RMSC4002 Tutorial 8

Chapter 5

November 9, 2017

1 Binary Logistic Regression

1.1 Model Description

Define $\pi_i = Pr(Y_i = 1|x_i)$ to be the probability of success given x_i. The assumption of logistic regression is that log-odd ratio of success probability is a linear function of x_i:

$$ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} = \mathbf{x}_i'\beta$$

where $\pi_i = \frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)}$

```
> summary(glm(HSI~EY+CFTP+ln_MV+DY+BTME+DTE,data=d,binomial))
Call:
glm(formula = d$HSI ~ d$EY + d$CFTP + d$ln_MV + d$DY + d$BTME +
    d$DTE, family = binomial)
Deviance Residuals:
       Min          1Q        Median          3Q         Max
    -8.490e+00   -2.107e-08  -2.107e-08   -2.107e-08  8.490e+00
Coefficients:
               Estimate Std. Error    z value Pr(>|z|)
(Intercept) -4.121e+15  1.066e+07 -386410689   <2e-16 ***
d$EY         1.516e+13  6.431e+05   23570628   <2e-16 ***
d$CFTP      -6.364e+13  1.483e+06  -42902735   <2e-16 ***
d$ln_MV      4.945e+14  1.625e+06  304287297   <2e-16 ***
d$DY        -1.144e+14  7.085e+05 -161536188   <2e-16 ***
d$BTME      -7.907e+12  3.155e+05  -25063060   <2e-16 ***
d$DTE        8.744e+12  7.168e+05   12198713   <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:  258.08  on 679  degrees of freedom
Residual deviance: 1153.40  on 673  degrees of freedom
AIC: 1167.4
```

```
> lreg<-glm(HSI~EY+CFTP+ln_MV+DY+BTME+DTE,data=d,binomial)
> names(lreg)        # display the items in lreg
 [1] "coefficients"      "residuals"         "fitted.values"
 [4] "effects"           "R"                 "rank"
 [7] "qr"                "family"            "linear.predictors"
[10] "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"
[16] "df.residual"       "df.null"           "y"
[19] "converged"         "boundary"          "model"
[22] "call"              "formula"           "terms"
[25] "data"              "offset"            "control"
[28] "method"            "contrasts"         "xlevels"

> pr<-(lreg$fitted.values>0.5)    # set pr=True if fitted >0.5 or otherwise
> table(pr,d$HSI)                 # Cross tabulation of pr and HSI

pr      0   1
  FALSE 634   2
  TRUE   14  30
```

1.2 Outliers Detection

To improve the result of logistic regression, the main method is to remove the outliers from the big sample size.

- Calculate the Mahalanobis distance of those in big sample size.
- Set the cut-off value to be the $[(1-\alpha)\times 100]$ th percentile of Chi-square with degree of freedom p, where p is the number of independent variables.
- Remove all data with Mahalanobis distance greater than the cut-off value.
- Put the remaining data into logistic regression model again.

(When doing your project, it is a good idea to clean the data first.)

```
mdist<-function(x) {
   t<-as.matrix(x)         # transform x to a matrix
   m<-apply(t,2,mean)      # compute column mean
   s<-var(t)               # compute sample covariance matrix
   mahalanobis(t,m,s)      # using built-in mahalanobis function
}
```

```
> d<-read.csv("fin-ratio.csv")     # read in dataset
> d0<-d[d$HSI==0,]                 # select HSI=0
> d1<-d[d$HSI==1,]                 # select HSI=1
> dim(d0)
[1] 648   7
> dim(d1)
[1] 32  7
```

```
> source("mdist.r")              # load the mdist function
> x<-d0[,1:6]                    # save d0 to x
> md<-mdist(x)                   # compute mdist
> plot(md)                       # plot md
```

```
> (c<-qchisq(0.99,df=6))  # p=6, and type-I error = 0.01
 [1] 16.81189

> d2<-d0[md<c,]           # select cases from d0 with md<c
> dim(d2)                 # we have throw away 648-626=22 cases
[1] 626   7
> d3<-rbind(d1,d2)        # combine d1 with d2 to form a cleaned dataset
> dim(d3)
[1] 658   7
#save the cleaned dataset to " fin-ratio1.csv"
> write.csv(d3,file="fin-ratio1.csv",row.names=F)
```

```
> summary(glm(HSI~CFTP+ln_MV+BTME,data=d3,binomial))

Call:
glm(formula = HSI ~ CFTP + ln_MV + BTME, family = binomial, data = d3)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.377e+00  -1.943e-04  -8.005e-06  -3.054e-07   1.738e+00

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -69.9309    21.3821  -3.271  0.00107 **
CFTP         -3.0376     1.2178  -2.494  0.01262 *
ln_MV         7.2561     2.2284   3.256  0.00113 **
BTME          1.3222     0.6418   2.060  0.03940 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

```
> lreg<-glm(HSI~CFTP+ln_MV+BTME,data=d3,binomial)   # save the output
> pr<-(lreg$fit>0.5)                                 # prediction
> table(pr,d3$HSI)                                 # classification table

pr      0   1
FALSE 624   3
TRUE    2  29
```

1.3 Lift Chart

```
ysort<-d3$HSI[order(lreg$fit,decreasing=T)]     # sort y according to lreg$fit
n<-length(ysort)                                # get length of ysort
perc1<-cumsum(ysort)/(1:n)                      # compute cumulative percentage
plot(perc1,type="l", col='blue' )                    # plot perc with line type
abline(h=sum(d3$HSI)/n)                    # add the baseline
yideal <- c(rep(1,sum(d3$HSI)),rep(0,length(d3$HSI)-sum(d3$HSI)))  # the ideal case
perc_ideal <- cumsum(yideal)/(1:n)                   # compute cumulative percentage
of ideal case
lines(perc_ideal, type="l", col="red")          # plot the ideal case in red line

perc2<-cumsum(ysort)/sum(ysort)            # cumulative perc. of success
pop<-(1:n)/n                          # x-coordinate
plot(pop,perc2,type="l")                   # plot
lines(pop,pop)                        # add the reference line
perc_ideal2 <- cumsum(yideal)/sum(yideal)  # cumulative perc. of success for ideal
case
lines(pop,perc_ideal2, type="l",col="red") # plot the ideal case in red line
```

1.4 Model Selection

```
> d<-read.csv("fin-ratio1.csv")              # read in data
> lreg<-glm(HSI~.,data=d,binomial)           # save the logistic reg
> step(lreg)                                 # perform stepwise selection
Start:  AIC=36.47
HSI ~ EY + CFTP + ln_MV + DY + BTME + DTE

        Df Deviance     AIC
- DTE    1   22.495  34.495
- DY     1   22.769  34.769
- EY     1   22.822  34.822
<none>       22.468  36.468
- BTME   1   27.628  39.628
- CFTP   1   30.586  42.586
- ln_MV  1  245.018 257.018

Step:  AIC=31.09
HSI ~ CFTP + ln_MV + BTME

        Df Deviance     AIC
<none>       23.087  31.087
- BTME   1   28.051  34.051
- CFTP   1   33.623  39.623
- ln_MV  1  246.700 252.700

Call:  glm(formula = HSI ~ CFTP + ln_MV + BTME, family = binomial, data = d)

Coefficients:
(Intercept)         CFTP         ln_MV          BTME
    -69.931       -3.038         7.256         1.322
```

1.5 Measure of Accuracy & Decision Threshold

Measure Accuracy:

Test \True	**True**	**False**
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

- $Precision = TP/(TP + FP)$
- $Recall = TP/(TP + FN)$
- $F1 = \frac{2}{1/Precision + 1/Recall}$

Decision Threshold:

Decision \True	**True**	**False**
Accept	correct	cost: c_2
Reject	cost: c_1	correct

When predicting the class given $\mathbf{x}$, we first compute the probability $p(\mathbf{x})$ of acceptance, then compare the two possible loss:

- Rejecting a true sample: $c_1(1 - p(\mathbf{x}))$
- Accepting a false sample: $c_2 p(\mathbf{x})$

Choose the smaller misclassification cost as the decision.

Exercise 2013-14 final Q4(a)(b) A logistic regression with stepwise selection is fitted to the dataset with the following output:

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.274855   0.343976 -18.242  < 2e-16 ***
Vmail          -0.023810   0.004535  -5.250 1.52e-07 ***
Day_Mins        0.012767   0.001045  12.214  < 2e-16 ***
Eve_Mins        0.006604   0.001091   6.052 1.43e-09 ***
CustServ_Calls  0.470659   0.038070  12.363  < 2e-16 ***
```

- Let $x_1 = (Vmail, DayMins, EveMins, CustServCalls) = (0, 255, 230, 3)$, compute $Pr(Change = 0|x_1)$ and $Pr(Change = 1|x_1)$, where $Change$ is the target variable.
- Suppose the cost of misclassifying a customer with $Change = 1$ to $Change = 0$ is 3 times as high as the cost of misclassifying a customer with $Change = 0$ to $Change = 1$. What is the classification rule based on this logistic regression. How would you predict the target variable of x_1 in the first question.