library(car)

library(alr4)

library(ResourceSelection)

heart <- read.csv("/home/thomas/git/datascience/MSDataSci/MATH550/data/heart.csv",header=TRUE)

attach(heart)

#Change Absent/Present to 0/1 in famhist

lookup <- c("Absent" = 0, "Present" = 1)

heart$famhist <- lookup[heart$famhist]

attach(heart)

heart

#1

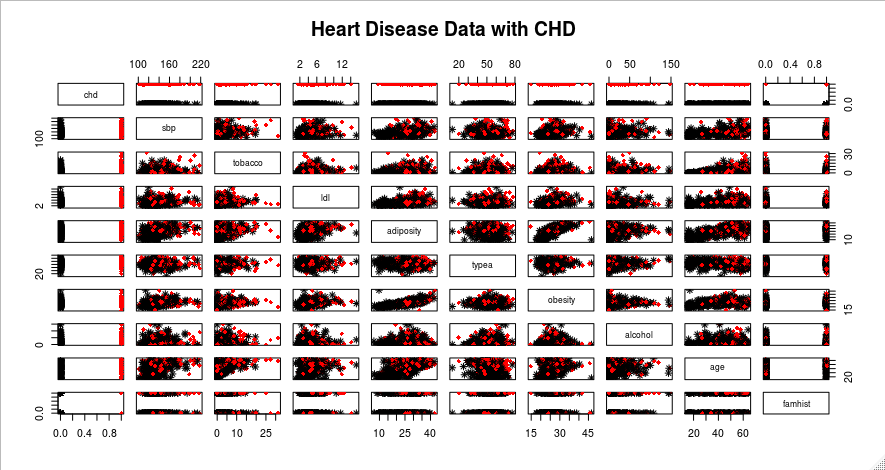
group <- NA

group[heart$chd == 0] <- 1

group[heart$chd == 1] <- 2

pairs(chd~sbp+tobacco+ldl+adiposity+typea+obesity+alcohol+age+famhist,heart,

col = c("black", "red")[group],pch = c(8, 18)[group],main = "Heart Disease Data with CHD")



#2

par(mfrow=c(3,3))

# Skewed

boxplot(sbp~chd, ylab="Systolic Blood Pressure",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Skewed

boxplot(tobacco~chd, ylab="Tobacco Use",xlab="CHD? (0=No, 1=Yes)",col=10, pch=19)

# Skewed

boxplot(ldl~chd, ylab="Low-Densty Lipid Level",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Looks OK

boxplot(adiposity~chd, ylab="Adiposity",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

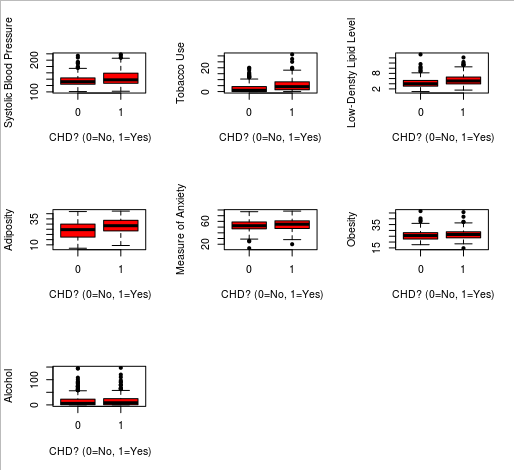
# Looks OK

boxplot(typea~chd, ylab="Measure of Anxiety",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Skewed

boxplot(obesity~chd, ylab="Obesity",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Badly Skewed

boxplot(alcohol~chd, ylab="Alcohol",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Add 0.001 to Tobacco and alcohol for log transforms

tobacco <- tobacco+0.001

alcohol <- alcohol+0.001

#Log Transform sbp,tobacco,ldl,obesity,alcohol

sbp<-log(sbp)

tobacco <- log(tobacco)

ldl<-log(ldl)

obesity<-log(obesity)

alcohol<-log(alcohol)

typea<-log(typea)

par(mfrow=c(3,3))

# Skewed

boxplot(sbp~chd, ylab="Systolic Blood Pressure",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Skewed

boxplot(tobacco~chd, ylab="Tobacco Use",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Skewed

boxplot(ldl~chd, ylab="Low-Densty Lipid Level",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Looks OK

boxplot(adiposity~chd, ylab="Adiposity",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Looks OK

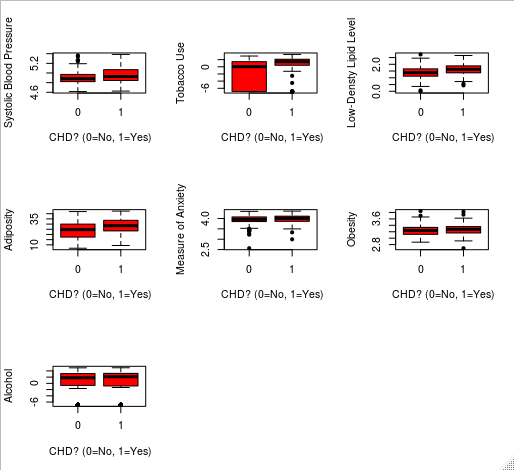
boxplot(typea~chd, ylab="Measure of Anxiety",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Skewed

boxplot(obesity~chd, ylab="Obesity",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)

# Skewed

boxplot(alcohol~chd, ylab="Alcohol",xlab=" CHD? (0=No, 1=Yes)",col=10, pch=19)



#3

# The transformations have visually improved the data and will be used in the model

m1 <- glm(chd~sbp+tobacco+ldl+adiposity+typea+obesity+alcohol+age,family=binomial(),data=heart)

summary(m1)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6364 -0.8659 -0.4484 0.9728 2.5906

Coefficients:

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

(Intercept) -9.41462 5.59133 -1.684 0.09222 .

sbp 0.85039 0.83102 1.023 0.30616

tobacco 0.11969 0.04031 2.969 0.00298 \*\*

ldl 0.94707 0.31127 3.043 0.00235 \*\*

adiposity 0.03174 0.02886 1.100 0.27132

typea 1.83847 0.59623 3.083 0.00205 \*\*

obesity -2.16328 1.19132 -1.816 0.06939 .

alcohol -0.00742 0.02805 -0.265 0.79136

age 0.04685 0.01189 3.939 8.19e-05 \*\*\*

---

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: 487.90 on 453 degrees of freedom

AIC: 505.9

Number of Fisher Scoring iterations: 5

#full model with all interactions

m1 <- glm(chd~sbp+tobacco+ldl+adiposity+typea+obesity+alcohol+age+famhist+

sbp\*ldl+

sbp\*adiposity+

sbp\*tobacco+

sbp\*typea+

sbp\*obesity+

sbp\*alcohol+

sbp\*age+

sbp\*famhist+

famhist\*ldl+

famhist\*adiposity+

famhist\*tobacco+

famhist\*typea+

famhist\*obesity+

famhist\*alcohol+

famhist\*age+

tobacco\*ldl+

tobacco\*adiposity+

tobacco\*typea+

tobacco\*obesity+

tobacco\*alcohol+

tobacco\*age+

ldl\*adiposity+

ldl\*typea+

ldl\*obesity+

ldl\*alcohol+

ldl\*age+

adiposity\*typea+

adiposity\*obesity+

adiposity\*alcohol+

adiposity\*age+

typea\*obesity+

typea\*alcohol+

typea\*age+

obesity\*alcohol+

obesity\*age+

alcohol\*age

,family=binomial(),data=heart)

summary(m1)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.9834 -0.8084 -0.3142 0.7835 2.8484

Coefficients:

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

(Intercept) -48.722888 211.446696 -0.230 0.8178

sbp 34.206018 38.653936 0.885 0.3762

tobacco 5.994471 2.724341 2.200 0.0278 \*

ldl -10.484103 18.181865 -0.577 0.5642

adiposity -1.159272 1.414712 -0.819 0.4125

typea -19.697809 34.574862 -0.570 0.5689

obesity 25.136141 70.895559 0.355 0.7229

alcohol -3.036374 1.703721 -1.782 0.0747 .

age -0.715393 0.649238 -1.102 0.2705

famhist -1.499332 13.648896 -0.110 0.9125

sbp:ldl 0.609688 2.609516 0.234 0.8153

sbp:adiposity 0.201383 0.251662 0.800 0.4236

sbp:tobacco -0.489254 0.406124 -1.205 0.2283

sbp:typea 0.429719 5.183888 0.083 0.9339

sbp:obesity -14.015275 12.783017 -1.096 0.2729

sbp:alcohol 0.513400 0.245233 2.094 0.0363 \*

sbp:age 0.074012 0.094994 0.779 0.4359

sbp:famhist 0.445469 1.934091 0.230 0.8178

ldl:famhist 1.660382 0.703348 2.361 0.0182 \*

adiposity:famhist -0.062812 0.069866 -0.899 0.3686

tobacco:famhist 0.131618 0.112346 1.172 0.2414

typea:famhist -1.180735 1.426754 -0.828 0.4079

obesity:famhist 0.872464 2.922345 0.299 0.7653

alcohol:famhist 0.010287 0.064437 0.160 0.8732

age:famhist 0.024131 0.028278 0.853 0.3935

tobacco:ldl -0.196034 0.141675 -1.384 0.1665

tobacco:adiposity 0.015438 0.013678 1.129 0.2590

tobacco:typea -0.634817 0.304785 -2.083 0.0373 \*

tobacco:obesity -0.254987 0.511437 -0.499 0.6181

tobacco:alcohol 0.010661 0.010737 0.993 0.3207

tobacco:age -0.005773 0.004744 -1.217 0.2236

ldl:adiposity 0.015699 0.081853 0.192 0.8479

ldl:typea -0.111528 2.055002 -0.054 0.9567

ldl:obesity 1.904586 3.844375 0.495 0.6203

ldl:alcohol 0.057241 0.096554 0.593 0.5533

ldl:age 0.031150 0.034213 0.910 0.3626

adiposity:typea -0.005035 0.180757 -0.028 0.9778

adiposity:obesity 0.136778 0.122812 1.114 0.2654

adiposity:alcohol -0.016996 0.009548 -1.780 0.0751 .

adiposity:age -0.004119 0.002959 -1.392 0.1639

typea:obesity 6.600330 8.180496 0.807 0.4198

typea:alcohol -0.147742 0.183518 -0.805 0.4208

typea:age -0.013936 0.070973 -0.196 0.8443

obesity:alcohol 0.490682 0.387352 1.267 0.2052

obesity:age 0.154539 0.142936 1.081 0.2796

alcohol:age -0.002721 0.003494 -0.779 0.4361

---

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: 435.84 on 416 degrees of freedom

AIC: 527.84

Number of Fisher Scoring iterations: 6

It is interesting to see that interactions might seem to matter somewhat. I am hesitant still. There could be a lot of multicollinearity going on with the interactions now. As far as a physical explanation for the significance of the interactions, I have none. I do not understand heart disease at such a level. However, What can be done is a back step model. At least that way, I will be able to see an optimal set of predictors across the span – which looks a little crowded right now.

#4

# Variable Reduction

n <- length(m1$residuals)

# BIC gives a score of 513.44 and 3 variables

backBIC <- step(m1,direction="backward", heart=heart, k=log(n))

# AIC gives us a score of 480.36 and the same 3 variables

backAIC <- step(m1,direction="backward", heart=heart)

# It is interesting to see that the significant interaction predictors are the ones

# that ended up in the back step models.

# Reduced Model

reducedModel <- glm(chd ~ age + tobacco:typea + ldl:famhist,family=binomial(),data=heart)

summary(reducedModel)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7756 -0.8808 -0.4354 0.9012 2.5915

Coefficients:

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

(Intercept) -3.267132 0.473419 -6.901 5.16e-12 \*\*\*

age 0.047889 0.009608 4.984 6.23e-07 \*\*\*

tobacco:typea 0.032341 0.009923 3.259 0.00112 \*\*

ldl:famhist 0.698333 0.133131 5.245 1.56e-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: 484.59 on 458 degrees of freedom

AIC: 492.59

Number of Fisher Scoring iterations: 5

This is a lot better! Now that the span is not so crowded, the remaining variables are much

more predictive than they were before.

#5

# MMPs and Hosmer-Lemeshow

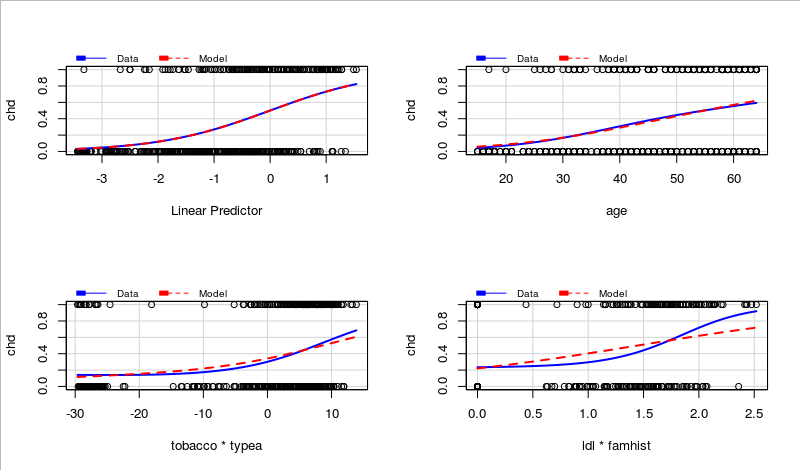
par(mfrow = c(2,2))

mmp(reducedModel)

mmp(reducedModel,age)

mmp(reducedModel,tobacco\*typea)

mmp(reducedModel,ldl\*famhist)



The first two plots are looking good. However, for the combined predictors, the model is looking a

little conservative on the MMP plots relative to the data.

hoslem.test(chd, fitted(reducedModel), g=10)

data: chd, fitted(reducedModel)

X-squared = 5.1677, df = 8, p-value = 0.7395

With a high p value, we see that the model is not fitting the data poorly.

#6

#Create a confusion matrix

y\_hat <- predict(reducedModel, type='response')

myPredict <- ifelse(y\_hat>0.5,1,0)

table(myPredict, heart$chd)

myPredict 0 1

0 259 82

1 43 78

Accuracy <- (78 + 259) / (78 + 259 + 82 + 43)

Precision <- 78 / (78 + 82)

Accuracy

[1] 0.7294372

Precision

[1] 0.4875