REGRESSION MODELS PROJECT

A.M

Monday, November 23, 2015

Using this dataset mtcars of a collection of cars:  
1.Is an automatic or manual transmission better for MPG?  
2.Quantify the MPG difference between automatic and manual transmissions.

## Executive Summary

Using this dataset of 32 cars with 11 variables and forward selection regression modelling, manual transmission is better for fuel efficiency. The final model uses 4 features for prediction: transmission, weight, interaction between transmission and weight, and 1/4 mile time.

Transmission type alone accounts for about 34% variability in fuel efficiency. The number of cylinders which has a good correlation with the miles consumed per gallon(-0.852162) and other features in the dataset do not improve the predictive power of the model. The elimination of few outliers in the final model increases the predictive power by 2%. Any further outlier deletion does not improve the prediction, and not advisable because of the size of the sample.

The MPG difference between manual and automatic transmission is **12.5** with 95% confidence interval between **5.56-19.57**

## Forward model selection

##   
## Call:  
## lm(formula = mpg ~ am, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3923 -3.0923 -0.2974 3.2439 9.5077   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.147 1.125 15.247 1.13e-15 \*\*\*  
## am 7.245 1.764 4.106 0.000285 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.902 on 30 degrees of freedom  
## Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385   
## F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

##   
## Call:  
## lm(formula = mpg ~ am \* wt, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.6004 -1.5446 -0.5325 0.9012 6.0909   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.4161 3.0201 10.402 4.00e-11 \*\*\*  
## am 14.8784 4.2640 3.489 0.00162 \*\*   
## wt -3.7859 0.7856 -4.819 4.55e-05 \*\*\*  
## am:wt -5.2984 1.4447 -3.667 0.00102 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.591 on 28 degrees of freedom  
## Multiple R-squared: 0.833, Adjusted R-squared: 0.8151   
## F-statistic: 46.57 on 3 and 28 DF, p-value: 5.209e-11

##   
## Call:  
## lm(formula = mpg ~ am \* wt + qsec, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5076 -1.3801 -0.5588 1.0630 4.3684   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.723 5.899 1.648 0.110893   
## am 14.079 3.435 4.099 0.000341 \*\*\*  
## wt -2.937 0.666 -4.409 0.000149 \*\*\*  
## qsec 1.017 0.252 4.035 0.000403 \*\*\*  
## am:wt -4.141 1.197 -3.460 0.001809 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.084 on 27 degrees of freedom  
## Multiple R-squared: 0.8959, Adjusted R-squared: 0.8804   
## F-statistic: 58.06 on 4 and 27 DF, p-value: 7.168e-13

Even though there is a high correlation between the number of cylinders and miles per gallon, it is not statistically significant when added to model 3.Also, addition of other variables in the dataset does not improve the adjusted r squared beyond that of model3.

## Diagnostics I

## Analysis of Variance Table  
##   
## Model 1: mpg ~ am  
## Model 2: mpg ~ am \* wt  
## Model 3: mpg ~ am \* wt + qsec  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 30 720.90   
## 2 28 188.01 2 532.89 61.342 9.089e-11 \*\*\*  
## 3 27 117.28 1 70.73 16.284 0.000403 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The addition of other variables(weight, weight and transmission interaction, and 1/4 mile time) to mtcars$am significantly improves the model.

The errors do not show any significant departure from normality, and there is no systematic variation in the residuals.

tail(sort(abs(round(dfbetas(model3)[,2],3))),5)

## Merc 240D Toyota Corolla Cadillac Fleetwood   
## 0.342 0.377 0.396   
## Maserati Bora Chrysler Imperial   
## 0.398 0.608

tail(sort(abs(round(hatvalues(model3),3))),5)

## Chrysler Imperial Lincoln Continental Lotus Europa   
## 0.285 0.320 0.326   
## Merc 230 Maserati Bora   
## 0.346 0.374

tail(sort(abs(round(dffits(model3),3))),5)

## Toyota Corolla Cadillac Fleetwood Maserati Bora   
## 0.720 0.756 0.758   
## Fiat 128 Chrysler Imperial   
## 0.939 1.100

Both Chrysler Imperial and Maserati Bora have a high leverage and influence on the regression model. Fiat and Chrysler Imperial exert the highest influnece on the predicted response

\*Remove the three cars from the dataset and run model3

mtcarsNew<-filter(mtcars, !mtcars$carname %in% c("Chrysler Imperial","Maserati Bora","Fiat 128"))

Run model4 with mtcarsNew

##   
## Call:  
## lm(formula = mpg ~ am \* wt + qsec, data = mtcarsNew)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7972 -1.1895 -0.3781 0.9665 3.7630   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.8451 5.4140 2.742 0.01136 \*   
## am 12.5648 3.3922 3.704 0.00111 \*\*   
## wt -3.6672 0.6630 -5.531 1.09e-05 \*\*\*  
## qsec 0.8745 0.2246 3.893 0.00069 \*\*\*  
## am:wt -4.1291 1.2145 -3.400 0.00236 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.805 on 24 degrees of freedom  
## Multiple R-squared: 0.915, Adjusted R-squared: 0.9008   
## F-statistic: 64.59 on 4 and 24 DF, p-value: 1.704e-12

Adjusted R squared increases by 2%

## Diagnostics II

The errors do not show any significant departure from normality and there is no systematic variation in the residuals. Infact, they align better on the line in the QQ plot

## 24 15 8 19 18   
## 0.208 0.278 0.368 0.420 0.641

## 27 9 26 15 16   
## 0.355 0.359 0.360 0.377 0.447

## 3 23 19 8 18   
## 0.609 0.626 0.651 0.859 1.184

Toyota Corolla seems to affect the predicted response by more than 1 point, but its removal does not affect the model's prediction positively.

Using the model4, the estimated difference in Miles Per Gallon between Automatic and Manual Transmission is **12.56** (**5.56-19.57**)

## **Apendix**









