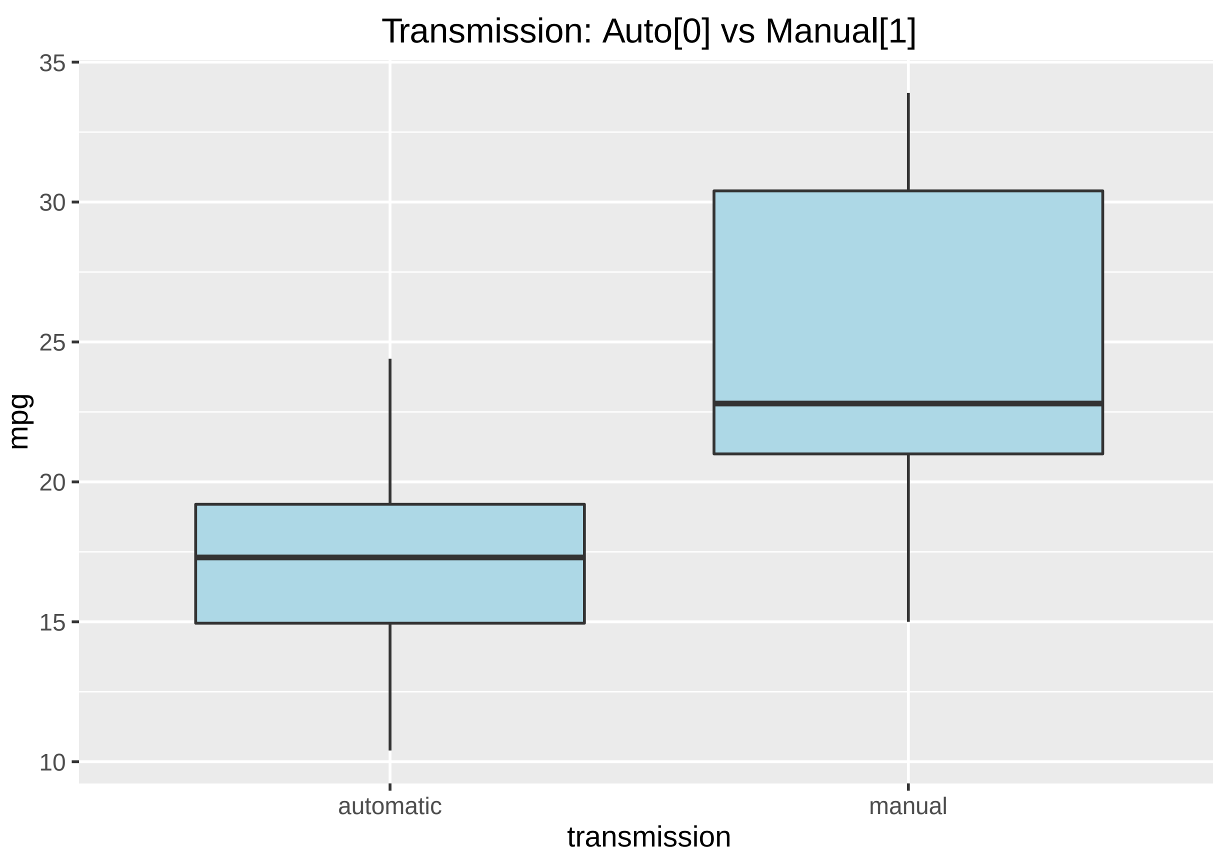
Motor Trend Car Road Tests: Automatic MPG vs Manual MPG

Executive Summary

The goal of this analysis is to specifically identify which type of transmission (automatic vs manual) is better for MPG. Using the dataset mtcars, a model that isolates the variables which predict mpg will be developed. This report details the exploratory analysis phase, generating various models, testing model variation for model selection, conducting model diagnostics and examining the variability of the outcome. Using this approach, my analysis shows that automatic transmission produces a lower mpg than manual transmission. The analysis shows that manual transmission has almost a 10% higher mpg on average however this estimation has a wide confidence interval.

Exploratory Analysis

```
##boxplot
ggplot(dat, aes(am, mpg)) +
  geom_boxplot(fill = "light blue") +
  ggtitle("Transmission: Auto[0] vs Manual[1]") + xlab("transmission")
```



Boxplot shows Manual Transmission on average has higher MPG.

Model Selection

```
## Analysis of Deviance Table
```

```
##
## Model 1: (mpg * 10) ~ am + wt - 1
## Model 2: (mpg * 10) ~ am + wt + qsec - 1
## Model 3: (mpg * 10) ~ am + wt + qsec + carb - 1
## Model 4: (mpg * 10) ~ am + wt + qsec + carb + drat - 1
## Model 5: (mpg * 10) ~ am + wt + qsec + carb + drat + gear - 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        29    104.490
## 2        28     64.470  1   40.021 2.513e-10 ***
## 3        27     63.034  1    1.435    0.2309
## 4        26     62.380  1    0.654    0.4186
## 5        25     60.123  1    2.257    0.1330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova shows that Model 2 (fit2.2) is the best model as far as minimizing residual deviation and deviance while keeping degrees of freedom higher. Model avoids bias by including necessary variables and avoids overfitting.

Coefficient Interpetation

On average, manual tramsissions have almost 10% higher mpg than automatic (please note that the confidence interval is wide). Additionally, lower bound and upper bound of the 95% confidence interval is higher.

```
##manual vs automatic
1 / (exp(coef(fit2.2))[1] / exp(coef(fit2.2))[2])
```

```
## amautomatic
##    1.094378
```

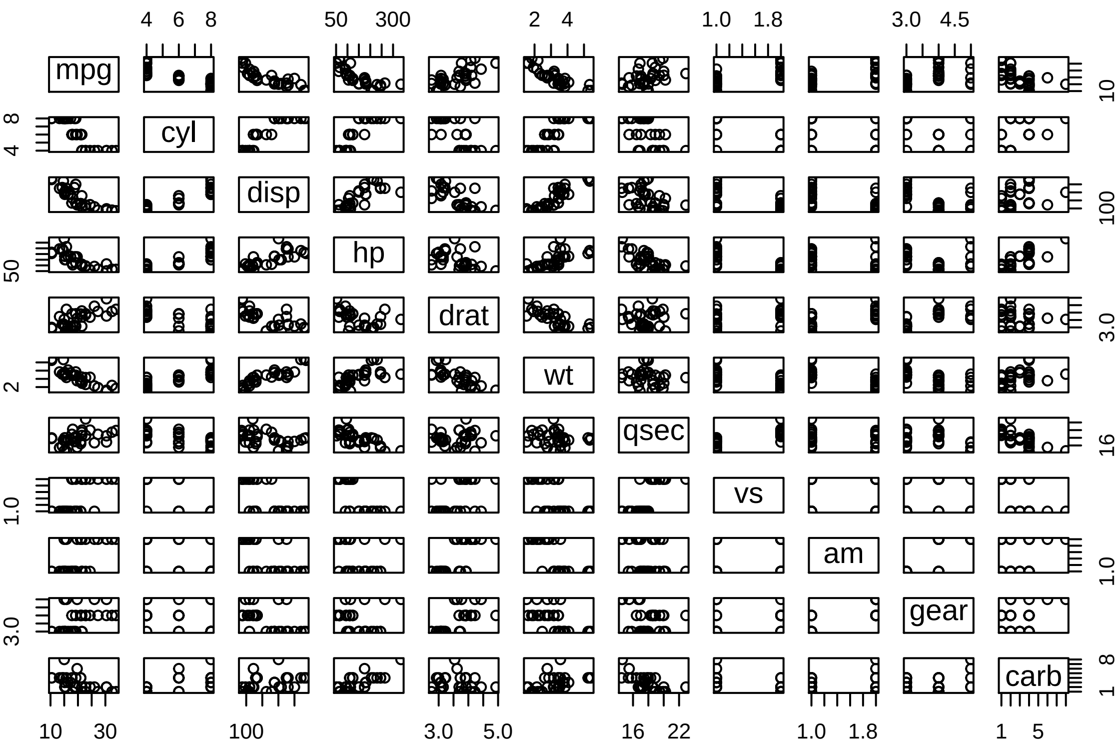
```
##create confidence interval
exp(confint(fit2.2))
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %      97.5 %
## amautomatic 100.4227248 227.9317834
## ammanual    116.2980276 235.7377225
## wt            0.7611586   0.8327665
## qsec          1.0371438   1.0714727
```

Appendix

```
##pairs plot
pairs(dat)
```



Pairs plot shows a matrix of scatterplots to help visualize correlation across predictors (hp, disp, drat, wt are correlated).

```r
##correlation matrix
cor(dat[,c(2:7,10:11)])
```

```
##             cyl       disp         hp        drat         wt        qsec
## cyl   1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958 -0.59124207
## disp  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788
## hp    0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339
## drat -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476
## wt    0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588
## qsec -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000
## gear -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870 -0.21268223
## carb  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059 -0.65624923
##            gear       carb
## cyl  -0.4926866  0.5269883
## disp -0.5555692  0.3949769
## hp   -0.1257043  0.7498125
## drat  0.6996101 -0.0907898
## wt   -0.5832870  0.4276059
## qsec -0.2126822 -0.6562492
## gear  1.0000000  0.2740728
## carb  0.2740728  1.0000000
```

```r
##calculate variance inflation for each variable (except vs and am)
sqrt(vif(glm((mpg*10) ~ . -am, "poisson", dat)))
```

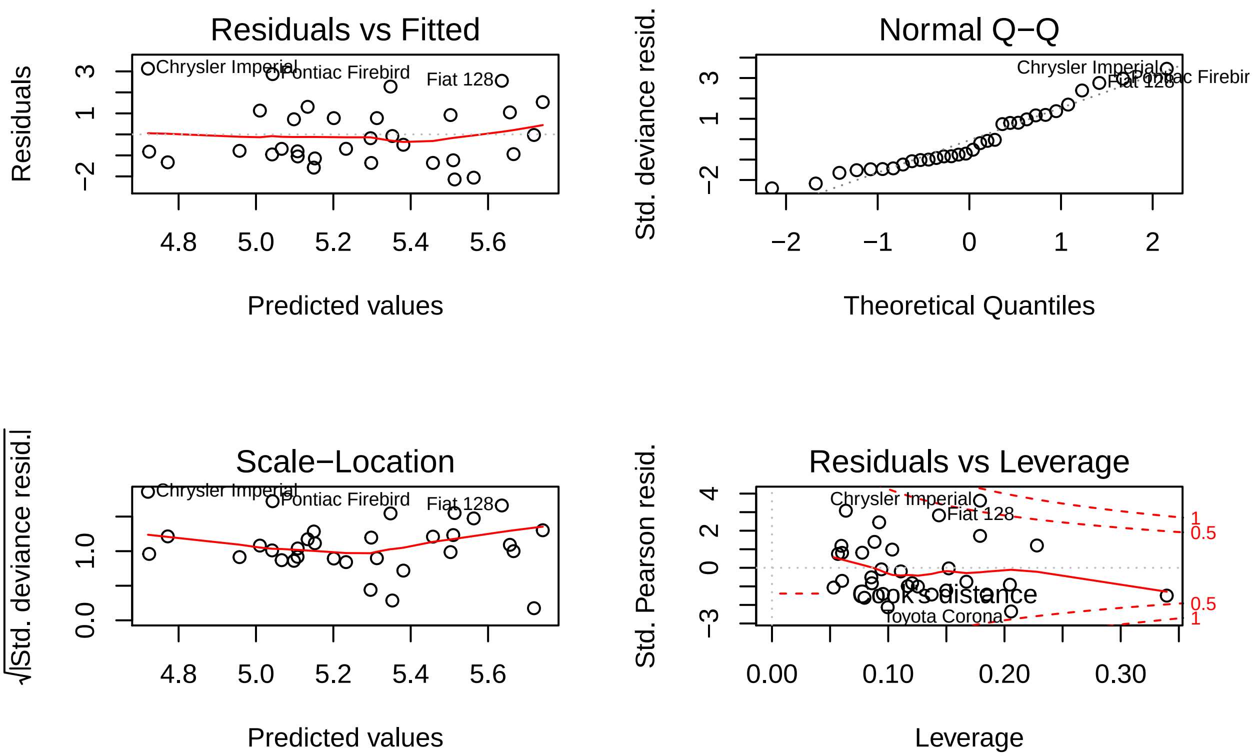
```
##      cyl     disp       hp     drat       wt     qsec       vs     gear
## 3.941103 4.670736 3.029959 1.790916 3.764505 2.526815 2.086419 2.149142
##     carb
## 2.706795
```

```r
##model summary
summary(fit2.2)
```

```
##
## Call:
## glm(formula = (mpg * 10) ~ am + wt + qsec - 1, family = "poisson",
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1455  -1.0738  -0.5904   0.9525   3.1287
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## amautomatic  5.018965   0.209094  24.003  < 2e-16 ***
## ammanual     5.109152   0.180245  28.346  < 2e-16 ***
## wt          -0.227778   0.022937  -9.931  < 2e-16 ***
## qsec         0.052773   0.008307   6.353 2.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 55935.13  on 32  degrees of freedom
## Residual deviance:    64.47  on 28  degrees of freedom
## AIC: 299.63
##
## Number of Fisher Scoring iterations: 4
```

Model Diagnostics

```r
##model diagnostics includes plot of Residuals vs Fitted
par(mfrow = c(2, 2))
plot(fit2.2)
```


Residuals vs Fitted
Residuals
Predicted values
Chrysler Imperial
Pontiac Firebird
Fiat 128
Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
Scale−Location
√|Std. deviance resid.|
Residuals vs Leverage
Std. Pearson resid.
Leverage
Cook's distance
Toyota Corona