Regression Models Course Project

Brian Hampton

2022-06-11

Executive Summary

This report addresses questions posed in the Johns Hopkins Coursera Regression Models course project:

Working for Moter Trend, a magazine about the automobile industry, analyze the collection of cars in the mtcars data set to answer:

1. Is an automatic or manual transmission better for MPG?

2. Quantify the MPG difference between automatic and manual transmissions.

Manual transmission cars have a better MPG rating. A univariate model, which only accounts for 36 percent of the observed variability, predicts manual transmission cars to be better than automatic transmission cars by 7.2 gallons. A multivariate model, which accounts for 85% of the observed variability predicts manual transmission cars to be better than automatic transmission cars by 2.9 mpg. Further investigation into the selection of cars within the mtcars data set may reveal additional factors that contributed to the quantified results.

Exploratory Data Analysis

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(knitr)
library(datasets)
library(dplyr)
data("mtcars")
```

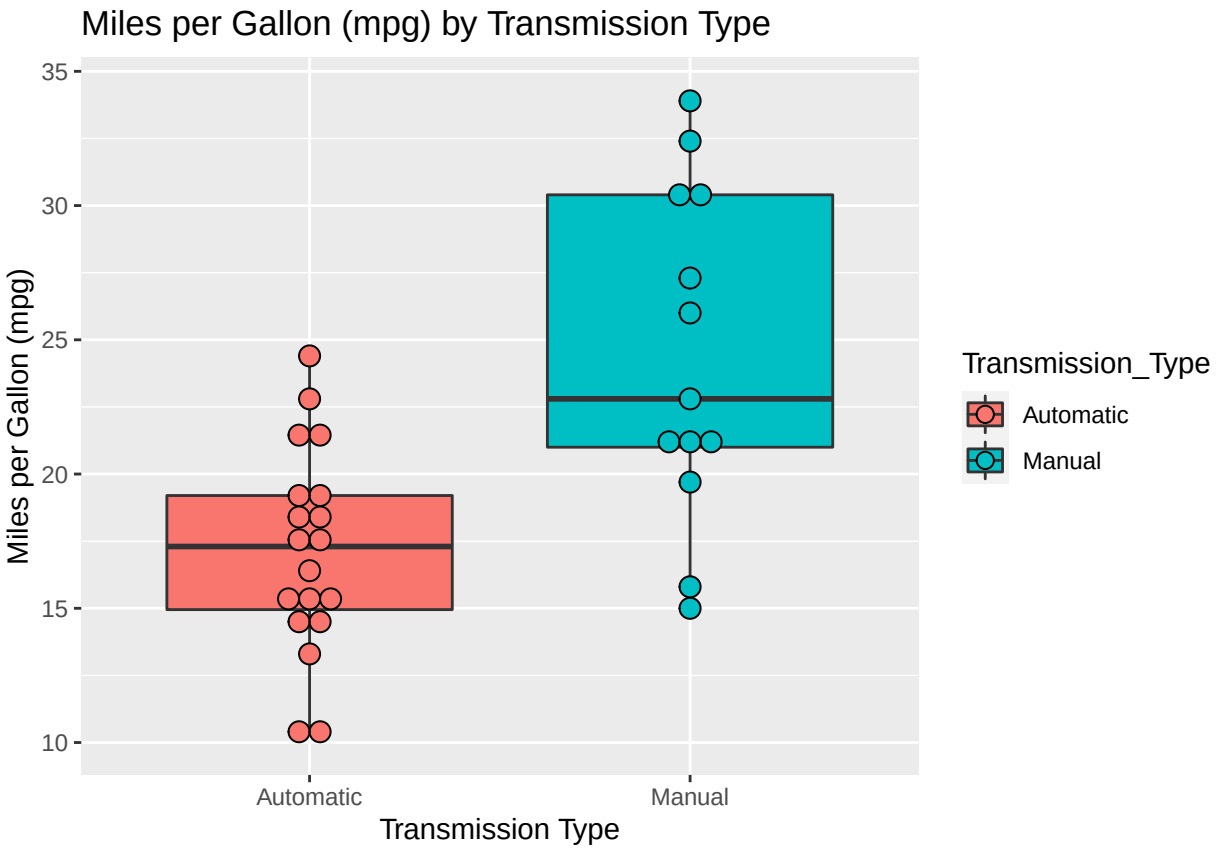
NOTE: code not shown in the report body is included in the appendix

After looking at the mtcars data set, for easier interpretation I added a variable *Transmission* with responses of *Automatic* or *Manual* based on the *am* variable responses.

Table 1. Summary of Miles per Gallon (mpg) by Transmission Type

Transmission_Type	n	min	q1	median	mean_mpg	q3	max	sd_mpg
Automatic	19	10.4	14.95	17.3	17.15	19.2	24.4	3.83
Manual	13	15.0	21.00	22.8	24.39	30.4	33.9	6.17

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with `binwidth`.
```



The summary table for mpg by transmission and the visual representation of that data suggests that manual transmission cars within the data set have a higher mpg.

Inference

Null Hypothesis

H_o: there is no difference in mean mpg between automatic and manual transmission cars.

H_a: the mean mpg for manual transmission cars is greater than the mean mpg for automatic transmission cars.

Univariate Regression Model

```
uv <- lm(mpg ~ Transmission_Type, data = df)
summary(uv)
```

```
##
## Call:
## lm(formula = mpg ~ Transmission_Type, data = df)
##
## 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 ***
## Transmission_TypeManual     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
```

Considering only the relationship between mpg and transmission type, manual transmission cars are more fuel efficient, by about 7.2 mpg. However, the R-squared value is 0.3598, meaning that only 36 percent of the observed variability is explained by this univariate regression model.

Multivariate Regression Model

```
mv <- step(lm(mpg ~ ., data=df),direction="both", trace=FALSE)
summary(mv)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + Transmission_Type, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                9.6178     6.9596   1.382 0.177915
## wt                        -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec                       1.2259     0.2887   4.247 0.000216 ***
## Transmission_TypeManual    2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

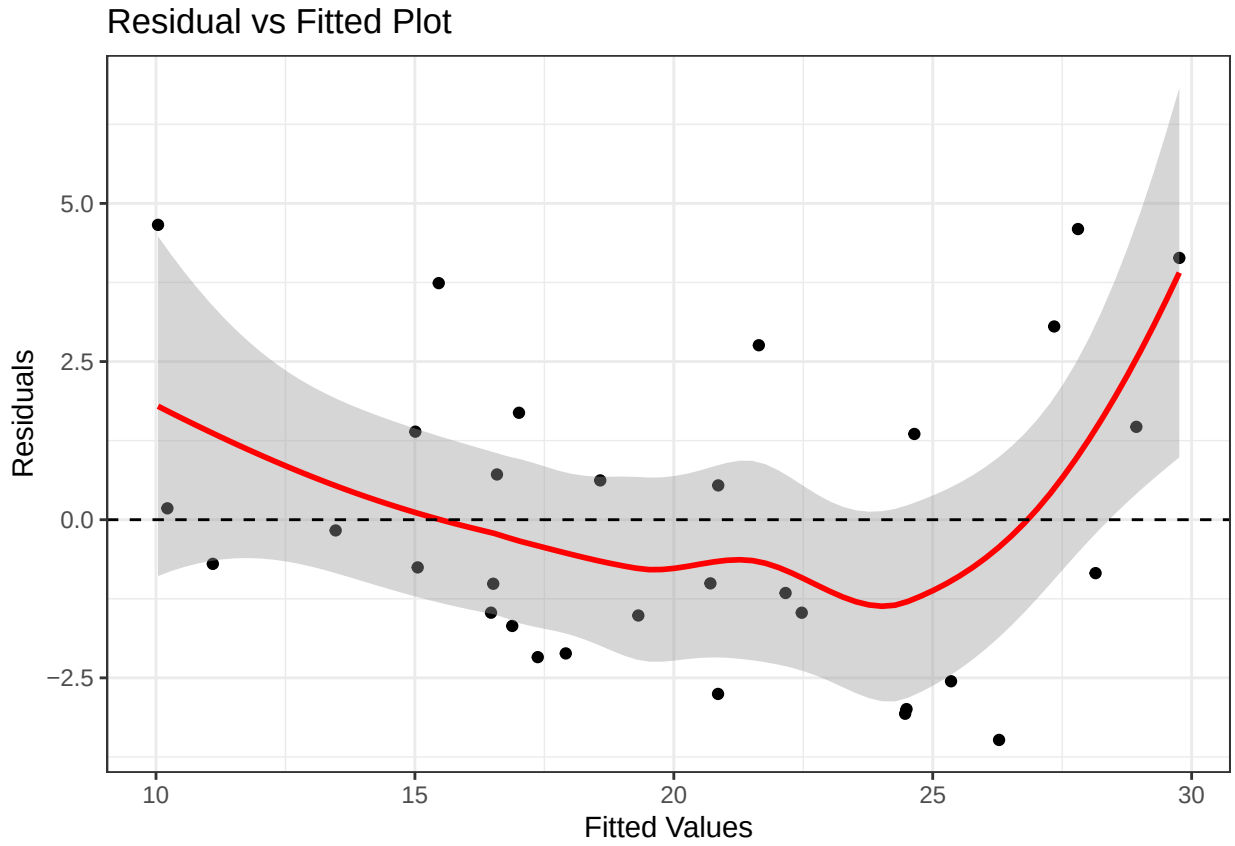
```
mv_a <-lm(mpg ~ I(wt - mean(wt)) + I(qsec - mean(qsec)) + Transmission_Type, data = df)
summary(mv_a)$coef
```

```
##                           Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)              18.897941  0.7193542 26.270704 2.855851e-21
## I(wt - mean(wt))         -3.916504  0.7112016 -5.506882 6.952711e-06
## I(qsec - mean(qsec))      1.225886  0.2886696  4.246676 2.161737e-04
## Transmission_TypeManual   2.935837  1.4109045  2.080819 4.671551e-02
```

The multivariate model indicates that weight (wt), and quarter-mile time (qt) are confounding variables to the transmission type and mpg relationship within this data set. With an R-squared value 0f 0.8497, 85% of the observed variability in the data set is explained by this multivariate model. With this model a manual transmission car is expected to have an mpg rating of 2.9 higher/better than an automatic transmission car. For a mean weight car with a mean quarter-mile time, the model expects a manual transmission car to have an mpg of 21.8 and an automatic transmission car to have an mpg of 18.9.

Residual Analysis

```
## `geom_smooth()` using formula 'y ~ x'
```



The residuals appear to "bounce randomly" around the 0 line, suggesting that it is reasonable to assume a linear relationship. The residuals roughly form a "horizontal band" around the 0 line except for the very highest observed mpg observations, suggesting that the majority of the error terms are equal. No one residual appears to stand out from the pattern of residuals, suggesting that there are no outliers.

Uncertainty in Results

Given report length constraints I recommend further investigation to better understand the data set. Specifically why are 7 of the 13 manual transmission cars lighter than the lightest manual transmission car. And conversely, why are 8 of the 19 automatic transmission cars heavier than the heaviest manual transmission car. When looking at a subset of mtcars bounded by the lightest automatic transmission car and the heaviest manual transmission car, the difference in mpg by transmission type is not as discernible.

Table 2. Summary of MPG by Transmission Type for Data Bounded by Lightest Automatic and Heaviest Manual Transmission Cars

Transmission_Type	n	min	q1	median	mean_mpg	q3	max	sd_mpg
Automatic	11	14.3	16.65	18.70	18.99	21.45	24.4	3.27
Manual	6	15.0	16.78	20.35	18.98	21.00	21.4	2.85

Conclusion

1. Manual transmission cars attained a higher mpg within the entire mtcars data set.
2. A model that considers weight and quarter-mile time as confounding variables over the entire mtcars data set accounts for more of the observed variability within the data set.

Other factors not included in the variables, such as manufacturing trends or people's preferences, may have played a role in the selection of the mtcars data set and influenced the results.

Appendix

```
knitr::opts_chunk$set(echo = TRUE, results = "hide")
### Initial look at mtcars
display <- head(mtcars, 2)
```

```
mtcars$Transmission_Type[mtcars$am==0] <- "Automatic"
mtcars$Transmission_Type[mtcars$am==1] <- "Manual"
df <- mtcars[,-9, drop=FALSE]
```

```
### Summary table 1 for mtcars mpg by Transmission
table_1 <- df %>%
    group_by(Transmission_Type) %>%
    summarise(n = n(),
              min = min(mpg),
              q1 = quantile(mpg, 0.25),
              median = median(mpg),
              mean_mpg = mean(mpg),
              q3 = quantile(mpg, 0.75),
              max = max(mpg),
              sd_mpg = sd(mpg))
kable(table_1, digits = 2)
```

```
### box plot 1 for mtcars mpg by Transmission
bp1 <- ggplot(df, aes(x=Transmission_Type, y=mpg, fill=Transmission_Type,
                          group=Transmission_Type)) + geom_boxplot()
bp1 <- bp1 + geom_dotplot(binaxis='y', stackdir='center', dotsize=1)
bp1 <- bp1 + labs(x="Transmission Type", y="Miles per Gallon (mpg)",
title="Miles per Gallon (mpg) by Transmission Type")
```

```
### Residual vs Fitted Plot for the multivariate (mv) model
ra <- ggplot(mv, aes(.fitted, .resid)) + geom_point()
ra <- ra + stat_smooth(method="loess", col="red") + geom_hline(yintercept=0, linetype="dashed")
ra <- ra + xlab("Fitted Values") + ylab("Residuals")
ra <- ra + ggtitle("Residual vs Fitted Plot") + theme_bw()
```

```r
### Summary table 2: subset mtcars bounded by lightest automatic and heaviest manual
### and publish summary table of mpg by transmission type
auto <- subset(df, Transmission_Type == "Automatic")
manual <- subset(df, Transmission_Type == "Manual")
df2 <- subset(df, wt >= min(auto$wt))
df2 <- subset(df2, wt <= max(manual$wt))
table_2 <- df2 %>%
    group_by(Transmission_Type) %>%
    summarise(n = n(),
              min = min(mpg),
              q1 = quantile(mpg, 0.25),
              median = median(mpg),
              mean_mpg = mean(mpg),
              q3 = quantile(mpg, 0.75),
              max = max(mpg),
              sd_mpg = sd(mpg))
kable(table_2, digits = 2)
```