Regression Fast

Jerad Acosta

December 21, 2014

Executive Summary

We are analyzing the **mtcars** dataset from the R library package **datasets** in search of an effect on fuel consumption as any relationship ofa product of variables.
In particularly, we are trying to decide whether or not an Automatic transmission **AT** has a statistically significant amount of impage on **mpg** over that of a manual transmission **MT**

The conclusion of the analysis was that Manual transmissions **MT** do infact have a statistically significant increase in Miles Per Gallon **mpg** when compared to an Automatic Transmission **AT** - all other variables held aside. We also discovered, however, that transmission alone could not account for whether one car would have an increase or decrease in **mpg**. This

Analysis

Data Exploration

Calculating the correlation coefficients among the features we see high correlation, both positive and negative between MPG and number of other features.

```
##      wt     cyl    disp      hp    carb    qsec    gear      am      vs    drat
## -0.8677 -0.8522 -0.8476 -0.7762 -0.5509  0.4187  0.4803  0.5998  0.6640  0.6812
```

[Refer Appendix A - Figure 1 for a visual representation.] This suggests that transmission type by itself may not be a good model to predict fuel efficiency.

Basic Model

First, we built a Basic linear model using transmission type as the only predictor of fuel efficiency. [Refer to Basic Model Summary in the Appendix B for model details.]

```
    fastFit <- lm(mpg ~ am, data=mtcars)
```

Examining this model we see that the average the fuel efficiency for a car with an AT is 17.15 **MPG** and that with a **MT** will get approximately 7.24 more **MPG**. [Refer to Appendix A - Figure 2 for a visual representation.] Both coefficients have significant p-values ($p < 0.05$) as we would expect, knowing there is a strong correlation between MPG and transmission type. The model F-statistic is significant ($p < 0.05$); however, the overall model fit is poor with an R-Squared value of 0.36.

Multivariate Model

Next, we built a more complex multivariate model which included all the variables that we found to be highly correlated to fuel efficiency during initial data exploration.

```
multiFit <- lm(mpg ~ cyl + disp + hp + wt + drat + vs + am + carb, data=mtcars)
```

From this model we see a much better fit with an R-squared value of 0.86; however, the model contains many features which are not significant ($p > 0.05$).

Optimized Model

Knowing this, we used the AIC Stepwise algorithm to build an an optimized model, using the multivariate model as the starting point. [Refer to Optimized Model Summary in the Appendix B for model details.]

```
bestFit <- step(multiFit, direction="backward")
```

Examining the optimized model we see that a significant number of variables have been removed and now includes only number of cylinders, horsepower, and weight as features. The optimized model exhibits a slightly better fit with an R-squared value of 0.843 and a more signicant F-statistic. Of the features remaining in the model, only wieght has a significant p-value ($p < 0.05$). The number of cylinders and horsepower are not significant ($p > 0.05$); however, their inclusion improves the fit of the model, i.e., higher R-squared and lower F-statistic p-value.

Final Model

Lastly, because the main objective of this analysis is to determine a relationship between transmission type and fuel efficiency we want our final model to include transmission type as a feature. To do this we examined the output of the AIC Stepwise Algorithm to determine the best predictive model including transmission type as a feature. [Refer to Final Model Summary in the Appendix for model details.]

```
finalFit <- lm(mpg ~ am + wt + cyl + hp, data=mtcars)
```

Examining our final model we see that weight remains as the only significant feature, and the R-squared value and F-statistic significance are only slightly worse than in the optimized model.

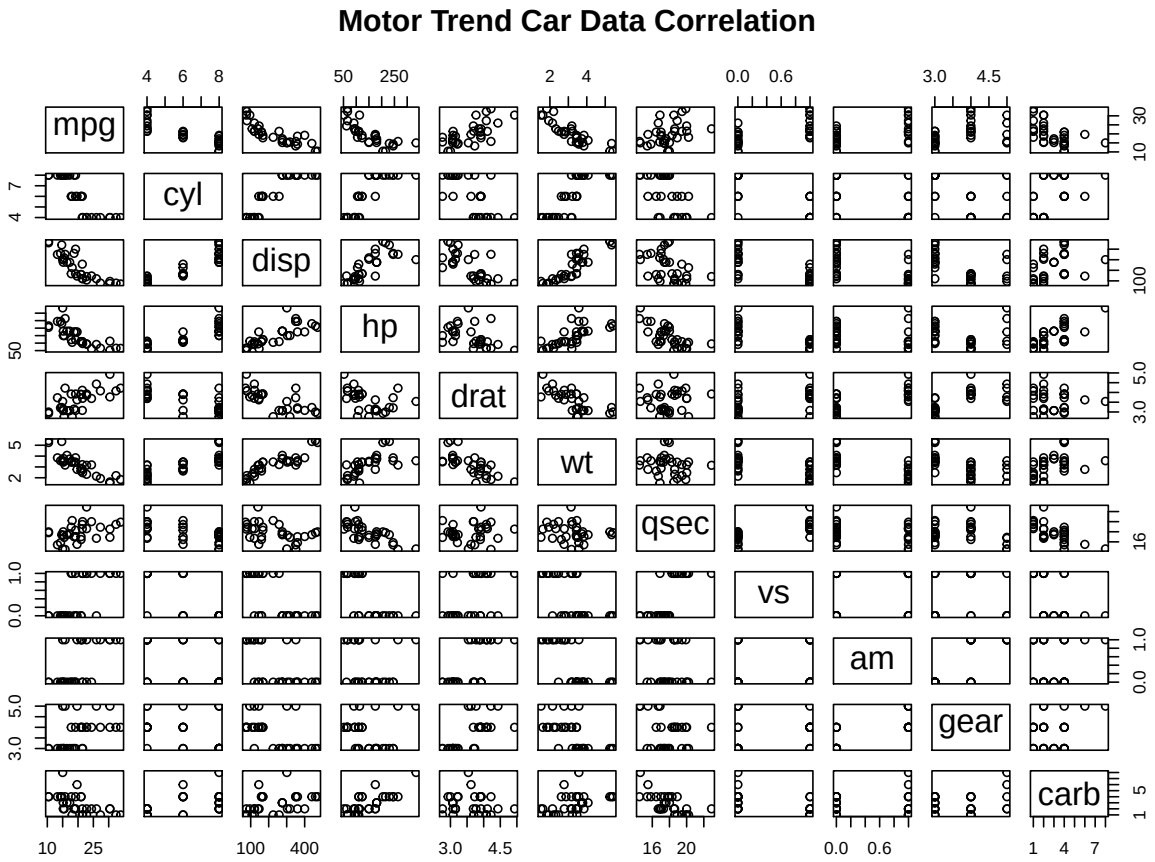
Havning finalized out model, we examined the residuals of the to verify the validity of the model. [Refer to Appendix A - Figure 3 visual summary.] The plots show that the model supports our assumptions of independence and normality of the data with no heteroskedasticity and no inflential outliers.

- The “Residuals vs Fitted” plot verifies the independance assumption as the points are randomly scattered above and below the zero line.
- The “Normal Q-Q”" plot verifies that the residuals are normally distributed as the points hug the line closely.
- The “Scale-Location” plot verifies the constant variance assumption as the points fall in a constant band displaying no heteroskedasticity.
- The “Residuals vs Leverage” plot indicates all points are within Cook’s distance, verifying there are no influential outliers.

Conclusion

MTs have greater fuel efficiency than ATs when ignoring the confounding features of weight, number of cyclinders, and horsepower. The average MT gets 7.24 more MPG than the average AT at 17.15 MPG.

Appendix A - Figures



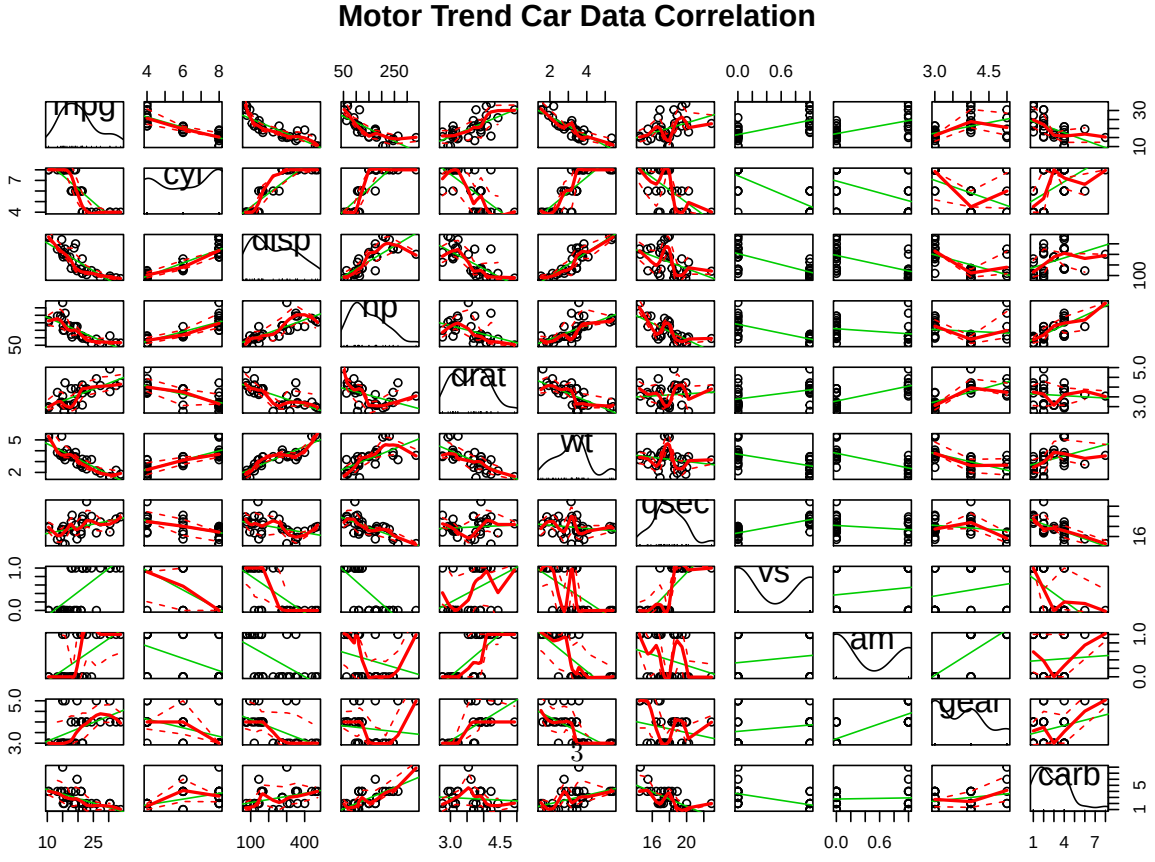


Figure 1: Feature Correlation

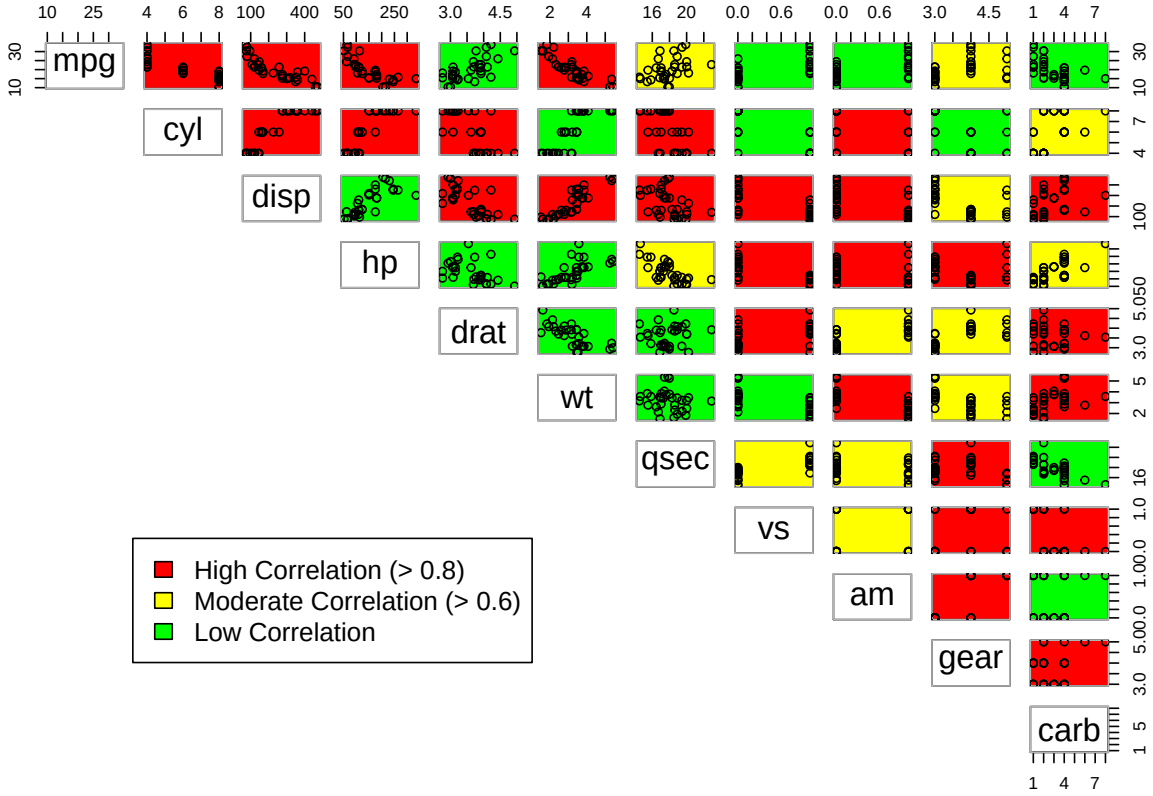


Figure 2: Basic Linear Model

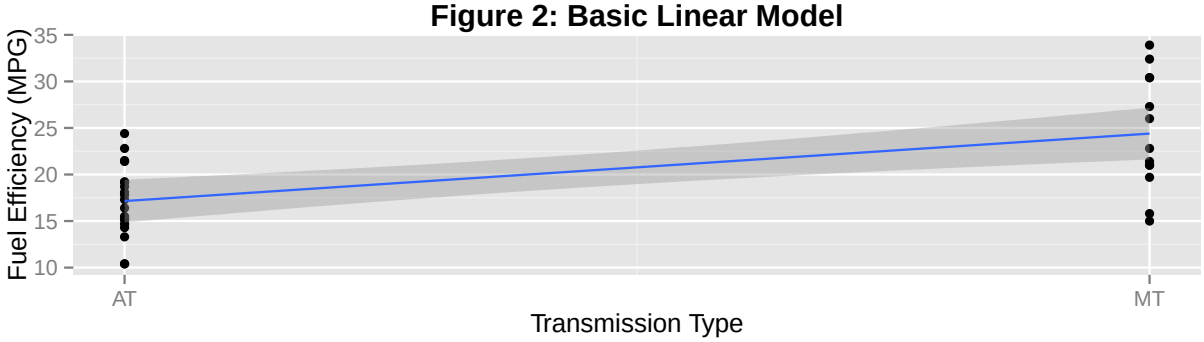
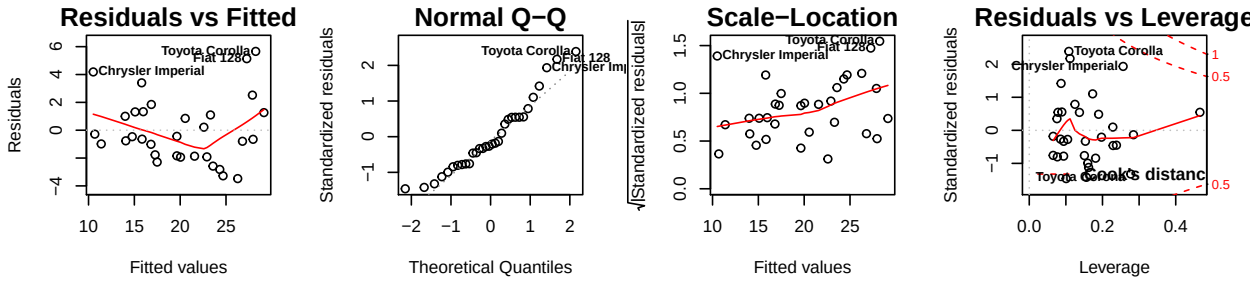


Figure 3: Residual Diagnostics



Appendix B - Models

Basic Model Summary

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.15       1.12   15.25  1.1e-15 ***
## am              7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

Optimized Model Summary

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.929 -1.560 -0.531  1.185  5.899
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.7518     1.7869   21.69   <2e-16 ***
## cyl          -0.9416     0.5509   -1.71   0.0985 .
## hp           -0.0180     0.0119   -1.52   0.1400
```

```
## wt           -3.1670     0.7406   -4.28   0.0002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.51 on 28 degrees of freedom
## Multiple R-squared:  0.843,  Adjusted R-squared:  0.826
## F-statistic: 50.2 on 3 and 28 DF,  p-value: 2.18e-11
```

Final Model Summary

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.476 -1.847 -0.554  1.276  5.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.1465     3.1048   11.64  4.9e-12 ***
## am            1.4780     1.4411    1.03   0.3142
## wt           -2.6065     0.9198   -2.83   0.0086 **
## cyl          -0.7452     0.5828   -1.28   0.2119
## hp           -0.0250     0.0136   -1.83   0.0786 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.51 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.827
## F-statistic:   38 on 4 and 27 DF,  p-value: 1.02e-10
```