Regression Models Course Project: MT Cars Analysis

Lei Fan

1/29/2020

Overview

The following analysis studies the `mtcars` dataset, explores the relationship between the miles-per-gallon (MPG) and various other factors of the cars in the dataset, and attempts to answer the following two questions:

- "Is an automatic or manual transmission better for MPG?"
- "Quantify the MPG difference between automatic and manual transmissions."

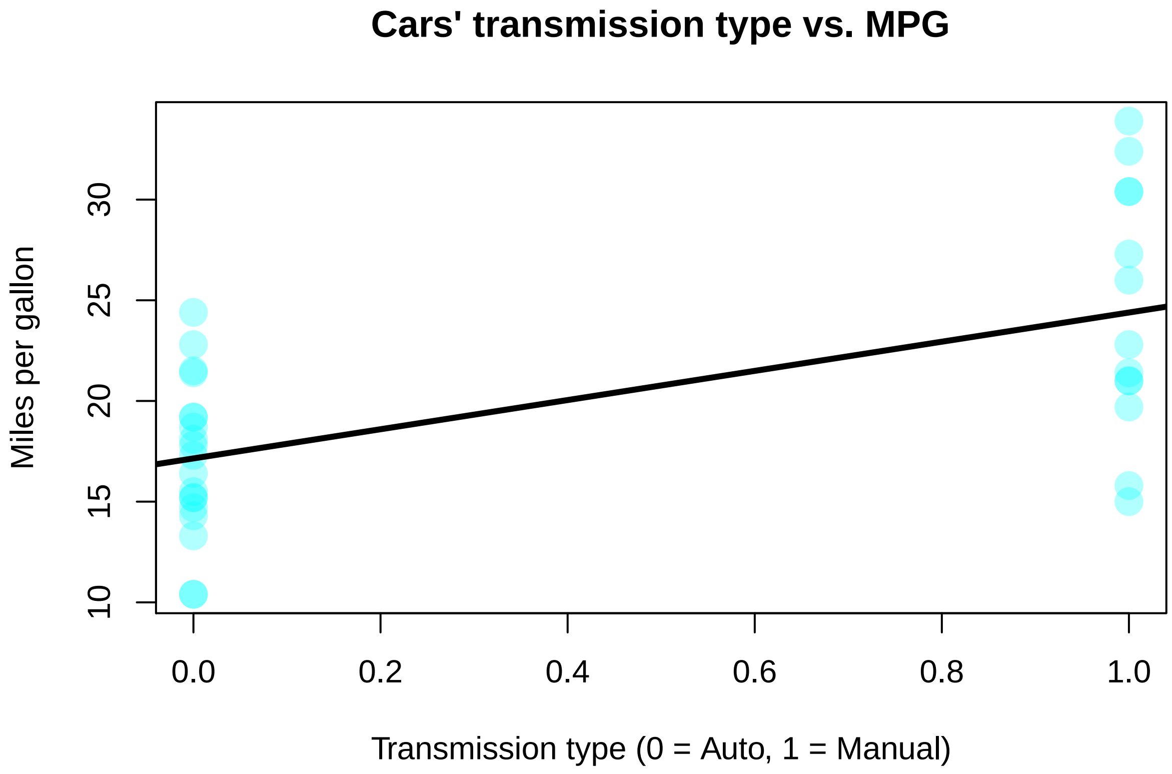
Exploratory Analysis

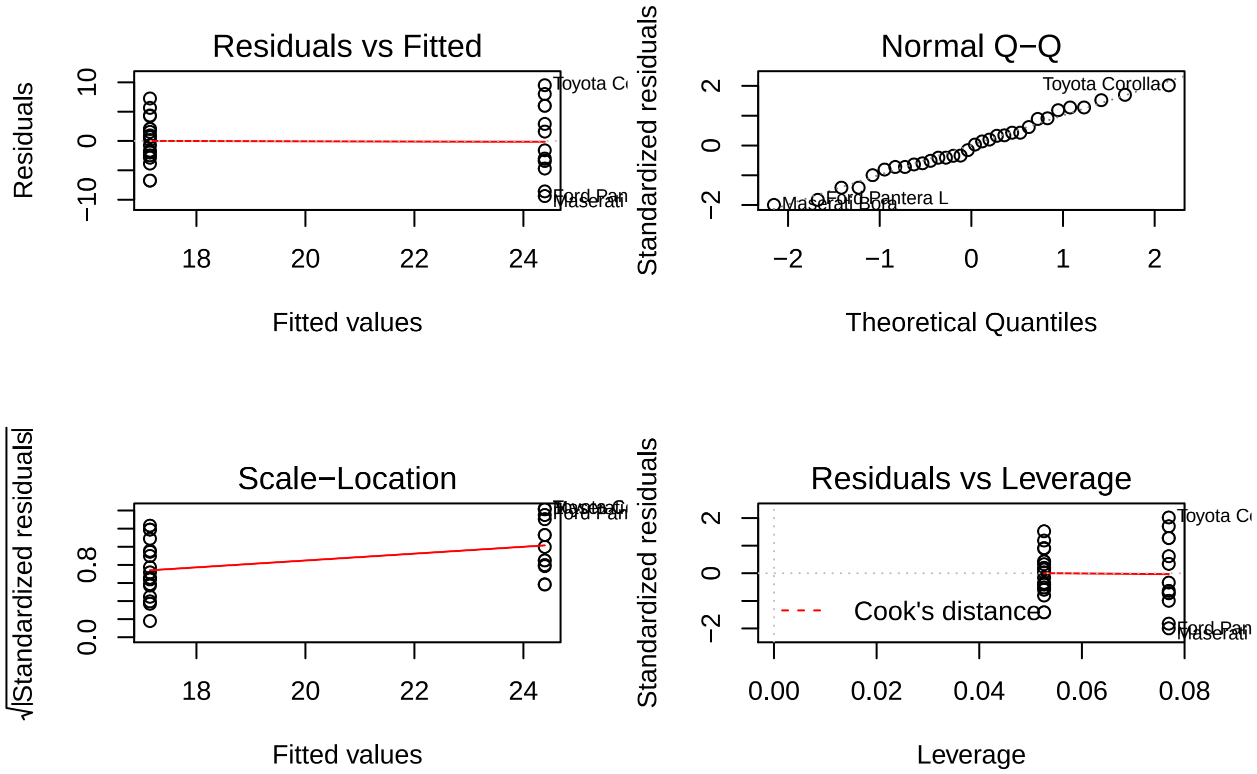
Looking at the average MPG value for cars with automatic transmission vs. manual transmission, it appears that we can reject the null hypothesis (that the two have the same average MPG) in favor of the alternative, that cars with automatic transmission have a lower average MPG than cars with manual transmission:

```
## 
##  Welch Two Sample t-test
## 
## data:  mtcars_auto$mpg and mtcars_manual$mpg
## t = -3.7671, df = 18.332, p-value = 0.0006868
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.913256
## sample estimates:
## mean of x mean of y 
##  17.14737  24.39231
```

If only the transmission type was taken into account, it appears that, on average, cars with manual transmission get ~7.2 more miles per gallon than cars with automatic transmission (see slope of the `am` regressor below):

```
## 
## 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
```





Although the am factor was clearly significant, and there's nothing particularly wrong with the model (no heteroscedasticity observed, the variance was roughly normally distributed, etc.), with an adjusted R-squared value of only 0.3385, this is a rather poor model.

Multiple Regressors

We will attempt to find other regressors that may also contribute to determining MPG:

```
cor(mtcars$mpg, mtcars)
```

```
##      mpg       cyl       disp         hp      drat         wt     qsec
## [1,]   1 -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684
##             vs        am      gear       carb
## [1,] 0.6640389 0.5998324 0.4802848 -0.5509251
```

As we can see, cyl, disp, hp, and wt are all quite highly correlated with the MPG value. A model that includes those 4 regressors, as well as am, should perform better than just am alone. Let's see it, along with the VIFs, since we have so many regressors, some of which, intuitively, seem correlated (e.g. cyl and hp).

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp + cyl + disp, data = mtcars)
##
## Residuals:
```

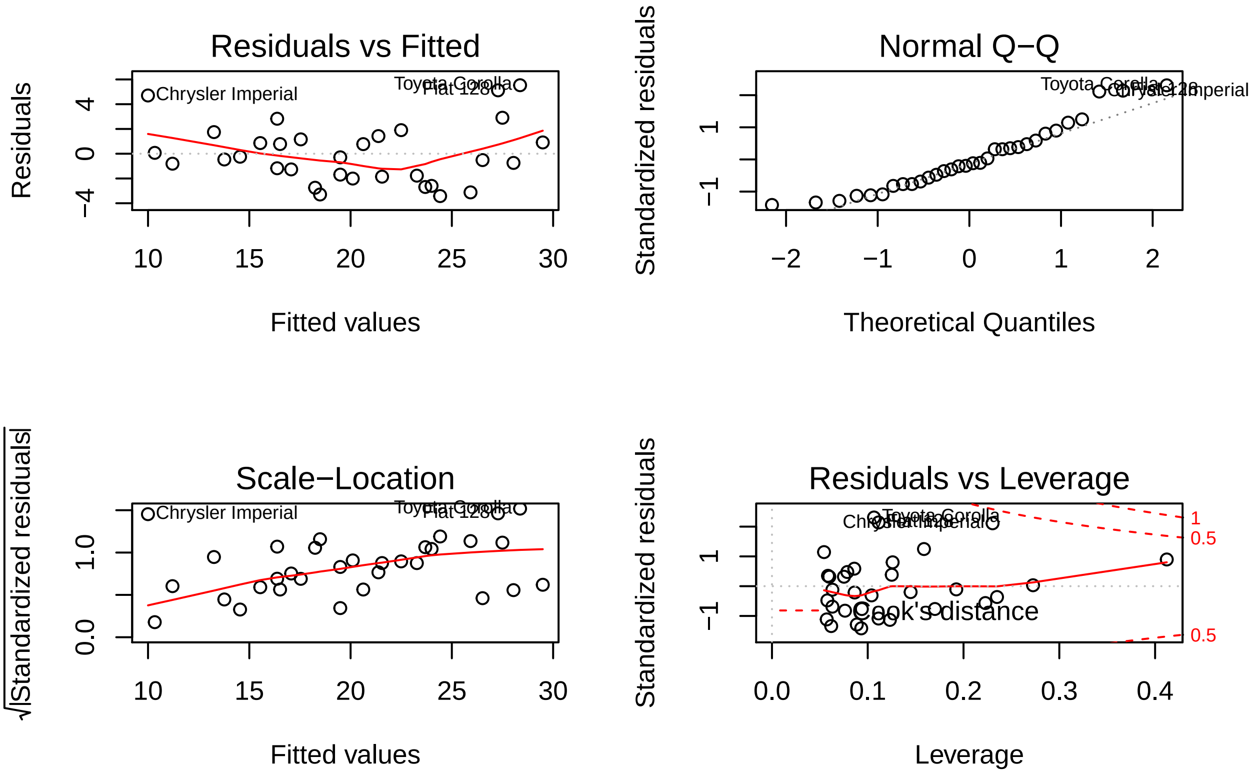
```
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## am           1.55649    1.44054   1.080  0.28984
## wt          -3.30262    1.13364  -2.913  0.00726 **
## hp          -0.02796    0.01392  -2.008  0.05510 .
## cyl         -1.10638    0.67636  -1.636  0.11393
## disp         0.01226    0.01171   1.047  0.30472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

```
##       am       wt       hp      cyl     disp
## 1.597831 2.465654 2.121758 2.685043 3.225123
```

The adjusted R-squared value is high (~83% of variation explained by model), but too many regressors are correlated with the others (especially `cyl` and `disp`). Let's try removing them from the model and look at the VIFs again.

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am           2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

```
##       am       wt       hp
## 1.507011 1.942894 1.445034
```



The model is pretty good at explaining the MPG (~82% variation explained) and has no significant issues (no heteroscedasticity observed, the variance was roughly normally distributed aside from a couple of outliers, no overly large VIFs). How does it compare to some other models?

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + hp
## Model 4: mpg ~ am + wt + hp + cyl
## Model 5: mpg ~ am + wt + hp + cyl + disp
## Model 6: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3     28 180.29  1     98.03 13.9571  0.001219 **
## 4     27 170.00  1     10.29  1.4655  0.239500
## 5     26 163.12  1      6.88  0.9793  0.333646
## 6     21 147.49  5     15.63  0.4449  0.812059
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As the ANOVA result shows, the 3rd model (only `am`, `wt` and `hp`) is perhaps our fit; the other, more complicated models cannot justify their inclusion of additional regressors at a loss of our precious DFs and their adjusted R-squared value is probably artificially high. Therefore, we conclude that **cars with manual transmission gets ~2 more miles per gallon than cars with automatic transmission**.