Car MPG Regression Models

Ellen Tworkoski

2/27/2022

Executive Summary

This document summarizes findings from an investigation into characteristics that are associated with car mileage-per-gallon (mpg). Specifically, we focus on the association between transmission type (automatic vs. manual) and mpg, and consider nine other factors that may confound this association including car weight and number of cylinders. A small dataset containing a sample of 32 cars was used for analysis. Several nested linear models were fit based on results from initial exploratory analyses, and ANOVA tests were used to assess model quality.

Associations between transmission type and mpg were found to be non-significant when holding car weight and number of cylinders constant. Therefore, we find no discernible difference in mpg between automatic and manual transmission when holding other factors constant.

Exploratory Analysis

Initial exploratory analysis shows that the median mpg, among cars in this 32-observation dataset, is higher for cars with a manual transmission (~23mpg) than for cars with an automatic transmission (~17mpg) (Figure 1).

However, additional investigations show that there are several other characteristics that are associated with both mpg and transmission type (Figure 2). For example, cars with a greater number of cylinders typically have a lower mpg. We also observe that most of the cars with higher cylinder counts have automatic transmissions, while most of the cars with lower cylinder counts have manual transmissions. Therefore, it is difficult to determine from these graphical investigations alone whether transmission type, number of cylinders, or some other associated factor is the driving force in car mpg.

Regression Model Selection

Linear models were used to quantify the relationship between mpg and ten car characteristics: transmission type, number of cylinders, displacement, gross horsepower, rear axle ratio, car weight, quarter mile time, engine type, number of forward gears, and number of carburetors. We focused on the relationship between mpg and transmission type with other variables considered as potential confounders.

Previously generated exploratory graphs (Figure 2) indicated that several factors (e.g., cylinder number, car weight) could confound the mpg/transmission association. However, they did not indicate any variation in transmission effect on mpg across different confounding variable values. Because of this, no interaction terms were included in the regression models.

Three models are presented in this document: 1) a model with transmission type (am) as the only independent variable, 2) a model with transmission type, number of cylinders (cyl), and car weight (wt) as independent variables, and 3) a model with all available variables included in the model. Additional models were also fit, but are not presented in this document due to space constraints.

Ultimately, the 'best' model was judged to be the one that a) explained a large amount of the variation in mpg, and b) did not include extraneous variables which inflated standard error estimates and provided

insignificant improvements in model accuracy. Model coefficients and results from ANOVA tests were used to evaluate these criteria. Code for model creation and ANOVA testing is provided below.

```
#Load necessary packages
library(tidyverse)
library(lattice)
library(reshape2)
library(car)
library(ggpubr)
```

```
#Load dataset
car_df <- mtcars

#Convert relevant variables into factor variables
factor_names <- c("cyl", "vs", "am", "gear", "carb")
car_df[,factor_names] <- lapply(car_df[,factor_names], factor)

#Model 1: Transmission type only
am_model <- lm(mpg~am, data=car_df)
#Model 2: Transmission, number of cylinders, and car weight
am_wt_cyl_model <- lm(mpg~am+wt+cyl, data = car_df)
#Model 3: Full model, all variables included
full_model <- lm(mpg~., data = car_df)

#ANOVA to compare models 1, 2, and 3
anova(am_model, am_wt_cyl_model, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + cyl
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     27 182.97  3    537.93 22.3387 8.689e-06 ***
## 3     15 120.40 12     62.57  0.6495    0.7715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test result shows that the additional variables contained in model #2 provide significant improvement in model fit compared to model #1 (p-value < 0.001). However, model #3 does not have a significantly better fit compared to model #2 (p-value $= 0.77$). Therefore, model #2 was selected as the 'best' model to use when drawing conclusions about the effects of transmission type on mpg, accounting for confounding variables.

Conclusions and Limitations

A summary of the parameters of the selected model (model #2), and relevant model diagnostics are presented below.

In brief, we conclude that, when holding car weight and number of cylinders constant, there is a non-significant association (p-value $= 0.91$) between transmission type and mpg. Going from an automatic to a manual transmission results in a 0.15 increase in mpg. This point estimate has a 95% CI of (-2.52, 2.82). In contrast, both car weight and number of cylinders show highly significant associations with mpg (p-values < 0.01).

There are several limitations to this analysis. First, the dataset only contains 32 observations, limiting the

generalizability of these findings. Second, the normal Q-Q diagnostics plot shows that there are some points at the outer edges of the data that deviate slightly from the diagonal pattern, indicating that perhaps the residuals are not entirely normally distributed. This may result in some bias in the standard error estimate and therefore in the calculation of the 95% CI. Finally, we note that the remaining diagnostic plots do not appear to show any outlying data points with particularly high leverage, nor do they indicate any substantial heteroskedasticity in the data.

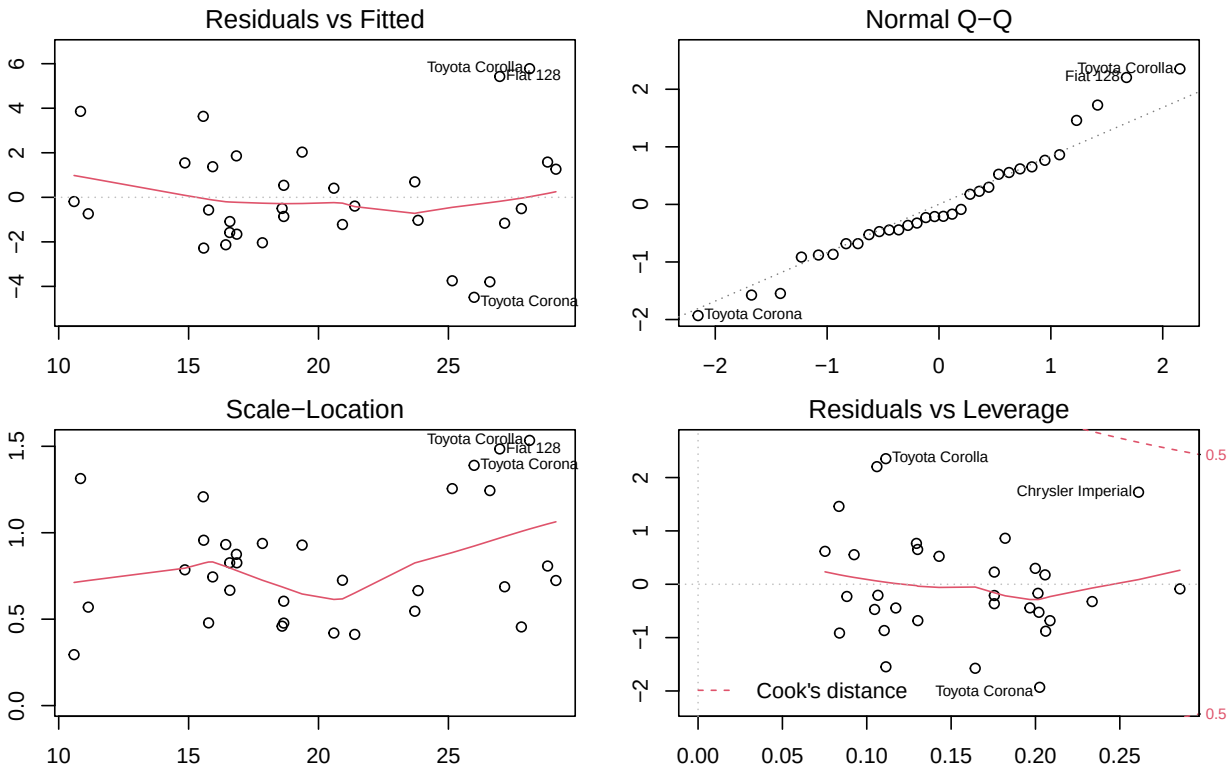
```
summary(am_wt_cyl_model)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl, data = car_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4898 -1.3116 -0.5039  1.4162  5.7758
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7536     2.8135  11.997  2.5e-12 ***
## am1           0.1501     1.3002   0.115  0.90895
## wt           -3.1496     0.9080  -3.469  0.00177 **
## cyl6         -4.2573     1.4112  -3.017  0.00551 **
## cyl8         -6.0791     1.6837  -3.611  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.603 on 27 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8134
## F-statistic: 34.79 on 4 and 27 DF,  p-value: 2.73e-10
```

```
confint(am_wt_cyl_model)
```

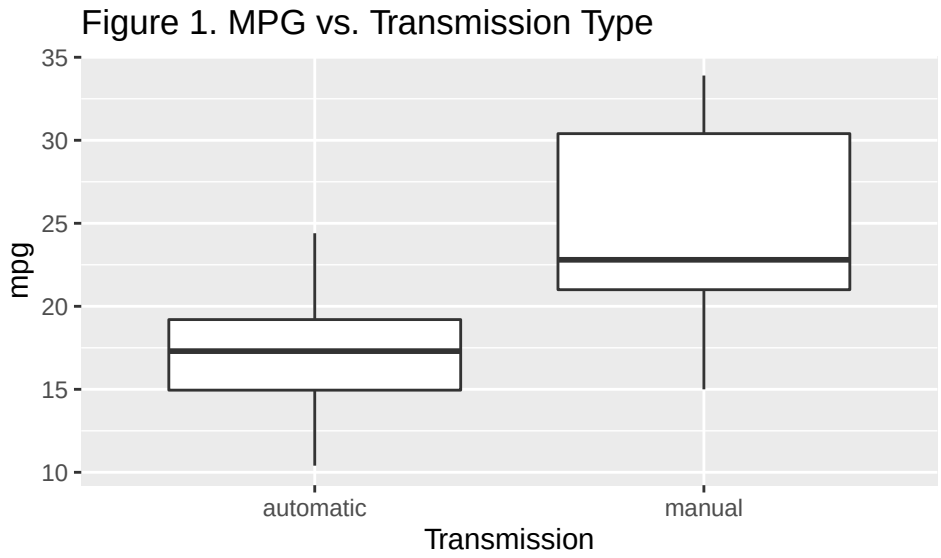
```
##                 2.5 %    97.5 %
## (Intercept) 27.980802 39.526382
## am1         -2.517734  2.817941
## wt          -5.012761 -1.286434
## cyl6        -7.152943 -1.361694
## cyl8        -9.533813 -2.624425
```

```
par(mfrow = c(2,2), mar = c(2,2,2,2))
plot(am_wt_cyl_model)
```



Appendix: Exploratory Code and Figures

```
#Plot transmission type vs. mpg
p1 <- ggplot(data = mtcars, aes(factor(am, levels = c(0,1), labels = c('automatic',
↪  'manual') ), mpg)) +
    geom_boxplot() +
    labs(x = "Transmission", title = "Figure 1. MPG vs. Transmission Type")
p1
```



```r
#Plot other variables vs. mpg, stratified by transmission type
car_df2 <- melt(mtcars, c("mpg", "am"))
p2 <- ggplot(data = car_df2, aes(value, mpg)) +
    geom_point(aes(color = factor(am, levels = c(0,1), labels = c('automatic',
    ↪  'manual'))), alpha = 0.3) +
    facet_wrap(.~variable, scales = "free") +
    labs(color = "Transmission", title = "Figure 2. MPG vs. Car Characteristics,
    ↪  Stratified by Transmission Type") +
    theme(legend.position = "bottom")
p2
```

Figure 2. MPG vs. Car Characteristics, Stratified by Transmission Type