# Logistic Regression

Code ▾

Hide

```
library(car)
library(mlogit)
```

Hide

```
df<- read.delim('/home/atrides/Desktop/R/statistics_with_R/08_LogisticRegression/Data_Files/eel.dat', header=TRUE)

# listing all columns in data frame
names(df)
```

```
[1] "Cured"        "Intervention" "Duration"
```

Hide

```
# checking whether the passed columns ae factor or not
print(is.factor(df$Cured))
```

```
[1] FALSE
```

Hide

```
print(is.factor(df$Intervention))
```

```
[1] FALSE
```

Hide

```
# converting column to a factor
df$Cured<- as.factor(df$Cured)
df$Intervention<- as.factor(df$Intervention)


# Default factors were not suitable. So refactoring the revels
df$Cured<- relevel(df$Cured, "Not Cured")
df$Intervention<- relevel(df$Intervention, "No Treatment")
```

Hide

```
# fitting the model
# newModel<-glm(outcome ~ predictor(s), data = dataFrame, family = name of a
distribution, na.action = an action)
m01<- glm(Cured~Intervention, data=df, family = binomial())
m02<- glm(Cured~Intervention+Duration, data=df, family = binomial())
m00<- glm(Cured~1, data=df, family = binomial())
```

Hide

```
# printing summary
print(summary(m00))
```

```


Call:
glm(formula = Cured ~ 1, family = binomial(), data = df)

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
-1.309  -1.309   1.052   1.052   1.052  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3032     0.1903   1.593    0.111

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 154.08  on 112  degrees of freedom
Residual deviance: 154.08  on 112  degrees of freedom
AIC: 156.08

Number of Fisher Scoring iterations: 4
```

Hide

```
print(summary(m01))
```

```

Call:
glm(formula = Cured ~ Intervention, family = binomial(), data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5940  -1.0579   0.8118   0.8118   1.3018

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.2877     0.2700  -1.065  0.28671
InterventionIntervention   1.2287     0.3998   3.074  0.00212 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 154.08  on 112  degrees of freedom
Residual deviance: 144.16  on 111  degrees of freedom
AIC: 148.16

Number of Fisher Scoring iterations: 4
```

Hide

```
print(summary(m02))
```

```

Call:
glm(formula = Cured ~ Intervention + Duration, family = binomial(),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6025  -1.0572   0.8107   0.8161   1.3095

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.234660   1.220563  -0.192  0.84754
InterventionIntervention  1.233532   0.414565   2.975  0.00293 **
Duration                 -0.007835   0.175913  -0.045  0.96447
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 154.08  on 112  degrees of freedom
Residual deviance: 144.16  on 110  degrees of freedom
AIC: 150.16

Number of Fisher Scoring iterations: 4
```

Hide

```
# Accessing some other statistics of our Logmodel
print(m01$null.deviance)
```

```
[1] 154.084
```

Hide

```
print(m01$deviance)
```

```
[1] 144.1578
```

Hide

```
print(m01$coefficients)
```

```
             (Intercept) InterventionIntervention
              -0.2876821                1.2286654
```

Hide

```
# to see what all statistics are there , we  could do as follows
names(m01)
```

```
 [1] "coefficients"      "residuals"         "fitted.values"     "effects"
"R"
 [6] "rank"              "qr"                "family"            "linear.pre
dictors" "deviance"
[11] "aic"               "null.deviance"     "iter"              "weights"
"prior.weights"
[16] "df.residual"       "df.null"           "y"                 "converged"
"boundary"
[21] "model"             "call"              "formula"           "terms"
"data"
[26] "offset"            "control"           "method"            "contrasts"
"xlevels"
```

Hide

```
# getting some critical statistics, model chi square and its significance
modelChi<- m01$null.deviance - m01$deviance
modelChi
```

```
[1] 9.926201
```

Hide

```
chidf<- m01$df.null-m01$df.residual
chidf
```

```
[1] 1
```

Hide

```
# feeding model chi square and its degree of freedom to calculate the p valu
e
chisq.prob<- 1-pchisq(modelChi , chidf)
chisq.prob
```

```
[1] 0.001629425
```

Hide

```
# Note: we reject the null model that our model 'm01' is not better than jus
t chance to predict outcome
```

Hide

```
# Now we will calculate various different R and R^2

R<- sqrt((3.074^2-2*1)/m01$null.deviance)
R
```

```
[1] 0.2198792
```

Hide

```
pseudoRsquared<- function(m){
  dev<- m$deviance
  nulldev<- m$null.deviance
  n<- length(m$fitted.values)
  R2_hl<- 1-dev/nulldev
  R2_cs<- 1-exp(-(nulldev-dev)/n)
  R2_n<-  R2_cs/(1-(exp(-(nulldev/n))))
  cat("Pseudo R^2 for logistic regression: \n")
  cat("Hosmer and Lemeshow R^2: ", round(R2_hl, 3), "\n")
  cat("Cox and Snell R^2: ", round(R2_cs ,3), "\n")
  cat("Nagelkerke R^2: ", round(R2_n, 3),"\n")
}

pseudoRsquared(m01)
```

```
Pseudo R^2 for logistic regression:
Hosmer and Lemeshow R^2:  0.064
Cox and Snell R^2:  0.084
Nagelkerke R^2:  0.113
```

Hide

```
# odds Ratio
exp(m01$coefficients)
```

```
             (Intercept) InterventionIntervention
                0.750000                 3.416667
```

Hide

```
# confidence interval of these odds, as it doesn't cross 1 , so it says as i
ntervention is done odds of
# being cured increases
exp(confint(m01))
```

```
Waiting for profiling to be done...
```

```
                             2.5 %   97.5 %
(Intercept)              0.4374531 1.268674
InterventionIntervention 1.5820127 7.625545
```

Hide

```
# Model 2 , Intervention and Duration as predictor
summary(m02)
```

```


Call:
glm(formula = Cured ~ Intervention + Duration, family = binomial(), 
    data = df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6025  -1.0572   0.8107   0.8161   1.3095  

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)   
(Intercept)              -0.234660   1.220563  -0.192  0.84754   
InterventionIntervention  1.233532   0.414565   2.975  0.00293 **
Duration                 -0.007835   0.175913  -0.045  0.96447   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 154.08  on 112  degrees of freedom
Residual deviance: 144.16  on 110  degrees of freedom
AIC: 150.16

Number of Fisher Scoring iterations: 4
```

Hide

```
modelChi<- m01$deviance - m02$deviance
chidf<- m01$df.residual - m02$df.residual
chisq.prob<- 1 - pchisq(modelChi, chidf)

chisq.prob
```

```
[1] 0.9644765
```

Hide

```
# from above chisq.prob , we can conclude that model 2 is not such an improv
ement over model 1
```

Hide

```
# also doing anova
anova(m01, m02)
```

```
Analysis of Deviance Table

Model 1: Cured ~ Intervention
Model 2: Cured ~ Intervention + Duration
  Resid. Df Resid. Dev Df  Deviance
1       111     144.16
2       110     144.16  1 0.0019835
```

Hide

```
# Doing casewise diagnostics
df$predicted.probablities<- fitted(m01)
df$standarized.residuals<- rstandard(m01)
df$studentized.residuals<- rstudent(m01)
df$dfbeta<- dfbeta(m01)
df$dffits<- dffits(m01)
df$leverage<- hatvalues(m01)

head(df)
```

Hide

```
# by seeing the residuals we can see that  none of the case to be seem an ou
tlier
head(df[order(-df$standarized.residuals),]$standarized.residuals, 10)
```

```
 [1] 1.313547 1.313547 1.313547 1.313547 1.313547 1.313547 1.313547 1.313547
1.313547 1.313547
```

Hide

```
# all cases have DFBetas less than 1, and leverage statistics are very close
to the calculated expected value of 0.018.
# All in all, this means that there are no influential cases having an effec
t on the model.
# The studentized residuals all have values of less than ±2 and so there see
ms to be very little here to concern us.
```

Hide

```
# Another Example
data<- read.delim('/home/atrides/Desktop/R/statistics_with_R/08_LogisticRegr
ession/Data_Files/penalty.dat', header=TRUE)
head(data)
```

Hide

```
# checking if Scored is a factor or not
is.factor(data$Scored)
```

```
[1] FALSE
```

Hide

```
# it  is not , so
data$Scored<- as.factor(data$Scored)

names(data)
```

```
[1] "PSWQ"     "Anxious"  "Previous" "Scored"
```

Hide

```
m01<- glm(Scored~PSWQ+Previous, data=data, family=binomial())
m02<- glm(Scored~PSWQ+Previous+Anxious, data=data, family=binomial())

anova(m01, m02)
```

```
Analysis of Deviance Table

Model 1: Scored ~ PSWQ + Previous
Model 2: Scored ~ PSWQ + Previous + Anxious
  Resid. Df Resid. Dev Df Deviance
1        72     48.662            
2        71     47.416  1   1.2463
```

Hide

```
print(summary(m01))
```

```

Call:
glm(formula = Scored ~ PSWQ + Previous, family = binomial(), 
    data = data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.2212  -0.3306   0.1038   0.5046   1.6067  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)   
(Intercept)  1.28031    1.67017   0.767  0.44333   
PSWQ        -0.23009    0.07983  -2.882  0.00395 **
Previous     0.06480    0.02209   2.934  0.00335 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 103.638  on 74  degrees of freedom
Residual deviance:  48.662  on 72  degrees of freedom
AIC: 54.662

Number of Fisher Scoring iterations: 6
```

Hide

```
print(summary(m02))
```

```

Call:
glm(formula = Scored ~ PSWQ + Previous + Anxious, family = binomial(), 
    data = data)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.31374  -0.35996   0.08334   0.53860   1.61380  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)   
(Intercept) -11.49256   11.80175  -0.974  0.33016   
PSWQ         -0.25137    0.08401  -2.992  0.00277 **
Previous      0.20261    0.12932   1.567  0.11719   
Anxious       0.27585    0.25259   1.092  0.27480   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 103.638  on 74  degrees of freedom
Residual deviance:  47.416  on 71  degrees of freedom
AIC: 55.416

Number of Fisher Scoring iterations: 6
```

Hide

```
modelChi1<- m01$null.deviance - m01$deviance
chidf1<- m01$df.null - m01$df.residual
chisq.prob1<- 1- pchisq(modelChi1, chidf1)

chisq.prob1
```

```
[1] 1.1533e-12
```

Hide

```
# the chisquare probability 'chisq.prob1' value is less than 0.05 which tell
s that this
# model was quite an improvement over a null model(i.e just chance)
```

Hide

```
# Now we will see whether model 2 is any improvement over model 1
modelChi2<- m01$deviance - m02$deviance
chidf2<- m01$df.residual - m02$df.residual
chisq.prob2<- 1-pchisq(modelChi2, chidf2)

chisq.prob2
```

```
[1] 0.2642667
```

Hide

```
# the chisquare probability 'chisq.prob2' value is greater than 0.05 , which
tells that this
# model(i.e m02) was a improvement over m01  , just by chance.
```

Hide

```
# dataframe of studentized residuals
df_resid<- rstudent(m01)
# printing the head, i.e top 10 residuals
head(df_resid[order(-df_resid)], 10)
```

```
        4        14        32        13         2         3        28
33         1        27
1.6882430 1.5949348 1.4174228 1.4170369 1.3540485 1.2509617 1.2509617 1.1185
592 0.8917579 0.8907728
```

Hide

```
# now , we will head to model m02, for assumption checking

# Testing for multicollinearity

# vif
vif(m02)
```

```
    PSWQ Previous  Anxious
  1.0898  35.2270  35.5820
```

Hide

```
# tolerance
1/vif(m02)
```

```
      PSWQ   Previous    Anxious
0.91759956 0.02838732 0.02810410
```

Hide

```
# from the output of  vif and tolerance , we can deduce that there is a high
multicollinearity in our model
```

Hide

```
# checking correlation between different independent variables
cor(data[, cbind('PSWQ', 'Anxious', 'Previous')])
```

```
               PSWQ    Anxious   Previous
PSWQ      1.0000000  0.6516416 -0.6435448
Anxious   0.6516416  1.0000000 -0.9928699
Previous -0.6435448 -0.9928699  1.0000000
```

Hide

```
# from the above table , the correlation b/w Anxious and Previous is very hi
gh,  thus leading to high multicollinearity
```

Hide

```
# Testing for linearity of logit
data$logPSWQ<- data$PSWQ * log(data$PSWQ)
data$logAnxious<- data$Anxious * log(data$Anxious)
data$logPrevious<- data$Previous * log(data$Previous)

head(data)
```

Hide

```
m03<- glm(Scored~PSWQ+logPSWQ+Anxious+logAnxious+Previous+logPrevious, data=
data, family=binomial())
summary(m03)
```

```

Call:
glm(formula = Scored ~ PSWQ + logPSWQ + Anxious + logAnxious + 
    Previous + logPrevious, family = binomial(), data = data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.0685  -0.3846   0.1116   0.5460   1.8272  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.87885   14.92410  -0.260    0.795
PSWQ        -0.42233    1.10267  -0.383    0.702
logPSWQ      0.04393    0.29675   0.148    0.882
Anxious     -2.64485    2.79702  -0.946    0.344
logAnxious   0.68077    0.65277   1.043    0.297
Previous     1.66601    1.48202   1.124    0.261
logPrevious -0.31855    0.31731  -1.004    0.315

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 97.283  on 70  degrees of freedom
Residual deviance: 45.909  on 64  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 59.909

Number of Fisher Scoring iterations: 7
```

Hide

```
# From the summary output , if any interaction term has significance less than 0.05 , it will mean that assumption
# of linearity has been violated. In our output we can conclude that the assumption of linearity has been met as all
# interaction term is non-significant
```