Multiple Regression Course Project

Jacob M. Lundeen

13 March 2019

Executive Summary

Through the use of central tendency analysis with the confirmation by multiple regression to the formula mpg ~ am + wt + hp + (am * wt), we find that, in 1974, manual transmissions tended to be more fuel efficient than automatic transmissions.

Introduction: Automatics versus Manuals

Which transmission type is more fuel efficient, automatics or manuals? This has long been a question posed by automotive enthusiasts for many decades. This analysis, using data collected by *Motor Trend* in 1974, will apply multiple regression models to that data in an attempt to answer that question.

We start off with looking over the data, looking to understand what exactly we are looking at, looking for anything that seems out of the ordinary or may need to be addressed and cleaned.

The mtcars data set is comprised of the following eleven numeric variables:

- MPG - Miles/(US) gallon
- CYL - Number of cylinders
- DISP- Displacement (cu.in.)
- HP - Gross horsepower
- DRAT - Rear axle ratio
- WT - Weight (1000 lbs)
- QSEC - 1/4 mile time
- VS - Engine (0 = V-shaped, 1 = straight)
- AM - Transmission (0 = automatic, 1 = manual)
- GEAR - Number of forward gears
- CARB - Number of carburetors

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

The initial look into the data shows no discrepencies or other issues to be concerned with. The only thing we would need to do is to change some of the variables over to factors: cyl, vs, am, gear and carb. From there, we will take a look at the central tendencies of the data.

Table 1: The central tendencies of the fuel efficiency data.

	Mean	Median	Variance	STD
Auto	17.15	17.3	14.70	3.83
Manual	24.39	22.8	38.03	6.17

So we see what we expected, the manual transmissions have a higher mean MPG than the automatics by 7.24 MPG. However, we also see a lower median value for the manuals (though still higher than the mean and median for the automatics), as well as a larger variance. So manual transmissions may, on average, have a higher MPG, but there seems to be a wider spread. A quick look tells us that while the manual data has a higher variance, none of the data points would constitute an outlier.

Models

Linear Regression

To continue our analysis, we now move to applying regression models. First, we will apply a standard linear model.

```
##
## Call:
## lm(formula = mpg ~ ., data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## amManual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

Table 2: Table Showing the Original MPG, LM Predicated Coefficients and the LM Residuals.

	mpg	predicted	residuals
Mazda RX4	21.0	21.37974	-0.3797416
Mazda RX4 Wag	21.0	20.43064	0.5693582
Datsun 710	22.8	26.30873	-3.5087294
Hornet 4 Drive	21.4	20.80728	0.5927215
Hornet Sportabout	18.7	17.42580	1.2741991
Valiant	18.1	18.78540	-0.6854022

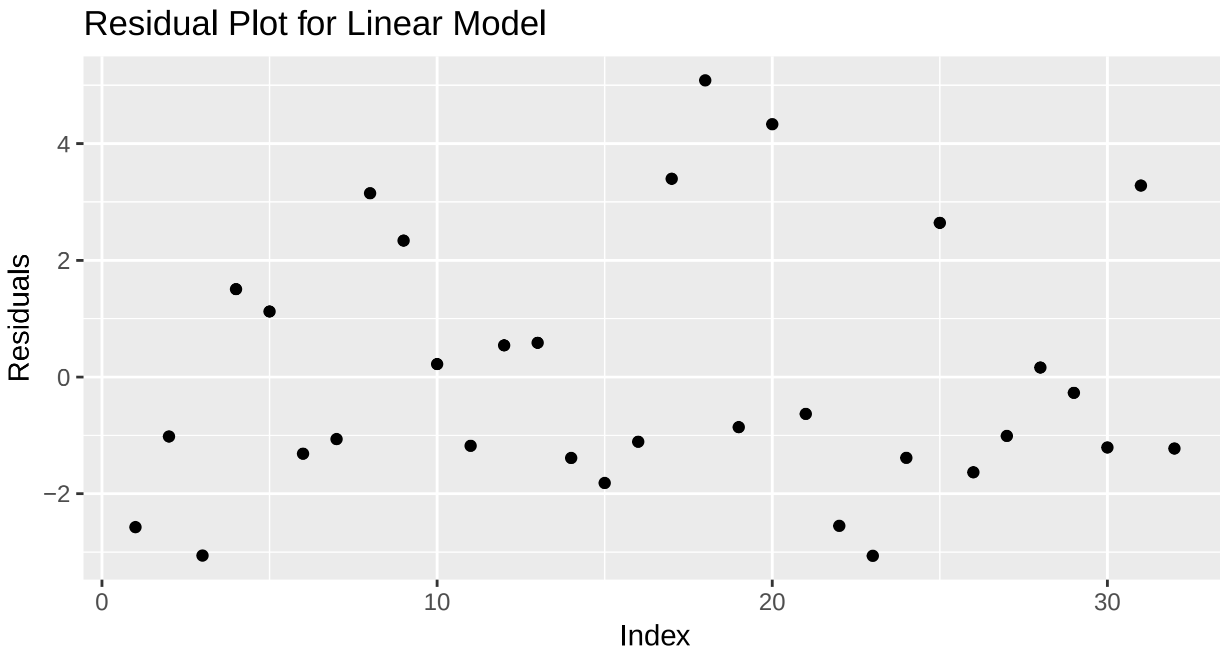
From these first two pieces of data, we can see our coefficients, p-values, predicted and residual values from the linear model. While the R^2 is showing a strong fit, our p-values for the coefficient estimates show that we have a lot of regressors that are not statistically signficant to our outcome. Let us remove the least significant regressor, cyl, and see how it affects the rest.

The model lm(mpg ~ . - cyl) has a similar R^2 and residuals, but the p-values for the remaining regressors have gotten more significant, with "wt" being the most significant at 0.102. I won't include all the iterations of model here, but the following model is what we found to fit the data the best.

```
##
## Call:
## lm(formula = mpg ~ wt + am + hp + (am * wt), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0639 -1.3315 -0.9347  1.2180  5.0822
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.947333   2.723411  11.363 8.55e-12 ***
## wt          -2.515586   0.844497  -2.979  0.00605 **
## am          11.554813   4.023277   2.872  0.00784 **
## hp          -0.026949   0.009796  -2.751  0.01048 *
## wt:am       -3.577910   1.442796  -2.480  0.01968 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.332 on 27 degrees of freedom
## Multiple R-squared:  0.8696, Adjusted R-squared:  0.8503
## F-statistic: 45.01 on 4 and 27 DF,  p-value: 1.451e-11
```

Table 3: Table Showing the Original MPG, LM Predicated Coefficients and the LM Residuals.

	mpg	predicted	residuals
Mazda RX4	21.0	23.57276	-2.572760
Mazda RX4 Wag	21.0	22.01892	-1.018919
Datsun 710	22.8	25.85895	-3.058948
Hornet 4 Drive	21.4	19.89530	1.504703
Hornet Sportabout	18.7	17.57758	1.122417
Valiant	18.1	19.41373	-1.313726



The chosen model includes four regressors: weight, transmission type, horsepower and an interaction term between transmission type and weight. To answer the initial question of this paper, we see our coefficient for transmission type is 11.55, meaning that manual transmissions are better for MPG. Of interest, our interaction term indicates that we can reject the null hypothesis and conclude that the mean mpg of manual transmission is significantly different from the mean mpg of automatic transmissions.

The residual plot shows no apparant pattern, which lends credence to the use of our model.

To keep the focus of this paper on the orignal question, the plots for the remaining regressors will be available in the appendix. Above, we have a nicer plot of the actual values connected to the predicted values with the residual values included as a color gradient. This is simply a more appealing way to show the relationship of the residuals to the actual values.

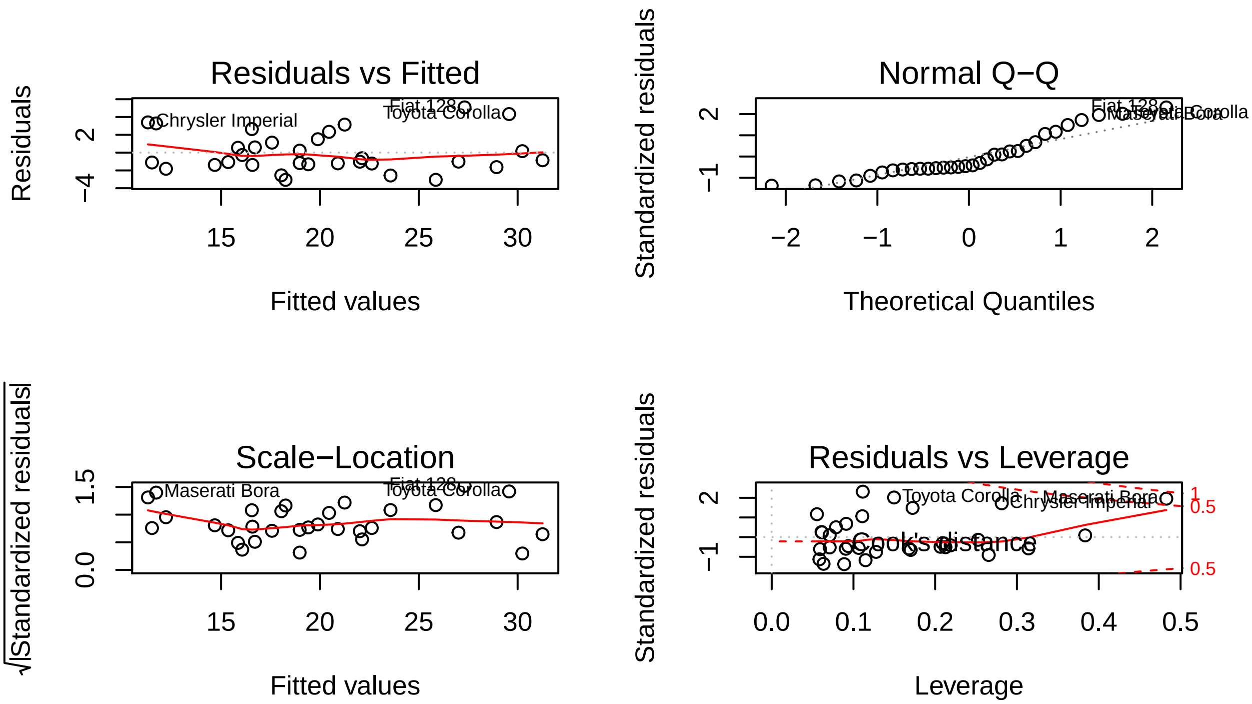
Inference

Our final analysis will be performing a Welch Two Sample t-test with the two transmission types.What we see here is that with a p-value of 0.001, We can reject the null hypothesis and conclude that the mean mpg of manual transmission is significantly different from the mean mpg of automatic transmissions.

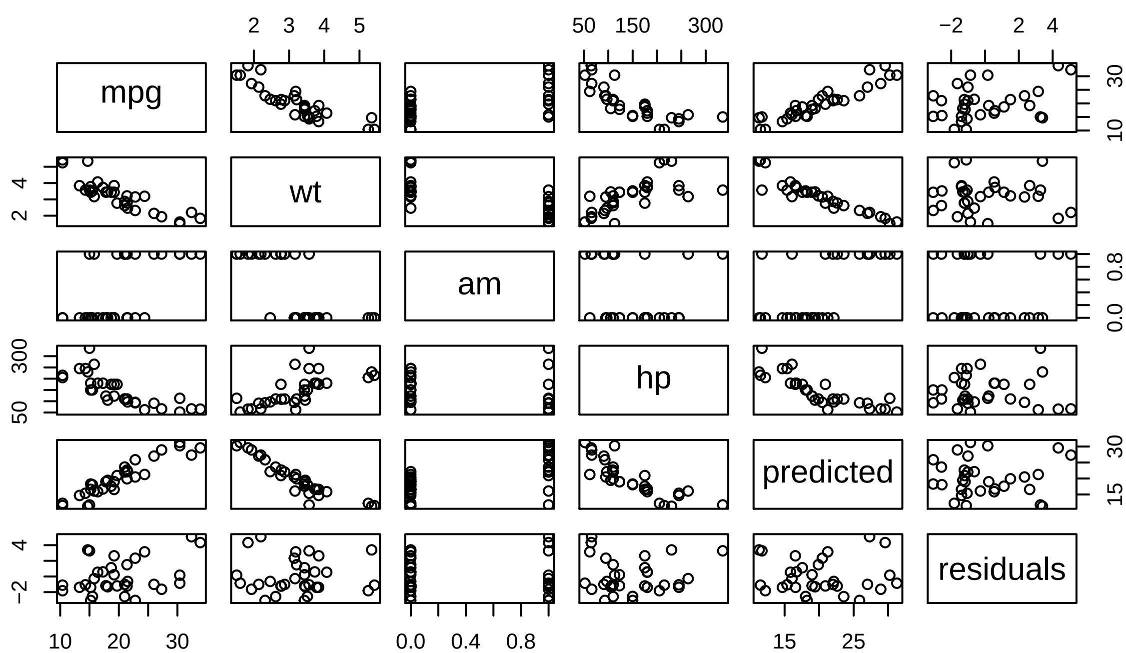
Summary

Through the use of central tendency evaluation, and then the application of multiple linear regression, we can conclude that, in 1974, manual tranmissions were more fuel efficient, on average, than automatic transmissions. However, we discovered that the weight of the vehicle was the most signficant factor to determine a cars mileage, and that horsepower and the interaction between weight and transmission type were all significant factors.

Appendix



Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Maserati Bora
Residuals vs Leverage
Leverage
Cook's distance



mpg
wt
am
hp
predicted
residuals