Regression: Evaluate effects of Transimission type on MPG

Executive Summary

We were asked by Motor Trend to help understand the relationship between a number of variables, predictors, and miles-per-gallon (MPG), outcome, Specifically, were are asked to answer the following two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions"

By performing regression analysis, we obtained a model indicating a 1.81 mpg increase with switching from an automatic to manual. The analysis provided below supports this conclusion.

Data Processing

Use data(mtcars) to load the R dataset into the current enviromnent.

```
data(mtcars)
mtcars2 <- mtcars
mtcars2$am <- as.factor(mtcars2$am)
mtcars2$cyl <- as.factor(mtcars2$cyl)
```

Exploratory Data Analysis

To obtain a high-level insight to the relationship of the different variables in the data set, we created a panel plot and correlation matrix (see Appendix: Figure 1 & Figure 2) to understand the relationships. We find that the variables cyl, disp, hp, and wt have the strongest relationship with mpg.

Exhibit 1: Correlation Factor of Each Variable with MPG

```
mtcars.cor <- round(cor(mtcars), 3)
mtcars.cor[1,]
```

```
##    mpg    cyl   disp     hp   drat     wt   qsec     vs     am   gear
##  1.000 -0.852 -0.848 -0.776  0.681 -0.868  0.419  0.664  0.600  0.480
##   carb
## -0.551
```

Regression Analysis

Here we look build and compare various regression models to identify a model that best fits data. We use several metrics to evaluate our model fit, including: adjusted r-square, residual squared error (sigma), and p-values. We also use ANOVA and an analysis of the residuals to evaluate the model. Regression 'fit4' determined to provide best fit.

Regression Models

```
fit1 <- lm(mpg ~ am, data = mtcars2)
fit10 <- lm(mpg ~ ., data = mtcars2)
bestfit <- step(fit10, direction = "both")   ## stepwise process to identify best fit

## Obtained by manually adding / removing variables (using correlations as guide)
fit3 <- lm(mpg ~ am + wt + cyl, data = mtcars2)
fit4 <- lm(mpg ~ am + wt + cyl + hp, data = mtcars2)
```

Comparison of Regression Models

Two comparisons performed to evaluate the best model. The first is a comparison of the r.square and p-value for each model. Then, we perform ANOVA test to evaluate whether model is significantly better.

Exhbit 2: R.Square, Sigma, and P-value Comparisons

```
##     model r.squared adj.r.squared  sigma statistic p.value df
## 1    fit1    0.3598        0.3385 4.9020   16.8603   3e-04  2
## 2    fit3    0.8375        0.8134 2.6032   34.7917   0e+00  5
## 3    fit4    0.8659        0.8401 2.4101   33.5712   0e+00  6
## 4 bestfit    0.8497        0.8336 2.4588   52.7496   0e+00  4
## 5   fit10    0.8816        0.8165 2.5819   13.5381   0e+00 12
```

Exhibit 3: ANOVA

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + cyl
## Model 3: mpg ~ am + wt + cyl + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     27 182.97  3    537.93 30.8692 1.008e-08 ***
## 3     26 151.03  1     31.94  5.4991   0.02693 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residuals & Diagnostics

Understanding residuals is critical to understanding how well the regression model fits the data. The different plots (see Appendix: Figure 3) provide some insight to how closely the regression line fits the data. Although a few “outlier” points noted, the results seem to validate model fit.

1. Residuals vs Fitted: Want to see that the residuals are fairly well-distributed and no particular pattern exists.
2. Normal Q-Q: Ideally, residuals are lined closely with the straight dashed line.
3. Scale-Location(homoscedasticity): Ideally, points are equally spread around a line along the entire range of predictors.
4. Cooks distance: Visually illustrates those points that are outliers and may influence coefficients.

High Leverage/Influential Data Points

We use the function influence.measures() to help identify points/observations that may need to be considered.

```
## Potentially influential observations of
##   lm(formula = mpg ~ am + wt + cyl + hp, data = mtcars2) :
##
##                     dfb.1_ dfb.am1 dfb.wt dfb.cyl6 dfb.cyl8 dfb.hp dffit
## Lincoln Continental  0.16  -0.09   -0.19   0.06     0.04     0.04  -0.22
## Maserati Bora       -0.18   0.04   -0.09  -0.14    -0.25     0.53   0.70
##                     cov.r   cook.d hat
## Lincoln Continental  1.74_*  0.01   0.29
## Maserati Bora        2.10_*  0.08   0.47
```

Statistical Inference

To provide additional perspective on the strength of the model in predicting, a comparison of the actual MPG to the fitted mpg and associated 95% confidence interval to look at how well the model estimated results. While there are a few instances (see Appendix: Figure 4) where the actual MPG falls outside the 95% confidence interval, a majority fit the model.

Conclusion

The regression model, fit4, provides the best fit, explaining 84% of the changes in MPG. Outline of coefficients:
1. Intercept [33.71] - estimate of MPG for an average wt & hp car with 4 cyl car and automatic transmission
2. am1 - MPG increases 1.81 mpg for switching to a manual transmission, all else equal
3. wt - reduces MPG by 2.5 for a 1 unit (1,000 lbs) change in the weight of a car, all else equal
4. cyl6 - a switch from a 4 cyl to 6 cyl decreases MPG by 3.03, all else equal
5. cyl8 - a switch from a 4 cyl to 8 cyl decreases MPG by 2.16, all else equal (less than 6 cyl?) 6. hp - a 1 unit change in hp reduces MPG by 0.03, all else equal

Regression Coefficients

```
## (Intercept)         am1          wt        cyl6        cyl8          hp
##       33.71        1.81       -2.50       -3.03       -2.16       -0.03
```

Although there may more that can be learned about interaction between cyl (cyl8) and hp, this model produces a fairly reliable approach to predicting the MPG of a car. With a larger population and/or more detailed look at values within a variable, it may be possible to create a better model. This creates the risk of overfitting to reduce residuals, but not necessarily improving the applicability of the model.

Appendix

Figure 1: Illustrate the relationship of each variable

A panel plot of the relationship of each variable to another within the ‘mtcars’ data set.

```
pairs(mtcars, panel = panel.smooth, main = "MT Cars Data", col = 3)
```

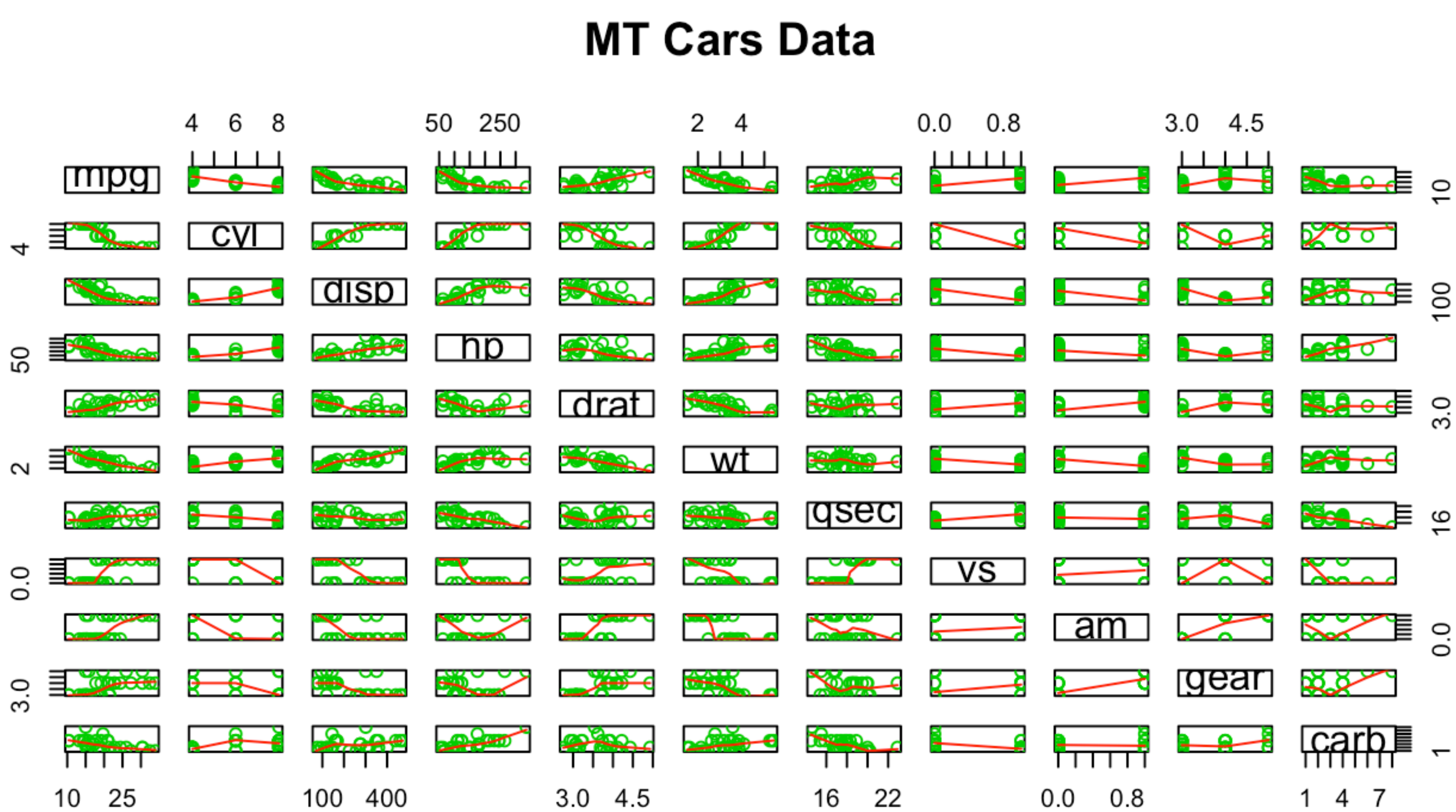

Figure 2: Regression Model Residuals

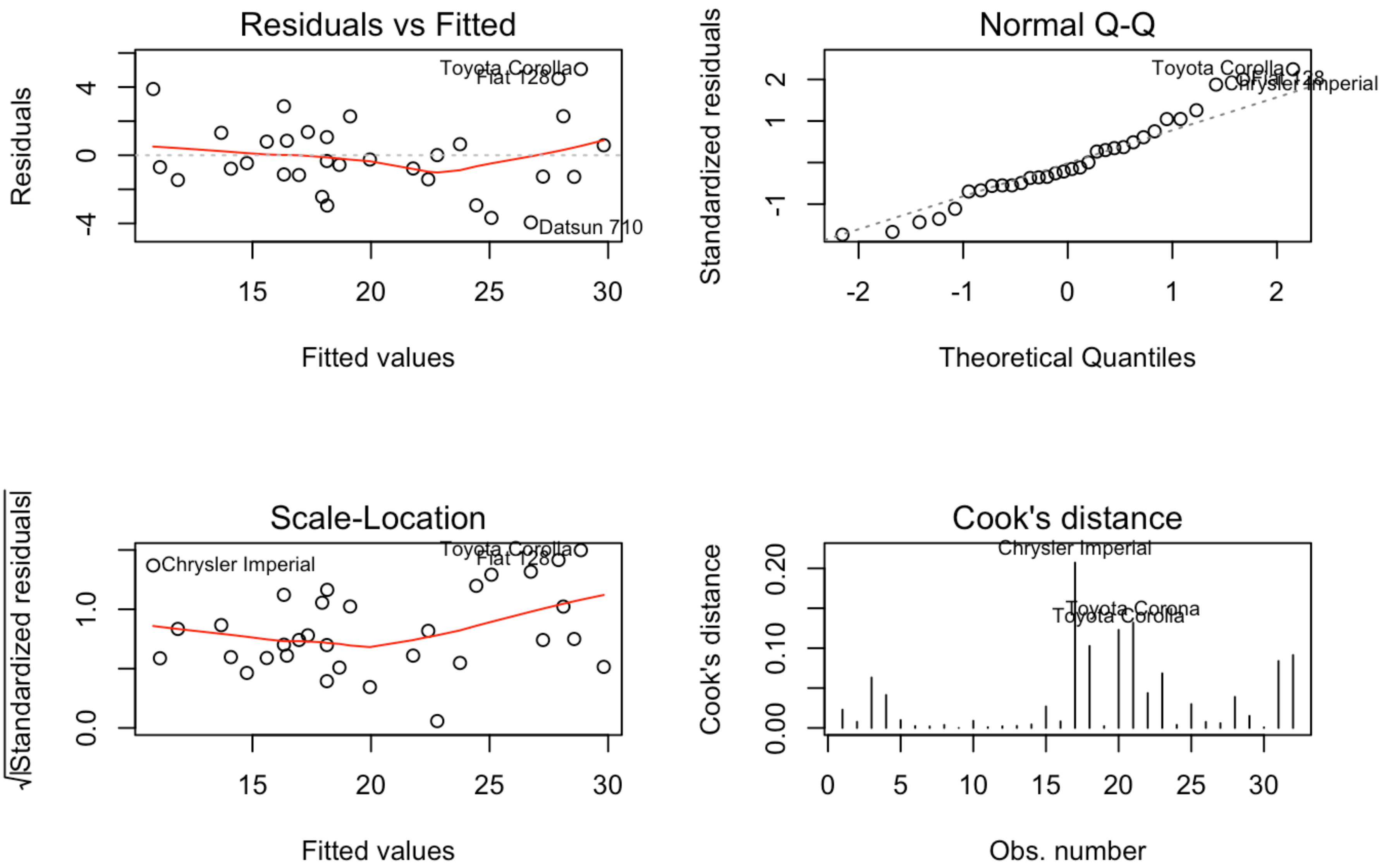

Figure 3: Regression Model Observations outside Confidence Interval

```
fit4.confint <- round(predict(fit4, mtcars2, interval = "confidence", level = 0.95),
2)
actual <- mtcars2[,1]
fit4.compare <- as.data.frame(cbind(actual, fit4.confint))
fit4.compare$outlier <- ifelse(fit4.compare$actual < fit4.compare$lwr | fit4.compare$
actual > fit4.compare$upr, "Y","N")
fit4.outlier <- subset(fit4.compare, outlier == "Y")
select(fit4.outlier, actual, fit, lwr, upr)
```

```
##                   actual   fit   lwr   upr
## Datsun 710          22.8 26.74 25.08 28.40
## Hornet 4 Drive      21.4 19.12 16.99 21.24
## Chrysler Imperial   14.7 10.81  8.28 13.35
## Fiat 128            32.4 27.91 26.10 29.71
## Toyota Corolla      33.9 28.85 27.08 30.62
## Toyota Corona       21.5 24.44 21.83 27.05
## Dodge Challenger    15.5 17.94 15.87 20.01
## AMC Javelin         15.2 18.15 16.03 20.27
## Pontiac Firebird    19.2 16.33 14.75 17.90
## Lotus Europa        30.4 28.11 26.03 30.19
## Volvo 142E          21.4 25.08 23.07 27.09
```