Regression Models: Motor Trend Analysis

Executive Summary

Motor Trend, a magazine about the automobile industry, is interested in exploring the relationship between the transmission type and miles per gallon (MPG). They are particularly interested in the following two questions:

1. "Is an automatic or manual transmission better for MPG"
2. "Quantify the MPG difference between automatic and manual transmissions"

Based on the analysis below we found that the mean for automatic transmission is 17.15 mpg and manual transmission is 24.39 mpg. Our model also shows that a manual transmission has a greater positive effect on MPG than an automatic transmission.

Exploratory Analysis

The data comes from the Motor Trend Car Road Tests (`mtcars`) dataset with 32 cars and the following 11 attributes:

1. mpg = Miles/(US) gallon
2. cyl = Number of cylinders
3. disp = Displacement (cu.in.)
4. hp = Gross horsepower
5. drat = Rear axle ratio
6. wt = Weight (lb/1000)
7. qsec = 1/4 mile time
8. vs = V/S
9. am = Transmission (0 = automatic, 1 = manual)
10. gear = Number of forward gears
11. carb = Number of carburetors

An initial boxplot comparing MPG and Transmission Type **(see Appendix Plot 1)** would seem to indicate an immediate relationship that manual transmissions get better MPG than Automatic.

When we perform a t-test of mpg for each transmission type the output confirms that this difference is statistically significant (p-value < 0.05). However we need to look at the rest of the attributes to see if there are any other variables we should include in the model.

```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

When we run a scatter plot matrix **(see Appendix Plot 2)** across all attributes we see that there are several that should be investigated.

Regression Analysis

Model Generation & Comparison

We will explore linear regression models based on the different variables. We need to build and compare multiple models to see which one is the best fit:

1. First we build an initial model with all of the variables as predictors.
2. Next we perform stepwise model selection to identify the significant predictors. We use the `step()` function which runs `lm` multiple times with different variables to build multiple regression models and select the best one. It uses both *forward selection* and *backward elimination* methods of the *AIC* algorithm. AIC is a goodness of fit measure that favours smaller residual error in the model, but penalises for including further predictors and helps avoiding overfitting. The results of `step()` shows that the best model contains the variables `cyl`, `wt`, `hp`, and `am` as most relevant. The *adjusted R-squared* value of 0.84 indicates that 84% of the variability/uncertianty is explained by this

3. Finally, we compare the candidate models ANOVA output and Adjusted R-Squared values to verify which model is *best*. The results show that the fit_best model has better p-value and adj.r.squared values.

```
#build multiple models
fit_all <- lm(mpg ~ ., data = mtcars)                    #model with ALL variables
fit_am <- lm(mpg ~ am, data = mtcars)                    #model with just mpg and transmission(am)
fit_best <- step(fit_all, direction = "both", trace=0)   #model generated from step() and chosen as
fit_best2 <- lm(mpg ~ cyl+hp+wt+am, data = mtcars)       #FYI, generating the 'best fit' model manual

#ANOVA compares multiple models and suggests the best fit model using sinificance codes and p-values
anova(fit_all, fit_am, fit_best)

# Adjusted R-Squared tells us how confident we are in the result.
#attributes(summary(fit_all))
summary(fit_all)$adj.r.squared
summary(fit_am)$adj.r.squared
summary(fit_best)$adj.r.squared
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am
## Model 3: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS  Df Sum of Sq       F    Pr(>F)
## 1     15 120.40
## 2     30 720.90 -15   -600.49  4.9874  0.001759 **
## 3     26 151.03   4    569.87 17.7489 1.476e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] 0.7790215
## [1] 0.3384589
## [1] 0.8400875
```

Residuals & Diagnostic Analysis

Lets look and see if we have any outliers in our data sets with high leverage or influence. After analyzing the residuals plots **(See Appendix Plot 3)** we find the following:

1. Residuals vs Fitted plot = points appear random, without any obvious pattern, which seems to verify the independence condition.
2. Normal Q-Q plot = the points mostly fall on the line (using the pencil test) indicating that the residuals are normally distributed.
3. Scale-Location plot = the points are in a constant band pattern, indicating constant variance.
4. Residuals vs Leverage plot = shows some outliers or leverage points are on the top and right edges

We can use a couple of diagnostics to isolate and identify the leverage points.

```
#hatvalues helps us identify the points with potential leverage
lev_best <- hatvalues(fit_best)
tail(sort(lev_best), 3)

#dfbetas helps us identify the points with influence (points that are using their leverage)
inf_best <- dfbetas(fit_best)
tail(sort(inf_best[, 6]), 3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

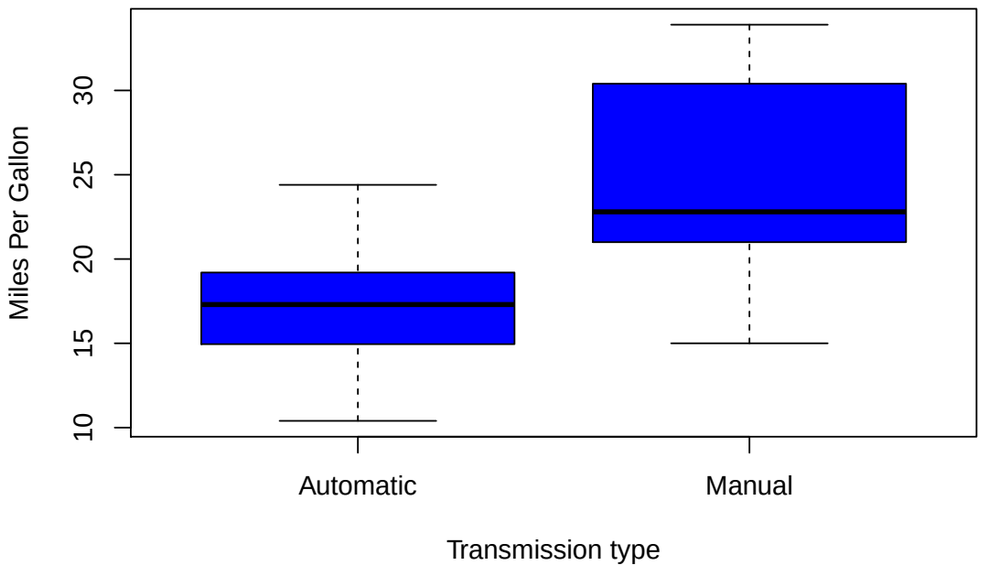
Conclusions

Based on the 'fit_best' model we conclude that:

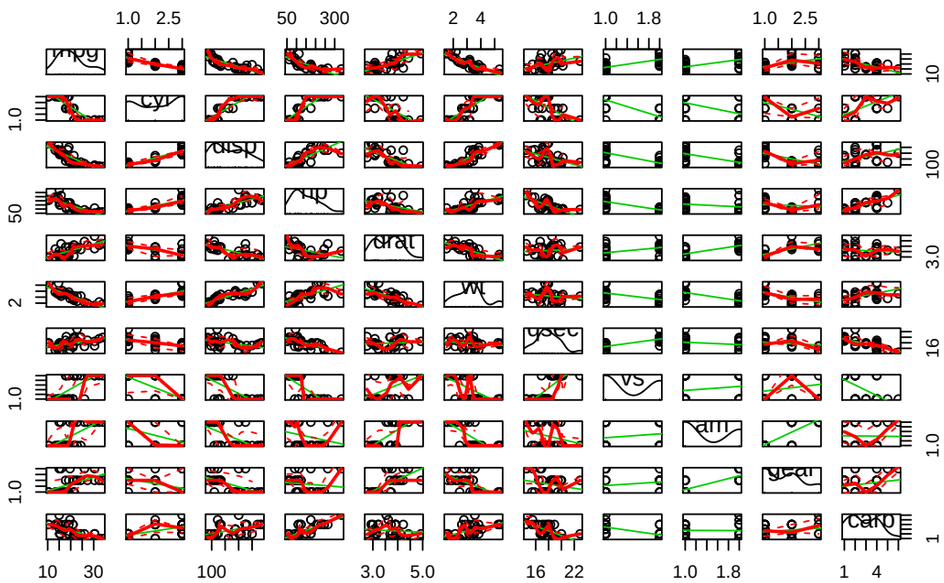
1. MPG will **increase** by 1.81 for cars with **Manual** transmissions.
2. MPG will decrease by 2.5 for every 1000 lb of increase in wt
3. MPG will decrease for 6cyl (by 3.03) and 8cyl (by 2.16) cars
4. MPG will decrease by .03 for each horsepower added.

Appendix / Plots

Plot 1: Miles per gallon by Transmission type

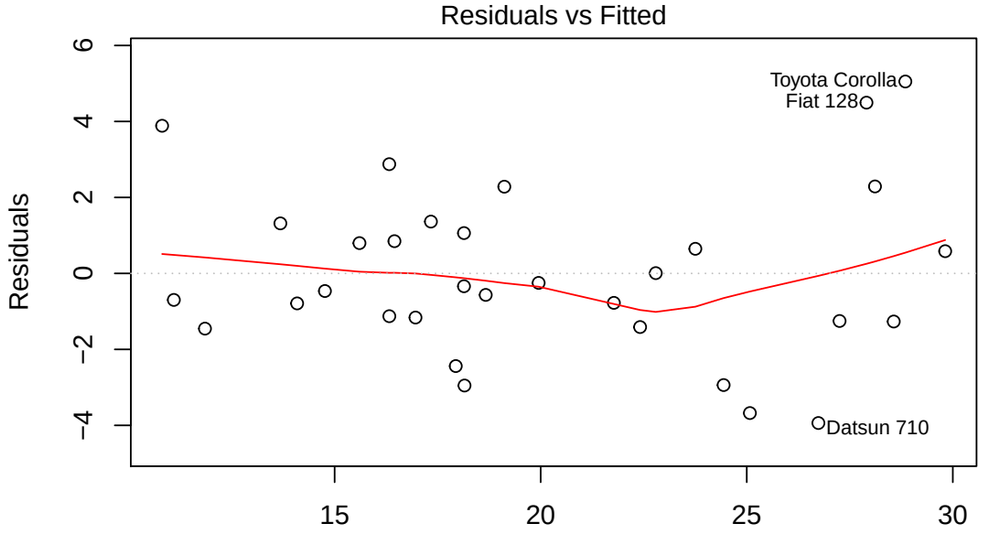


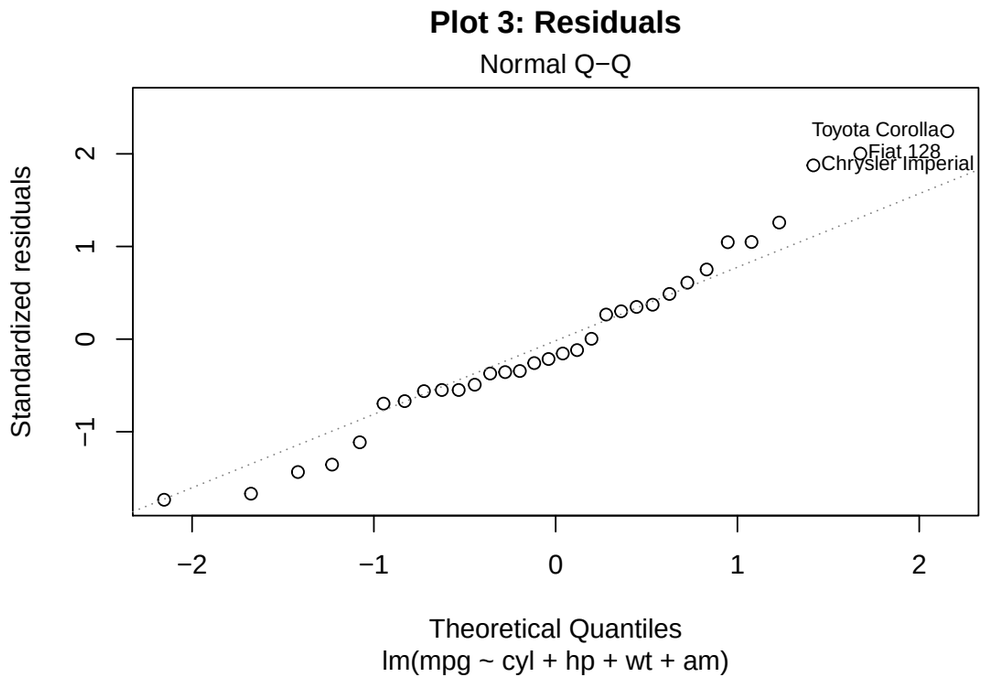
Plot 2: Scatter Plot for mtcars Data

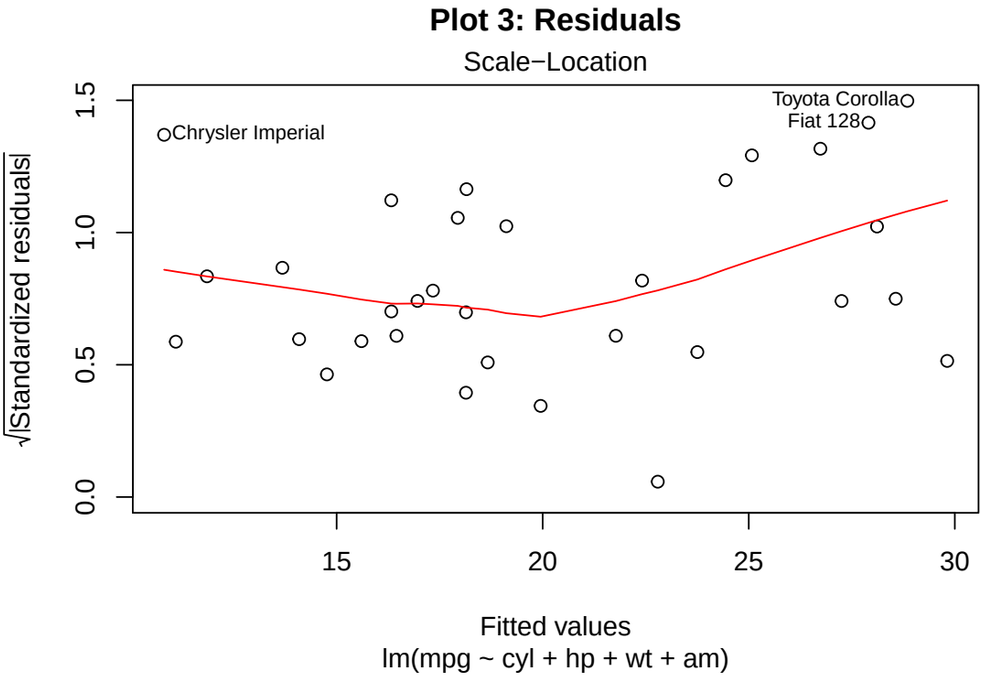


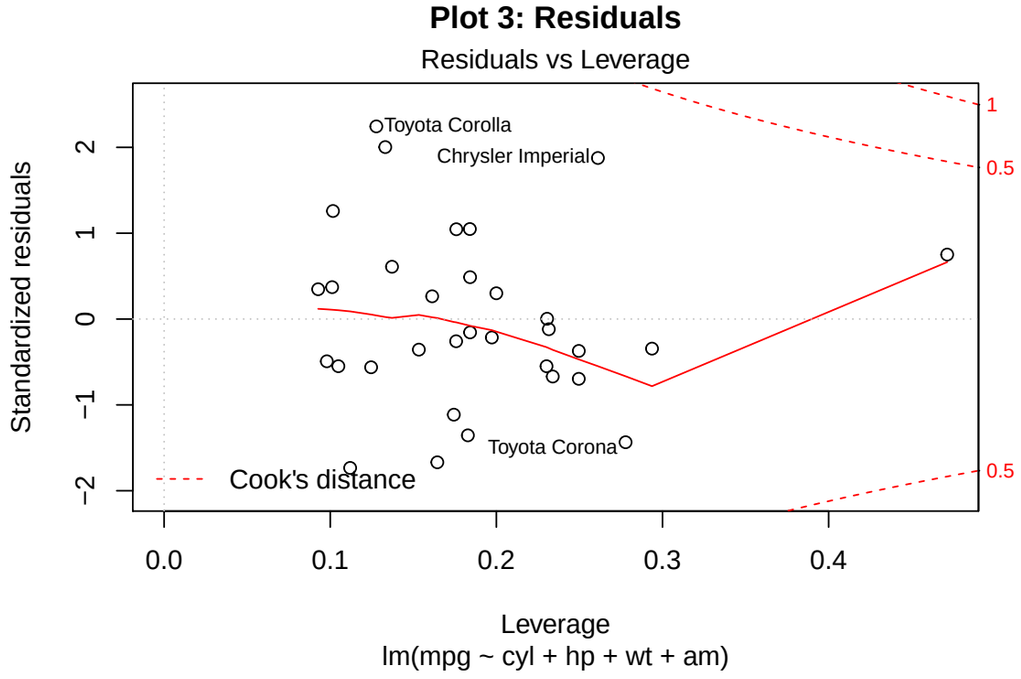
Plot 3: Residuals

Residuals vs Fitted




Plot 3: Residuals
Normal Q−Q
Toyota Corolla
Fiat 128
Chrysler Imperial
Standardized residuals
Theoretical Quantiles
lm(mpg ~ cyl + hp + wt + am)


Plot 3: Residuals
Scale−Location
Chrysler Imperial
Toyota Corolla
Fiat 128
√|Standardized residuals|
Fitted values
lm(mpg ~ cyl + hp + wt + am)


Plot 3: Residuals
Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
Standardized residuals
Leverage
lm(mpg ~ cyl + hp + wt + am)