Assignment10-.2

Joshua Burden

2022-05-22

Fit a Logistic Regression Model to Thoracic Surgery Binary Dataset

- 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.

```
TS_df$DGN <- as.factor(TS_df$DGN)
TS_df$PRE14 <- as.factor(TS_df$PRE14)
thoracicSurgery <- glm(Risk1Yr ~  DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 + PRE9 + PRE10 + PRE11 + PRE14
summary(thoracicSurgery)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 +
##     PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 +
##     PRE32 + AGE, family = binomial(), data = TS_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
## PRE7TRUE     7.153e-01  5.556e-01   1.288  0.19788
## PRE8TRUE     1.743e-01  3.892e-01   0.448  0.65419
## PRE9TRUE     1.368e+00  4.868e-01   2.811  0.00494 **
## PRE10TRUE    5.770e-01  4.826e-01   1.196  0.23185
## PRE11TRUE    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 **
## PRE17TRUE    9.266e-01  4.445e-01   2.085  0.03709 *
## PRE19TRUE   -1.466e+01  1.654e+03  -0.009  0.99293
## PRE25TRUE   -9.789e-02  1.003e+00  -0.098  0.92227
## PRE30TRUE    1.084e+00  4.990e-01   2.172  0.02984 *
## PRE32TRUE   -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
```

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

PRE9, PRE14OC14 has the greatest effect on the survival rate.

- 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?

```
predict_data <- predict(thoracicSurgery,TS_df,type="response") > .5
totalCorrect <- sum( (TS_df$Risk1Yr == "T") == (predict_data) )
totalPercent <- round(sum( (TS_df$Risk1Yr == "T") == (predict_data) ) / nrow(TS_df),2)*100
totalRows <- nrow(TS_df)
totalWrong <- totalRows - totalCorrect
print(paste0("The model predicted ",totalCorrect," successful outcomes. Out of ",totalRows," elements, 
```

```
## [1] "The model predicted 457 successful outcomes. Out of 470 elements, 13 elements were found incorr
```

```
print(paste0("Accruacy score: ",round(totalPercent,digits = 2),"%"))
```

```
## [1] "Accruacy score: 97%"
```

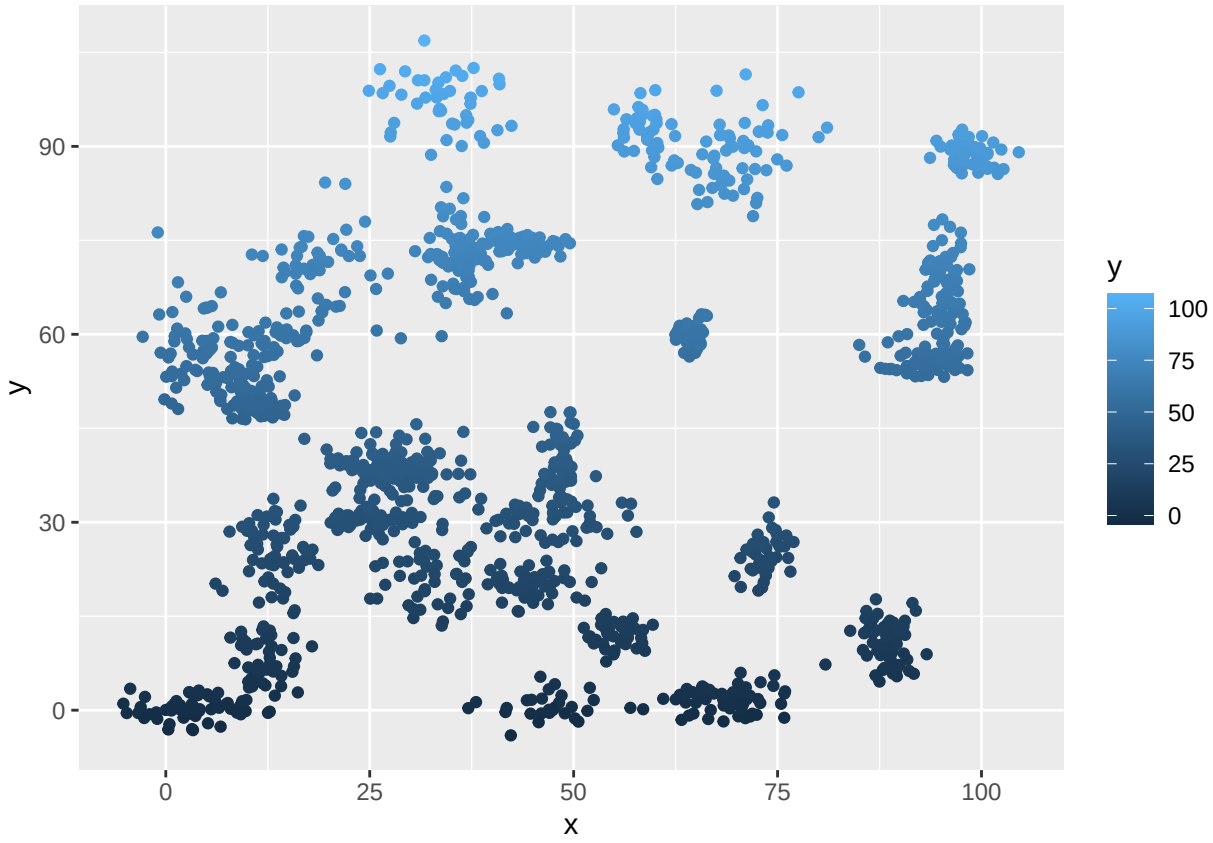
Fit a Logistic Regression Model

- Fit a logistic regression model to the binary-classifier-data.csv dataset
- The dataset (found in binary-classifier-data.csv) contains three variables; label, x, and y.
- The label variable is either 0 or 1 and is the output we want to predict using the x and y variables.

```
setwd("/Users/joshua/Documents/PERSONAL_GITHUB_REPOS/dsc520")
bc_df <- read_csv("data/binary-classifier-data.csv")
```

```
## Rows: 1498 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): label, x, y
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ggplot(bc_df, mapping = aes(x = x, y = y)) + geom_point(mapping = aes(colour = y))
```



What is the accuracy of the logistic regression classifier?

```
model1_xy <- glm(label ~ x + y, data = bc_df, family = "binomial")
model1_x  <- glm(label ~ x, data = bc_df, family = "binomial")

summary(model1_xy)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = "binomial", data = bc_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
```

```
summary(model1_x)
```

```
##
## Call:
## glm(formula = label ~ x, family = "binomial", data = bc_df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.246  -1.159  -1.065   1.184   1.293
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.137369   0.095119   1.444   0.1487
## x           -0.004119   0.001775  -2.321   0.0203 *
## ---
## 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: 2070.4  on 1496  degrees of freedom
## AIC: 2074.4
##
## Number of Fisher Scoring iterations: 3
```

The data model at the moment produces a very inaccurate summary.

Keep this assignment handy, as you will be comparing your results from this week to next week.