Homework5

Anshuman Narayan

11/14/2018

Q1

```
mod=glm(formula=Class~Adhes+BNucl+Chrom+Epith+Mitos+NNucl+Thick+UShap+USize,data=train,family=binomial)
summary(mod)
```

```
##
## Call:
## glm(formula = Class ~ Adhes + BNucl + Chrom + Epith + Mitos +
##     NNucl + Thick + UShap + USize, family = binomial, data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.60729  -0.01151   0.04030   0.09199   2.63572
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.31706    1.58508   7.140 9.35e-13 ***
## Adhes       -0.39454    0.13588  -2.904 0.003689 **
## BNucl       -0.36282    0.10591  -3.426 0.000613 ***
## Chrom       -0.60080    0.19718  -3.047 0.002312 **
## Epith       -0.10083    0.17022  -0.592 0.553608
## Mitos       -0.26079    0.48317  -0.540 0.589369
## NNucl       -0.26075    0.12682  -2.056 0.039771 *
## Thick       -0.73732    0.19449  -3.791 0.000150 ***
## UShap       -0.20434    0.26356  -0.775 0.438154
## USize        0.02325    0.24192   0.096 0.923437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 701.04  on 560  degrees of freedom
## Residual deviance:  76.86  on 551  degrees of freedom
## AIC: 96.86
##
## Number of Fisher Scoring iterations: 8
```

Q1a

The model assumes the following about the data, Let y_i be the observed sample proportion of the tumor being benign for the ith subject in one Bernoulli trial where the 9 explanatory variables had values $x_{i1}.....x_{i9}$. The model assumes that y_i is a realization of Y_i where $Y_i = Binom(1, exp(\beta_0 + \sum_{j=1}^{9} \beta_j x_{ij})/1 + exp(\beta_0 + \sum_{j=1}^{9} \beta_j x_{ij}))$ for i = $1 \ldots 561$. $(\beta_0, ...\beta_9)$ are unkown parameters, and $Y_1, Y_2.....Y_{561}$ are independent. We estimate and make inference for the 10 unkown parameters $\beta_0, \beta_1....\beta_9$, using maximum likelihood.

Q1b

We cannot use deviance on this fitted model, which has ungrouped subjects, to assess it's goodness of it. This is since deviance is treated as a random variable d that is approximately Chi-squared with n-(p+1) degrees of freedom that improves in this approximation as the number of trials for each subject increases. Since we have only 1 trial for each subject, we have a poor approximation for our variable d and we cannot use it to assess the goodness of fit.

Q1c

We have β_2=-0.36282. For a unit increase in BNucl, the estimated log odds of the tumor being benign decreases by 0.36282.

Q1d

The 99% Wald approximate confidence intervals for all parameters of the model are given below.

```
confint.default(mod,level=0.99)
```

```
##                  0.5 %      99.5 %
## (Intercept)  7.2341746 15.39994963
## Adhes       -0.7445389 -0.04453177
## BNucl       -0.6356322 -0.09000248
## Chrom       -1.1087049 -0.09289179
## Epith       -0.5393010  0.33763280
## Mitos       -1.5053524  0.98377100
## NNucl       -0.5874037  0.06590769
## Thick       -1.2382948 -0.23635472
## UShap       -0.8832138  0.47453620
## USize       -0.5999048  0.64640530
```

Q1e

We predict the probabilty of benign status using the predict function. Which returns to us the log-odds, the probability is then calculated using the ilogit function that we define.

```
pred=predict(mod,newdata=data.frame(Adhes=1,BNucl=1,Chrom=2,Epith=3,Mitos=1,NNucl=1,Thick=4,
                                    UShap=1,USize=1),se.fit=TRUE)
pred$fit
```

```
##        1
## 5.403684
```

```
pred$se.fit
```

```
## [1] 0.7220334
```

```
z.prec=qnorm(0.995)
LB=pred$fit - z.prec*pred$se.fit
UB=pred$fit + z.prec*pred$se.fit
ilogit=function(u) return ( exp(u)/(1+exp(u)) )
ilogit(cbind(LB,UB))
```

```
##        LB        UB
## 1 0.97191 0.9992999
```

Q1f

```
cat("The total subjects used to fit the model",nrow(train))
```

```
## The total subjects used to fit the model 561
```

```
cat("\nTotal subjects that were malignant",nrow(train[train$Class==0, ]))
```

```
##
## Total subjects that were malignant 178
```

Q1g

```
logpredt=predict(mod,newdata=train[ ,2:10],se.fit=TRUE)
predt=ilogit(logpredt$fit)
i=1
acc=0
for(p in predt)
{
  prediction =0
  if(p>0.5)
  {
    prediction=1
  }

  if(prediction == train[i,1])
  {
    acc=acc+1;

  }
  i=i+1

}
cat("\nThe accuracy of the model on the training set is",acc/nrow(train))
```

```
##
## The accuracy of the model on the training set is 0.9679144
```

Q1h

```
logpredt=predict(mod,newdata=test[ ,2:10],se.fit=TRUE)
predt=ilogit(logpredt$fit)
i=1
acc=0
for(p in predt)
{
  prediction =0
  if(p>0.5)
  {
    prediction=1
  }

  if(prediction == test[i,1])
  {
    acc=acc+1;
```

```
  }
  i=i+1

}
cat("\nThe accuracy of the model on the test set is",acc/nrow(test))
```

```
##
## The accuracy of the model on the test set is 0.9833333
```

Q2

Q2a

THe methodlogy involved in using the likelihoos ratio test "backward elimination" approach where we call drop1 function on the mod and looking at the largest p-value for LRT, we remove that parameter from the model and repeat till all parameters are significant to the model.

```
drop1(mod,test="Chi")
```

```
## Single term deletions
##
## Model:
## Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
##     UShap + USize
##        Df Deviance     AIC     LRT  Pr(>Chi)
## <none>      76.860  96.860
## Adhes   1   85.712 103.712  8.8522 0.0029274 **
## BNucl   1   90.250 108.250 13.3902 0.0002529 ***
## Chrom   1   87.849 105.849 10.9893 0.0009164 ***
## Epith   1   77.204  95.204  0.3443 0.5573567
## Mitos   1   77.187  95.187  0.3277 0.5670083
## NNucl   1   81.441  99.441  4.5814 0.0323213 *
## Thick   1   98.292 116.292 21.4324 3.665e-06 ***
## UShap   1   77.418  95.418  0.5587 0.4547890
## USize   1   76.869  94.869  0.0091 0.9238875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\nThe variable Usize has the larget p-value, we remove it from the model and re-test")
```

```
##
## The variable Usize has the larget p-value, we remove it from the model and re-test
```

```
modn<-update(mod,~.-USize)
drop1(modn,test="Chi")
```

```
## Single term deletions
##
## Model:
## Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
##     UShap
##        Df Deviance     AIC     LRT  Pr(>Chi)
## <none>      76.869  94.869
## Adhes   1   85.973 101.973  9.1041 0.0025503 **
## BNucl   1   90.297 106.297 13.4284 0.0002478 ***
## Chrom   1   88.007 104.007 11.1384 0.0008456 ***
```

```
## Epith   1   77.207  93.207  0.3383 0.5607914
## Mitos   1   77.207  93.207  0.3386 0.5606509
## NNucl   1   81.455  97.455  4.5862 0.0322295 *
## Thick   1   98.299 114.299 21.4304 3.669e-06 ***
## UShap   1   78.046  94.046  1.1774 0.2778858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\nThe variable Epith has the larget p-value, we remove it from the model and re-test")
```

```
##
## The variable Epith has the larget p-value, we remove it from the model and re-test
```

```
modn<-update(modn,~.-Epith)
drop1(modn,test="Chi")
```

```
## Single term deletions
##
## Model:
## Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap
##        Df Deviance     AIC     LRT  Pr(>Chi)
## <none>      77.207  93.207
## Adhes   1   86.899 100.899  9.6921 0.0018506 **
## BNucl   1   90.956 104.956 13.7485 0.0002090 ***
## Chrom   1   89.328 103.328 12.1211 0.0004985 ***
## Mitos   1   77.519  91.519  0.3118 0.5765994
## NNucl   1   82.073  96.073  4.8663 0.0273862 *
## Thick   1   98.511 112.511 21.3036  3.92e-06 ***
## UShap   1   78.808  92.808  1.6005 0.2058394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\nThe variable Mitos has the larget p-value, we remove it from the model and re-test")
```

```
##
## The variable Mitos has the larget p-value, we remove it from the model and re-test
```

```
modn<-update(modn,~.-Mitos)
drop1(modn,test="Chi")
```

```
## Single term deletions
##
## Model:
## Class ~ Adhes + BNucl + Chrom + NNucl + Thick + UShap
##        Df Deviance     AIC     LRT  Pr(>Chi)
## <none>      77.519  91.519
## Adhes   1   87.890  99.890 10.3710 0.0012801 **
## BNucl   1   91.521 103.521 14.0019 0.0001826 ***
## Chrom   1   89.690 101.690 12.1711 0.0004854 ***
## NNucl   1   82.195  94.195  4.6759 0.0305886 *
## Thick   1  102.443 114.443 24.9240 5.963e-07 ***
## UShap   1   79.402  91.402  1.8829 0.1700051
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\nThe variable UShap has the larget p-value, we remove it from the model and re-test")
```

```
##
```

```
## The variable UShap has the larget p-value, we remove it from the model and re-test
```

```
modn<-update(modn,~.-UShap)
drop1(modn,test="Chi")
```

```
## Single term deletions
##
## Model:
## Class ~ Adhes + BNucl + Chrom + NNucl + Thick
##        Df Deviance     AIC    LRT  Pr(>Chi)
## <none>      79.402  91.402
## Adhes   1   92.460 102.460 13.058  0.000302 ***
## BNucl   1  100.865 110.865 21.464 3.606e-06 ***
## Chrom   1   97.187 107.187 17.785 2.473e-05 ***
## NNucl   1   87.686  97.686  8.284  0.004000 **
## Thick   1  122.612 132.612 43.211 4.916e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\nThis model has no insignificant parameters and therefore we stop backwards elimination")
```

```
##
## This model has no insignificant parameters and therefore we stop backwards elimination
```

```
cat("\nThe fitted model using these paramters is given below")
```

```
##
## The fitted model using these paramters is given below
```

```
modfit=glm(formula=Class~Adhes+BNucl+Chrom+NNucl+Thick,data=train,family=binomial)
summary(modfit)
```

```
##
## Call:
## glm(formula = Class ~ Adhes + BNucl + Chrom + NNucl + Thick,
##     family = binomial, data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.42081  -0.01218   0.03524   0.09403   2.76344
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.43889    1.52247   7.513 5.76e-14 ***
## Adhes       -0.44085    0.13102  -3.365 0.000766 ***
## BNucl       -0.41554    0.09975  -4.166 3.10e-05 ***
## Chrom       -0.66629    0.18015  -3.699 0.000217 ***
## NNucl       -0.32294    0.11871  -2.721 0.006518 **
## Thick       -0.87696    0.17618  -4.978 6.44e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 701.037  on 560  degrees of freedom
## Residual deviance:  79.402  on 555  degrees of freedom
## AIC: 91.402
```

```
##
## Number of Fisher Scoring iterations: 8
```

Q2b

The 99% Wald confidence interval of the parameters of the updated model are the following

```
confint.default(modn,level=0.99)
```

```
##                   0.5 %      99.5 %
## (Intercept)  7.5172807 15.36050642
## Adhes       -0.7783378 -0.10336804
## BNucl       -0.6724843 -0.15859824
## Chrom       -1.1303112 -0.20226079
## NNucl       -0.6287025 -0.01717445
## Thick       -1.3307652 -0.42315069
```

Q2c

```
pred=predict(modn,newdata=data.frame(Adhes=1,BNucl=1,Chrom=2,Epith=3,Mitos=1,NNucl=1,Thick=4,
                                     UShap=1,USize=1),se.fit=TRUE)
pred$fit
```

```
##        1
## 5.419157
```

```
pred$se.fit
```

```
## [1] 0.6989492
```

```
z.prec=qnorm(0.995)
LB=pred$fit - z.prec*pred$se.fit
UB=pred$fit + z.prec*pred$se.fit
ilogit=function(u) return ( exp(u)/(1+exp(u)) )
ilogit(cbind(LB,UB))
```

```
##         LB        UB
## 1 0.973885 0.9992684
```

Q2d

```
logpredt=predict(modn,newdata=train[ ,2:10],se.fit=TRUE)
predt=ilogit(logpredt$fit)
i=1
acc=0
for(p in predt)
{
  prediction =0
  if(p>0.5)
  {
    prediction=1
  }

  if(prediction == train[i,1])
  {
    acc=acc+1;
```

```
  }
  i=i+1

}
cat("\nThe accuracy of the updated model on the training set is",acc/nrow(train))
```

```
##
## The accuracy of the updated model on the training set is 0.9661319
```

Q2e

```
logpredt=predict(modn,newdata=test[ ,2:10],se.fit=TRUE)
predt=ilogit(logpredt$fit)
i=1
acc=0
for(p in predt)
{
  prediction =0
  if(p>0.5)
  {
    prediction=1
  }

  if(prediction == test[i,1])
  {
    acc=acc+1;

  }
  i=i+1

}
cat("\nThe accuracy of the model on the test set is",acc/nrow(test))
```

```
##
## The accuracy of the model on the test set is 0.9833333
```

Q3

```
step(mod,direction="backward")
```

```
## Start:  AIC=96.86
## Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
##     UShap + USize
##
##         Df Deviance     AIC
## - USize  1   76.869  94.869
## - Mitos  1   77.187  95.187
## - Epith  1   77.204  95.204
## - UShap  1   77.418  95.418
## <none>       76.860  96.860
## - NNucl  1   81.441  99.441
## - Adhes  1   85.712 103.712
```

```
## - Chrom  1   87.849 105.849
## - BNucl  1   90.250 108.250
## - Thick  1   98.292 116.292
##
## Step:  AIC=94.87
## Class ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick +
##     UShap
##
##         Df Deviance     AIC
## - Epith  1   77.207  93.207
## - Mitos  1   77.207  93.207
## - UShap  1   78.046  94.046
## <none>       76.869  94.869
## - NNucl  1   81.455  97.455
## - Adhes  1   85.973 101.973
## - Chrom  1   88.007 104.007
## - BNucl  1   90.297 106.297
## - Thick  1   98.299 114.299
##
## Step:  AIC=93.21
## Class ~ Adhes + BNucl + Chrom + Mitos + NNucl + Thick + UShap
##
##         Df Deviance     AIC
## - Mitos  1   77.519  91.519
## - UShap  1   78.808  92.808
## <none>       77.207  93.207
## - NNucl  1   82.073  96.073
## - Adhes  1   86.899 100.899
## - Chrom  1   89.328 103.328
## - BNucl  1   90.956 104.956
## - Thick  1   98.511 112.511
##
## Step:  AIC=91.52
## Class ~ Adhes + BNucl + Chrom + NNucl + Thick + UShap
##
##         Df Deviance     AIC
## - UShap  1   79.402  91.402
## <none>       77.519  91.519
## - NNucl  1   82.195  94.195
## - Adhes  1   87.890  99.890
## - Chrom  1   89.690 101.690
## - BNucl  1   91.521 103.521
## - Thick  1  102.443 114.443
##
## Step:  AIC=91.4
## Class ~ Adhes + BNucl + Chrom + NNucl + Thick
##
##         Df Deviance     AIC
## <none>       79.402  91.402
## - NNucl  1   87.686  97.686
## - Adhes  1   92.460 102.460
## - Chrom  1   97.187 107.187
## - BNucl  1  100.865 110.865
## - Thick  1  122.612 132.612
```

```
##
## Call:  glm(formula = Class ~ Adhes + BNucl + Chrom + NNucl + Thick,
##     family = binomial, data = train)
##
## Coefficients:
## (Intercept)        Adhes        BNucl        Chrom        NNucl
##     11.4389      -0.4409      -0.4155      -0.6663      -0.3229
##       Thick
##     -0.8770
##
## Degrees of Freedom: 560 Total (i.e. Null);  555 Residual
## Null Deviance:       701
## Residual Deviance: 79.4  AIC: 91.4
```

The paramters that remain are the same as the ones from our likelihood ratio test "backward approach" We now fit this model.

```
modafit<-glm(formula = Class~Adhes+BNucl+Chrom+NNucl+Thick,data = train, family=binomial)
summary(modafit)
```

```
##
## Call:
## glm(formula = Class ~ Adhes + BNucl + Chrom + NNucl + Thick,
##     family = binomial, data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.42081  -0.01218   0.03524   0.09403   2.76344
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 11.43889    1.52247   7.513 5.76e-14 ***
## Adhes       -0.44085    0.13102  -3.365 0.000766 ***
## BNucl       -0.41554    0.09975  -4.166 3.10e-05 ***
## Chrom       -0.66629    0.18015  -3.699 0.000217 ***
## NNucl       -0.32294    0.11871  -2.721 0.006518 **
## Thick       -0.87696    0.17618  -4.978 6.44e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 701.037  on 560  degrees of freedom
## Residual deviance:  79.402  on 555  degrees of freedom
## AIC: 91.402
##
## Number of Fisher Scoring iterations: 8
```

Both the models fitted after removing paramters using the two methods are the same.