ASSIGNMENT 7.1

Blaine Blasdell

2020-07-19

Assignment 7.1

Question 7.1a

a. Fit a binary logistic regression model to the data set that predicts whether or not the patient survived for one year (the Risk1Y variable) after the surgery. Use the glm() function to perform the logistic regression. See Generalized Linear Models for an example. Include a summary using the summary() function in your results.

Answer - See output.

```
##
## Call:
## glm(formula = thor_df$Risk1Y ~ DGN, family = binomial(), data = thor_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1774  -0.5128  -0.5128  -0.5128   2.0464
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.557e+01  1.455e+03  -0.011    0.991
## DGNDGN2      1.436e+01  1.455e+03   0.010    0.992
## DGNDGN3      1.360e+01  1.455e+03   0.009    0.993
## DGNDGN4      1.382e+01  1.455e+03   0.009    0.992
## DGNDGN5      1.543e+01  1.455e+03   0.011    0.992
## DGNDGN6      3.159e-08  1.627e+03   0.000    1.000
## DGNDGN8      1.557e+01  1.455e+03   0.011    0.991
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.61  on 469  degrees of freedom
## Residual deviance: 379.79  on 463  degrees of freedom
## AIC: 393.79
##
## Number of Fisher Scoring iterations: 14
```

```
##
## Call:
## glm(formula = thor_df$Risk1Y ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 +
##     PRE8 + PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 +
##     PRE30 + PRE32 + AGE, family = binomial(), data = thor_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6084  -0.5439  -0.4199  -0.2762   2.4929
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.655e+01  2.400e+03  -0.007  0.99450
## DGNDGN2      1.474e+01  2.400e+03   0.006  0.99510
## DGNDGN3      1.418e+01  2.400e+03   0.006  0.99528
## DGNDGN4      1.461e+01  2.400e+03   0.006  0.99514
## DGNDGN5      1.638e+01  2.400e+03   0.007  0.99455
## DGNDGN6      4.089e-01  2.673e+03   0.000  0.99988
## DGNDGN8      1.803e+01  2.400e+03   0.008  0.99400
## PRE4        -2.272e-01  1.849e-01  -1.229  0.21909
## PRE5        -3.030e-02  1.786e-02  -1.697  0.08971 .
## PRE6PRZ1    -4.427e-01  5.199e-01  -0.852  0.39448
## PRE6PRZ2    -2.937e-01  7.907e-01  -0.371  0.71030
## PRE7T        7.153e-01  5.556e-01   1.288  0.19788
## PRE8T        1.743e-01  3.892e-01   0.448  0.65419
## PRE9T        1.368e+00  4.868e-01   2.811  0.00494 **
## PRE10T       5.770e-01  4.826e-01   1.196  0.23185
## PRE11T       5.162e-01  3.965e-01   1.302  0.19295
## PRE14OC12    4.394e-01  3.301e-01   1.331  0.18318
## PRE14OC13    1.179e+00  6.165e-01   1.913  0.05580 .
## PRE14OC14    1.653e+00  6.094e-01   2.713  0.00668 **
## PRE17T       9.266e-01  4.445e-01   2.085  0.03709 *
## PRE19T      -1.466e+01  1.654e+03  -0.009  0.99293
## PRE25T      -9.789e-02  1.003e+00  -0.098  0.92227
## PRE30T       1.084e+00  4.990e-01   2.172  0.02984 *
## PRE32T      -1.398e+01  1.645e+03  -0.008  0.99322
## AGE         -9.506e-03  1.810e-02  -0.525  0.59944
## ---
## 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


##
## Call:
## glm(formula = thor_df$Risk1Y ~ PRE4 + PRE5 + PRE6 + PRE7 + PRE8 +
##     PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE30, family = binomial(),
##     data = thor_df)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7038  -0.5550  -0.4585  -0.3410   2.6126
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.70657    0.75776  -3.572 0.000355 ***
## PRE4        -0.13926    0.17045  -0.817 0.413909
## PRE5        -0.02224    0.01697  -1.311 0.189958
## PRE6PRZ1    -0.44887    0.50751  -0.884 0.376451
## PRE6PRZ2    -0.29050    0.75465  -0.385 0.700279
## PRE7T        0.61027    0.53182   1.148 0.251168
## PRE8T        0.28484    0.36849   0.773 0.439520
## PRE9T        1.20136    0.47398   2.535 0.011258 *
## PRE10T       0.46073    0.47195   0.976 0.328951
## PRE11T       0.51119    0.38198   1.338 0.180809
## PRE14OC12    0.42699    0.31784   1.343 0.179138
## PRE14OC13    1.28126    0.59121   2.167 0.030221 *
## PRE14OC14    1.69374    0.57811   2.930 0.003392 **
## PRE17T       0.95634    0.42852   2.232 0.025632 *
## PRE30T       0.80779    0.44757   1.805 0.071100 .
## ---
## 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: 360.51  on 455  degrees of freedom
## AIC: 390.51
##
## Number of Fisher Scoring iterations: 5
```

Question 7.1b

b. According to the summary, which variables had the greatest effect on the survival rate?

Answer - PRE9 (Dyspnoea before surgery) with a value of True, PRE14 (size of the original tumor) with a diagnosis of OC14 (the largest), PRE17 (type 2 diabetes) with a diagnosis of True, and PRE30 (Smoking) with a value of true.

Question 7.1c

c. To compute the accuracy of your model, use the dataset to predict the outcome variable. The percent of correct predictions is the accuracy of your model. What is the accuracy of your model?

Answer - The accuracy of my initial model with all variables is 83.6 %. By playing with the model I took several less significant variables, and surprisingly including DGN and Age and I got my model to be 84.3% accurate.

```
##              Predicted_Value
## Actual_Value FALSE
##            F   400
##            T    70
```

```
##              Predicted_Value
## Actual_Value FALSE TRUE
##            F   390   10
##            T    67    3
```

```
##              Predicted_Value
## Actual_Value FALSE TRUE
##            F   395    5
##            T    69    1
```

```
## [1] 83.61702
```

```
## [1] 84.25532
```

References