Week 10

Stewart Wilson

2022-05-18

Thoracic Surgey Binary Dataset

i. Fitting Logistic Regression Model

```
##
## Call:
## glm(formula = Risk1Yr ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 +
##     PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 +
##     PRE32 + AGE, family = binomial, data = thoracic)
##
## 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
```

ii. Best Predictors

We should narrow down the predictors to those that likely have a statistically significant effect on the outcome variable. That is we should observe the variables with a p-value $< .05$. Those are:

- PRE9T (Dyspnoea before surgery)
- PRE14OC14 (largest size of the original tumor)
- PRE17T (Type 2 diabetes)
- PRE30T (Smoking)

We can also convert the coefficients by taking the inverse logit to better understand the coefficients.

```
##  (Intercept)      DGNDGN2      DGNDGN3      DGNDGN4      DGNDGN5      DGNDGN6
## 6.481697e-08 9.999996e-01 9.999993e-01 9.999995e-01 9.999999e-01 6.008129e-01
##      DGNDGN8         PRE4         PRE5     PRE6PRZ1     PRE6PRZ2        PRE7T
## 1.000000e+00 4.434320e-01 4.924247e-01 3.910942e-01 4.270981e-01 6.715802e-01
##        PRE8T        PRE9T       PRE10T       PRE11T    PRE14OC12    PRE14OC13
## 5.434741e-01 7.970918e-01 6.403671e-01 6.262543e-01 6.081074e-01 7.648053e-01
##    PRE14OC14       PRE17T       PRE19T       PRE25T       PRE30T       PRE32T
## 8.392924e-01 7.163837e-01 4.317674e-07 4.755459e-01 7.472496e-01 8.455357e-07
##          AGE
## 4.976236e-01
```

Some variables have higher coefficients than the ones listed above and thus could be interpreted to have a greater effect on the survival rate. However, the z-statistic, which tells us how far the b coefficient is from 0 is very small for those with an otherwise high coefficient.

As such, we can conclude that the variables with the greatest effect on the survival rate (and in this case having these conditions lowers your survival rate) are

1. PRE9T (Dyspnoea before surgery) b = .64, z= 2.81, $p < .005$
2. PRE14OC14 (largest size of the original tumor) b = .84, z = 2.71 $p < .007$
3. PRE30T (Smoking) b = .75, z = 2.17, $p < .03$
4. PRE17T (Type 2 diabetes) b = .72, z = 2.09, $p < .04$

in roughly that order.

iii. Computing Accuracy

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
## [13]  TRUE FALSE  TRUE FALSE  TRUE
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
##             Predicted_value
## Actual_Value FALSE TRUE
##            F   307    6
##            T    43    5
```

```
## [1] 0.8642659
```

Binary Classifier Dataset

a. Fit a Logistic Regression Model

```
##
## Call:
## glm(formula = label ~ x + y, family = binomial, data = binary_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3728  -1.1697  -0.9575   1.1646   1.3989
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.424809   0.117224   3.624  0.00029 ***
## x           -0.002571   0.001823  -1.411  0.15836
## y           -0.007956   0.001869  -4.257 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2075.8  on 1497  degrees of freedom
## Residual deviance: 2052.1  on 1495  degrees of freedom
## AIC: 2058.1
##
## Number of Fisher Scoring iterations: 4
```

b. Accuracy

```
## [1]  TRUE  TRUE FALSE
```

```
##             Predicted_Value
## Actual_Value FALSE TRUE
##            0   429  338
##            1   286  445
```

```
## [1] 0.5834446
```

Accuracy is 58.34%