Regression Models - Motor Trend Project

Bill Dusch

March 10, 2017

Executive Summary

Motor Trend is interested in a certain collection of cars within the dataset `mtcars`. This study will examine and explore how miles per gallon (MPG) is affected by different variables. In particular, the following two questions will be answered: (1) Is an automatic or manual transmission better for MPG, and (2) Quantify the MPG difference between automatic and manual transmissions.

Exploratory Data Analysis

```
library(ggplot2) # for plots
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
# Transform certain variables into factors
mtcars$cyl  <- factor(mtcars$cyl)
mtcars$vs   <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
```

We need to build exploratory plots to understand the data. Appendix - Plot 1 shows that Automatic transmissions have a lower MPG than Manual transmissions.

Regression Analysis

```
aggregate(mpg ~ am, data=mtcars, mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

Let's determine if there is a statistically significant difference by doing a t-test.

```
automatic <- mtcars[mtcars$am == "Automatic",]
manual <- mtcars[mtcars$am == "Manual",]
t.test(automatic$mpg, manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  automatic$mpg and manual$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

The p-value is significant at the 0.05 level; thus the difference between automatic and manual is statistically significant from zero. Let's quantify this through linear regression.

```
unifit <- lm(mpg ~ am, data=mtcars)
sum1 <- summary(unifit)
print(sum1$coef)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

```
print(sum1$r.squared)
```

```
## [1] 0.3597989
```

```
confint(unifit)
```

```
##                2.5 %   97.5 %
## (Intercept) 14.85062 19.44411
## amManual     3.64151 10.84837
```

The average MPG for automatic is 17.1 MPG, while manual is 7.2 MPG higher. The R^2 value is 0.36, which means the model explains only 36% of the variance.

We'll create a new multivariate regression model to make it more accurate. To determine which variables to pick, we will use the `bestglm` package to automatically determine the best subset. Appendix: Analysis 1 determines that the variables we should select are `am`, `qsec` (1/4 mile time), and `wt` (weight in 1000 lbs). The new model will include these variables and determine the significance of the three regressors, using nested model testing using the `anova` function.

```
bifit <- update(unifit, mpg ~ am + wt)
multifit <- update(bifit, mpg ~ am + wt + qsec)
anova(unifit, bifit, multifit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + qsec
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 73.203 2.673e-09 ***
## 3     28 169.29  1    109.03 18.034 0.0002162 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This nested model test demonstrates that all three regressions are significant at alpha = 0.05. Appendix:

Plot3 checks the assumptions of our regression model, and checks the residuals for non-normality. The residuals are homoskedastic but deviate from normality after a standard deviation. The summary of the full model is as follows:

```
sum2 <- summary(multifit)
print(sum2$coef)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## amManual     2.935837  1.4109045  2.080819 4.671551e-02
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
```

```
print(sum2$r.squared)
```

```
## [1] 0.8496636
```

```
confint(multifit)
```

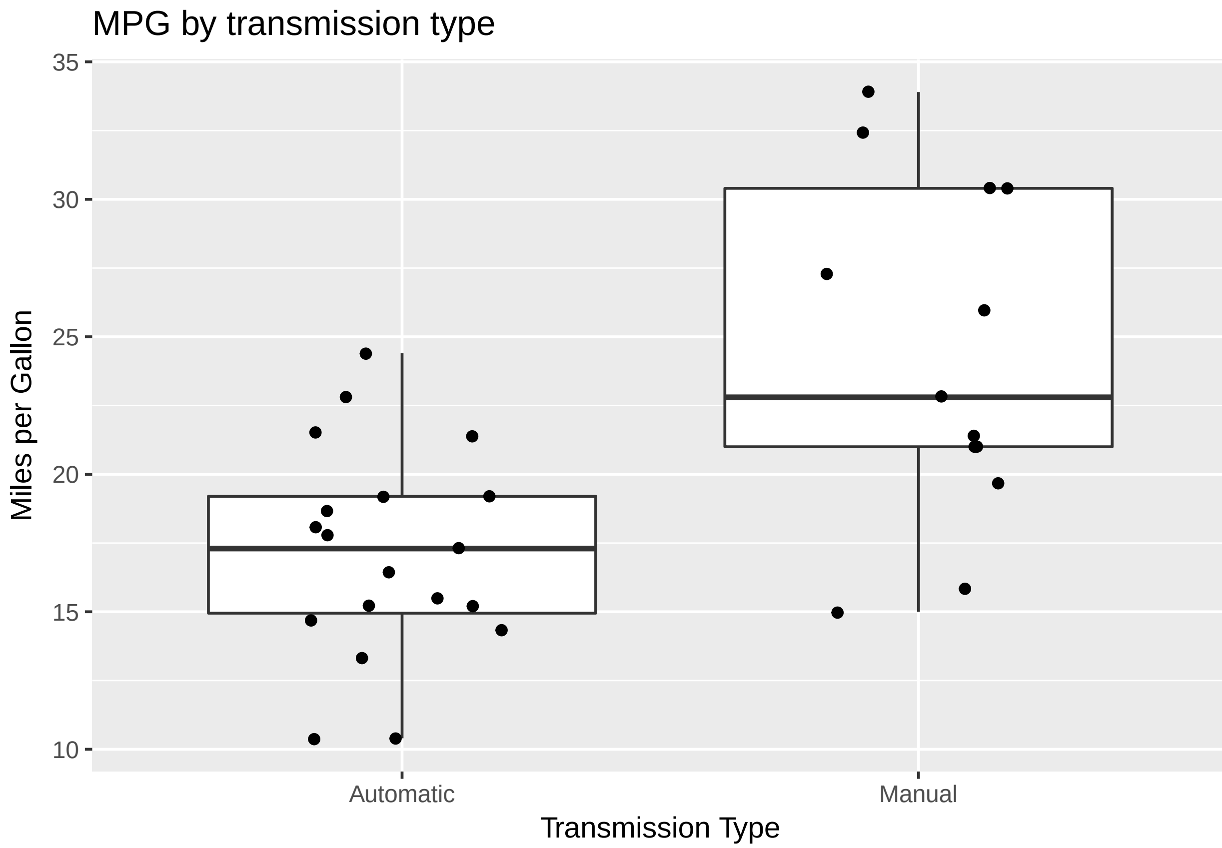
```
##                   2.5 %    97.5 %
## (Intercept) -4.63829946 23.873860
## amManual     0.04573031  5.825944
## wt          -5.37333423 -2.459673
## qsec         0.63457320  1.817199
```

This model explains 84.97% of the variance of the result; the other variables affected the correlation between `mpg` and `am`. The difference between automatic and manual transmissions, correcting for 1/4 mile time and weight is **2.94 MPG**. All of the variables' coefficients are statistically significant from zero at the 0.05 level; however, the intercept's confidence interval includes zero and is not statistically significant from zero.

Appendix

Plot 1: Plot of MPG by transmission type

```
g <- ggplot(data=mtcars, aes(y=mpg, x=am))
g <- g + geom_boxplot()
g <- g + geom_point(position = position_jitter(width = 0.2))
g <- g + xlab("Transmission Type")
g <- g + ylab("Miles per Gallon")
g <- g + ggtitle("MPG by transmission type")
print(g)
```



Analysis 1: Best Subset Analysis

```
library(bestglm)
cars <- within(mtcars, {y <- mpg; mpg <- NULL})
res.bestglm <- bestglm(Xy=cars, family=gaussian, IC="AIC", method="exhaustive")
```

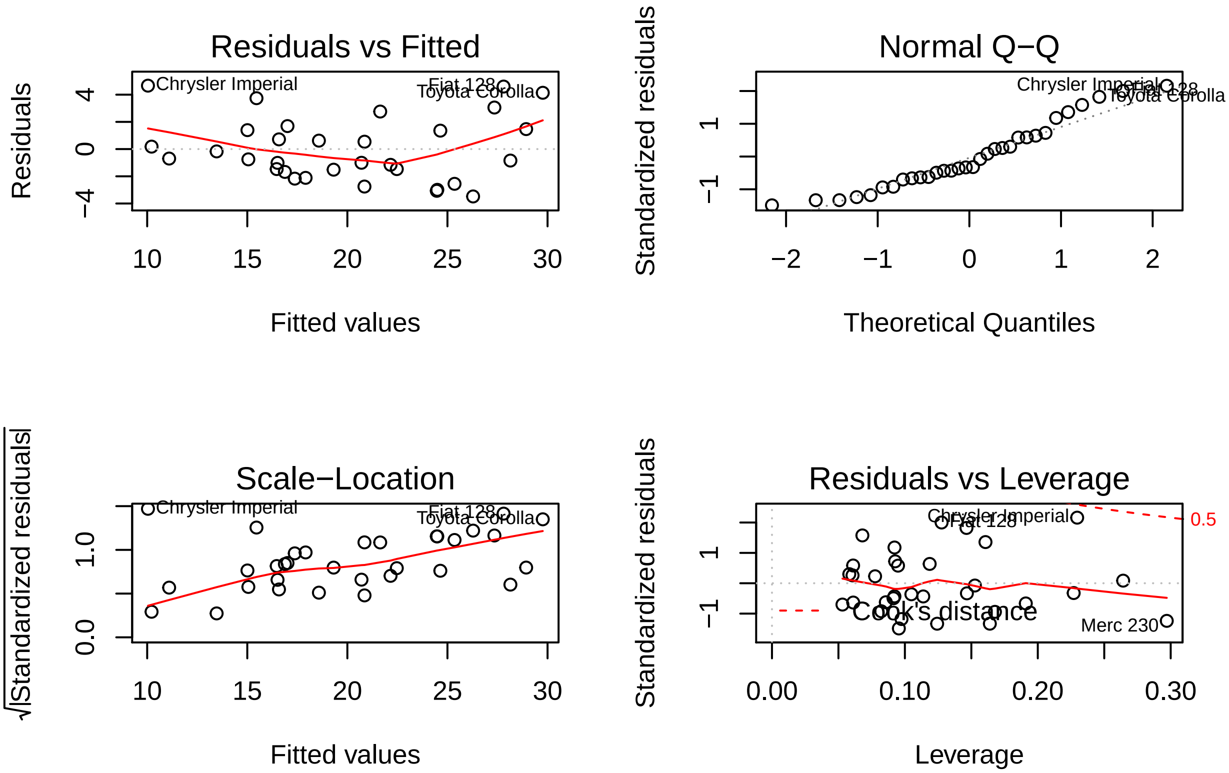
```
## Morgan-Tatar search since factors present with more than 2 levels.
```

```
res.bestglm$BestModels
```

```
##     cyl  disp    hp  drat   wt  qsec    vs    am  gear  carb Criterion
## 1 FALSE FALSE FALSE FALSE TRUE  TRUE FALSE  TRUE FALSE FALSE  59.30730
## 2 FALSE FALSE  TRUE FALSE TRUE  TRUE FALSE  TRUE FALSE FALSE  59.51530
## 3  TRUE FALSE  TRUE FALSE TRUE FALSE FALSE  TRUE FALSE FALSE  59.65483
## 4  TRUE FALSE  TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE  59.65716
## 5  TRUE FALSE  TRUE FALSE TRUE FALSE  TRUE  TRUE FALSE FALSE  60.05921
```

Plot 2: Residual Analysis

```
par(mfrow = c(2,2))
plot(multifit)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
Merc 230
0.5