Effects of various car features on fuel mileage using the mtcars dataset.

Zoheb Nensey

December 27, 2015

Executive Summary

Fuel mileage in cars became a subject of interest in recent years as gas prices in the United States began to rapidly increase in the last decade (though they have since declined preciptiously). Increasing fuel mileage, as a result, has become a primary focus of automakers as a major selling point.

In this analysis, our main question is whether the type of transmission in a car (automatic or manual) has an effect on fuel mileage, as well as trying to assess how much of an impact it has. My analysis here indicates that on it's own, the variable appears to have a significant effect on mileage, but has little impact when other variables are held constant to account for confounding.

Analyses

Hypotheses

The null hypothesis is that transmission type does not have a coefficient different than zero when trying to explain fuel mileage. The alternative hypothesis is that the coefficient is different than zero.

Exploratory Analysis

To start, correlations were checked to see what variables might have the most impact on mileage. The correlations displayed below are the correlations between mileage and other variables.

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

Looking here farther down, it appears that the variables with the larges amount of impact are cylinders, the displacement, horsepowers, the rear axle ratio, weight, and the transmission. A plot of the correlation between these variables can be found in the appendix.

Regression Analyses

To start, we present a simple linear regression analysis with the transmission as a predictor.

```
simple_linear <- lm(mpg ~ am, data=mtcars)
summary(simple_linear)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

```
summary(simple_linear)$r.squared
```

```
## [1] 0.3597989
```

The simple linear regression indicates that transmission is a significant variable, and the R-squared value indicates that 35.98% of variance is explained by this version of the model, but we still need to account for other variables. So we'll move to a multivariate model using the variables above.

```
multivariate_linear_1 <- lm(mpg ~ cyl + disp + hp + drat + wt + am,
data=mtcars)
anova(multivariate_linear_1)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value   Pr(>F)
## cyl        1 817.71  817.71 125.8534    3e-11 ***
## disp       1  37.59   37.59   5.7861 0.023879 *
## hp         1   9.37    9.37   1.4423 0.241027
## drat       1  16.47   16.47   2.5345 0.123950
## wt         1  77.48   77.48  11.9242 0.001985 **
## am         1   4.99    4.99   0.7684 0.389058
## Residuals 25 162.43    6.50
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ANOVA for this model, there appear to be several highly insignificant variables, so these will be removed (with the exception of transmission.)

```
multivariate_linear_2 <- lm(mpg ~ cyl + disp + wt + am, data=mtcars)
summary(multivariate_linear_2)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 40.898313414 3.60154037 11.3557837 8.677574e-12
## cyl         -1.784173258 0.61819218 -2.8861142 7.581533e-03
## disp         0.007403833 0.01208067  0.6128661 5.450930e-01
## wt          -3.583425472 1.18650433 -3.0201537 5.468412e-03
## am           0.129065571 1.32151163  0.0976651 9.229196e-01
```

```r
anova(multivariate_linear_2)
```

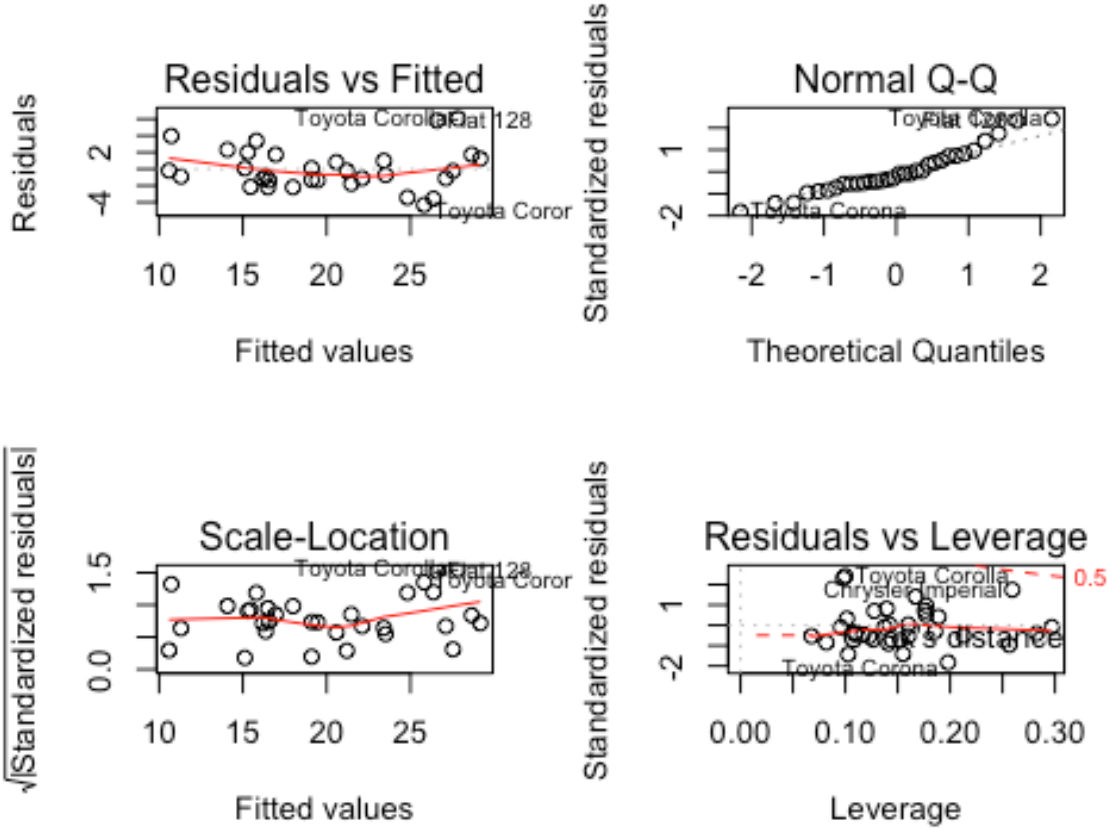
```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## cyl        1 817.71  817.71 117.1721 2.526e-11 ***
## disp       1  37.59   37.59   5.3869   0.02808 *
## wt         1  82.25   82.25  11.7855   0.00194 **
## am         1   0.07    0.07   0.0095   0.92292
## Residuals 27 188.43    6.98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
rsqr <- summary(multivariate_linear_2)$r.squared
```

After creating a regression with just these variables, we can recheck the coefficients and use an ANOVA to test significance. Looking at this model, it appears that the transmission variable is more or less insignificant compared to the rest of the predictor variables. The R-squared for this model is 0.8326661, meaning that this version of the model explains 83.26% of the variance. Despite the r-squared, we cannot reject the null hypothesis since this null hypothesis is testing if the coefficient of transmission is non-zero.

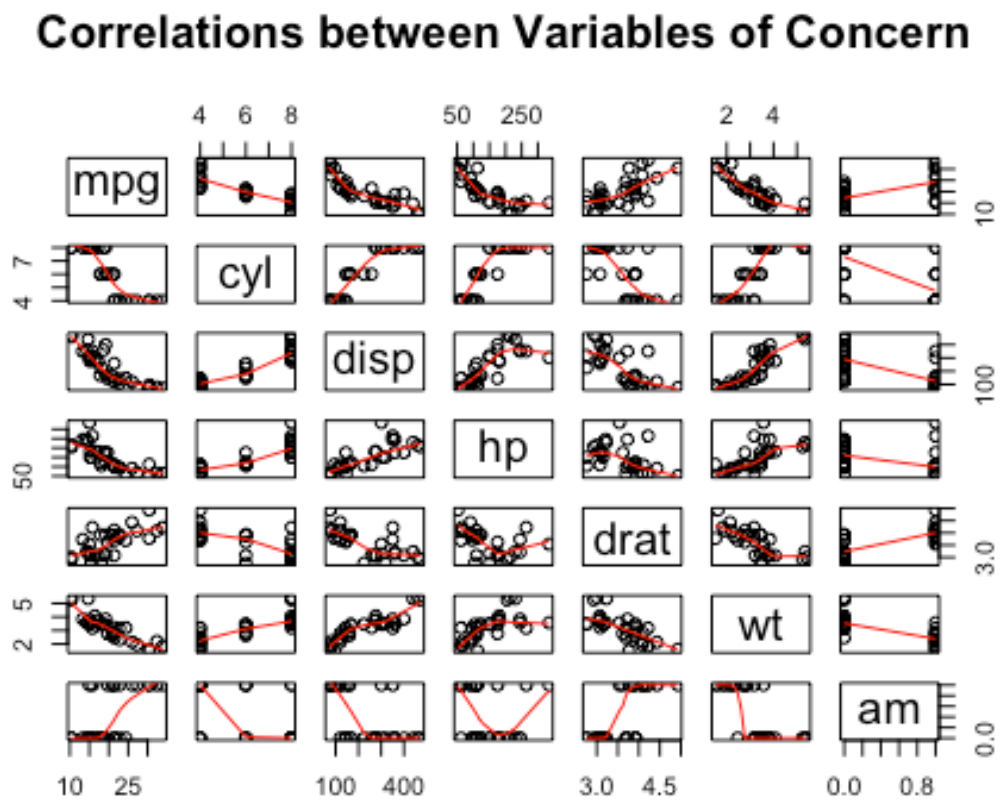
Residuals

The other thing we need to check is residuals. This can be performed by plotting.



The residuals are apprxoimately normal, based on the Q-Q plot. though based on the residuals vs fitted plot it appears that some vehicles may be considered outliers and could distort the data.

Appendix


Correlations between Variables of Concern
mpg
cyl
disp
hp
drat
wt
am
4 6 8
50 250
2 4
10 25
100 400
3.0 4.5
0.0 0.8
4 7
50
2 5
10
100
3.0
0.0