Tuan Pham - STAT 4310 - Project

Tuan Pham

2023-11-06

```
#install.packages('AER')
library(AER)
data("CreditCard")
head(CreditCard)
```

```
##   card reports      age income       share expenditure owner selfemp dependents
## 1  yes       0 37.66667 4.5200 0.033269910  124.983300   yes      no          3
## 2  yes       0 33.25000 2.4200 0.005216942    9.854167    no      no          3
## 3  yes       0 33.66667 4.5000 0.004155556   15.000000   yes      no          4
## 4  yes       0 30.50000 2.5400 0.065213780  137.869200    no      no          0
## 5  yes       0 32.16667 9.7867 0.067050590  546.503300   yes      no          2
## 6  yes       0 23.25000 2.5000 0.044438400   91.996670    no      no          0
##   months majorcards active
## 1     54          1     12
## 2     34          1     13
## 3     58          1      5
## 4     25          1      7
## 5     64          1      5
## 6     54          1      1
```

About the data

- Dimension: 12 variables, 1319 observations
- Author : Greene, W.H. (2003).
- Published in ***Econometric Analysis***, 5th edition. Upper Saddle River, NJ: Prentice Hall.
- Time published: 2003

Key features/Terminology

- `card`: Categorical variable with two factors, either Yes or No.

  The variable is about whether the application for credit card was accepted.

- `report`: integer numerical variable (0-14)

  Number of major derogatory reports. (bad credit report)

- `age:` decimal numerical variable (0-85) `attention why 0 here?`

  Age of the applicants.

- `income:` decimal numerical variable in $10,000 US Dollar (0-14) > Yearly income.
- `share:` numerical variable (0-1)

  Ratio of monthly credit card expenditure to yearly income.

  The bigger the number the more credit card spending over the income.

Key features/Terminology (cont.)

- `expenditure:` numerical variable (0-3200)

  Average monthly credit card expenditure.

- `owner:` Categorical variable with two factors, either Yes or No.

  Whether the applicant owns a house.

- `selfemp:` Categorical variable with two factors, either Yes or No.

  Whether the applicant is self-employed.

- `dependents:` numerical variable (0-6)

  Number of dependents.

- `months:` numerical variable (0-550)

  Months living at current address.

Key features/Terminology (cont.)

- `majorcards:` numerical variable (0-1)

  Number of major credit cards held.

- `active:` numerical variable (0-1)

  Number of active credit accounts.

  Number of loans/credit debts besides the credit card like mortgage loans, car loans, expenditure loans, student loans, etc…

Analysis Approach

```r
str(CreditCard)
```

```
## 'data.frame':    1319 obs. of  12 variables:
##  $ card       : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ reports    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ age        : num  37.7 33.2 33.7 30.5 32.2 ...
##  $ income     : num  4.52 2.42 4.5 2.54 9.79 ...
##  $ share      : num  0.03327 0.00522 0.00416 0.06521 0.06705 ...
##  $ expenditure: num  124.98 9.85 15 137.87 546.5 ...
##  $ owner      : Factor w/ 2 levels "no","yes": 2 1 2 1 2 1 1 2 2 1 ...
##  $ selfemp    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ dependents : num  3 3 4 0 2 0 2 0 0 0 ...
##  $ months     : num  54 34 58 25 64 54 7 77 97 65 ...
##  $ majorcards : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ active     : num  12 13 5 7 5 1 5 3 6 18 ...
```

The dataset has twelve variables with three categorial variables and nine numerical variables. Since we would like to know whether the credit card is approved, we want to build a model to justify the decision based on the factors that was represented by the data.

According to that, we would build a logistic model to analyze whether there are a strong relationship among the predictors to the response.

We will construct a logistic model with the response variable, `card`, and eleven other predictors.

Binomial Logistic Model

```r
model <- glm(card  ~ ., family = 'binomial', CreditCard)
summary(model)
```

```
##
## Call:
## glm(formula = card ~ ., family = "binomial", data = CreditCard)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -8.49    0.00    0.00    0.00    8.49
##
## Coefficients:
##               Estimate Std. Error    z value Pr(>|z|)
## (Intercept)  6.610e+14  8.418e+06   78529581   <2e-16 ***
## reports     -5.411e+14  1.436e+06 -376913075   <2e-16 ***
## age          1.015e+12  2.208e+05    4595721   <2e-16 ***
## income      -4.840e+13  1.470e+06  -32917418   <2e-16 ***
## share        1.732e+16  4.353e+07  397938311   <2e-16 ***
## expenditure -7.184e+11  1.568e+04  -45826822   <2e-16 ***
## owneryes     1.262e+11  4.361e+06      28936   <2e-16 ***
## selfempyes   2.868e+14  7.375e+06   38894429   <2e-16 ***
## dependents   2.155e+13  1.619e+06   13313846   <2e-16 ***
## months      -1.194e+12  3.140e+04  -38025935   <2e-16 ***
```

```
## majorcards   4.265e+13  4.858e+06    8780075   <2e-16 ***
## active       8.221e+12  3.175e+05   25897563   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance:  1404.6  on 1318  degrees of freedom
## Residual deviance: 13336.2  on 1307  degrees of freedom
## AIC: 13360
##
## Number of Fisher Scoring iterations: 25
```

Summary Output Explanation

According to the model, we can spot four negative and seven positive betas.

The explanation to the negative and positive coefficients:

- If it is a positive sign, that means increasing the variable according to its positive coefficient will increase the probability of the response or increase the likeliness of the outcome.
- If it is a negative sign, that means increasing the variable of the negative coefficient will decrease the probability of the response.

The explanation to the values of the coefficients

Unlike the regression model where the coefficients corresponding to each variable are the actual coefficient, the logistic model shows the transformation of the coefficient in the model summary output.

Therefore, to get the actual value change in response by changing a variable's value, we implement the transformation of the coefficient according to that variable by simply taking the exponential of the coefficient according to a variable.

Summary Output Explanation(cont.)

The summary output explanation

> The `reports` variable has its coefficient estimation at -5.411e+14.

According to this negative value, we can tell that with **one-unit increasing** in `report`, there's **exp(5.411e+14) decreasing** in the probability of the response, `cards`. In other words, the more bad credit reports, the less chance the applicant will get approved for the credit card.

> The `active` variable has its coefficient estimation at 8.221e+12.

According to this positive value, we can tell that with **one-unit increasing** in `active`, there's **exp(8.221e+12) increasing** in the probability of the response, `cards`. In other words, the more open credits such as expenditure loans, student debts, mortgage loans, car loans etc., the more chance the applicant will get approved for the credit card.

How adequate is the model?

Compute the difference between the null and residual deviance

```
Dev1 <- 13336.2 - 1404.6
df1 <- 1319 - 11
Dev1
```

```
## [1] 11931.6
```

```
qchisq(1- .05, 1308)
```

```
## [1] 1393.251
```

> Since the deviance difference between the null and full model is within the 95% Confidence Interval, then support the null or the model is not adequate.

Reduced Model

```
summary(model.reduced)
```

```
##
## Call:
## glm(formula = card ~ reports + expenditure + dependents + active,
##     family = "binomial", data = CreditCard)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.767   0.000   0.000   0.000   2.869
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.95984    0.31673  -6.188  6.1e-10 ***
## reports      -2.43860    1.02346  -2.383   0.0172 *
## expenditure  31.98104  160.58361   0.199   0.8421
## dependents   -0.65784    0.29795  -2.208   0.0273 *
## active        0.09958    0.03479   2.863   0.0042 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1404.57  on 1318  degrees of freedom
## Residual deviance:  118.34  on 1314  degrees of freedom
## AIC: 128.34
##
## Number of Fisher Scoring iterations: 25
```

Reduced Model(cont.)

The step model suggests the reduced model with four variables, but the p-value of `expenditure` variable is greater than the 0.05 significance level, which means the contribution of `expenditure` is not significant, so we take it out of the reduced model.

We implement the "new" reduced model with the last three variables including `reports`, `dependents`, and `active`.

```
model.reduced2 <- glm(formula = card ~ reports + dependents + active,
    family = "binomial", data = CreditCard)
summary(model.reduced2)
```

```
##
## Call:
## glm(formula = card ~ reports + dependents + active, family = "binomial",
##     data = CreditCard)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5086   0.1938   0.4501   0.6346   2.8626
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.27908    0.12287  10.410   <2e-16 ***
## reports     -1.75066    0.13694 -12.784   <2e-16 ***
## dependents  -0.12935    0.06190  -2.090   0.0367 *
## active       0.14770    0.01778   8.307   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1404.6  on 1318  degrees of freedom
## Residual deviance: 1019.0  on 1315  degrees of freedom
## AIC: 1027
##
## Number of Fisher Scoring iterations: 6
```

Reduced Model Summary Output Explanation

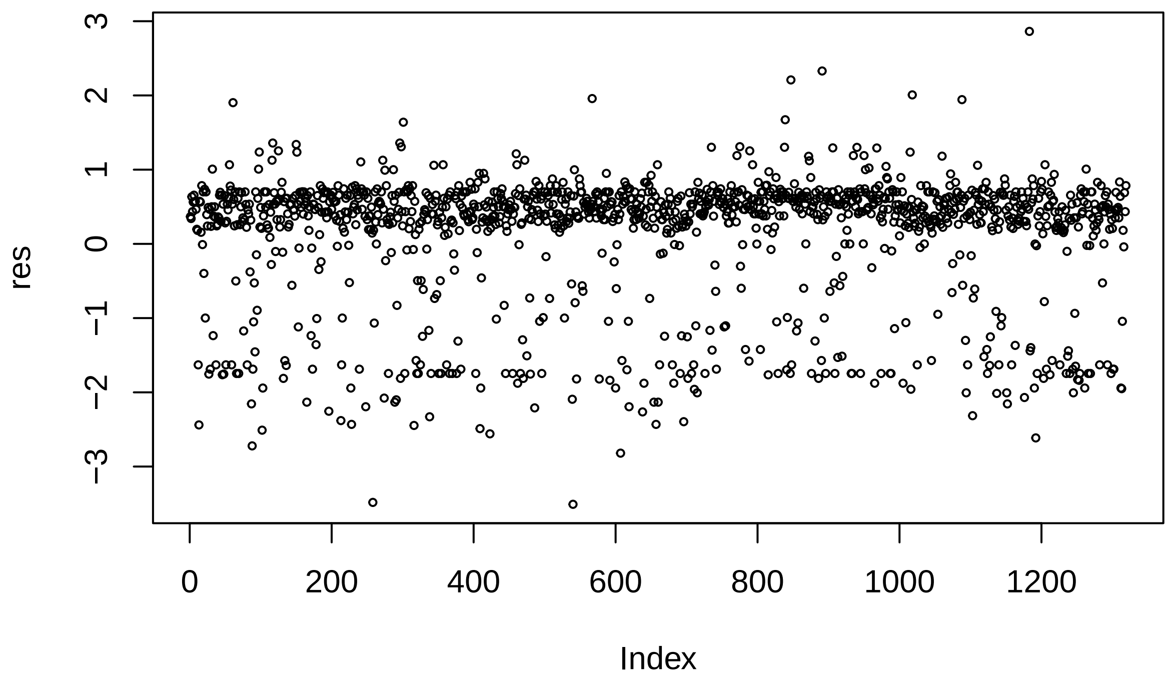
- In this reduced model, we have two negative coefficients corresponding to `reports`, `dependents` and one positive coefficient corresponding to `active`.
- All the chosen variables are significant since their p-values are less than the 0.05 significance level.

Residuals

```
# Residuals
res <- residuals(model.reduced2)
```

There is no Normal Distribution checking needed

```r
plot(res, cex = .5)
```



Prediction

```r
fitted<- model.reduced2$fitted.values
head(fitted)
```

```
##         1         2         3         4         5         6
## 0.9348331 0.9432736 0.8176004 0.9099458 0.8530687 0.8063990
```

```r
# Prediction in terms of {0,1}
summary(fitted)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.7661  0.8484  0.7756  0.9213  0.9979
```

```r
# Chose predictions, using probability > 0.8484
prediction<- ifelse(fitted>= 0.8484, 1, 0)
table(prediction)
```

```
## prediction
##   0   1
## 658 661
```

Prediction(cont.)

```
Construct the Confuse Matrix
```

```
table(CreditCard$card, prediction)
```

```
##      prediction
##         0   1
##   no  250  46
##   yes 408 615
```

```
true.positive <- 615
true.negative <- 250
total <- 615 + 408 + 46 + 250
accuracy = (true.positive + true.negative) / total
accuracy
```

```
## [1] 0.6557998
```

The accuracy of model in predicting whether the credit card applicants will be approved is 65.59%