# Motor Trend: Automatic or Manual?

SR Stevenson

2022-07-26

## Executive Summary

Is automatic or manual transmission better for miles per galon (mpg)? While at first glance, manual cars appear to have higher mpg values, this association is potentially confounded by the weight of the vehicles. Manual cars in this study are on average lighter than automatic cars and weight has a strong negative relationship with mpg. Given this, model building was begun from the starting model "mpg ~ am + wt" with ANOVA testing plus coefficient inspection used to decide on inclusion of further regressors. The final model (mpg ~ am + wt + qsec) however, is able to confidently say that manual transmission is associated with a 2.94 increase in mpg (95% confidence interval: 0.04 - 5.83) and with an adjusted $R^2$ of

## Exploratory Analysis

Before we start building the model, we will first plot the data (see Appendix, Figure 2) to see what kind of data we are dealing with (factors, continuous or discrete), whether any variables look like good predictors of mpg and whether there appears to be correlation between any of the variables. From this plot we can see that a few variables are discrete (cyl (number of cylinders), gear (number of gears) and carb (number of carburetors)) and some are factors including the variable of interest, am (Transmission (0 = automatic, 1 = manual)) and vs (Engine (0 = V-shaped, 1 = straight)) with the remaining variables being continuous. We can see that cyl, disp, hp and wt appear to have a negative correlations with mpg while drat and qsec appear to have positive correlations with mpg. Our variable of interest, am, suggests that manual (encoded as 1) cars have a higher mpg. The correlation matrix (Appendix, Table 6) also shows that a number of variables are strongly correlated such as cyl with disp, hp and wt.

As am is our variable of interest, it is worth looking at the relationships am has with other variables focusing on those with strong correlations (positive or negative based on Appendix, Table 6).

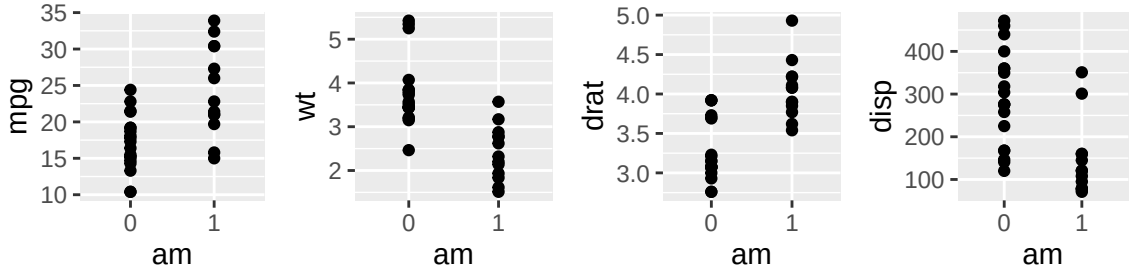


Figure 1: Relationship between am with mpg, wt, drat and disp. 1 represents manual transmission and is associated with higher mpg, lower weight (wt), higher rear axle ratio (drat) and lower displacement (disp).

There appear to be some strong associations between am and not just mpg but some other potentially important variables. We must therefore bear in mind the confounding effects that these relationships will have on modelling mpg with am.

## Initial naive Model building

As we want to answer the question of whether automatic or manual cars are better for mpg, we must build the best model we can that includes am. To start off, we build a simple linear model with only mpg ~ am.

Table 1: Model coefficients for mpg ~ am.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 17.147368 | 1.124602 | 15.247492 | 0.000000 |
| am | 7.244939 | 1.764422 | 4.106127 | 0.000285 |

This suggests that a switch from automatic to manual increases mpg by 7.24. The $R^2$ is however pretty poor at 0.36 suggesting a better model is possible. We've seen that other variables are likely important in predicting mpg and including them should improve the $R^2$ however, some of these possibly confound am. Before being selective about which variables to include, we will build a model with all variables which will likely introduce variation inflation but provides a benchmark against which we can compare the adjusted $R^2$. This model dramatically improves our adjusted $R^2$ to 0.81. However, none of the slopes are actually significant and by including so many variables, variance inflation is likely an issue and may partly explain this. Any increase in variance increases a given confidence interval reducing our ability to find significant coefficients although the overall model can perform better (higher $R^2$). Calculating the variance inflation factors (VIF) on this model with all variables reveals high VIF values (>20x for disp for example) confirming that variance inflation is an issue and we should be selective of correlated variables to avoid it.

## Additive Model Building

We will now start our first iteration of the proper model with mpg ~ am + wt (leave out disp, hp and cyl for now given their correlation with wt). We need to keep am in and the weight of the car is very likely to negatively effect the mpg given basic physics and we have seen it is a possible confounding variable with am so important to include. We will then compare this model using ANOVA with the simple linear model mpg ~ am to ensure it is significantly better.

Table 2: Model coefficients for mpg ~ am + wt.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 37.3215513 | 3.0546385 | 12.2179928 | 0.0000000 |
| am | -0.0236152 | 1.5456453 | -0.0152786 | 0.9879146 |
| wt | -5.3528114 | 0.7882438 | -6.7908072 | 0.0000002 |

Table 3: ANOVA test for mpg ~ am + wt compared to mpg ~ am.

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|
| 30 | 720.8966 | NA | NA | NA | NA |
| 29 | 278.3197 | 1 | 442.5769 | 46.11506 | 2e-07 |

The inclusion of wt not only improves the model significantly (R^2 goes from 0.36 to adjusted R^2 of 0.74) but changes the slope coefficient of am to close to 0 suggesting that the potential confounding effect seen with manual cars being lighter may be important for our interpretation of am and its effect on mpg. As might be expected, wt strongly negatively affects mpg (slope coefficient of -5.4).

We will now add in only one extra variable from the following: drat, qsec, gear and carb to our model of mpg ~ am + wt and see whether any improve the model significantly using anova tests.

ANOVA p-value for mpg ~ am + wt + drat: 0.2849371.

ANOVA p-value for mpg ~ am + wt + qsec: $2.1617371 \times 10^{-4}$.

ANOVA p-value for mpg ~ am + wt + gear: 0.6623234.

ANOVA p-value for mpg ~ am + wt + carb: 0.0080462.

ANOVA p-value for mpg ~ am + wt + vs: 0.0084542.

From these, it appears that qsec, carb and vs improve the model significantly and while this model has a high adjusted $R^2$ of 0.84, neither of the coefficients for carb or vs are significant and so they are removed meaning we will only keep qsec. This model, with coefficients shown below, still has a high adjusted $R^2$ of 0.83.

Table 4: Model coefficients for mpg ~ am + wt + qsec

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.617781 | 6.9595930 | 1.381946 | 0.1779152 |
| am | 2.935837 | 1.4109045 | 2.080819 | 0.0467155 |
| wt | -3.916504 | 0.7112016 | -5.506882 | 0.0000070 |
| qsec | 1.225886 | 0.2886696 | 4.246676 | 0.0002162 |

To this model (mpg ~ am + wt + qsec) we will carry out further ANOVA tests with the variables that showed correlation with wt (cyl, disp and hp) to see if the model is significantly improved by any of their inclusion.

Cyl: slope coefficient p-value is 0.6270601 and the anova test p-value is 0.0010785

Disp: slope coefficient p-value is 0.4717085 and the anova test p-value is $9.342523 \times 10^{-4}$

Hp: slope coefficient p-value is 0.2230879 and the anova test p-value is $5.7109896 \times 10^{-4}$

Each extra variable improves the model significantly, however each ones slope coefficient is non-significant and so we will not include them in the final model.

## Final Model Analysis

Our best model that includes am is mpg ~ am + wt + qsec. This model therefore includes the transmission alongside two variables that logically should affect mpg all else being equal. A greater mass requires more energy (fuel) to move and the shorter the qsec (1/4 mile time), the more fuel is required to propel a given mass. Our coefficients support these interpretations but also do suggest that the transmission of the cars has a significant effect on mpg with a change from automatic to manual associated with a 2.9 increase in mpg. Table 5 shows the 95% confidence intervals for our model coefficients showing that the am coefficient does get close to 0.

Table 5: 95% confidence intervals for final model coefficients.

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -4.6382995 | 23.873860 |
| am | 0.0457303 | 5.825944 |
| wt | -5.3733342 | -2.459673 |
| qsec | 0.6345732 | 1.817199 |

Finally, we should inspect the residual plots for our model (Appendix, Figure 3). Residuals vs Fitted show that there is no clear pattern in the residuals. Often this would indicate that a variable is missing as our model leaves some patterns unexplained.

## Appendix

Here are some of the extra plots and tables referenced in the text for data exploration.

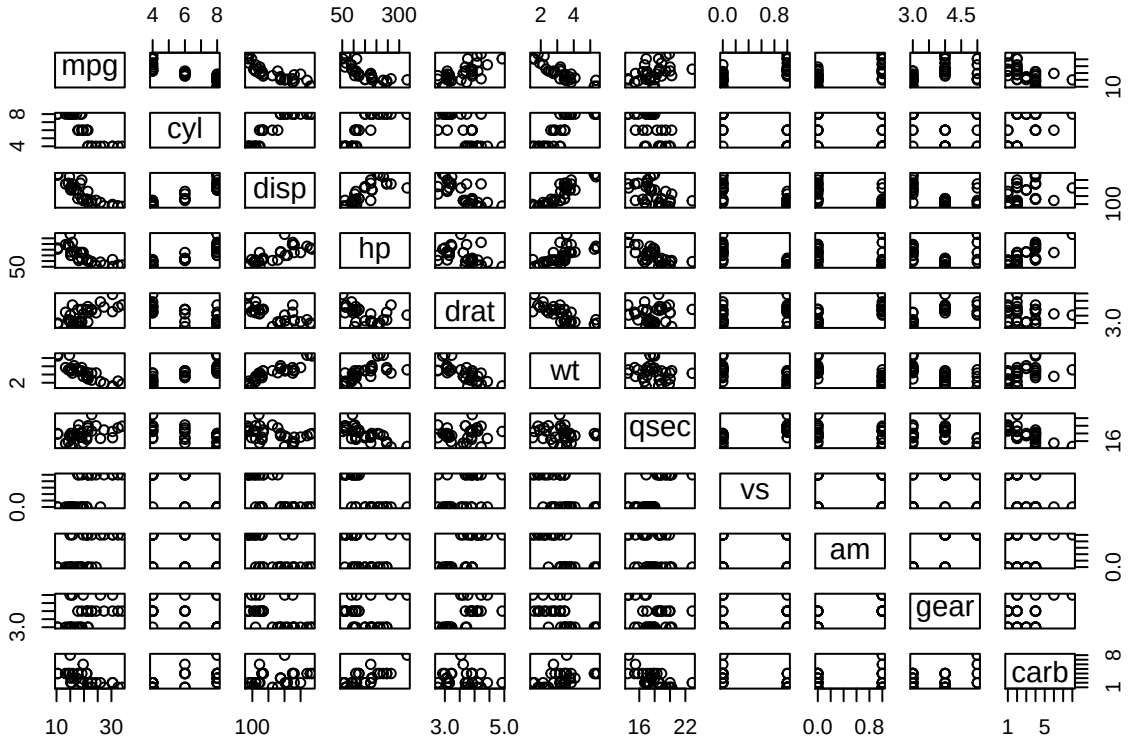


Figure 2: Scatter plots of all variables in the mtcars dataframe.

Table 6: Table of correlations for all variables compared to our dependent variable mpg and the regressor of interest am as well as a potential confounder, wt.

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mpg | 1.00 | -0.85 | -0.85 | -0.78 | 0.68 | -0.87 | 0.42 | 0.66 | 0.60 | 0.48 | -0.55 |
| wt | -0.87 | 0.78 | 0.89 | 0.66 | -0.71 | 1.00 | -0.17 | -0.55 | -0.69 | -0.58 | 0.43 |
| am | 0.60 | -0.52 | -0.59 | -0.24 | 0.71 | -0.69 | -0.23 | 0.17 | 1.00 | 0.79 | 0.06 |

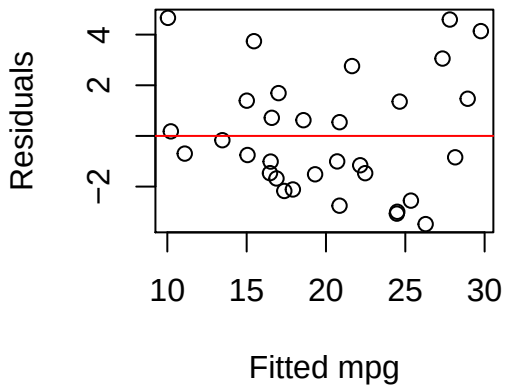


Figure 3: Residual plot of the model mpg + am + wt + qsec