Activity 2.10

Facundo Bautista Barbera

Activity 2.10: Exploring multiple Logistic Regression in Regression

Setup

Libraries

```r
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Load dataset and transform data

```r
# Load the dataset
data(mtcars)

# Convert 'am' to factor (0 = auto, 1 = manual)
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
```

Inspect dataset

```r
# Show first rows
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs        am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0    Manual    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0    Manual    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1    Manual    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1 Automatic    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0 Automatic    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1 Automatic    3    1
```

```r
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##          am          gear            carb      
##  Automatic:19   Min.   :3.000   Min.   :1.000  
##  Manual   :13   1st Qu.:3.000   1st Qu.:2.000  
##                 Median :4.000   Median :2.000  
##                 Mean   :3.688   Mean   :2.812  
##                 3rd Qu.:4.000   3rd Qu.:4.000  
##                 Max.   :5.000   Max.   :8.000
```

```r
# How many cars are automatic vs manual?
table(mtcars$am)
```

```
## 
## Automatic    Manual 
##        19        13
```

Fit a Multiple Logistic Regression

```r
model1 <- glm(am ~ mpg + hp + wt, data = mtcars, family = "binomial")

# With the summary we can see what the predictors are
summary(model1)
```

```
## 
## Call:
## glm(formula = am ~ mpg + hp + wt, family = "binomial", data = mtcars)
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -15.72137   40.00281  -0.393   0.6943  
## mpg           1.22930    1.58109   0.778   0.4369  
## hp            0.08389    0.08228   1.020   0.3079  
## wt           -6.95492    3.35297  -2.074   0.0381 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.2297  on 31  degrees of freedom
## Residual deviance:  8.7661  on 28  degrees of freedom
## AIC: 16.766
##
## Number of Fisher Scoring iterations: 10
```

The model has a strong Intercept, with a p-value of 0.6943, while the other variables have a p-value lower than 0.5, which reduces the significance of them. The variable `mpg` has a coefficient value of: 1.22930.

Model Evaluation

How much deviance is reduced from the null model?

```r
null_dev <- model1$null.deviance
residual <- model1$deviance

dev_reduction <- null_dev - residual
print(dev_reduction)
```

```
## [1] 34.46362
```

```r
AIC(model1)
```

```
## [1] 16.76611
```

The AIC generally is a good way to measure the difference between models, in this case the value of 16.76611 is slightly high, so we want to compare it against another model.

Predict probabilities

```r
probabilities <- predict(model1, type = "response")
predictions <- ifelse(probabilities > 0.5, 1, 0)
predictions <- factor(predictions, levels = c(0, 1), labels = c("Automatic", "Manual"))
head(predictions)
```

```
##         Mazda RX4     Mazda RX4 Wag        Datsun 710    Hornet 4 Drive
##            Manual         Automatic            Manual         Automatic
## Hornet Sportabout           Valiant
##         Automatic         Automatic
## Levels: Automatic Manual
```

```r
table(Predicted = predictions, Actual = mtcars$am)
```

```
##            Actual
## Predicted   Automatic Manual
##   Automatic        18      1
##   Manual            1     12
```

Almost all of the predictions where good.

Add more predictors

```r
model2 <- glm(am ~ mpg + hp + wt + disp, data = mtcars, family = "binomial")
summary(model2)
```

```
##
## Call:
## glm(formula = am ~ mpg + hp + wt + disp, family = "binomial",
##     data = mtcars)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -18.48207   40.90451  -0.452    0.651
## mpg           1.13503    1.55720   0.729    0.466
## hp            0.10871    0.09837   1.105    0.269
## wt           -4.80560    3.97978  -1.208    0.227
## disp         -0.02588    0.04087  -0.633    0.527
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.230  on 31  degrees of freedom
## Residual deviance:  8.162  on 27  degrees of freedom
## AIC: 18.162
##
## Number of Fisher Scoring iterations: 9
```

```r
AIC(model2)
```

```
## [1] 18.16197
```

For this model, the AIC value is higher, so we determine that this model is comparably worst than the first model.