Regression Models Course Project

Motor Trend

September 02, 2018

Executive Summary

In this edition of ***Motor Trend***, we look at a data set of a collection of cars from our 1974 archives. Using modern day data science and visualization tools, we explore the relationship between a set of variables and miles per gallon (MPG) and then answer the following questions for you: 1. Is an automatic or manual transmission better for MPG? 2. What is the difference in MPG for automatic and manual transmissions?

Exploratory Analysis

The `mtcars` data set comprises fuel consumption and 10 aspects of automobile design and performance for 32 autombiles (1973-74 models). Here's a glance at the first few rows of the data:

```
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
```

Our key variables of interest are `am`: Transmission type (0 = automatic, 1 = manual) and `mpg`: Miles/US gallon. Let's take a look at the difference in means of MPG for automatic and manual transmission:

```
aggregate(mpg ~ am, data = mtcars, FUN = mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

It appears that cars with manual transmission give about 7mpg more compared to cars with automatic transmission. This looks like a significant difference and we will use a *t-test* to find out:

```
t.test(mtcars$mpg[mtcars$am == 0], mtcars$mpg[mtcars$am == 1], conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[mtcars$am == 0] and mtcars$mpg[mtcars$am == 1]
## 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 of x mean of y
##  17.14737  24.39231
```

Since the *p-value is less than 0.05*, we reject the null hypothesis at 95% confidence level and conclude that the MPG differences are indeed statistically significant.

Regression Analysis

Let's have a closer look at the correlations of `mpg` to the other variables in `mtcars`:

```
round(cor(mtcars)[, 1], 3)
```

```
##    mpg    cyl   disp     hp   drat     wt   qsec     vs     am   gear 
##  1.000 -0.852 -0.848 -0.776  0.681 -0.868  0.419  0.664  0.600  0.480 
##   carb 
## -0.551
```

From the correlation data, there is a strong positive and inverse correlation between MPG and the other variables. Next, we will take a look at the regression model for MPG and transmission.

```
fit1 <- lm(mpg ~ am, data = mtcars)
summary(fit1)
```

```
## 
## 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
```

This regression model shows that the average MPG for automatic transmission is 17.1, and the MPG increases by 7.2 for manual transmission. However, the R^2 value is 0.36, which indicates this model only explains 36% of the variance. Thus we build a few multivariate linear regression models.

```
fit2 <- lm(mpg ~ am + cyl, data = mtcars)
fit3 <- lm(mpg ~ am + cyl + disp, data = mtcars)
fit4 <- lm(mpg ~ am + cyl + disp + hp, data = mtcars)
```

```r
fit5 <- lm(mpg ~ am + cyl + disp + hp + wt, data = mtcars)
fit6 <- lm(mpg ~ am + cyl + disp + hp + wt + drat, data = mtcars)
fit7 <- lm(mpg ~ am + cyl + disp + hp + wt + drat + vs, data = mtcars)
anova(fit1, fit2, fit3, fit4, fit5, fit6, fit7)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + disp
## Model 4: mpg ~ am + cyl + disp + hp
## Model 5: mpg ~ am + cyl + disp + hp + wt
## Model 6: mpg ~ am + cyl + disp + hp + wt + drat
## Model 7: mpg ~ am + cyl + disp + hp + wt + drat + vs
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 68.0021 1.837e-08 ***
## 3     28 252.08  1     19.28  2.9167  0.100574
## 4     27 216.37  1     35.71  5.4025  0.028894 *
## 5     26 163.12  1     53.25  8.0549  0.009086 **
## 6     25 162.43  1      0.69  0.1038  0.750087
## 7     24 158.65  1      3.78  0.5717  0.456945
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the probabilities in the above anova table, we can say that `am`, `cyl`, `hp` and `wt` are the significant predictors of `mpg` at 95% confidence interval.

```r
better_model <- lm(mpg ~ am + cyl + hp + wt, data = mtcars)
summary(better_model)
```

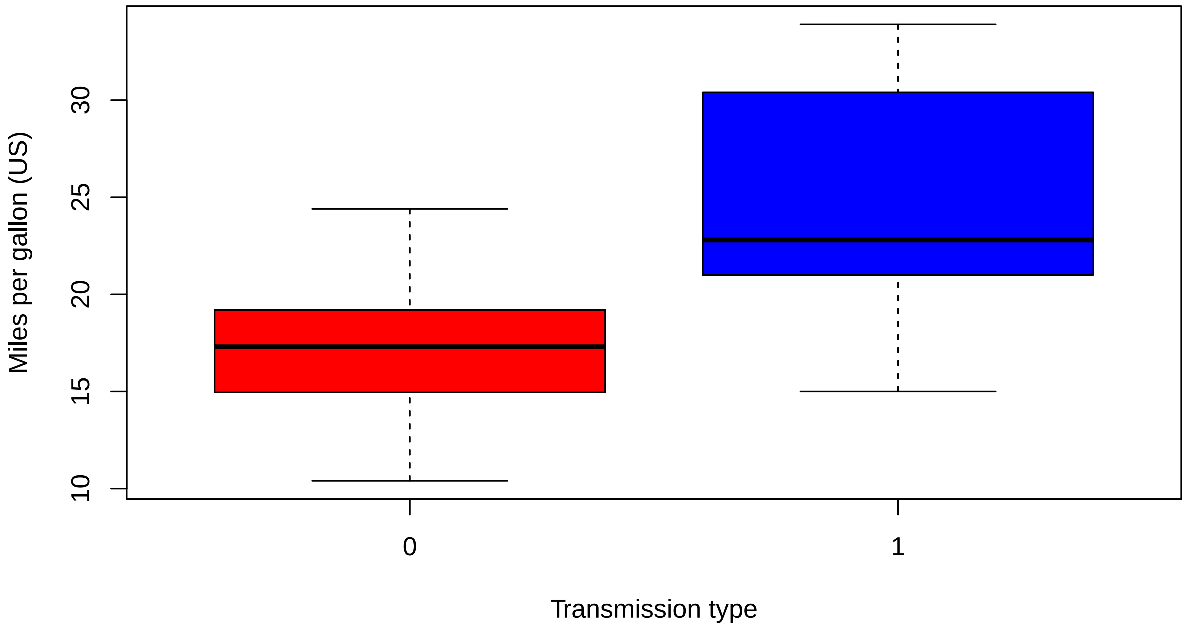
```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## am           1.47805    1.44115   1.026   0.3142
## cyl         -0.74516    0.58279  -1.279   0.2119
## hp          -0.02495    0.01365  -1.828   0.0786 .
## wt          -2.60648    0.91984  -2.834   0.0086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

Conclusions

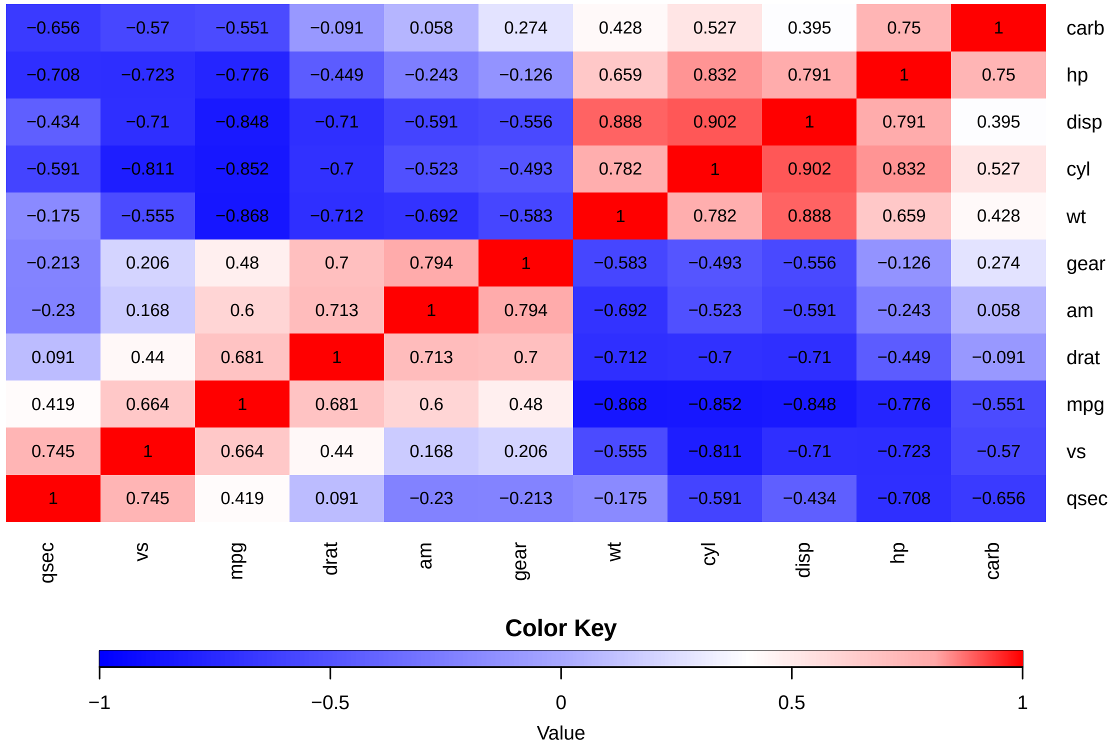
1. The above multivariate model shows that cars with manual transmission give better MPG compared to automatic transmission.
2. In contrast to our initial regression model which shows a huge difference in mileage for manual and automatic cars, the multivariate model indicates that cars with manual transmission give 1.48mpg higher than automatic transmission cars when other variables like number of cylinders (`cyl`), gross horsepower (`hp`) and weight (`wt`) are also taken into account. Moreover, this is a better regression model as it explains 85% of the variance.

Appendix

Plot - 1: Boxplot of MPG by transmission type

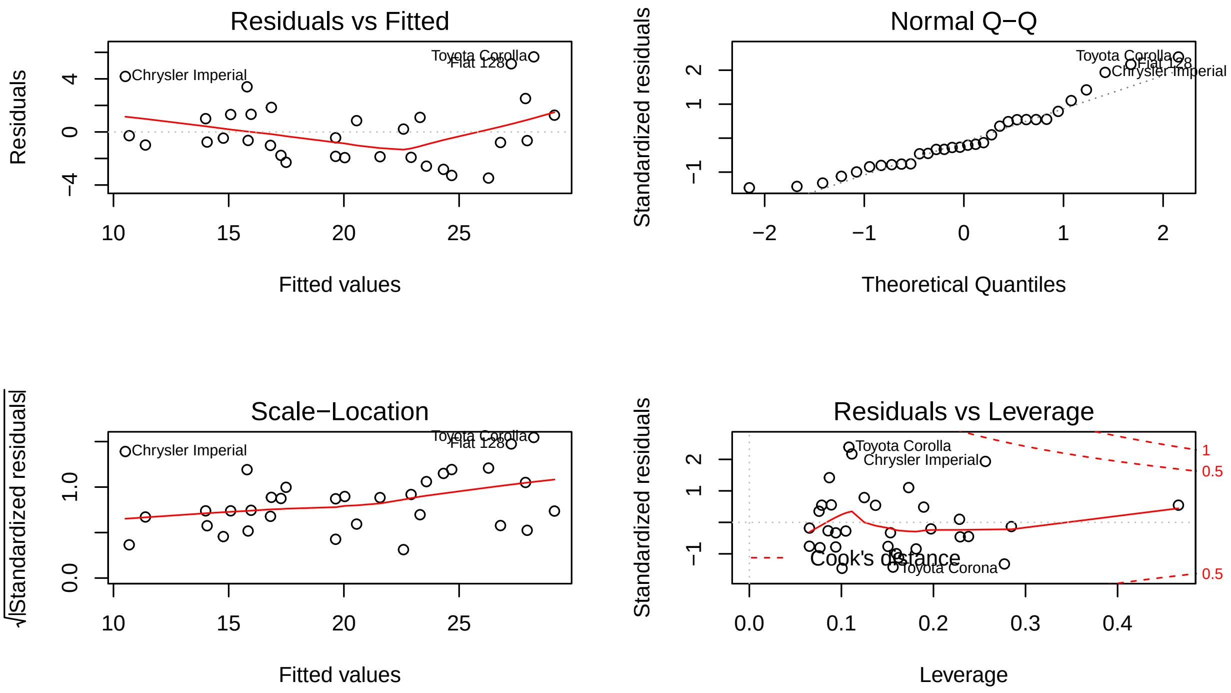


Plot - 2: Correlation heatmap of all variables



	qsec	vs	mpg	drat	am	gear	wt	cyl	disp	hp	carb
carb	−0.656	−0.57	−0.551	−0.091	0.058	0.274	0.428	0.527	0.395	0.75	1
hp	−0.708	−0.723	−0.776	−0.449	−0.243	−0.126	0.659	0.832	0.791	1	0.75
disp	−0.434	−0.71	−0.848	−0.71	−0.591	−0.556	0.888	0.902	1	0.791	0.395
cyl	−0.591	−0.811	−0.852	−0.7	−0.523	−0.493	0.782	1	0.902	0.832	0.527
wt	−0.175	−0.555	−0.868	−0.712	−0.692	−0.583	1	0.782	0.888	0.659	0.428
gear	−0.213	0.206	0.48	0.7	0.794	1	−0.583	−0.493	−0.556	−0.126	0.274
am	−0.23	0.168	0.6	0.713	1	0.794	−0.692	−0.523	−0.591	−0.243	0.058
drat	0.091	0.44	0.681	1	0.713	0.7	−0.712	−0.7	−0.71	−0.449	−0.091
mpg	0.419	0.664	1	0.681	0.6	0.48	−0.868	−0.852	−0.848	−0.776	−0.551
vs	0.745	1	0.664	0.44	0.168	0.206	−0.555	−0.811	−0.71	−0.723	−0.57
qsec	1	0.745	0.419	0.091	−0.23	−0.213	−0.175	−0.591	−0.434	−0.708	−0.656

Plot - 3: Analysis of residuals



Residuals vs Fitted
Residuals
Fitted values
−4
0
4
10
15
20
25
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
2
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
10
15
20
25
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
−1
1
2
0.0
0.1
0.2
0.3
0.4
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5