

Comparing Fuel Efficiency of Automatic and Manual Transmission Cars (1973–74 models)

Coursera Regression Models Final Project

Isaac Lello-Smith

2025-12-24

Executive Summary

Based on my regression analysis, among 1973-1974 models included in the 1974 Motor Trend US magazine, manual transmissions were better for fuel-efficiency / MPG. On average, manual transmissions among this set of cars got 14.1 more miles per gallon than automatic transmissions. This impact on MPG was significant at the 0.001 level.

Our final linear model has the formula: $\text{mpg} \sim \text{am} * \text{wt} + \text{qsec}$. Modeling mpg as a function of transmission type, vehicle weight and acceleration speed. The model has an adjusted R-squared of 0.88, and all coefficients were significant at the 0.01 level.

I selected the model after comparing a set of candidates using an ANOVA. I started from the most basic model of $\text{mpg} \sim \text{am}$, then added variables in order of intuitive importance for explaining mpg (while considering collinearity among the variables in the dataset). I also tested models using all variables. The chosen model of $\text{mpg} \sim \text{am} * \text{wt} + \text{qsec}$ delivered some of the best explanatory power without including too many of the available variables.

Analysis

Exploratory Data Analysis

Preview Data

cyl, vs, am and gear will be better represented as categorical variables with 2-3 unique values each.

```
## mpg cyl disp hp drat wt qsec vs am gear carb
## 25 3 27 22 22 29 30 2 2 3 6
```

(Helpful resource: Analysis Factor Categorical or Continuous)

We can likely leave some variables out of our final model as they are correlated and contain similar information. (see Fig. 1 in the Appendix)

- disp, wt, cyl are all correlated with each other.
- hp is also correlated with disp and cyl.

Model Fitting

Logical Preparation

Before we begin fitting models, let's think about the variables we want to hold constant to isolate the impact of transmission.

- wt comes to mind as the most intuitive inclusion. Heavier cars are going to have lower MPG, so we want to make sure we're not biasing our findings if the average automatic transmission vehicle is heavier than manual or vice versa. wt also seems to capture a lot of the same information as disp and cyl.
- qsec seems like a sensible addition, as it accounts for the type of performance a car delivers. If most manual cars are sports cars, we'd expect lower MPG as a trade-off for fast acceleration.
- hp may deliver a similar signal to qsec, though it could account for high hp vehicles that are slow, like farm or other work vehicles. (This would likely already be accounted for by wt to some degree).
- gear and carb have pretty low correlations across the board, so we may want to include them for explanatory power. (See Fig 1 in Appendix)

Least and Most Inclusive Models

The simplest model to get at our insights is comparing average mpg for automatic vs. manual. This model shows manual transmissions have an average MPG that is 7.35 MPG higher than automatic transmissions. With an R-squared of 0.34, this model does not explain the system very well. The residual plot also has more curvature than I'd like.

Including all variables in the model offers better explanatory power, but may muddy the signal around automatic vs. manual transmissions. With all variables included, manual transmissions have an average MPG that is 3.35 MPG higher than automatic transmissions, though the coefficient is not significant. With an R-squared of 0.81 this model explains a lot more of the variability, but we are spamming variables. There is also still some curvature in the residuals plot.

Model Iteration

Let's now build out progressively more complicated models, and use an ANOVA to compare the models.

Including wt removes the impact of automatic or manual transmission, with an R-squared of .74. I don't buy that transmission type doesn't impact MPG, so I added an interaction term. Including the interaction term isolates a major increase in MPG for manual over automatic. Heavier cars reduce MPG in both types of transmissions, with a larger drop for manual vehicles (R-squared 0.82).

In line with my thinking in the Logical Preparation section, I added select variables - qsec; qsec and hp; qsec, gear and carb; and all variables with an interaction term between am and wt. I then compared my various models using an ANOVA.

Based on the ANOVA output (next page), I am going to use Model 3: $\text{mpg} \sim \text{am} * \text{wt} + \text{qsec}$ for this analysis. While the model including all variables and an am*wt interaction term has the smallest RSS, the improvement isn't enough to justify adding 7 more variables in my opinion.

```

## Analysis of Variance Table
##
## Model 1: mpg ~ am + wt
## Model 2: mpg ~ am * wt
## Model 3: mpg ~ am * wt + qsec
## Model 4: mpg ~ am * wt + qsec + hp
## Model 5: mpg ~ am * wt + qsec + gear + carb
## Model 6: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 7: mpg ~ am * wt + (cyl + disp + hp + drat + wt + qsec + vs + am +
##           gear + carb)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     29 278.32
## 2     28 188.01  1    90.312 14.7956 0.001182 **
## 3     27 117.28  1    70.731 11.5878 0.003163 **
## 4     26 117.24  1     0.033  0.0054 0.942339
## 5     24 112.59  2     4.648  0.3808 0.688720
## 6     19 130.05  5    -17.456
## 7     18 109.87  1     20.180  3.3060 0.085709 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Final Model Analysis

The chosen model looks face valid. Heavier cars on average have lower MPG. Cars with slower acceleration tend to have higher MPG. Manual transmission cars have a higher MPG on average (14.1 MPG higher) holding weight and acceleration speed constant.

R-squared is 0.88, and our interaction term for automatic vs. manual transmissions is significant at the 0.001 level.

```

##
## Call:
## lm(formula = mpg ~ am * wt + qsec, data = .)
##
## Residuals:
##       Min     1Q Median     3Q    Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.723     5.899   1.648 0.110893
## am1         14.079     3.435   4.099 0.000341 ***
## wt          -2.937     0.666  -4.409 0.000149 ***
## qsec         1.017     0.252   4.035 0.000403 ***
## am1:wt      -4.141     1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13

```

The diagnostic plots for the chosen model look ok (See Fig 2 in Appendix). The residual plots are generally pretty flat. The QQ plot is a bit messy at the extremes, suggesting more extreme values than the normal distribution. But I'm not too worried about the degree of deviation from 1:1.

(Helpful resource: UVA diagnostic plots)

(Helpful resource: UVA QQ plots)

Appendix

Figure 1

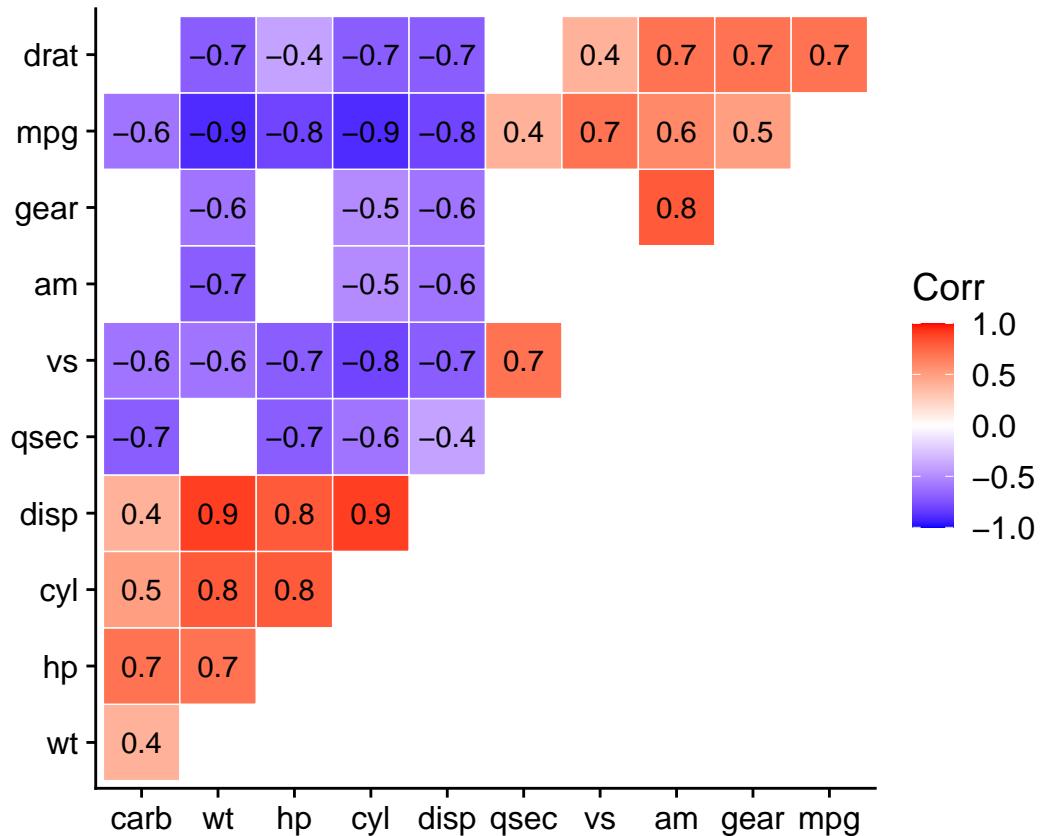


Figure 2

