> #install packeges: RSQLite

>

> library(RSQLite)

> # Load the RSQLite Library

>

> #!!! Please point to you directory where are the csv files are saved!!!

> setwd("C:\\Users\\pkorzystko\\Downloads")

>

> carslist <- read.csv("auto-mpg.csv")

> mpgreal <- read.csv("mpg\_real.csv")

>

> colnames(mpgreal)[1] <- 'id'

>

> head(mpgreal)

id mpg\_real

1 380 50.584564

2 156 16.910975

3 363 27.664862

4 63 8.751536

5 244 28.440181

6 40 17.892034

> conn <- dbConnect(RSQLite::SQLite(), "CarsDB.db")

>

> dbWriteTable(conn,"refcars", carslist, overwrite = TRUE)

> dbWriteTable(conn,"mpgreal", mpgreal, overwrite = TRUE)

>

> ---------------------------

+ #A)

+

+ # base on 2 data set , one with car attributes and other with fuel consumation reading

+ # we will fit a model to identify if a car american, Origin 1 or not, origin 2 or 3

+ # our explanatory variables are avg(mpg\_real) , cylinders,displacement,horsepower

+ # weight, acceleration

+

+

+ #check if the tables exist

+ dbListTables(conn)

Error in -dbListTables(conn) : invalid argument to unary operator

>

> regdata <- dbGetQuery(conn, "SELECT case origin when 1 then 0 else 1 end origin ,

+ avg(mpg\_real) mpg, cylinders,displacement,horsepower

+ weight, acceleration

+ FROM refcars

+ JOIN mpgreal on refcars.id = mpgreal.id

+ GROUP BY refcars.id")

>

> #obtain train / test data seta

>

> summary(regdata)

origin mpg cylinders displacement weight acceleration

Min. :0.0000 Min. : 3.892 Min. :3.000 Min. : 68.0 Min. : 46.0 Min. : 8.00

1st Qu.:0.0000 1st Qu.:16.630 1st Qu.:4.000 1st Qu.:103.2 1st Qu.: 75.0 1st Qu.:13.78

Median :0.0000 Median :21.978 Median :4.000 Median :148.5 Median : 92.5 Median :15.50

Mean :0.3699 Mean :23.370 Mean :5.457 Mean :193.7 Mean :104.1 Mean :15.53

3rd Qu.:1.0000 3rd Qu.:29.201 3rd Qu.:8.000 3rd Qu.:263.2 3rd Qu.:125.0 3rd Qu.:17.02

Max. :1.0000 Max. :60.851 Max. :8.000 Max. :455.0 Max. :230.0 Max. :24.60

> set.seed(1)

>

> sample <- sample(c(TRUE, FALSE), nrow(regdata), replace=TRUE, prob=c(0.8,0.2))

> train <- regdata[sample, ]

> test <- regdata[!sample, ]

>

> #1 model include all variables

> # Y is mp, we will explain origin frommgp n.cylinders displacement weight and acceleration

> model1 <- glm( origin ~ mpg + cylinders + displacement + weight + acceleration , data=train)

> summary(model1)

Call:

glm(formula = origin ~ mpg + cylinders + displacement + weight +

acceleration, data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.78531 -0.19548 0.03913 0.28847 0.69568

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.2179955 0.2806209 0.777 0.43784

mpg 0.0094823 0.0029627 3.201 0.00151 \*\*

cylinders 0.0688731 0.0374093 1.841 0.06655 .

displacement -0.0058390 0.0007404 -7.886 5.15e-14 \*\*\*

weight 0.0071059 0.0013970 5.087 6.27e-07 \*\*\*

acceleration -0.0036676 0.0101240 -0.362 0.71739

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1183422)

Null deviance: 73.801 on 320 degrees of freedom

Residual deviance: 37.278 on 315 degrees of freedom

AIC: 233.83

Number of Fisher Scoring iterations: 2

>

> model\_final <- glm( origin ~ mpg + cylinders + weight , data=train)

> summary(model\_final)

Call:

glm(formula = origin ~ mpg + cylinders + weight, data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.77246 -0.22906 0.04785 0.33624 0.82144

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.774981 0.154529 5.015 8.84e-07 \*\*\*

mpg 0.011890 0.003203 3.712 0.000243 \*\*\*

cylinders -0.175169 0.023632 -7.412 1.14e-12 \*\*\*

weight 0.002597 0.001075 2.417 0.016201 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1414822)

Null deviance: 73.801 on 320 degrees of freedom

Residual deviance: 44.850 on 317 degrees of freedom

AIC: 289.19

Number of Fisher Scoring iterations: 2

>

> ---------------------------

+ #B)

+

+ summary(model\_final)

Error in -summary(model\_final) : invalid argument to unary operator

> # By the interpretation of the result at significance of 5%, mgp and acelaration are not significant

>

> #removing no significant variable displacement iteratively we obtain:

> #as the correlation between cylinders and displacement is high we can only use one or another

> model2 <- glm( origin ~ mpg + cylinders , data=train, family=binomial)

>

> summary(model2)

Call:

glm(formula = origin ~ mpg + cylinders, family = binomial, data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8206 -0.4488 -0.1111 0.8453 2.3023

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 4.15680 1.18134 3.519 0.000434 \*\*\*

mpg 0.05113 0.02125 2.406 0.016118 \*

cylinders -1.24707 0.19942 -6.253 4.02e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 418.85 on 320 degrees of freedom

Residual deviance: 258.49 on 318 degrees of freedom

AIC: 264.49

Number of Fisher Scoring iterations: 6

> # the best model based on AIC criterium is origin ~ mpg + cylinder

>

> #All remaining varaibles are significant

> cor(regdata)

origin mpg cylinders displacement weight acceleration

origin 1.0000000 0.5085393 -0.6030285 -0.6496571 -0.4852385 0.2441314

mpg 0.5085393 1.0000000 -0.6477016 -0.6706126 -0.6566669 0.3758771

cylinders -0.6030285 -0.6477016 1.0000000 0.9511001 0.8414184 -0.5120941

displacement -0.6496571 -0.6706126 0.9511001 1.0000000 0.8991050 -0.5515411

weight -0.4852385 -0.6566669 0.8414184 0.8991050 1.0000000 -0.7041587

acceleration 0.2441314 0.3758771 -0.5120941 -0.5515411 -0.7041587 1.0000000

>

> ---------------------------

+ #C)

+

+ #predict result based on model2

+ predicted <- predict(model\_final, test, type="response")

Error in ---------------------------predicted <- predict(model\_final, :

could not find function "-<-"

>

> head(predicted)

4 6 7 18 21 29

-0.04077199 0.09166079 0.11720694 0.19824473 0.67192361 0.78657792

> summary(predicted)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-0.1269 0.1591 0.5352 0.4092 0.6582 0.8197

>

> ---------------------------

+ #D)

+

+ #array(confusion matrix)

+

+ #!!!please inatall caret library

+

+ library(caret)

Error in -library(caret) : invalid argument to unary operator

> confusionMatrix(as.factor(ifelse(predicted < 0.5,0,1)) , as.factor(test$origin))

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 29 5

1 12 25

Accuracy : 0.7606

95% CI : (0.6446, 0.8539)

No Information Rate : 0.5775

P-Value [Acc > NIR] : 0.001003

Kappa : 0.5242

Mcnemar's Test P-Value : 0.145610

Sensitivity : 0.7073

Specificity : 0.8333

Pos Pred Value : 0.8529

Neg Pred Value : 0.6757

Prevalence : 0.5775

Detection Rate : 0.4085

Detection Prevalence : 0.4789

Balanced Accuracy : 0.7703

'Positive' Class : 0

>

> #Acurracy obtained is .7606