GLM Model in mtcars dataset

Niraj

2023-03-10

# GLM model is a stastical model that extends the linear regression model to handle non-normality distributed data, including binary, count and categorical response variable.

data(mtcars)

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 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 8 360 175 3.15 3.440 17.02 0 0 3 2  
## Valiant 18.1 6 225 105 2.76 3.460 20.22 1 0 3 1

# Summary of the dataset

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   
## Min. :0.0000 Min. :3.000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000   
## Median :0.0000 Median :4.000 Median :2.000   
## Mean :0.4062 Mean :3.688 Mean :2.812   
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :1.0000 Max. :5.000 Max. :8.000

## The variable vs indicates the type of engine a car has, 0 for V-shaped, and 1 for straight. This will be our response variable. We want to create a model that helps us to predict the probability of a vehicle having a V-shaped engine or a straight engine. In order to do this, we’ll use two predictors, wt (Weight — 1000 lbs) and disp (Displacement — cubic inches)

create the model

glm1 <- glm(formula= vs ~ wt + disp, data=mtcars, family=binomial)

# Summary of the model

summary(glm1)

##   
## Call:  
## glm(formula = vs ~ wt + disp, family = binomial, data = mtcars)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.67506 -0.28444 -0.08401 0.57281 2.08234   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.60859 2.43903 0.660 0.510   
## wt 1.62635 1.49068 1.091 0.275   
## disp -0.03443 0.01536 -2.241 0.025 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 43.86 on 31 degrees of freedom  
## Residual deviance: 21.40 on 29 degrees of freedom  
## AIC: 27.4  
##   
## Number of Fisher Scoring iterations: 6

# Quality of Fit Metrics

## ways to determine the quality of fit for a GLM .

## Deviance — Null deviance and Residual deviance.

## Deviance is a quality of fit measurement for a GLM where larger values indicate a poorer fit.

## The Null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean of all the groups). For our example, we have a value of 43.9 on 31 degrees of freedom. Subsequently including the independent variables, wt and disp, serve to decrease the deviance to 21.4 on 29 degrees of freedom, a significant reduction in deviance. Similarly, the Residual deviance has reduced by 22.46 with a loss of two degrees of freedom.

## Fisher Scoring — The line in the GLM output shows how many iterations it took to complete the process (convergence). For glm1 we see that Fisher Scoring needed six iterations to perform the fit.

## The Akaike Information Criterion (AIC) — It’s based on the Deviance metric, but penalizes you for making the model more complicated. Much like adjusted R-squared, So it’s useful for comparing models, but isn’t interpretable on its own.

## Hosmer-Lemeshow — The glm1 model appears to fit well because there is no significant difference between the model and the observed data (i.e. the p-value is above 0.05). As with other model fit metrics, we can use this test as just one piece of information in deciding how well the model fits. It doesn’t work well in very large or very small data sets but is often useful nevertheless.

## predict vs given a weight of 3,000 lbs. and engine displacement of 150 cubic inches

newdata = data.frame(wt = 3.0, disp = 150)

predict(glm1, newdata, type="response")

## 1   
## 0.7896128

# The predicted probability is ~ 79% i.e. engine will be straight shape.

# The End