Quantifying the Effect of Transmission Type on Gas Mileage Using Linear Regression

Connor Claypool

29 June 2018

Executive Summary

This analysis used data from the 1974 *Motor Trend* magazine and a linear regression model to estimate the effect of transmission type on gas mileage (MPG), while taking into account the effect of other relevant variables. The estimate of the model is that, given a specific 1/4 mile time and weight,

$$\text{manual MPG} = \text{automatic MPG} + 14.1 - 4.1 \times \text{weight}$$

where weight is given in thousands of pounds. According to this rule, manual cars have better MPG at weights lower than about 3400 lbs, while automatic cars perform better at greater weights. Additionally, this difference is greater the farther the weight is from 3400 lbs. However, there is a fair amount of uncertainty in the model's estimate - the margin of error for the number 14.1 is around 7, and the margin of error for the number 4.1 is around 2.5. This means that the relationship as estimated by the model is far from definite.

1. Introduction

The aim of this analysis was to use a linear regression model to quantify the effect of transmission type on gas mileage (MPG) while adjusting for the effect of other variables. The data used was sourced from the `mtcars` dataset, details of which are given in the **Data** section.

2. Data

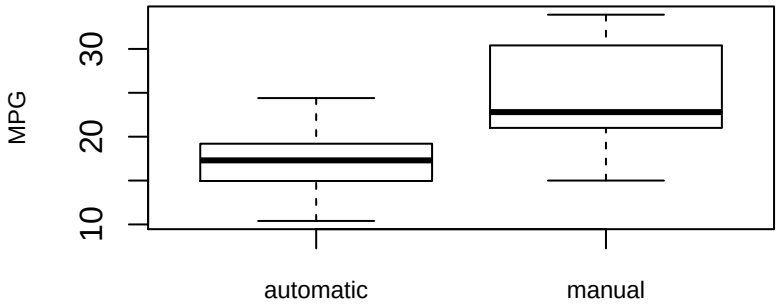
The `mtcars` dataset contains data from the 1974 *Motor Trend* magazine. It consists of the following observations for 32 1973-74 car models:

- `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 (seconds)
- `vs`: V/S
- `am`: transmission (0 = automatic, 1 = manual)
- `gear`: number of forward gears
- `carb`: number of carburetors

3. Analysis

The difference between the distributions of MPG values based on transmission type is clearly demonstrated in the box plot below.

Figure 1: MPG by Transmission Type



However, this plot conveys no information as to the effect of confounding variables on this difference. To determine the influence transmission type has after adjusting for the impact of other relevant variables, a linear regression model was fit using transmission type, weight, the interaction of these, and quarter-mile time as predictors of MPG. These features were selected using a backwards elimination model selection strategy which is detailed in Appendix A. The coefficients of this model and their corresponding p-values are shown in the table below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.7231	5.8990	1.6482	0.1109
am	14.0794	3.4353	4.0985	0.0003
wt	-2.9365	0.6660	-4.4090	0.0001
qsec	1.0170	0.2520	4.0354	0.0004
am:wt	-4.1414	1.1968	-3.4603	0.0018

Table 1: Model Coefficients

The fitted model has an adjusted R-squared of 0.8804, indicating a fairly high goodness of fit. Additionally, the table below, which summarizes the distribution of hat values for this model, shows that the model is not distorted by any extreme outliers.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05	0.09	0.13	0.16	0.22	0.37

Table 2: Distribution of Hat Values

Finally, plotting the residuals against the fitted values and the theoretical normal quantiles shows that they do not follow any clear pattern and are approximately normally distributed, in line with the assumptions of linear regression.

Figure 2: Residuals vs Fitted

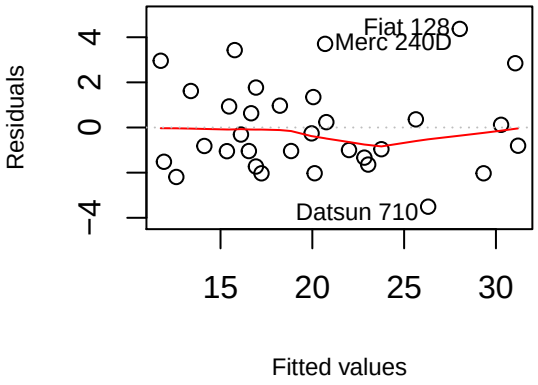
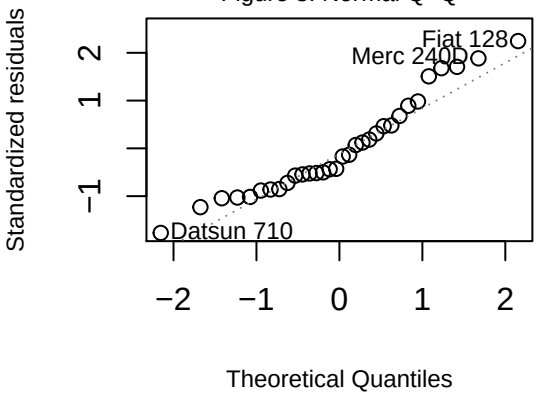


Figure 3: Normal Q−Q



4. Results

According to the coefficients estimated by the fitted model, the difference in mean MPG based on transmission type, for a specific weight and quarter mile time, can be represented by the equation

$$\text{MPG}_{manual} = \text{MPG}_{auto} + 14.0794 - 4.1414 \times \text{weight}$$

This equation indicates that manual transmissions are associated with higher MPG at weights lower than approximately 3400 lbs, with the difference becoming less pronounced the closer the weight is to this value. At greater weights, automatic transmissions are associated with higher MPG, with this difference becoming larger as weight increases. However, the 95% confidence intervals for the two model coefficients used in the above equation are wide; the first is between 7.0309 and 21.128, and the second is between -6.597 and -1.6857. This means there is a significant amount of uncertainty in this estimated relationship between transmission type and gas mileage.

Appendix A: Model Selection Procedure

To decide which variables to include as predictors, a simple backwards elimination technique was used. Firstly, a model was fit using all available predictors. Ignoring the intercept, the variable with the least significant p-value was noted, and a new model was fit without this variable. These steps were repeated until the highest p-value was less than the chosen threshold of 0.05. This was achieved with the following R code.

```
feature_select <- function() {
    data <- mtcars
    repeat {
        model <- lm(mpg ~ ., data)
        coef <- coefficients(summary(model))[-1,]
        max.p <- max(coef[,4])
        if(max.p < 0.05) {
            print(paste("Selected features:",
                        paste(rownames(coefficients(summary(model)))[-1],
                              collapse=" ")))
            return(model)
        }
        var.max.p <- rownames(coef)[which.max(coef[,4])]
        data <- data[setdiff(names(data), var.max.p)]
    }
}
```

```
model1 <- feature_select()
```

```
## [1] "Selected features: wt qsec am"
```

Next, this model was tested against models involving interactions. Three new models were fit, in which transmission type (`am`) interacts with weight (`wt`), quarter mile time (`qsec`) or both. The R code below was used to fit these models and create a table to compare them with the model which includes no interaction terms.

```
models <- list(
    both = lm(mpg ~ am * wt + am * qsec, mtcars),
    qsec = lm(mpg ~ am * qsec + wt, mtcars),
    wt = lm(mpg ~ am * wt + qsec, mtcars),
    neither = model1
)

varnames <- rownames(coefficients(summary(models$both)))
adj.r.squared <- sapply(models,
                        function(model) summary(model)$adj.r.squared)
p.value.anova <- sapply(models,
                        function(model) anova(model1, model)[["Pr(>F)"]][-1])
p.values <- t(sapply(models,
                     function(model) coefficients(summary(model))[,4][varnames]))
comparison_table <- cbind(adj.r.squared, p.value.anova, p.values)
print(xtable(comparison_table, digits = 4), comment = F)
```

	adj.r.squared	p.value.anova	(Intercept)	am	wt	qsec	am:wt	am:qsec
both	0.8767	0.0077	0.1197	0.4872	0.0002	0.0048	0.0200	0.6756
qsec	0.8532	0.0384	0.0311	0.0814	0.0000	0.0198		0.0384
wt	0.8804	0.0018	0.1109	0.0003	0.0001	0.0004	0.0018	
neither	0.8336		0.1779	0.0467	0.0000	0.0002		

This table gives the adjusted R-squared, the p-value calculated by the `anova` function (an indication of whether the addition of the interaction term(s) is statistically significant), and the p-values of each relevant coefficient. The model in which transmission type interacts with weight only was chosen as the final model for various reasons. It has the highest adjusted R-squared value, a very low p-value as calculated by `anova`, and is the only model which includes interaction whose coefficients' p-values are all below the 0.05 significance threshold.

Appendix B: Code for Reproducing the Figures and Tables

Figure 1: MPG by Transmission Type

```
data(mtcars)

boxplot(mpg ~ am, mtcars, xaxt = "n", ylab = "MPG", cex = 0.75, cex.lab = 0.75)
axis(side = 1, at = c(1, 2), labels = c("auto", "manual"), cex.axis = 0.75)
title(main = "Figure 1: MPG by Transmission Type",
      line = 0.75, cex.main = 0.75, font.main = 1)
```

Table 1: Model Coefficients

```
require(xtable)
data(mtcars)

model <- lm(mpg ~ am * wt + qsec, mtcars)

print(xtable(model, digits = 4, caption = "Model Coefficients"), comment = F)
```

Table 2: Distribution of Hat Values

```
require(xtable)
data(mtcars)

model <- lm(mpg ~ am * wt + qsec, mtcars)
hatvalue_summary <- t(as.matrix(summary(hatvalues(model))))

print(xtable(hatvalue_summary, caption = "Distribution of Hatvalues"),
      comment = F, include.rownames = F)
```

Figure 2: Residuals vs Fitted and Figure 3: Normal Q-Q

```
data(mtcars)

model <- lm(mpg ~ am * wt + qsec, mtcars)

par(mfrow = c(1, 2))
plot(model, which = c(1,2), caption = c("Figure 2: Residuals vs Fitted",
                                        "Figure 3: Normal Q-Q"),
     font.main = 1, cex.lab = 0.75, cex.caption = 0.75)
```