BUDA530Module2

## Problem 1 :

We want to understand if the tumor is cancerous or not based on measurements of the cell. The task is to create insights on these measurements and find if any of the measurements are useful and report on them. Using the wbca data in the faraway package. Fit a binary logistic regression model with Class as the response and the other variables as predictors. Comment on the model deviance and tests for the coefficients. Attempt model selection using the step function and comment on any reduction that takes place. (HINT THIS IS SIMLAR TO USING STEP with lm).

library(faraway)

## Warning: package 'faraway' was built under R version 3.6.2

data("wbca")  
head(wbca)

## Class Adhes BNucl Chrom Epith Mitos NNucl Thick UShap USize  
## 1 1 1 1 3 2 1 1 5 1 1  
## 2 1 5 10 3 7 1 2 5 4 4  
## 3 1 1 2 3 2 1 1 3 1 1  
## 4 1 1 4 3 3 1 7 6 8 8  
## 5 1 3 1 3 2 1 1 4 1 1  
## 6 0 8 10 9 7 1 7 8 10 10

#pairs(wbca,col=(wbca$Class))  
  
mod1 <- glm(Class ~ ., data=wbca, family="binomial")  
summary(mod1)

##   
## Call:  
## glm(formula = Class ~ ., family = "binomial", data = wbca)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.48282 -0.01179 0.04739 0.09678 3.06425   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 11.16678 1.41491 7.892 2.97e-15 \*\*\*  
## Adhes -0.39681 0.13384 -2.965 0.00303 \*\*   
## BNucl -0.41478 0.10230 -4.055 5.02e-05 \*\*\*  
## Chrom -0.56456 0.18728 -3.014 0.00257 \*\*   
## Epith -0.06440 0.16595 -0.388 0.69795   
## Mitos -0.65713 0.36764 -1.787 0.07387 .   
## NNucl -0.28659 0.12620 -2.271 0.02315 \*   
## Thick -0.62675 0.15890 -3.944 8.01e-05 \*\*\*  
## UShap -0.28011 0.25235 -1.110 0.26699   
## USize 0.05718 0.23271 0.246 0.80589   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 881.388 on 680 degrees of freedom  
## Residual deviance: 89.464 on 671 degrees of freedom  
## AIC: 109.46  
##   
## Number of Fisher Scoring iterations: 8

mod2 <- step(mod1)

## Start: AIC=109.46  
## Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +   
## UShap + USize  
##   
## Df Deviance AIC  
## - USize 1 89.523 107.52  
## - Epith 1 89.613 107.61  
## - UShap 1 90.627 108.63  
## <none> 89.464 109.46  
## - Mitos 1 93.551 111.55  
## - NNucl 1 95.204 113.20  
## - Adhes 1 98.844 116.84  
## - Chrom 1 99.841 117.84  
## - BNucl 1 109.000 127.00  
## - Thick 1 110.239 128.24  
##   
## Step: AIC=107.52  
## Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +   
## UShap  
##   
## Df Deviance AIC  
## - Epith 1 89.662 105.66  
## - UShap 1 91.355 107.36  
## <none> 89.523 107.52  
## - Mitos 1 93.552 109.55  
## - NNucl 1 95.231 111.23  
## - Adhes 1 99.042 115.04  
## - Chrom 1 100.153 116.15  
## - BNucl 1 109.064 125.06  
## - Thick 1 110.465 126.47  
##   
## Step: AIC=105.66  
## Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap  
##   
## Df Deviance AIC  
## <none> 89.662 105.66  
## - UShap 1 91.884 105.88  
## - Mitos 1 93.714 107.71  
## - NNucl 1 95.853 109.85  
## - Adhes 1 100.126 114.13  
## - Chrom 1 100.844 114.84  
## - BNucl 1 109.762 123.76  
## - Thick 1 110.632 124.63

summary(mod2)

##   
## Call:  
## glm(formula = Class ~ Adhes + BNucl + Chrom + Mitos + NNucl +   
## Thick + UShap, family = "binomial", data = wbca)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.44161 -0.01119 0.04962 0.09741 3.08205   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 11.0333 1.3632 8.094 5.79e-16 \*\*\*  
## Adhes -0.3984 0.1294 -3.080 0.00207 \*\*   
## BNucl -0.4192 0.1020 -4.111 3.93e-05 \*\*\*  
## Chrom -0.5679 0.1840 -3.085 0.00203 \*\*   
## Mitos -0.6456 0.3634 -1.777 0.07561 .   
## NNucl -0.2915 0.1236 -2.358 0.01837 \*   
## Thick -0.6216 0.1579 -3.937 8.27e-05 \*\*\*  
## UShap -0.2541 0.1785 -1.423 0.15461   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 881.388 on 680 degrees of freedom  
## Residual deviance: 89.662 on 673 degrees of freedom  
## AIC: 105.66  
##   
## Number of Fisher Scoring iterations: 8

model3 <- glm(Class ~Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap, data=wbca, family="binomial" )  
summary(model3)

##   
## Call:  
## glm(formula = Class ~ Adhes + BNucl + Chrom + Mitos + NNucl +   
## Thick + UShap, family = "binomial", data = wbca)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.44161 -0.01119 0.04962 0.09741 3.08205   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 11.0333 1.3632 8.094 5.79e-16 \*\*\*  
## Adhes -0.3984 0.1294 -3.080 0.00207 \*\*   
## BNucl -0.4192 0.1020 -4.111 3.93e-05 \*\*\*  
## Chrom -0.5679 0.1840 -3.085 0.00203 \*\*   
## Mitos -0.6456 0.3634 -1.777 0.07561 .   
## NNucl -0.2915 0.1236 -2.358 0.01837 \*   
## Thick -0.6216 0.1579 -3.937 8.27e-05 \*\*\*  
## UShap -0.2541 0.1785 -1.423 0.15461   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 881.388 on 680 degrees of freedom  
## Residual deviance: 89.662 on 673 degrees of freedom  
## AIC: 105.66  
##   
## Number of Fisher Scoring iterations: 8

### Interpretation :

As per the summary above for Model considering all the predictors, the residual deviance is 89.662 and corresponding degree of freedom are 673. So if we look at the summary of this model, as the tuning parameters (predictors) increases by one unit, the log odds will decrease by that unit. For example, if marginal adhesion (adhes) increases by -0.3984 and all others are set to zero than response variable Class will decrease by 0.3884 unit ( 11.03 - 0.3884). If predictors are set to zero, log odds are 11.033.

However, p-values of almost all coefficents are less than 0.05 except for predictors Mitos, NNucl and Ushap. So for predictors Mitos, NNucl and Ushap, we fail to reject null hypothesis as p > 0.05. So we will use step fuction to check the AIC and reduction of predictors if any required.

Based on the results of step function, predictors Epith and Usize are no longer in the model(mod2) and AIC is 105.66 compared to 109.46 for model with all the predictors. Thus, mod2 with the lowest AIC value can be considered because p-values of both these predictors are more than 0.05 indicating no significant effect on the response Class and fail to reject null hypothesis. MOdel 2 is the fit model for further analysis as per step function because removing other predictors do not give smaller AIC and simple model.

## Problem 2 :

At the same organization we are now evaluating the toxicity of a new chemical, so we want to see what exposure to the chemical does to animals. Using the aflatoxin data in the faraway package. Fit a logistic regression model for the number of animals with liver cancer as a function of the dose. Comment on the statistical significance of the model. Calculate the predicted probability of liver cancer for a dose of 25 ppb. Can you find a 95% confidence interval for this value. If so report it and give any insights you may have.

library(faraway)  
data("aflatoxin")  
head(aflatoxin)

## dose total tumor  
## 1 0 18 0  
## 2 1 22 2  
## 3 5 22 1  
## 4 15 21 4  
## 5 50 25 20  
## 6 100 28 28

summary(aflatoxin)

## dose total tumor   
## Min. : 0.00 Min. :18.00 Min. : 0.000   
## 1st Qu.: 2.00 1st Qu.:21.25 1st Qu.: 1.250   
## Median : 10.00 Median :22.00 Median : 3.000   
## Mean : 28.50 Mean :22.67 Mean : 9.167   
## 3rd Qu.: 41.25 3rd Qu.:24.25 3rd Qu.:16.000   
## Max. :100.00 Max. :28.00 Max. :28.000

which(aflatoxin$tumor!=0)

## [1] 2 3 4 5 6

aflatoxin$no\_tumor <- (aflatoxin$total - aflatoxin$tumor)  
aflatoxin$tumor

## [1] 0 2 1 4 20 28

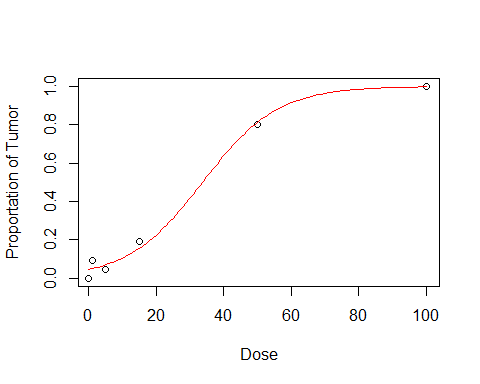
aflatoxin$total

## [1] 18 22 22 21 25 28

model1 <- glm(cbind(tumor,no\_tumor) ~ dose, family="binomial", data=aflatoxin)  
summary(model1)

##   
## Call:  
## glm(formula = cbind(tumor, no\_tumor) ~ dose, family = "binomial",   
## data = aflatoxin)  
##   
## Deviance Residuals:   
## 1 2 3 4 5 6   
## -1.2995 0.7959 -0.4814 0.4174 -0.1629 0.3774   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.03604 0.48226 -6.295 3.07e-10 \*\*\*  
## dose 0.09009 0.01456 6.189 6.04e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 116.524 on 5 degrees of freedom  
## Residual deviance: 2.897 on 4 degrees of freedom  
## AIC: 17.685  
##   
## Number of Fisher Scoring iterations: 5

plot(tumor/total ~ dose,aflatoxin,xlim=c(1,100),ylim=c(0,1), ylab="Proportation of Tumor", xlab="Dose")  
x<-seq(0,100,1)  
lines(x,ilogit(predict(model1,newdata=data.frame(dose=x))),col=2)



## Predict probablity for dose of 25 ppb   
ilogit(predict(model1, newdata=data.frame(dose=25)))

## 1   
## 0.3134978

confint(model1)

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) -4.10594724 -2.1868361  
## dose 0.06440816 0.1217981

### Interpretation :

As per the summary of Model1 with liver cancer as a function of the dose, animals that received dose developed tumors. The proportation that developed tumor increases with the dosage. The log-odds of a tumor increase by 0.09009 for every unit of dosage increase. p-value of co-efficent is <0.05 so we fail to reject null hypothesis.

The results of confindence intervals looks good as the main value of dose and intercept which are 0.09009 and -3.03604 respoective are beween CI range. The upper and lower CI remains well behaved.