Logistic Regression

Priyank Thakkar

10/11/2021

Import Data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.
3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
setwd("D:\\SJSU_HW\\GitHubSJSU\\RStudio_Learning\\Data_set")
AP_data <- read.csv("AutoPurchaseData.csv")
as_tibble(AP_data)
```

```
## # A tibble: 20 x 3
##    Income   Age Purchased
##     <dbl> <int>     <int>
##  1  45000     2         0
##  2  40000     4         0
##  3  60000     3         1
##  4  50000     2         1
##  5  55000     2         0
##  6  50000     5         1
##  7  35000     7         1
##  8  65000     2         1
##  9  53000     2         0
## 10  48000     1         0
## 11  37000     5         1
## 12  31000     7         1
## 13  40000     4         1
## 14  75000     2         0
## 15  43000     9         1
## 16  49000     2         0
## 17  37500     4         1
## 18  71000     1         0
```

```
## 19  34000     5         0
## 20  27000     6         0
```

```
# Let's get some Summary of this data.
summary(AP_data)
```

```
##      Income            Age          Purchased
##  Min.   :27000   Min.   :1.00   Min.   :0.0
##  1st Qu.:37375   1st Qu.:2.00   1st Qu.:0.0
##  Median :46500   Median :3.50   Median :0.5
##  Mean   :47275   Mean   :3.75   Mean   :0.5
##  3rd Qu.:53500   3rd Qu.:5.00   3rd Qu.:1.0
##  Max.   :75000   Max.   :9.00   Max.   :1.0
```

About Data

A study was performed to investigate new automobile purchases. A sample of 20 families was selected. Each family was surveyed to determine the age of their oldest vehicle and their total family income. A followup survey was conducted 6 months later to determine if they had actually purchased a new vehicle during that time period (y= 1 indicates yes and y=0 indicates no)

- [,1] Income : Total family income in the Dollar($)
- [,2] Age : Represent the Age of their Oldest Vehicle in Year
- [,3] Purchased : 0 = No purchased, 1 = yes purchased (in last 6 months)

Now let's add some Libraries.

```
library(tidyverse)
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
str(AP_data)
```

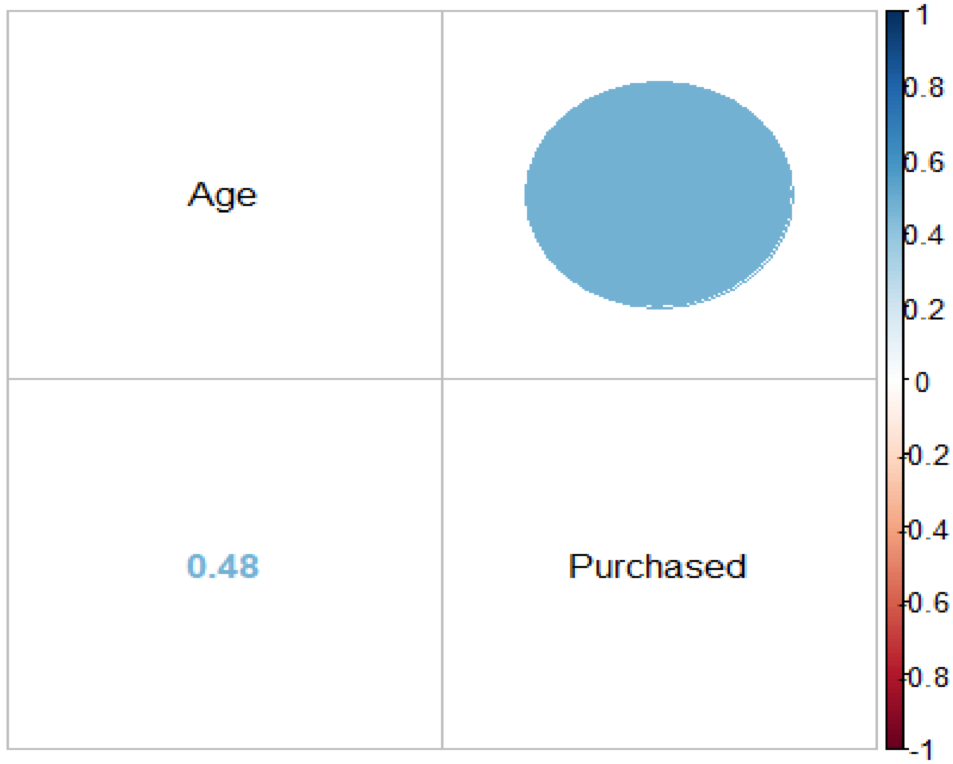
```
## 'data.frame':    20 obs. of  3 variables:
##  $ Income   : num  45000 40000 60000 50000 55000 50000 35000 65000 53000 4
8000 ...
##  $ Age      : int  2 4 3 2 2 5 7 2 2 1 ...
##  $ Purchased: int  0 0 1 1 0 1 1 1 0 0 ...
```

Graphs to Understand Data

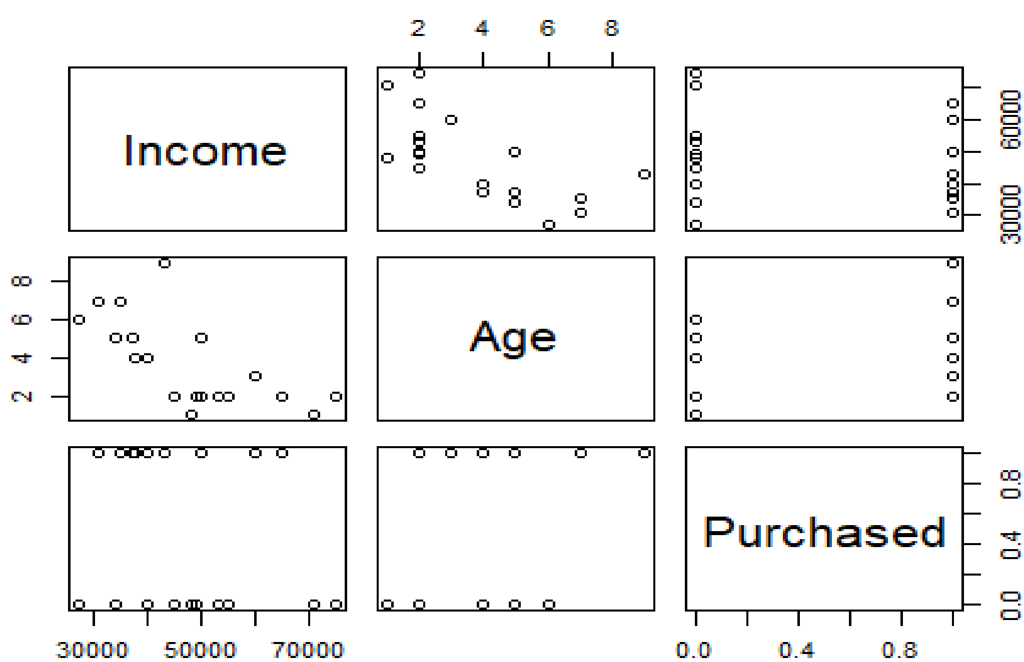
```
cor(AP_data)
```

```
##               Income        Age  Purchased
## Income     1.0000000 -0.6749662 -0.1901818
## Age       -0.6749662  1.0000000  0.4798825
## Purchased -0.1901818  0.4798825  1.0000000
```

```
corrplot.mixed(cor(AP_data[,-1]), order="hclust", tl.col="black")
```




```
pairs(AP_data)
```



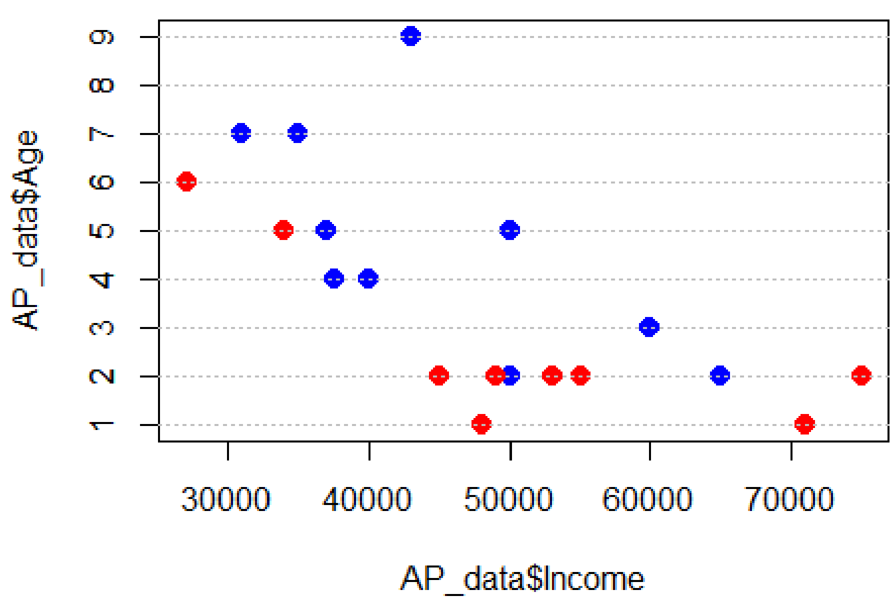
Scatterplot of Factoring

Here, we will factor our **Purchased** columns in categorical output. And will show them on the plot. And will use them in Prediction and in Odds results.

```
AP_data$Purchased <- factor(AP_data$Purchased, levels=c(0,1))
str(AP_data)

## 'data.frame':    20 obs. of  3 variables:
##  $ Income   : num  45000 40000 60000 50000 55000 50000 35000 65000 53000 4
8000 ...
##  $ Age      : int  2 4 3 2 2 5 7 2 2 1 ...
##  $ Purchased: Factor w/ 2 levels "0","1": 1 1 2 2 1 2 2 2 1 1 ...

plot(AP_data$Income, AP_data$Age, col=c("red","blue")[AP_data$Purchased],pch=
20, cex=2)
axis(side=1, at=c(0:9))
axis(side=2, at=c(0:9))
abline(h=0:9,v=0:9, col="gray", lty=3)
```



QUESTION 1 : Creating and Fitting the Model

```
mymodel = glm(Purchased ~ ., data = AP_data, family = binomial)
summary(mymodel)

##
## Call:
```

```
## glm(formula = Purchased ~ ., family = binomial, data = AP_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5635  -0.8045  -0.1397   0.9535   1.7915
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.047e+00  4.674e+00  -1.508    0.132
## Income       7.382e-05  6.371e-05   1.159    0.247
## Age          9.879e-01  5.274e-01   1.873    0.061 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 21.082  on 17  degrees of freedom
## AIC: 27.082
##
## Number of Fisher Scoring iterations: 5
```

This means, "Use the general linear model function to create a model that predicts Purchase from Age and Income, using the data in AP_data, with a logistic regression equation. here, *glm* function is a general-purpose prediction model maker. The family="binomial" parameter creates a logistic regression prediction model.

QUESTION 3 : interpret the model coefficients β1 and β2

The summary function displays a lot of information. The key information is in the coefficients section:

Coefficients:

Estimate

- (Intercept) -7.047e+00 **(β0)**
- Income 7.382e-05 **(β1)**
- Age 9.879e-01 **(β2)**

QUESTION 4 : What is the estimated probability that a family with an income of $45,000 and a car that is 5 year old will purchase a new vehicle in the next 6 months?

The values **-7.047e+00, 7.382e-05 and 9.879e-01** define a prediction equation that's best explained by an example. Suppose you want to predict the Purchased for a person with **Income = 45000** and **Age = 5**. First, you compute an intermediate z-value using the coefficients and input values:

- $Z = \beta 0 + \beta 1(Income) + \beta 2(Age)$

```
z = -7.047e+00 + (7.382e-05)*(45000) + (9.879e-01)*(5)
z

## [1] 1.2144
```

And then you compute a p-value : The p-value will always be a value between 0 and 1 and is a probability. If p <= 0.5, the predicted value is the first of the two possible values, **"red = 0,"** in the demo. If p > 0.5, the predicted value is the second possible value, **"blue = 1,"** in the demo.

- Equation is $p = 1/(1 + e^{-}z)$

```
e = 2.71828
p = 1 / (1 + e^-z)
p

## [1] 0.7710764

if (p <= 0.5) { cat("predicted party = red \n") } else { cat("predicted party
= blue \n") }

## predicted party = blue
```

So, this value **0.77** says its **blue = 1**, that means the Auto Mobile purchased by them in 6 months.

Now predict same thing form the function. And let's see.

```
nd = with(AP_data, data.frame(Income=45000, Age=5))
nd$pred = predict(mymodel, newdata=nd, type="response")
nd

##   Income Age      pred
## 1  45000   5 0.7710279
```

Let's predic all the values from the current model.

```
predict(mymodel, AP_data, type="response")

##          1          2          3          4          5          6
7
## 0.14810602 0.46434922 0.58555120 0.20093678 0.26671264 0.82965843 0.920687
```

```
78
##          8          9         10         11         12         13 
14 
## 0.43212218 0.23884915 0.07474622 0.65103464 0.89627104 0.46434922 0.614192
53 
##         15         16         17         18         19         20 
## 0.99342599 0.18934581 0.41887642 0.30614944 0.59920166 0.70543363
```

QUESTION 2 : Is the logistic regression model in part a adequate?

From the reference of this : https://en.wikipedia.org/wiki/Logistic_regression

The odds of the dependent variable equaling a case (given some linear combination x of the predictors) is equivalent to the exponential function of the linear regression expression.Given that the logit ranges between negative and positive infinity, it provides an **adequate** criterion upon which to conduct linear regression and the logit is easily converted back into the odds.

So we define odds of the dependent variable equaling a case (given some linear combination x of the predictors) as follows:

- $Odds = e^(\beta 0 + \beta 1(Income) + \beta 2(Age))$
- $Odds = e^z$

```
Odds = e^z
Odds

## [1] 3.36827
```

Following the way we can check the Adequacy. * Bigger the value of Odds better the Adequacy. * More over from the summary function information the AIC: 27.082. * Residuals from the summary function information also shows the adequacy, for the residuals we really needs all the values near to 0 or around zero.

Deviance Residuals:

- 1st quantile = -0.8045
- Median = -0.1397
- 3rd quantile = 0.9535

all are around 0 line.

But, we will compare these parameter with our new model and will say which model is good with all of these Adequacy checks.

QUESTION 5 : Expand the linear predictor to include an interaction term (, i.e. include a 3rd predictor that is $x1 \times x2$). Is there an evidence that this term is required in the model?

- New Model: $x3 = Income * Age$

```
n_mymodel = glm(Purchased ~ Income + Age + (Income*Age), data = AP_data, fami
ly = binomial)
summary(n_mymodel)

##
## Call:
## glm(formula = Purchased ~ Income + Age + (Income * Age), family = binomial
,
##     data = AP_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.63981  -0.62754  -0.05642   0.66213   1.85666
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.144e-01  6.394e+00   0.049    0.961
## Income      -1.411e-04  1.412e-04  -0.999    0.318
## Age         -2.462e+00  2.081e+00  -1.183    0.237
## Income:Age   1.014e-04  6.297e-05   1.610    0.107
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 16.551  on 16  degrees of freedom
## AIC: 24.551
##
## Number of Fisher Scoring iterations: 6

nz = 3.144e-01 + (-1.411e-04)*(45000) + (-2.462e+00)*(5) + (1.014e-04)*(45000
)*(5)
nz

## [1] 4.4699

e = 2.71828
np = 1 / (1 + e^-nz)
np

## [1] 0.9886811

if (np <= 0.5) { cat("predicted party = red \n") } else { cat("predicted part
y = blue \n") }
```

```
## predicted party = blue

new_Odd = e^nz
new_Odd

## [1] 87.34773

predict(n_mymodel, AP_data, type="response")

##           1           2           3           4           5           6 
## 0.137613105 0.739328883 0.937592954 0.178424263 0.228136783 0.998153927 
##           7           8           9          10          11          12 
## 0.951617675 0.353779687 0.207158167 0.017052791 0.823207946 0.669325997 
##          13          14          15          16          17          18 
## 0.739328883 0.503482378 0.999987884 0.169566947 0.594247207 0.006909235 
##          19          20 
## 0.608501186 0.136584101
```

Conclusion :

- AIC : 24.55, Which is very low, so this model is better then previous one.
- New Odd : 87.34, which is greater then previous model Odd.
- P-value : 0.98, This is good probability accuracy, again better then previous model.

We, can use second Logistic Regression model for our future purposes.