Regression Model - Automobile Design and Performance

Shalini Ruppa Subramanian

2015-09-27

Executive Summary

We will be analysing the **mtcars** dataset, extracted from the 1974 Motor Trend US magazine, and comprises of aspects of automobile design and performance for 32 automobiles. Our objective is to see if automatic or manual transmission yiels more miles per gallon (MPG) and quantify the difference between automatic and manual transmission type.

Exploratory Data Analysis

We are changing the numeric variables (vs,am,gear and carb) to factor variables in `mtcars` dataset so that it can be represented well in the linear regression model.

A pair plot, **Figure 1** shown in **Appendix**, is constructed to get an overall picture of the variables affecting mpg. The variables *'cyl'*,*'disp'*,*'hp'*,*'wt'*,*'vs'* and *'am'* shows higher correlations to *'mpg'* variable. A boxplot, **Figure 2** shown in **Appendix**, shows that manual transmission gives a higher mileage than automatic transmission.

Statistical Inference

A two sample t-test between '*mpg*' and '*am*' shows that the difference in means between automatic and manual transmission does not equal to zero and it is statistically significant at a p-value of 0.001374. This means that the automatic and manual transmissions are from different populatios. The mean MPG for automatic cars is about 7 MPG higher than that of manually transmitted cars. The results from the model is shown in the **Appendix**.

```
result <- t.test(mpg~am,mtcars)
```

Regression Analysis

We start with a model with '*mpg*' vs all other variables. We use the `step` function in R to select a subset of predictor variables in both directions from the large `mtcars` dataset.

```
fitall <- lm(mpg~.,mtcars)
stepFit <- step(fitall, direction = "both")
summary(stepFit)
```

The summary from the model derived using the `step` function shows the adjusted R-squared value is 0.83 and 83% of the variability in MPG is explained by this model. Now, we proceed to do a comparison of all the possible models.

```r
amFit <- lm(mpg~am,mtcars) #model that addresses the objective of comparing mpg and am
stepFit <- lm(mpg~wt+qsec+am,mtcars) #model derived using the step function
pairFit <- lm(mpg~disp+cyl+hp+wt+vs+am,mtcars) #confounder variables inferred from the pair plot
pair2Fit <- lm(mpg~am+hp+wt,mtcars)#confounder variables revised looking at the pairFit summary based o
anova(amFit,stepFit,pair2Fit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ am + hp + wt
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## 3     28 180.29  0    -11.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the Analysis of Variance results, we select the *stepFit* model to be appropriate. The p-value obtained for the *stepFit* model is highly significant to reject the null hypothesis that confounder variables *wt*, *qsec* don't contribute to the accuracy of the model.

Residuals and Diagnostics

The residual plots of the regression model and some regression diagnostics are studied. The plots and the diagnostic results are shown in **Appendix**.

```r
summary(stepFit)$coef
tail(sort(hatvalues(stepFit)),3) #most leverage in the fit
tail(sort(dfbetas(stepFit)[,4]),3) #most influential points
```

- The points in the `Residuals vs Fitted` plot are randomly scattered on the plot verifying the independence condition
- The `Normal Q-Q` plot shows that the model can be fitted as a normal distribution with some outliers.
- The `Scale-Location` plot consists of points scattered in a constant band variance.
- The distinct outliers are shown in the `Residuals vs Leverage` plot. The top three highest hatvalues are obtained for Merc 230, Chrysler Imperial and Lincoln Continental. The top three highest dfbetas are obtained for Chrysler Imperial, Fiat 128 and Toyoto Corona.

Conclusions

- Cars with manual transmission get 2.9 miles more per gallon compared to cars with automatic transmision (2.9 after being adjusted for wt, qsec).
- For every 1000 lb increase in weight (wt), the mpg will decrease by 3.9 with the other variables constant.
- For every 1/4 mile time qsec, the mpg will increase by 1.2 with the other variables constant.

Appendix

The t-test results are shown below.

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

The summary from the pairFit model is shown below and the variables to use in pair2Fit model is decided based on the probabilities from the pairFit model.

```
##
## Call:
## lm(formula = mpg ~ disp + cyl + hp + wt + vs + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8319 -1.7327 -0.4034  1.3154  5.3430
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.96611    5.68062   6.155 1.95e-06 ***
## disp         0.01311    0.01186   1.105  0.27964
## cyl         -0.73198    0.84488  -0.866  0.39452
## hp          -0.02926    0.01415  -2.068  0.04911 *
## wt          -3.27739    1.14376  -2.865  0.00832 **
## vs1          1.36178    1.81343   0.751  0.45970
## am1          2.14088    1.64807   1.299  0.20579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.526 on 25 degrees of freedom
## Multiple R-squared:  0.8583, Adjusted R-squared:  0.8243
## F-statistic: 25.25 on 6 and 25 DF,  p-value: 1.817e-09
```

The summary of the best fit model and diagnostic results are shown below.

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am1           2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
##   Chrysler Imperial Lincoln Continental            Merc 230
##           0.2296338           0.2642151           0.2970422
```

```
##     Toyota Corona          Fiat 128 Chrysler Imperial
##         0.4050410         0.4765680         0.5626418
```

The figures are shown below.

Figure 1: Pair Graph of Motor Trend Car Road Tests

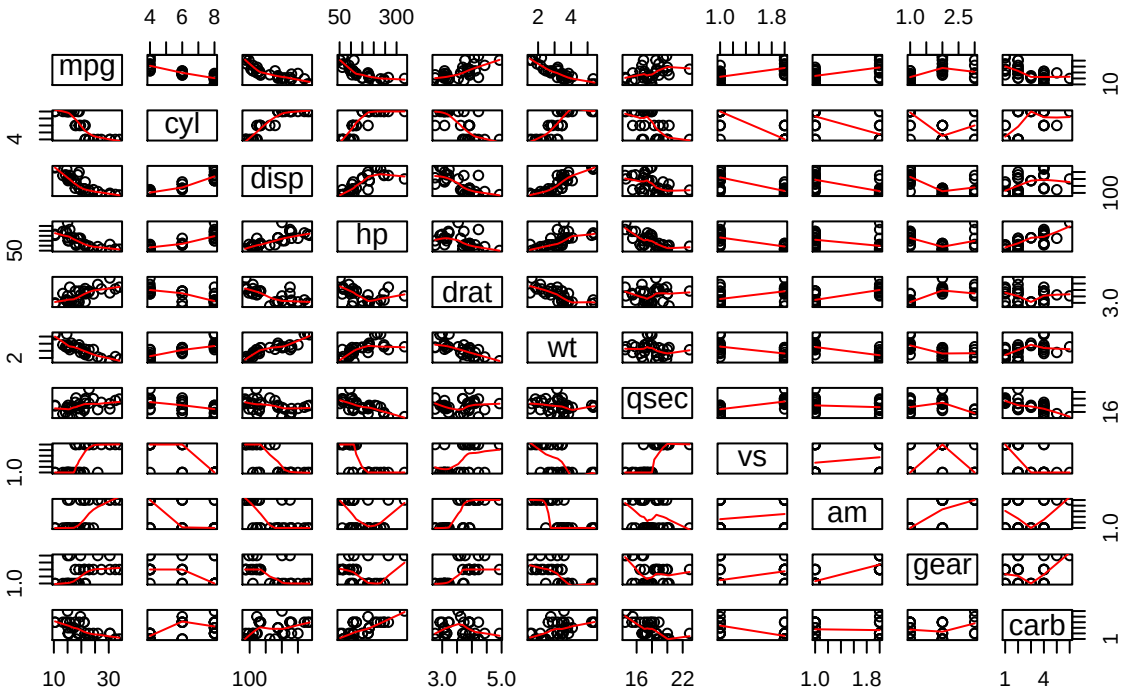

Figure 2: Boxplot of Miles per Gallon vs Transmission Type

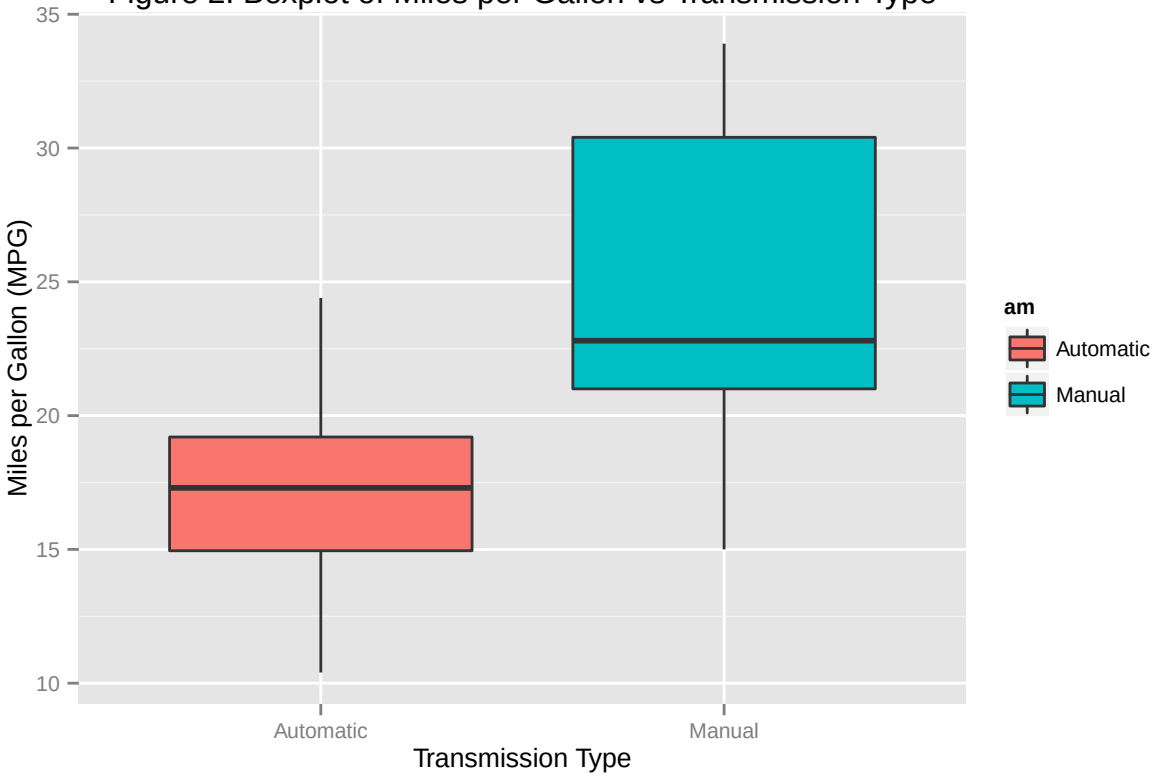

Figure 3: Regression model plots