**DATA 590 | Homework 2 | Praveen Kumar Neelappa**

**1. Using logistic regression, see how well you can predict CHD based on all the quantitative inputs. Examine the model and remove some inputs that do not seem to help classify one at a time (as we did using multiple regression in the spring), and come up with the simplest model that classifies nearly as well as, or better than, using all inputs.**

#All required library

install.packages("caret")

library(caret)

install.packages("e1071")

library("e1071")

> chdfit <- glm(as.factor(chd)~

+ sbp+tobacco+ldl+adiposity+famhist+typea+obesity+alcohol+age,data=sa,family=binomial())

>

> summary(chdfit)

Call:

glm(formula = as.factor(chd) ~ sbp + tobacco + ldl + adiposity +

famhist + typea + obesity + alcohol + age, family = binomial(),

data = sa)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7781 -0.8213 -0.4387 0.8889 2.5435

Coefficients:

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

(Intercept) -6.1507209 1.3082600 -4.701 2.58e-06 \*\*\*

sbp 0.0065040 0.0057304 1.135 0.256374

tobacco 0.0793764 0.0266028 2.984 0.002847 \*\*

ldl 0.1739239 0.0596617 2.915 0.003555 \*\*

adiposity 0.0185866 0.0292894 0.635 0.525700

famhistPresent 0.9253704 0.2278940 4.061 4.90e-05 \*\*\*

typea 0.0395950 0.0123202 3.214 0.001310 \*\*

obesity -0.0629099 0.0442477 -1.422 0.155095

alcohol 0.0001217 0.0044832 0.027 0.978350

age 0.0452253 0.0121298 3.728 0.000193 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 596.11 on 461 degrees of freedom

Residual deviance: 472.14 on 452 degrees of freedom

AIC: 492.14

Number of Fisher Scoring iterations: 5

>

> exp(coef(chdfit))

(Intercept) sbp tobacco ldl

0.002131944 1.006525214 1.082611790 1.189965004

adiposity famhistPresent typea obesity

1.018760374 2.522802582 1.040389357 0.939028105

alcohol age

1.000121670 1.046263608

> chdfac <- fitted(chdfit)

> thresh <- 0.5

> gchd <- cut(chdfac, breaks=c(-Inf, thresh, Inf), labels=c(0,1))

>

> confusionMatrix(gchd, sa$chd)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 256 77

1 46 83

Accuracy : 0.7338

95% CI : (0.691, 0.7735)

No Information Rate : 0.6537

P-Value [Acc > NIR] : 0.0001366

Kappa : 0.3839

Mcnemar's Test P-Value : 0.0068303

Sensitivity : 0.8477

Specificity : 0.5188

Pos Pred Value : 0.7688

Neg Pred Value : 0.6434

Prevalence : 0.6537

Detection Rate : 0.5541

Detection Prevalence : 0.7208

Balanced Accuracy : 0.6832

'Positive' Class : 0

> chdfit <- glm(as.factor(chd)~

+ sbp+tobacco+ldl+adiposity+famhist+typea+obesity+age,data=sa,family=binomial())

>

> summary(chdfit)

Call:

glm(formula = as.factor(chd) ~ sbp + tobacco + ldl + adiposity +

famhist + typea + obesity + age, family = binomial(), data = sa)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7795 -0.8207 -0.4391 0.8882 2.5427

Coefficients:

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

(Intercept) -6.150058 1.308008 -4.702 2.58e-06 \*\*\*

sbp 0.006524 0.005685 1.148 0.251149

tobacco 0.079515 0.026114 3.045 0.002327 \*\*

ldl 0.173770 0.059393 2.926 0.003436 \*\*

adiposity 0.018631 0.029245 0.637 0.524079

famhistPresent 0.925831 0.227266 4.074 4.63e-05 \*\*\*

typea 0.039604 0.012316 3.216 0.001302 \*\*

obesity -0.062957 0.044216 -1.424 0.154489

age 0.045191 0.012061 3.747 0.000179 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 596.11 on 461 degrees of freedom

Residual deviance: 472.14 on 453 degrees of freedom

AIC: 490.14

Number of Fisher Scoring iterations: 5

>

> exp(coef(chdfit))

(Intercept) sbp tobacco ldl

0.002133358 1.006545077 1.082761603 1.189782126

adiposity famhistPresent typea obesity

1.018805509 2.523964064 1.040398705 0.938984162

age

1.046227384

>

> chdfac <- fitted(chdfit)

> thresh <- 0.5

> gchd <- cut(chdfac, breaks=c(-Inf, thresh, Inf), labels=c(0,1))

>

> confusionMatrix(gchd, sa$chd)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 256 77

1 46 83

Accuracy : 0.7338

95% CI : (0.691, 0.7735)

No Information Rate : 0.6537

P-Value [Acc > NIR] : 0.0001366

Kappa : 0.3839

Mcnemar's Test P-Value : 0.0068303

Sensitivity : 0.8477

Specificity : 0.5188

Pos Pred Value : 0.7688

Neg Pred Value : 0.6434

Prevalence : 0.6537

Detection Rate : 0.5541

Detection Prevalence : 0.7208

Balanced Accuracy : 0.6832

'Positive' Class : 0

> chdfit <- glm(as.factor(chd)~

+ sbp+tobacco+ldl+famhist+typea+obesity+age,data=sa,family=binomial())

>

> summary(chdfit)

Call:

glm(formula = as.factor(chd) ~ sbp + tobacco + ldl + famhist +

typea + obesity + age, family = binomial(), data = sa)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8245 -0.8189 -0.4415 0.8892 2.5530

Coefficients:

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

(Intercept) -6.416927 1.240101 -5.175 2.28e-07 \*\*\*

sbp 0.006780 0.005683 1.193 0.23286

tobacco 0.079886 0.026157 3.054 0.00226 \*\*

ldl 0.182102 0.058077 3.136 0.00172 \*\*

famhistPresent 0.924464 0.227061 4.071 4.67e-05 \*\*\*

typea 0.038966 0.012266 3.177 0.00149 \*\*

obesity -0.042200 0.029437 -1.434 0.15169

age 0.048927 0.010556 4.635 3.57e-06 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 596.11 on 461 degrees of freedom

Residual deviance: 472.55 on 454 degrees of freedom

AIC: 488.55

Number of Fisher Scoring iterations: 5

>

> exp(coef(chdfit))

(Intercept) sbp tobacco ldl

0.001633669 1.006802723 1.083163654 1.199736086

famhistPresent typea obesity age

2.520517883 1.039734647 0.958678333 1.050143197

>

> chdfac <- fitted(chdfit)

> thresh <- 0.5

> gchd <- cut(chdfac, breaks=c(-Inf, thresh, Inf), labels=c(0,1))

>

> confusionMatrix(gchd, sa$chd)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 258 75

1 44 85

Accuracy : 0.7424

95% CI : (0.7, 0.7817)

No Information Rate : 0.6537

P-Value [Acc > NIR] : 2.562e-05

Kappa : 0.404

Mcnemar's Test P-Value : 0.005958

Sensitivity : 0.8543

Specificity : 0.5312

Pos Pred Value : 0.7748

Neg Pred Value : 0.6589

Prevalence : 0.6537

Detection Rate : 0.5584

Detection Prevalence : 0.7208

Balanced Accuracy : 0.6928

'Positive' Class : 0

> chdfit <- glm(as.factor(chd)~

+ tobacco+ldl+famhist+typea+obesity+age,data=sa,family=binomial())

>

> summary(chdfit)

Call:

glm(formula = as.factor(chd) ~ tobacco + ldl + famhist + typea +

obesity + age, family = binomial(), data = sa)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.8941 -0.8164 -0.4329 0.8966 2.5442

Coefficients:

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

(Intercept) -5.70273 1.07640 -5.298 1.17e-07 \*\*\*

tobacco 0.07999 0.02598 3.079 0.00208 \*\*

ldl 0.18372 0.05818 3.158 0.00159 \*\*

famhistPresent 0.91610 0.22645 4.046 5.22e-05 \*\*\*

typea 0.03827 0.01222 3.133 0.00173 \*\*

obesity -0.03760 0.02910 -1.292 0.19638

age 0.05211 0.01024 5.087 3.63e-07 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 596.11 on 461 degrees of freedom

Residual deviance: 473.98 on 455 degrees of freedom

AIC: 487.98

Number of Fisher Scoring iterations: 5

>

> exp(coef(chdfit))

(Intercept) tobacco ldl famhistPresent

0.003336857 1.083275322 1.201673625 2.499512668

typea obesity age

1.039015387 0.963100903 1.053493084

>

> chdfac <- fitted(chdfit)

> thresh <- 0.5

> gchd <- cut(chdfac, breaks=c(-Inf, thresh, Inf), labels=c(0,1))

>

> confusionMatrix(gchd, sa$chd)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 257 73

1 45 87

Accuracy : 0.7446

95% CI : (0.7023, 0.7838)

No Information Rate : 0.6537

P-Value [Acc > NIR] : 1.641e-05

Kappa : 0.4117

Mcnemar's Test P-Value : 0.01294

Sensitivity : 0.8510

Specificity : 0.5437

Pos Pred Value : 0.7788

Neg Pred Value : 0.6591

Prevalence : 0.6537

Detection Rate : 0.5563

Detection Prevalence : 0.7143

Balanced Accuracy : 0.6974

'Positive' Class : 0

> chdfit <- glm(as.factor(chd)~

+ tobacco+ldl+famhist+typea+age,data=sa,family=binomial())

>

> summary(chdfit)

Call:

glm(formula = as.factor(chd) ~ tobacco + ldl + famhist + typea +

age, family = binomial(), data = sa)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9165 -0.8054 -0.4430 0.9329 2.6139

Coefficients:

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

(Intercept) -6.44644 0.92087 -7.000 2.55e-12 \*\*\*

tobacco 0.08038 0.02588 3.106 0.00190 \*\*

ldl 0.16199 0.05497 2.947 0.00321 \*\*

famhistPresent 0.90818 0.22576 4.023 5.75e-05 \*\*\*

typea 0.03712 0.01217 3.051 0.00228 \*\*

age 0.05046 0.01021 4.944 7.65e-07 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 596.11 on 461 degrees of freedom

Residual deviance: 475.69 on 456 degrees of freedom

AIC: 487.69

Number of Fisher Scoring iterations: 5

>

> exp(coef(chdfit))

(Intercept) tobacco ldl famhistPresent

0.001586152 1.083693731 1.175850406 2.479793436

typea age

1.037812583 1.051755195

>

> chdfac <- fitted(chdfit)

> thresh <- 0.5

> gchd <- cut(chdfac, breaks=c(-Inf, thresh, Inf), labels=c(0,1))

>

> confusionMatrix(gchd, sa$chd)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 256 73

1 46 87

Accuracy : 0.7424

95% CI : (0.7, 0.7817)

No Information Rate : 0.6537

P-Value [Acc > NIR] : 2.562e-05

Kappa : 0.4076

Mcnemar's Test P-Value : 0.01715

Sensitivity : 0.8477

Specificity : 0.5437

Pos Pred Value : 0.7781

Neg Pred Value : 0.6541

Prevalence : 0.6537

Detection Rate : 0.5541

Detection Prevalence : 0.7121

Balanced Accuracy : 0.6957

'Positive' Class : 0

**Describe the inputs as having a positive effect on wellness (no CHD), ie, if the increase in the value is associated with a decrease in the probability of CHD.?**

As per the below table

Coefficients:

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

(Intercept) -6.44644 0.92087 -7.000 2.55e-12 \*\*\*

tobacco 0.08038 0.02588 3.106 0.00190 \*\*

ldl 0.16199 0.05497 2.947 0.00321 \*\*

famhistPresent 0.90818 0.22576 4.023 5.75e-05 \*\*\*

typea 0.03712 0.01217 3.051 0.00228 \*\*

age 0.05046 0.01021 4.944 7.65e-07 \*\*\*

We see that family history has the highest positive effect on the having CD with an value of 0.9 followed by ldl, tabacco, age and typea.

**2. Using LDFA, see how well you can predict CHD based on all the quantitative inputs.**

> salda <- lda(as.factor(chd)~

+ sbp+tobacco+ldl+adiposity+famhist+typea+obesity+alcohol+age,data=sa, CV=T)

> confusionMatrix(salda$class, sa$chd)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 256 79

1 46 81

Accuracy : 0.7294

95% CI : (0.6865, 0.7694)

No Information Rate : 0.6537

P-Value [Acc > NIR] : 0.0002962

Kappa : 0.372

Mcnemar's Test P-Value : 0.0042076

Sensitivity : 0.8477

Specificity : 0.5062

Pos Pred Value : 0.7642

Neg Pred Value : 0.6378

Prevalence : 0.6537

Detection Rate : 0.5541

Detection Prevalence : 0.7251

Balanced Accuracy : 0.6770

'Positive' Class : 0

**3. How does family history alone help predict CHD?**

> sa$famhist[sa$famhist == "Present"] <- 1

> sa$famhist[sa$famhist == "Absent"] <- 0

> cor.test(sa$chd, as.numeric(as.character(sa$famhist)))

Pearson's product-moment correlation

data: sa$chd and as.numeric(as.character(sa$famhist))

t = 6.0713, df = 460, p-value = 2.658e-09

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.1857596 0.3547858

sample estimates:

cor

0.2723727

**What other one or two inputs with age help predict CHD best?**

> chdfit <- glm(as.factor(chd)~ age+famhist+ldl,data=sa,family=binomial())

> chdfac <- fitted(chdfit)

> thresh <- 0.5

> gchd <- cut(chdfac, breaks=c(-Inf, thresh, Inf), labels=c(0,1))

> confusionMatrix(gchd, sa$chd)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 259 80

1 43 80

Accuracy : 0.7338

95% CI : (0.691, 0.7735)

No Information Rate : 0.6537

P-Value [Acc > NIR] : 0.0001366

Kappa : 0.3782

Mcnemar's Test P-Value : 0.0011703

Sensitivity : 0.8576

Specificity : 0.5000

Pos Pred Value : 0.7640

Neg Pred Value : 0.6504

Prevalence : 0.6537

Detection Rate : 0.5606

Detection Prevalence : 0.7338

Balanced Accuracy : 0.6788

'Positive' Class : 0

The accuracy and positive prediction value is high with age, family history and ldl