Assignment_10.1_HillZach

Zach Hill

May 13, 2019

A: Fitting a binary logistic regression model

```
# summary(data)

#glm.RA <- glm(Risk1Yr ~ AGE + DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 + PRE9 + PRE10 + PRE11 + PRE14 +

glm.RA <- glm(Risk1Yr ~ ., data = data, family = binomial)

glm.RA
```

```
##
## Call:  glm(formula = Risk1Yr ~ ., family = binomial, data = data)
##
## Coefficients:
## (Intercept)      DGNDGN2      DGNDGN4      DGNDGN6      DGNDGN5
##   26.039791    -0.555724    -0.427777    13.771698    -2.200769
##     DGNDGN8      DGNDGN1         PRE4         PRE5     PRE6PRZ1
##   -3.852310    14.180552     0.227245     0.030304     0.149014
##    PRE6PRZ0        PRE7F        PRE8F        PRE9F       PRE10F
##   -0.293701     0.715341     0.174337     1.368216     0.576958
##      PRE11F    PRE14OC14    PRE14OC12    PRE14OC13       PRE17F
##    0.516181    -1.652973    -0.439364    -1.179207     0.926593
##      PRE19F       PRE25F       PRE30F       PRE32F          AGE
##  -14.655378    -0.097894     1.083997   -13.983295     0.009506
##
## Degrees of Freedom: 469 Total (i.e. Null);  445 Residual
## Null Deviance:       395.6
## Residual Deviance: 341.2     AIC: 391.2
```

```
summary(glm.RA)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ ., family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4929   0.2762   0.4199   0.5439   1.6084
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.604e+01  2.333e+03   0.011 0.991093
## DGNDGN2     -5.557e-01  4.128e-01  -1.346 0.178199
## DGNDGN4     -4.278e-01  4.733e-01  -0.904 0.366122
## DGNDGN6      1.377e+01  1.178e+03   0.012 0.990671
## DGNDGN5     -2.201e+00  6.113e-01  -3.600 0.000318 ***
## DGNDGN8     -3.852e+00  1.550e+00  -2.485 0.012959 *
## DGNDGN1      1.418e+01  2.400e+03   0.006 0.995285
## PRE4         2.272e-01  1.849e-01   1.229 0.219094   
## PRE5         3.030e-02  1.786e-02   1.697 0.089715 . 
## PRE6PRZ1     1.490e-01  5.783e-01   0.258 0.796647   
## PRE6PRZ0    -2.937e-01  7.907e-01  -0.371 0.710303   
## PRE7F        7.153e-01  5.556e-01   1.288 0.197884   
## PRE8F        1.743e-01  3.892e-01   0.448 0.654188   
## PRE9F        1.368e+00  4.868e-01   2.811 0.004942 **
## PRE10F       5.770e-01  4.826e-01   1.196 0.231855   
## PRE11F       5.162e-01  3.965e-01   1.302 0.192948   
## PRE14OC14   -1.653e+00  6.094e-01  -2.713 0.006675 **
## PRE14OC12   -4.394e-01  3.301e-01  -1.331 0.183177   
## PRE14OC13   -1.179e+00  6.165e-01  -1.913 0.055799 . 
## PRE17F       9.266e-01  4.445e-01   2.085 0.037092 * 
## PRE19F      -1.466e+01  1.654e+03  -0.009 0.992928   
## PRE25F      -9.789e-02  1.003e+00  -0.098 0.922273   
## PRE30F       1.084e+00  4.990e-01   2.172 0.029840 * 
## PRE32F      -1.398e+01  1.645e+03  -0.008 0.993219   
## AGE          9.506e-03  1.810e-02   0.525 0.599442   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 395.61  on 469  degrees of freedom
## Residual deviance: 341.19  on 445  degrees of freedom
## AIC: 391.19
## 
## Number of Fisher Scoring iterations: 15
```

B: Most Valuable Variables

It appears that the diagnosis (and from the host site, that would be the classification given to the typee of cancer) had the largest effect, but this variable is not actually a cause of cancer. Of the variables which should be relevant to the likelihood of survival, PRE9 appears to hold the most significance based on its' standardized coefficients. This is the presence of dyspnoea. Original tumor size is also highly relevant, as is smoking.

C: Model Accuracy

```
stepAIC(glm.RA)
```

```
## Start:  AIC=391.19
## Risk1Yr ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 + PRE9 + PRE10 + 
##     PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 + PRE32 + AGE
## 
##          Df Deviance    AIC
## - PRE6    2   342.00 388.00
## - PRE25   1   341.20 389.20
## - PRE8    1   341.38 389.38
## - AGE     1   341.46 389.46
## - PRE32   1   341.49 389.49
## - PRE19   1   341.75 389.75
## - PRE10   1   342.67 390.67
## - PRE4    1   342.73 390.73
## - PRE7   1   342.76 390.76
## - PRE11  1   342.82 390.82
## <none>       341.19 391.19
## - PRE17  1   345.17 393.17
## - PRE5   1   345.22 393.22
## - PRE14  3   350.04 394.04
## - PRE30  1   346.88 394.88
## - PRE9   1   348.35 396.35
## - DGN    6   359.28 397.28
##
## Step:  AIC=388
## Risk1Yr ~ DGN + PRE4 + PRE5 + PRE7 + PRE8 + PRE9 + PRE10 + PRE11 +
##     PRE14 + PRE17 + PRE19 + PRE25 + PRE30 + PRE32 + AGE
##
##         Df Deviance    AIC
## - PRE25  1   342.03 386.03
## - AGE    1   342.18 386.18
## - PRE8   1   342.22 386.22
## - PRE32  1   342.27 386.27
## - PRE19  1   342.57 386.57
## - PRE10  1   342.79 386.79
## - PRE7   1   343.18 387.18
## - PRE4   1   343.36 387.36
## - PRE11  1   343.82 387.82
## <none>       342.00 388.00
## - PRE5   1   345.45 389.45
## - PRE17  1   345.82 389.82
## - PRE14  3   351.24 391.24
## - PRE30  1   347.46 391.46
## - PRE9   1   349.08 393.08
## - DGN    6   360.26 394.26
##
## Step:  AIC=386.03
## Risk1Yr ~ DGN + PRE4 + PRE5 + PRE7 + PRE8 + PRE9 + PRE10 + PRE11 +
##     PRE14 + PRE17 + PRE19 + PRE30 + PRE32 + AGE
##
##         Df Deviance    AIC
## - AGE    1   342.22 384.22
## - PRE8   1   342.24 384.24
## - PRE32  1   342.31 384.31
## - PRE19  1   342.60 384.60
## - PRE10  1   342.83 384.83
## - PRE7   1   343.24 385.24
## - PRE4   1   343.38 385.38
## - PRE11  1   343.85 385.85
## <none>       342.03 386.03
## - PRE5   1   345.45 387.45
## - PRE17  1   345.82 387.82
## - PRE14  3   351.33 389.33
## - PRE30  1   347.46 389.46
## - PRE9   1   349.11 391.11
## - DGN    6   360.33 392.33
##
## Step:  AIC=384.22
## Risk1Yr ~ DGN + PRE4 + PRE5 + PRE7 + PRE8 + PRE9 + PRE10 + PRE11 +
##     PRE14 + PRE17 + PRE19 + PRE30 + PRE32
##
##         Df Deviance    AIC
## - PRE8   1   342.43 382.43
## - PRE32  1   342.49 382.49
## - PRE19  1   342.77 382.77
## - PRE10  1   342.96 382.96
## - PRE4   1   343.39 383.39
## - PRE7   1   343.41 383.41
## - PRE11  1   343.88 383.88
## <none>       342.22 384.22
## - PRE5   1   345.53 385.53
## - PRE17  1   345.93 385.93
## - PRE14  3   351.58 387.58
## - PRE30  1   347.67 387.67
## - PRE9   1   349.14 389.14
## - DGN    6   360.39 390.39
##
## Step:  AIC=382.43
## Risk1Yr ~ DGN + PRE4 + PRE5 + PRE7 + PRE9 + PRE10 + PRE11 + PRE14 +
##     PRE17 + PRE19 + PRE30 + PRE32
##
##         Df Deviance    AIC
## - PRE32  1   342.71 380.71
## - PRE19  1   342.99 380.99
## - PRE10  1   343.23 381.23
## - PRE4   1   343.76 381.76
## - PRE7   1   343.97 381.97
## - PRE11  1   344.14 382.14
## <none>       342.43 382.43
## - PRE5   1   345.64 383.64
## - PRE17  1   346.09 384.09
## - PRE14  3   351.67 385.67
## - PRE30  1   347.83 385.83
## - PRE9   1   349.65 387.65
## - DGN    6   360.92 388.92
##
## Step:  AIC=380.71
## Risk1Yr ~ DGN + PRE4 + PRE5 + PRE7 + PRE9 + PRE10 + PRE11 + PRE14 +
##     PRE17 + PRE19 + PRE30
##
##         Df Deviance    AIC
## - PRE19  1   343.27 379.27
## - PRE10  1   343.55 379.55
## - PRE4   1   344.00 380.00
## - PRE7   1   344.27 380.27
## - PRE11  1   344.44 380.44
## <none>       342.71 380.71
## - PRE5   1   345.91 381.91
## - PRE17  1   346.41 382.41
## - PRE14  3   351.95 383.95
## - PRE30  1   348.18 384.18
## - PRE9   1   349.97 385.97
## - DGN    6   361.29 387.29
##
## Step:  AIC=379.27
## Risk1Yr ~ DGN + PRE4 + PRE5 + PRE7 + PRE9 + PRE10 + PRE11 + PRE14 +
##     PRE17 + PRE30
##
##         Df Deviance    AIC
## - PRE10  1   344.10 378.10
## - PRE4   1   344.55 378.55
## - PRE7   1   344.84 378.84
## - PRE11  1   344.91 378.91
## <none>       343.27 379.27
## - PRE5   1   346.48 380.48
## - PRE17  1   347.03 381.03
## - PRE14  3   352.55 382.55
## - PRE30  1   348.73 382.73
## - PRE9   1   350.58 384.58
## - DGN    6   361.97 385.97
##
## Step:  AIC=378.1
## Risk1Yr ~ DGN + PRE4 + PRE5 + PRE7 + PRE9 + PRE11 + PRE14 + PRE17 +
##     PRE30
##
##         Df Deviance    AIC
## - PRE4   1   345.34 377.34
## - PRE7   1   345.45 377.45
## <none>       344.10 378.10
## - PRE11  1   346.23 378.23
## - PRE5   1   347.38 379.38
## - PRE17  1   347.91 379.91
## - PRE14  3   354.48 382.48
## - PRE30  1   350.56 382.56
## - PRE9   1   351.71 383.71
## - DGN    6   362.31 384.31
##
## Step:  AIC=377.34
## Risk1Yr ~ DGN + PRE5 + PRE7 + PRE9 + PRE11 + PRE14 + PRE17 +
##     PRE30
##
##         Df Deviance    AIC
## - PRE7   1   346.61 376.61
## <none>       345.34 377.34
## - PRE11  1   347.70 377.70
## - PRE5   1   348.76 378.76
## - PRE17  1   349.78 379.78
## - PRE14  3   355.48 381.48
## - PRE30  1   351.81 381.81
## - PRE9   1   352.79 382.79
## - DGN    6   362.99 382.99
##
## Step:  AIC=376.61
## Risk1Yr ~ DGN + PRE5 + PRE9 + PRE11 + PRE14 + PRE17 + PRE30
##
##         Df Deviance    AIC
## <none>       346.61 376.61
## - PRE11  1   348.73 376.73
## - PRE5   1   349.39 377.39
## - PRE17  1   351.34 379.34
## - PRE30  1   352.69 380.69
## - PRE14  3   357.50 381.50
## - PRE9   1   354.07 382.07
## - DGN    6   364.16 382.16
```

```
##
## Call:  glm(formula = Risk1Yr ~ DGN + PRE5 + PRE9 + PRE11 + PRE14 + PRE17 +
##     PRE30, family = binomial, data = data)
##
## Coefficients:
## (Intercept)      DGNDGN2      DGNDGN4      DGNDGN6      DGNDGN5
##    -0.37123     -0.51022     -0.34251     13.02342     -2.15878
##     DGNDGN8      DGNDGN1         PRE5        PRE9F       PRE11F
##    -3.43514     13.47962      0.02428      1.35551      0.50303
##   PRE14OC14    PRE14OC12    PRE14OC13       PRE17F       PRE30F
##    -1.77128     -0.45340     -1.31605      0.98455      1.10136
##
## Degrees of Freedom: 469 Total (i.e. Null);  455 Residual
## Null Deviance:       395.6
## Residual Deviance: 346.6     AIC: 376.6
```

I used the Stepwise AIC function to find the best fit model. AIC is the Akaike information criterion, m etric assigned to each model relative to other models. The function uses a stepwise process to find the model with the best AIC.

```r
glm.FIT <- glm(formula = Risk1Yr ~ DGN + PRE5 + PRE9 + PRE11 + PRE14 + PRE17 + PRE30, family = binomial

summary(glm.FIT)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ DGN + PRE5 + PRE9 + PRE11 + PRE14 + PRE17 +
##     PRE30, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5340   0.2863   0.4617   0.5583   1.4667
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.37123    0.71516  -0.519 0.603706
## DGNDGN2      -0.51022    0.40310  -1.266 0.205610
## DGNDGN4      -0.34251    0.46053  -0.744 0.457030
## DGNDGN6      13.02342  719.21661   0.018 0.985553
## DGNDGN5      -2.15878    0.59442  -3.632 0.000282 ***
## DGNDGN8      -3.43514    1.51159  -2.273 0.023055 *
## DGNDGN1      13.47962 1455.39755   0.009 0.992610
## PRE5          0.02428    0.01731   1.403 0.160590
## PRE9F         1.35551    0.46854   2.893 0.003816 **
## PRE11F        0.50303    0.33762   1.490 0.136241
## PRE14OC14    -1.77128    0.59355  -2.984 0.002843 **
## PRE14OC12    -0.45340    0.32471  -1.396 0.162613
## PRE14OC13     -1.31605    0.60232  -2.185 0.028890 *
## PRE17F         0.98455    0.43089   2.285 0.022316 *
## PRE30F         1.10136    0.49490   2.225 0.026054 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.61  on 469  degrees of freedom
## Residual deviance: 346.61  on 455  degrees of freedom
## AIC: 376.61
##
## Number of Fisher Scoring iterations: 14
```

```r
train(Risk1Yr~.,data=data ,trControl = trainControl(method = "cv"), method = "svmRadial")
```

```
## Support Vector Machines with Radial Basis Function Kernel
##
## 470 samples
##  16 predictor
##   2 classes: 'T', 'F'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 423, 423, 423, 423, 423, 423, ...
## Resampling results across tuning parameters:
##
##   C     Accuracy   Kappa
##   0.25  0.8510638  0.000000000
##   0.50  0.8510638  0.000000000
##   1.00  0.8468085  0.008282129
##
## Tuning parameter 'sigma' was held constant at a value of 0.04906806
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were sigma = 0.04906806 and C = 0.25.
```

```r
# predict(glm.FIT, data, type = "response")
```

The model appears to offer around 85% accuracy.