction so that the users can have their mail already sorted based on prior knowledge of mail classification.

3. Stock Analysis - Classifiying the stocks as buy, sell or hold is very important in stock market analysis. The predictor variables can be performance of the stock yesterday, market capital of the stock, P/E

ratio, dividend yield, one-month high, one-month low. The goal of the application is prediction as the stock traders want to beat the market to make maximum profits.

(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.

1. To find the GPA for students currently in Data Science: Theory class - The response variable would be the GPA of students for Data Science:Theory class.The predictors could be a number of factors such as the past performance in other Data Science elective classes, number of online Data science courses taken, number of projects done in Data science. The goal in this application would be to predict the GPA of the students.

2. Does median income affect hospitalizations in Washington state? The response variable is the number of hospitalizations. The predictor variable is the median income of the patient. The goal in this application is inference which tells us whether there is a relationship between median income and the hospitalizations in Washington state.

3. How does weather affect the football result of Seattle Seahawks when playing in Seattle? The response variable is outcome(winning or lossing) of the football match. The predictor variable is temperature of the match day, precipitation in inches on the match day, wind speed on the match day. The goal of the application is inference as we want to understand the affect of temperature, precipitation and wind speed (weather) on the outcome of the football match.

(c) What are the advantages and disadvantages of a very fexible (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?

The advantage of a very flexible approach for regression or classification is that bias will decrease and we can obtain a better fit for the training data. The disadvantage of a very flexible approach is that variance will increase and there is a risk of overfitting the training data and increasing the test error. When we are interested in interpretation or inference, we might prefer a less flexible approach. When we are interested in prediction, we might prefer a more flexible approach even though the intepretability might be less.