PSTAT 127 HMWK 2

Kevin Ayala

1/31/2019

```
library(faraway)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#Question 1
#a
binreg <- glm(Class~., data = wbca, family = "binomial")
summary(binreg)
```

```
##
## 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
```

```
 # residual deviance is 89.464  on 671  degrees of freedom
```

For part B,

$$\phi = \frac{residualdeviance}{df}$$

```
#b
#from part a summary, residuals = 89.464 and degrees of freedom = 671
#pearson chisquare statistics residual is
estimate <- 89.464/671
estimate
```

```
## [1] 0.1333294
```

```
#our estimate is .1333294, which is rather poor compared to 1, thus model may need to be refined
#does not seem a plausable model


#c
AICselection <- step(binreg, direction = "backward")
```

```
## 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
```

```
AICselection
```

```
##
## Call:  glm(formula = Class ~ Adhes + BNucl + Chrom + Mitos + NNucl +
##     Thick + UShap, family = "binomial", data = wbca)
##
## Coefficients:
## (Intercept)        Adhes        BNucl        Chrom        Mitos
##     11.0333      -0.3984      -0.4192      -0.5679      -0.6456
##       NNucl        Thick        UShap
##     -0.2915      -0.6216      -0.2541
##
## Degrees of Freedom: 680 Total (i.e. Null);  673 Residual
## Null Deviance:       881.4
## Residual Deviance: 89.66     AIC: 105.7
```

```
# best model has a min AIC score of 105.66
#with predictors thick, BNucl, Chrom, Adhes, NNuc1,Mitos, UShap

#d
x<-matrix(data=NA, nrow = 1,ncol = 7)
x[]<-c(4,1,3,1,1,1,1)
x <- as.data.frame(x)
names(x)<- c("Thick", "BNucl", "Chrom", "Adhes", "NNucl", "Mitos", "UShap")

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

```
##
## Call:
## glm(formula = Class ~ Thick + BNucl + Chrom + Adhes + NNucl +
##     Mitos + 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 ***
## Thick        -0.6216     0.1579  -3.937 8.27e-05 ***
## BNucl        -0.4192     0.1020  -4.111 3.93e-05 ***
## Chrom        -0.5679     0.1840  -3.085  0.00203 **
## Adhes        -0.3984     0.1294  -3.080  0.00207 **
## NNucl        -0.2915     0.1236  -2.358  0.01837 *
## Mitos        -0.6456     0.3634  -1.777  0.07561 .
## 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
```

```
tumorAprob<-predict.glm(object = reducedmodel, newdata = x, type = "response")
tumorAlogodds <-predict.glm(object = reducedmodel, newdata = x, type = "link")

InfoA <- matrix(ncol = 2, nrow = 1, data = NA)
InfoA[1] <- tumorAprob
InfoA[2] <- tumorAlogodds
InfoA = as.data.frame(InfoA)
names(InfoA) = c("Probability","Log Odds")
InfoA
```

```
##   Probability Log Odds
## 1   0.9921115 4.834428
```

```
#info regarding probability and log odds for Tumor A being benign

#e
y<-matrix(data=NA, nrow = 1,ncol = 7)
y[]<-c(3,1,3,1,1,1,1)
y <- as.data.frame(y)
names(y)<- c("Thick", "BNucl", "Chrom", "Adhes", "NNucl", "Mitos", "UShap")
y
```

```
##   Thick BNucl Chrom Adhes NNucl Mitos UShap
## 1     3     1     3     1     1     1     1
```

```
tumorBprob<-predict.glm(object = reducedmodel, newdata = y, type ="response")
tumorBlogodds <- predict.glm(object = reducedmodel, newdata = y, type ="link")
InfoB <- matrix(ncol = 2, nrow = 1, data = NA)
InfoB[1] <- tumorBprob
InfoB[2] <- tumorBlogodds
InfoB = as.data.frame(InfoB)
names(InfoB) = c("Probability","Log Odds")
InfoB
```

```
##   Probability Log Odds
## 1   0.9957478 5.456056
```

```
InfoB-InfoA #differences
```

```
##   Probability  Log Odds
## 1 0.003636304 0.6216276
```

```
#tumor B is higher in log odds than tumor A by .6216

-.81489 - (.8529*1.96)
```

```
## [1] -2.486574
```

```r
#f
tumorA_errors<-predict.glm(object = reducedmodel, newdata = wbca, type = "response")
errors_tumorA <- ifelse(tumorA_errors < .5, 0 ,1)
tumorA_mislassified<-length(which(errors_tumorA != wbca$Class))
tumorA_mislassified #20 total subjects have been misclassified under the reduced model for tumor A
```

```
## [1] 20
```

```r
test<-cbind(True=wbca$Class, Predicted=errors_tumorA)
test <- as.data.frame(test)
errors<-filter(test, test$True != test$Predicted)
filter(errors, True == "1")
```

```
##   True Predicted
## 1    1         0
## 2    1         0
## 3    1         0
## 4    1         0
## 5    1         0
## 6    1         0
## 7    1         0
## 8    1         0
## 9    1         0
```

```r
# 9 cases of tumors that are benign have been misclassified
#20-9 = 11, thus 11 cases of malignant have been misclassified
```