Motor Trend Data Regression Analysis

Akhil Kota

2022-07-29

Executive Summary

In this paper, we attempt to answer the following questions: (1) "Is an automatic or manual transmission better for MPG?" and (2) "Quantify the MPG difference between automatic and manual transmissions.". The mtcars dataset is first explored for the relationship between mpg and am variables, along with their correlations to other variables. Then, we find an appropriate model controlling for confounders using ANOVA (with a series of linear regression model fits using nested predictor sets). We also carry out some diagnostics on our fit to confirm its validity and find points with high leverage and influence. Lastly, using this appropriate model, we find that the average difference in MPG between transmission types is insignificant (not significantly different from 0) with a p-value of 0.89. The difference in MPG between transmission types can be quantified as 0.1765 MPG with a standard deviation of 1.3045 MPG, or in the range (-2.496 MPG, 2.849 MPG) with 95% confidence. Thus, we cannot conclude that there is a difference in MPG between auto and manual transmission cars in the population.

Exploratory Analysis

In this section, we explore the data and variables in the mtcars dataset. Please refer to the Appendix for descriptions of each variable in the set. Since we are setting out to analyze the relationship between mpg and am (auto vs manual transmission), we initially plot some figures relating to mpg and am (see Figure 1). Note that automatic transmission is represented by 0, and manual by 1. Thus, we see a slightly greater mpg in manual transmission cars than automatic transmission cars within this data set.

The correlations of am and mpg with the other variables were also analyzed (see Table 1). We see that am has a strong correlation with the drat and gear variables, while mpg has a stronger correlation with cyl, disp, hp, drat, wt and vs than it does with am. We now set out to fit an appropriate model to our data to make this relationship clearer and answer our questions about the population.

Model Selection and Fitting

Selection

Having analyzed correlations between variables in the Exploratory Analysis section above, we now delineate the strategy used for fitting a linear model to the data while accounting for confounders. We used a series of nested predictor sets for our regressions, first performing the regression with just am as a predictor of mpg, and then adding predictors based on their individual correlations with mpg. The predictors with higher absolute correlations with mpg were added into the predictor set first. Adding to the predictor set stops when the next predictor to add has a lower absolute correlation with mpg than am itself. These nested regression models were then be tested using ANOVA to determine which variables were necessary in our model (using the associated p-values from the ANOVA test). The summary of the ANOVA analysis is available in Table 2 in the appendix section. Using an $\alpha = 0.05$ for the ANOVA test, it appears to be necessary to add wt and cyl as confounders to account for in our model. The reason we don't simply include every variable is to avoid multicollinearity and the resulting increased variance in our model coefficients/predictions.

Fitting and Interpretation

Now, we fit mpg vs. am (see Table 3) and mpg vs. the set of am, wt, and cyl (see Table 4). The impact of the confounders becomes clear when we interpret the coefficients. Whereas in Table 3, we see that there is a

predicted average difference of 7.245 mpg between manual and automatic (manual having the higher mpg), accounting for confounders gives us the result that, holding car weight and cylinder count constant, there is a predicted average difference of only 0.17 mpg between manual and automatic transmission (manual still having higher mpg). The intercept in the first model indicates a predicted average mpg of 17.147 mpg for an automatic car, but the intercept in the second model does not have much interpretive value (predicts an average mpg of 39.4 for a car with 0 weight, 0 cyl, and automatic transmission). Thus, when accounting for key confounders, we see a very stark decrease in contribution of transmission to mpg, compared to the other variables.

Inferential Analysis

To make the observations outlined above a bit more robust, we can interpret the p-values given in each of the models, which come from Student's t-tests on the coefficients of the models using a null hypothesis that the coefficient is equal to 0. We will use a standard $\alpha = 0.05$. The first model (in Table 3) without confounders shows that the difference between mpg of manual transmission and automatic transmission is significant, with a p-value of $0.000285 < \alpha$. However, the adjusted R^2 is only 0.3385, suggesting that not much variation in mpg is actually explained by the am values, and other variables need to be accounted for. We have done this in the second model (in Table 4) with confounders, and we see that the difference between mpg between manual and automatic transmissions immediately becomes insignificant and indistinguishable from 0, with a p-value of $0.89 > \alpha$. The variability in mpg has been mostly explained away by the confounders wt and cyl, and in such a drastic fashion that, even holding them constant, the difference in mpg between transmission is not enough to reject the null hypothesis that am coefficient equals 0. This conclusion is also easy to see with the fact that the standard deviation in the am coefficient is significantly larger than the estimated coefficient itself, which leads to the very small t-statistic and large p-value.

From here onwards, we will use the second regression model with the confounders accounted for.

Diagnostics

Residual plots for our model are provided in Figure 2. We can see that there is no visible pattern in the residuals vs. fitted values plot, which is a good omen for our model. Although the residuals are not perfectly normally distributed, they do lie close enough to the identity line. Lastly, we see that the data points with the highest residuals tend to have somewhat lower leverage, such as the Toyota Corolla. There is nothing odd in any of the residual plots to suggest that the model we have used may not be a good fit to predict mpg from am, wt, and cyl.

Some other diagnostics have also been carried out. Hat values and df-beta's for each point were calculated, and points with the top 10 highest absolute values for each df beta (except the intercept dfbeta, which is somewhat irrelevant here) and hat value are listed in Table 5. In other words, these points are the points with highest leverage (high hat value implies high leverage), which also have exerted their leverage to influence our model in a significant way (high df beta's imply high influence). From our diagnostics, we find that the Chrysler Imperial, the Toyota Corona, and the Volvo 142E have both high leverage and high influence. The former two are also easily seen in some of our residual plots as potential outliers. The main takeaway is that removing any of these 3 points could have a significant impact on our coefficient estimates.

Results

Using our linear model with confounders, we may now answer our questions.

1. It is difficult to say whether automatic or manual transmission are better for MPG, after accounting for confounding variables: weight and number of cylinders. From our Inferential Analysis section, we see that the difference in MPG by transmission type is insignificant, with a p-value of 0.89 using a null hypothesis of no difference. Thus, we fail to reject that there is no difference between the two transmissions in terms of MPG.
2. If we try to quantify our results, we may say that the difference between manual transmission MPG and automatic transmission MPG is, on average, 0.1765 MPG, with a standard deviation of 1.3045 MPG. A 95% confidence interval of the same difference would be (-2.496 MPG, 2.849 MPG), which comfortably contains 0.

Appendix

Variables in mtcars Data Set

All variable definitions below come from R Documentation.

1. mpg: Miles/gallon (US)
2. cyl: Cylinders (number)
3. disp: Displacement (in^3)
4. hp: Horsepower (gross)
5. drat: Rear axle ratio
6. wt: Weight (US half-tons)
7. qsec: Quarter mile time
8. vs: Engine (0 = V-shaped, 1 = straight)
9. am: Transmission (0 = auto, 1 = manual)
10. gear: Forward gears (number)
11. carb: Carburetors (number)

Table 1: Correlations with am and mpg

```
##               am        mpg
## mpg   0.59983243  1.0000000
## cyl  -0.52260705 -0.8521620
## disp -0.59122704 -0.8475514
## hp   -0.24320426 -0.7761684
## drat  0.71271113  0.6811719
## wt   -0.69249526 -0.8676594
## qsec -0.22986086  0.4186840
## vs    0.16834512  0.6640389
## am    1.00000000  0.5998324
## gear  0.79405876  0.4802848
## carb  0.05753435 -0.5509251
```

Table 2: ANOVA for Confounders

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + cyl
## Model 4: mpg ~ am + wt + cyl + disp
## Model 5: mpg ~ am + wt + cyl + disp + hp
## Model 6: mpg ~ am + wt + cyl + disp + hp + drat
## Model 7: mpg ~ am + wt + cyl + disp + hp + drat + vs
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 66.9496 2.113e-08 ***
## 3     28 191.05  1     87.27 13.2019  0.001323 **
## 4     27 188.43  1      2.62  0.3965  0.534840
## 5     26 163.12  1     25.31  3.8281  0.062133 .
## 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
```

Table 3: Model Fit without Confounders

```
##
```

```
## 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
```

Table 4: Model Fit with Significant Confounders

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## am            0.1765     1.3045   0.135  0.89334
## wt           -3.1251     0.9109  -3.431  0.00189 **
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

Table 5: Model Diagnostics, Top 10 Leverage and Influence

```
##                   dfbeta_intercept  dfbeta_am  dfbeta_wt dfbeta_cyl    hatval
## Chrysler Imperial       -0.6177622  0.3953935  0.9470405 -0.4495640 0.2557513
## Toyota Corona           -0.8295668  0.6741522  0.2650010  0.3404518 0.1901075
## Volvo 142E               0.0494444 -0.3687057 -0.3967186  0.4197177 0.1544287
```

Figure 1: Exploring Miles/Gallon vs. Transmission

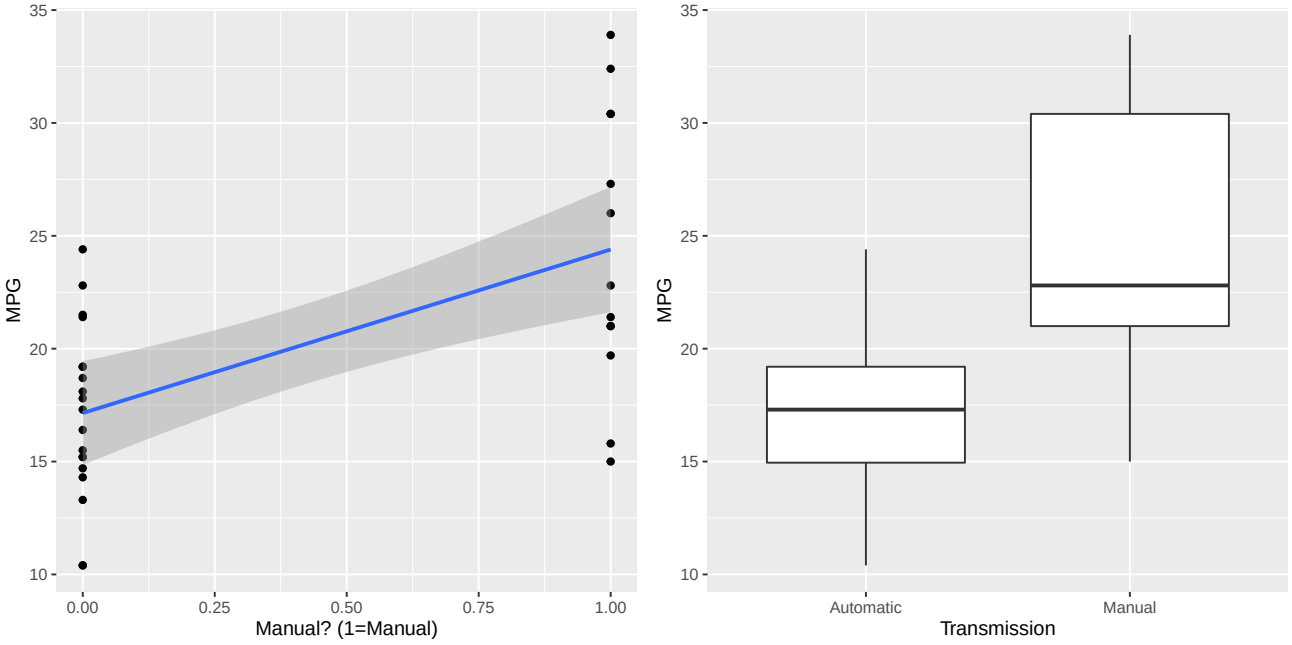

Figure 2: Residual Plots