Pstat 127 Homework 5

Marissa Santiago

```r
#install.packages('faraway')
library(faraway)
data("wbca")
```

```r
wbca <- within(wbca, Class <- factor(Class, labels=c("Malignant","Benign")))
head(wbca)
```

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

Part a)

```r
lm_fit <- glm(Class ~ .,family = binomial,data = wbca)
summary(lm_fit)
```

```
## 
## 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
```

→ The residual deviance is 89.5 with 671 degrees of freedom.

Part b)

```r
step_fit <- step(lm_fit,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
```

```
summary(step_fit)
```

```
##
## 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
```

→ When using AIC only 2 variables have been eliminated, Epith and USize. Most of the remaining variables are statistically significant but not all looking at the significance codes. AIC has been reduced from 109.46 to 105.66.

Part c)

```
newrow <- wbca[1,]
newrow[1,-1] <-  c(1, 1, 3, 2, 1, 1, 4, 1, 1)
newpred <- predict.glm(step_fit,newdata = newrow,type = "link",se.fit = TRUE)
newpred
```

```
## $fit
##        1
## 4.834428
##
## $se.fit
## [1] 0.5815185    (sd)
##
## $residual.scale
## [1] 1
```

```
newVals <- newpred$fit + 1.96*c(0,-1,1)*newpred$se.fit
newVals
```

```
## [1] 4.834428 3.694652 5.974204
```

```
probVals <- round(1/(1+exp(-newVals)),3)
probVals
```

```
## [1] 0.992 0.976 0.997
```

The estimated probability of there being a benign tumour for new patient given above is 0.992 with 95% confidence interval (0.976, 0.997).

Part d)

```
predBenign <- factor(predict.glm(step_fit,type="response") > .5, labels=c("pred_Mal","pred_Ben"))
xtab <- table(predBenign,wbca$Class)
knitr::kable(xtab)
```

	Malignant	Benign
pred_Mal	227	9
pred_Ben	11	434

Of the 443 patients with benign tumors, 9 or 2.0% of the patients recieve a false positive and were incorrectly classified to be malignant. This is considered a Type II error. Of the 238 patients with malignant tumours, 11 or 4.6% of the patients recieve a false negative and were incorrectly classified to be benign. This is considered a Type I error and must try to be fixed.