Regression Models Project

Howard Murray

August 29, 2024

Executive Summary

As part of my assignment for *Motor Trend* magazine, I was tasked with reviewing a data set on a collection of cars and exploring the relationship between a set of variables and their impact on miles per gallon (MGP). Of particular interest are the following two questions:

1. Is an automatic or manual transmission better for MPG?
2. What is the average difference in MPG between automatic and manual transmissions?

Using a combination of inferential statistics and linear regression I determined that:

- **Manual Transmission is better for MPG.**
- **The average difference in MPG between vehicles observed with manual and automatic transmissions is 7.245 mi/gal.**

Further analysis to support my conclusions are listed in the remaining pages below.

Exploratory Analysis of the mtcars Data Set

```
data("mtcars")
str(mtcars)
```

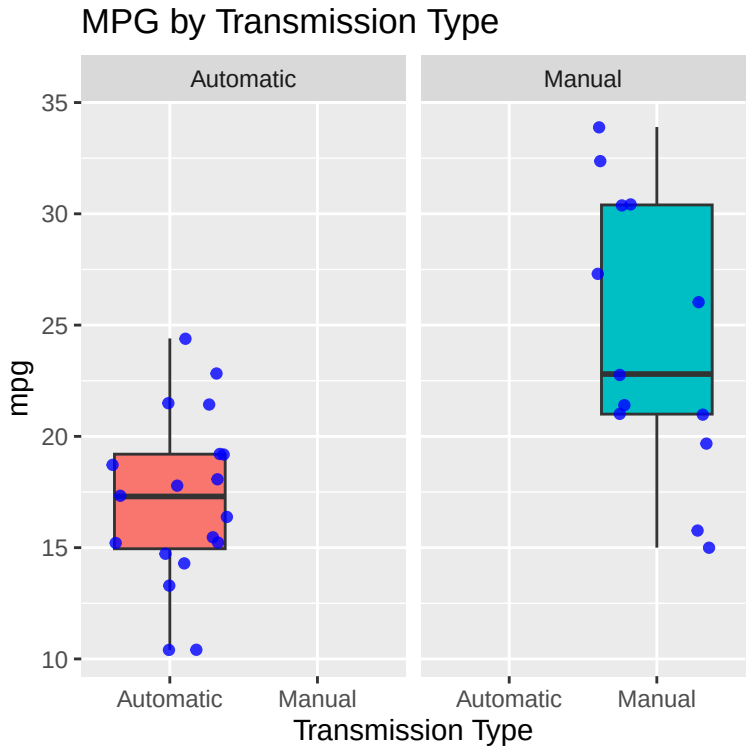
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
# Modify some of the numerical variables into factors to make for meaningful analysis.
mtcars <- within(mtcars, {
  vs <- factor(vs, labels = c("V", "S"))
  am <- factor(am, labels = c("Automatic", "Manual"))
  cyl <- ordered(cyl)
  gear <- ordered(gear)
  carb <- ordered(carb)
})
```

Initially, all of the variables were numeric. For variables like hp (horsepower), qsec (1/4 mile time in seconds) and wt (weight), this makes sense, but for other variables like 'vs', 'am', 'cyl', 'gear', and 'card', it makes more sense to consider them factors. From the boxplot, we can see that the vehicles with manual transmission have higher MPG on average.

```
library(tidyverse)
library(hrbrthemes)
library(viridis)

# Generate boxplot to see relationship between MPG and Automatic vs. Manual transmission
mtcars %>%
  ggplot(aes(am, mpg, fill = am)) +
  geom_boxplot() +
  #scale_fill_viridis(discrete = TRUE, alpha = 0.5) +
  geom_jitter(color = "blue", alpha = 0.8) +
  #theme_ipsum() +
  facet_wrap(~am) +
  theme(
    legend.position = "none") +
  ggtitle("MPG by Transmission Type") +
  xlab("Transmission Type")
```




```
autoData <- mtcars[mtcars$am == "Automatic",]
manualData <- mtcars[mtcars$am == "Manual",]
amdiff <- round(mean(manualData$mpg) - mean(autoData$mpg), 3)
```

The average difference in MPG between cars with manual transmissions versus those with automatic transmission is given by $MPG_{manual} - MPG_{auto} = 7.245$. Is this difference statistically significant? I employed the **t.test** function to validate this difference. The resulting **p-value = 0.0006868** is less than 0.05, therefore we reject the null hypothesis and confirm that the difference in means of the two populations

(automatic and manual transmissions) is statistically significant.

```
t.test(manualData$mpg, autoData$mpg, alternative = c("greater"))
```

```
##
##  Welch Two Sample t-test
##
## data:  manualData$mpg and autoData$mpg
## t = 3.7671, df = 18.332, p-value = 0.0006868
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.913256      Inf
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

Regression Analysis

It was just proven that that a relationship exists between mpg and transmission (automatic vs. manual). This is displayed via linear regression using the **lm** function expressed in the form: $MPG = \beta_0 + \beta_1 * am$ or $MPG = 17.147 + 7.245am$.

```
library(ggplot2)

# Linear regression model relating mpg to the variable am.
fit <- lm(mtcars$mpg ~ mtcars$am)
summary(fit)
```

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$am)
##
## 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 ***
## mtcars$amManual    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
```

```
# Gather residuals
e <- resid(fit)
# Sum of residuals should equal or be nearly zero
resid.sum <- sum(e)
resid.sum
```

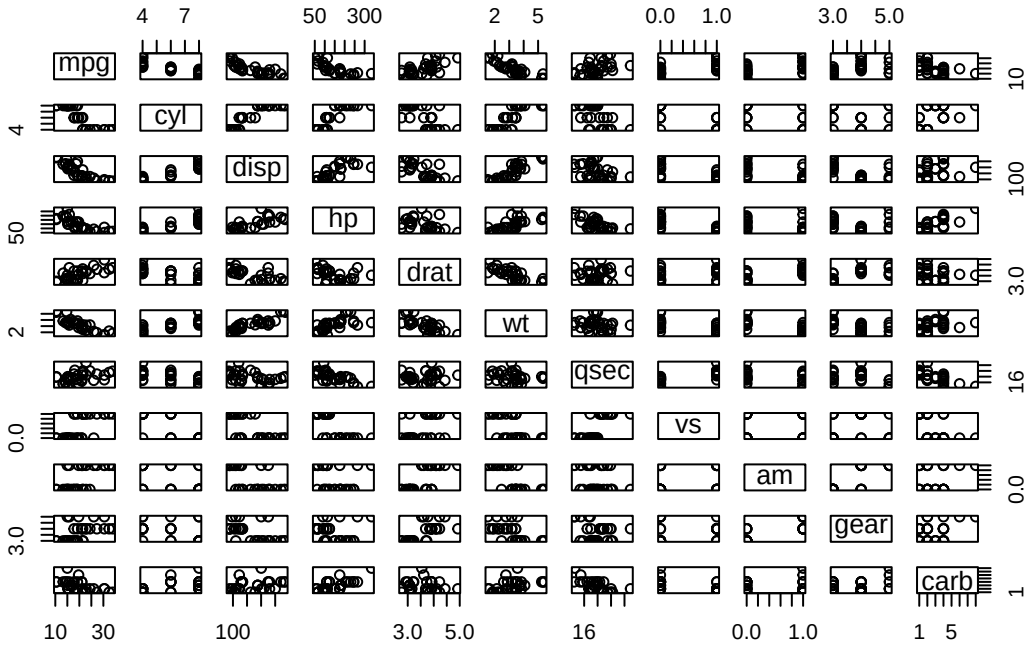
```
## [1] -7.882583e-15
```

One property of residuals is that the sum of the residuals equals zero if an intercept is included. This model

passes this check as the sum of residuals equals $-7.8825835 \times 10^{-15}$ (very, very near zero). $Pvalue_{am} = 0.000285$ says that the transmission type (Automatic vs. Manual) is a significant variable in the MPG model. While this model shown to provide some statistical support for MPG performance, an $R^2 = 0.3598$ says that only 36% of the MPG's variability is explained by this model. We need a better model that uses more of the variables on the data set.

Similar to the first model, we should plot the data to see if we can visualize any relationships between the output (MPG) and the various inputs variables. **pairs** is a good way to see multiple relationships simultaneously.

```
data("mtcars")
pairs(mtcars)
```




```
# Compute the correlation for variables in mtcars to see how they align with the visualization.
sort(cor(mtcars)[1,])
```

```
##         wt        cyl       disp         hp       carb       qsec       gear
## -0.8676594 -0.8521620 -0.8475514 -0.7761684 -0.5509251  0.4186840  0.4802848
##         am         vs       drat        mpg
##  0.5998324  0.6640389  0.6811719  1.0000000
```

Multiple Regression

Looking at the plot, one can observe some potential negative and positive relationships between MPG and the other variables. For example: cyl, disp, hp, and wt all appear to have a negative correlation to MPG. While drat and qsec appear to have a positive correlation to MPG. There are also intuitive relationships, like as horsepower increases, qsec decreases. The correlation values support what is displayed in the graph. Start with a model using all of the provided variables as inputs to explain the output, MPG.

```
library(car)
fit2 <- lm(mpg ~., mtcars)
```

```r
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```r
vif(fit2)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs        am
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487
##      gear      carb
##  5.357452  7.908747
```

When I use all of the variables in the data set, I get a very non-functional model as all of the variable p-values are much greater than 0.05. It is apparent that there was a high degree of correlation between some of the predictor variables as mentioned in the paragraph above. Computing the Variable Inflation Factor (VIF) of each variable confirmed this suspicion. The rule of thumb is that any variable with VIF > 5 is red flag pointing to potentially severe correlation between predictor variables. **cyl**, **disp**, **hp**, **wt**, **qsec**, **gear**, and **carb** each have VIF above this threshold. Let's remove them from our model and see how the new model stacks up.

```r
#New model after removing the highly correlated predictor variables
fit3 <- lm(mpg ~ drat + vs + am, mtcars)

#Compute summary statistics of new model
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ drat + vs + am, data = mtcars)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -5.9892 -2.6090  0.2629  2.1127  6.2924
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.327      6.017   1.384 0.177316
## drat           1.985      1.883   1.054 0.300772
## vs             6.235      1.421   4.387 0.000148 ***
## am             4.669      1.838   2.540 0.016898 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.485 on 28 degrees of freedom
## Multiple R-squared:  0.6981, Adjusted R-squared:  0.6657
## F-statistic: 21.58 on 3 and 28 DF,  p-value: 1.922e-07
```

```
#Compute VIF of new model
vif(fit3)
```

```
##     drat       vs       am
## 2.587587 1.310339 2.146837
```

This model shows much more promise than the model with all of the variables included. The p-values for **vs** and **am** are well below 0.05 and the **Adjusted R-squared** value of 0.6657 is much improved even over the previous model. Also all of the VIF's in this model are less than 5. The **F-statistic is equal to 21.58** with **28 degrees of freedom** and a **p-value of 1.922e-07**. As a final check, I will perform some residual analysis.

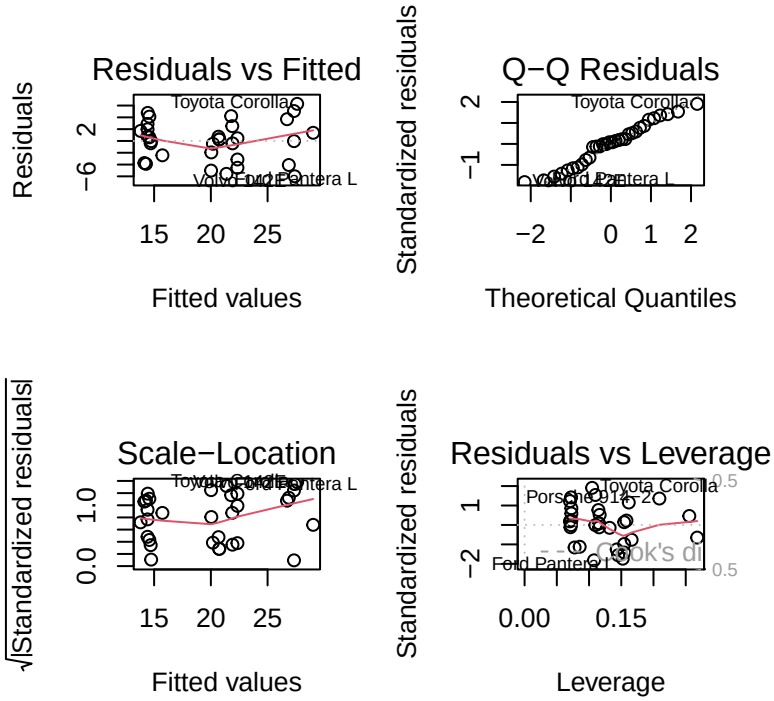
Residual Analysis

Now that the best fitting model has been selected, residual analysis needs to be run to ensure that the residuals are normally distributed, there are no signs heteroskedasticity, outliers or influential data points. I performed these diagnostic checks by plotting the residuals. Observations are listed below.

- **Residuals vs Fitted** - The residuals appear to follow a linear pattern. **GOOD**
- **Q-Q Residuals** - Residuals follow a normal distribution. **GOOD**
- **Scale-Location** - The residuals appear to demonstrate constant variance (no pattern or signs of heteroskedasticity). **GOOD**

- **Residuals vs. Leverage** - All points are within the dashed lines. No signs of influential data points or outliers. **GOOD**

```
par(mfrow = c(2,2))
# Generate residual plots of fit3
plot(fit3)
```



Based on the above, I conclude that this is an acceptable model to describe MPG performance for the mtcars data set.