

An Analysis of Fuel Efficiency

Clyde Tressler

December 18, 2015

Executive Summary

MotorTrend magazine data from observations of 32 models of cars were analyzed to help understand the characteristics that account for fuel efficiency. Of particular interest is whether there is an effect associated with automatic versus manual transmission. After we account for the effect of a strong relationship between vehicle weight and transmission type, we can say that there is a statistically significant increase in gas mileage associated with manual transmissions, which we report as a mean increase of **9.23 mpg**, with a maximum **P-value of 0.00378**.

Data Processing and Initial Exploratory Analyses

This report omits some code for brevity. The entire R markdown file can be found [here](#).

Here we show the first several rows of the dataset.

```
##           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
```

In the Appendix we present a table explaining the meaning of variable names. Several binary and discrete variables were transformed from numeric vectors to factor variables for this analysis.

Also in the Appendix we present a box and whisker plot of the two transmission types demonstrating that the median mpg of the automatic group lies below that of the manual group. We use hypothesis testing to determine if the difference in the means is statistically significant.

We report a p-value for a t-test with unequal variances as 0.001374 and therefor reject the null hypothesis at the 95% confidence level, to conclude that the difference between the means, 7.24 mpg, is significant.

Developing a Multivariate Linear Model

Next we turn to multivariate regression to develop a model that will help us understand whether confounding variables are contributing **omitted variable bias**. We tabulate the correlation of the numeric variables, as many of them measure closely-related properties.

Correlation Table of mtcars Data

	mpg	disp	hp	drat	wt	qsec
mpg	1.0000000	-0.8475514	-0.7761684	0.6811719	-0.8676594	0.4186840
disp	-0.8475514	1.0000000	0.7909486	-0.7102139	0.8879799	-0.4336979
hp	-0.7761684	0.7909486	1.0000000	-0.4487591	0.6587479	-0.7082234
drat	0.6811719	-0.7102139	-0.4487591	1.0000000	-0.7124406	0.0912048
wt	-0.8676594	0.8879799	0.6587479	-0.7124406	1.0000000	-0.1747159
qsec	0.4186840	-0.4336979	-0.7082234	0.0912048	-0.1747159	1.0000000

We now consider the apriori exclusion of certain variables. First, the quarter second mile time is influenced by driver skill and is heavily biased in favor of manual transmission, but adds little meaning to ordinary driving conditions, so we choose not to include it.

We also know that ‘drat,’ the rear axle ratio, is a downstream variable determined by the upstream drive train, including the transmission, so we choose not to include it when analyzing fuel efficiency.

From the correlation table we select weight as our first predictor, since it is the variable most highly correlated with mpg.

Looking at the data sorted by weight below, we see that 9 of the top 10 heaviest cars have automatic transmissions, while 9 of lightest 10 have manual. This leads us to the addition of an interaction term between weight and transmission type to the model. In the Appendix we show a visualization of this relationship.

Sorting the Data by Vehicle Weight

rownames	wt	am	hp	mpg
Lincoln Continental	5.424	0	215	10.4
Chrysler Imperial	5.345	0	230	14.7
Cadillac Fleetwood	5.250	0	205	10.4
Merc 450SE	4.070	0	180	16.4
Pontiac Firebird	3.845	0	175	19.2
Camaro Z28	3.840	0	245	13.3
Merc 450SLC	3.780	0	180	15.2
Merc 450SL	3.730	0	180	17.3
Duster 360	3.570	0	245	14.3
Maserati Bora	3.570	1	335	15.0

rownames	wt	am	hp	mpg
Ferrari Dino	2.770	1	175	19.7
Mazda RX4	2.620	1	110	21.0
Toyota Corona	2.465	0	97	21.5
Datsun 710	2.320	1	93	22.8
Fiat 128	2.200	1	66	32.4
Porsche 914-2	2.140	1	91	26.0
Fiat X1-9	1.935	1	66	27.3
Toyota Corolla	1.835	1	65	33.9
Honda Civic	1.615	1	52	30.4
Lotus Europa	1.513	1	113	30.4

We systematically examine the inclusion of other variables while seeking to maintain significant P-values for their coefficients and maximize the fit as measured by R-squared. We choose to include horsepower, which will therefor serve as a proxy for the remaining collinear engine performance variables and refer the reader to the full R markdown file for the code.

Diagnostics Now we use ANOVA testing of nested models to choose which terms to include in our selection.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + hp
## Model 3: mpg ~ wt + hp + am
```

```
## Model 4: mpg ~ wt + hp + am + wt:am
## Model 5: mpg ~ wt + hp + am + wt:am + hp:am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1      30 278.32
## 2      29 195.05  1    83.274 15.9318 0.000478 ***
## 3      28 180.29  1    14.757  2.8232 0.104888
## 4      27 146.84  1    33.446  6.3988 0.017820 *
## 5      26 135.90  1    10.945  2.0940 0.159833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Once we satisfy the required inclusion of transmission type, we see that the addition of the interaction term (wt*am) is necessary.

In the Appendix we plot residuals vs hat-values for the model **mpg~wt+hp+am+wt*am** and identify the Maserati Bora as a high influence point. This is a high-performance vehicle and is the only vehicle in the data set with 8 carburetors. Furthermore, the dfbeta of this point for the coefficient of transmission type, -2.2929, lends motivation to our choice to remove it.

Now we note that the response for mpg vs horsepower (see Appendix) looks a bit quadratic but the removal of the Maserati data point has lessens this effect.

The residuals vs Fit and Q-Q plots shown in the Appendix indicate that the residuals are close to normally-distributed, which is further supported by the R-squared value of the model.

```
##
## Call:
## lm(formula = mpg ~ wt + hp + am + wt * am, data = mtcars2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.015 -1.452 -0.393  1.341  4.897
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.79263    2.57244  11.970 4.43e-12 ***
## wt          -2.09632    0.82265   -2.548  0.01708 *
## hp          -0.03584    0.01020   -3.515  0.00163 **
## am1         13.84775    3.95675    3.500  0.00170 **
## wt:am1      -4.61832    1.45195   -3.181  0.00378 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.202 on 26 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8677
## F-statistic: 50.18 on 4 and 26 DF,  p-value: 7.415e-12

##              2.5 %      97.5 %
## (Intercept) 25.5049130 36.08035486
## wt          -3.7873067 -0.40533831
## hp          -0.0568056 -0.01488208
## am1         5.7145332 21.98095771
## wt:am1      -7.6028370 -1.63380182
```

Conclusion

We have shown that a linear model using weight, horsepower, transmission type, and an interaction term between weight and transmission fits observations in the data set well, with an R-squared value of 0.868, indicating 86.8% of the total variance in mpg has been explained. The 95% confidence intervals for the coefficients show no instances where the intervals cross zero and the residuals plots show no discernible patterns. The Q-Q plot supports the assertion that the residual

Clearly fuel efficiency is adversely impacted by weight, as shown by the negative slope in the linear model. Likewise, increases in horsepower, which we take to be a proxy for other engine size variables such as displacement and numbers of cylinders, also reduce fuel efficiency.

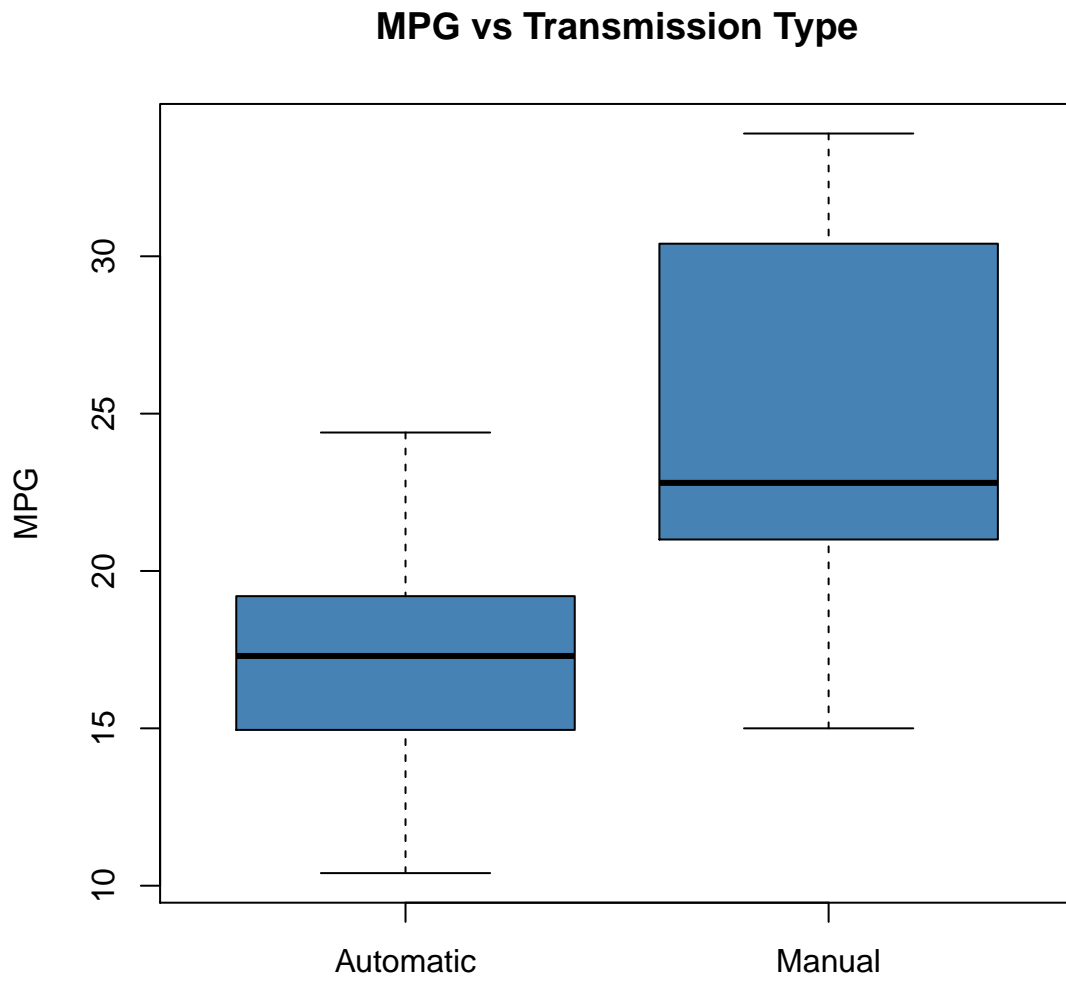
We can report that the mean increase in miles per gallon associated with a manual transmission, as calculated from the values of the transmission type coefficient and its interaction term with weight, is shown to be 9.23 mpg, with a maximum P-value of 0.00378 as given by the interaction term.

Appendix

Table 1: Variables

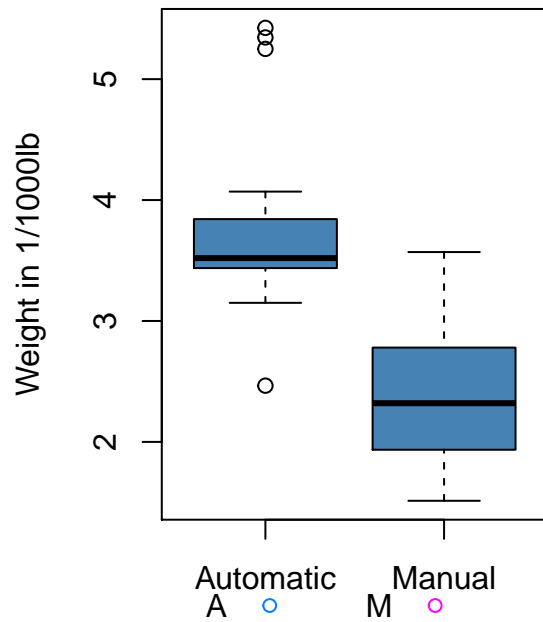
Variable Name	Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight(lb/1000)
qsec	1/4 mile time
vs	Cylinder alignment, V or Straight
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Box and Whisker plot of MPG vs Transmission Type

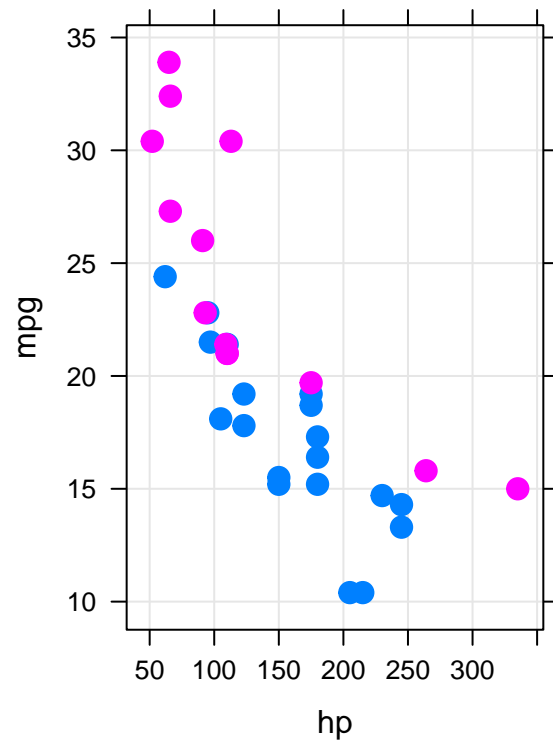
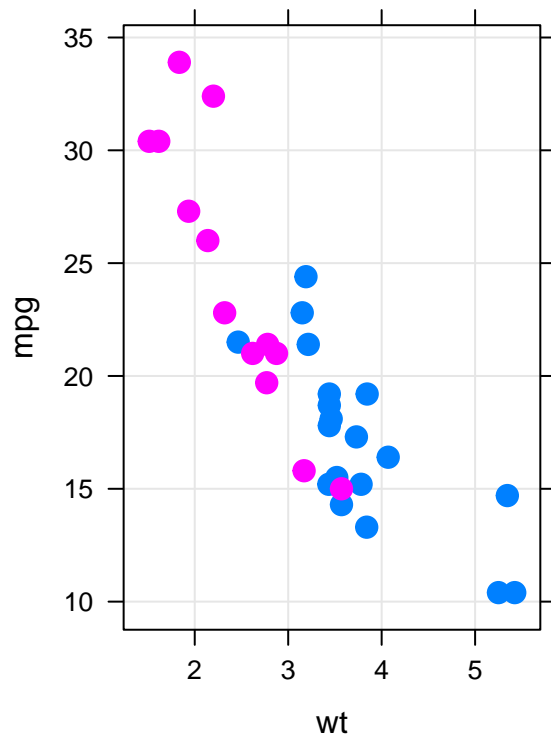
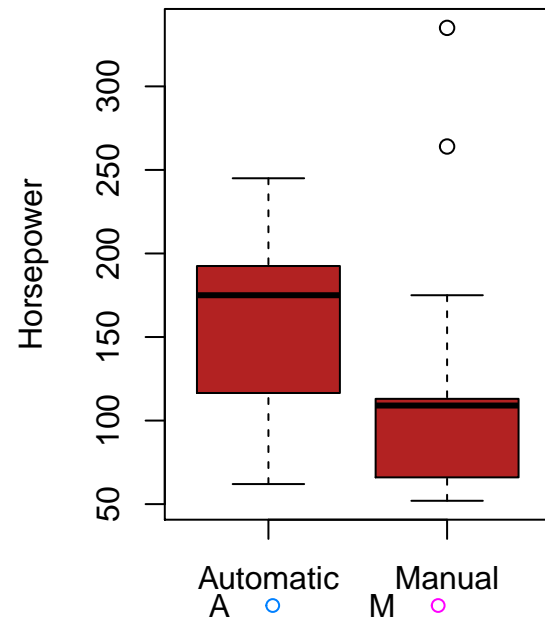


Interactions for Weight and Transmission and for Horsepower and Transmission

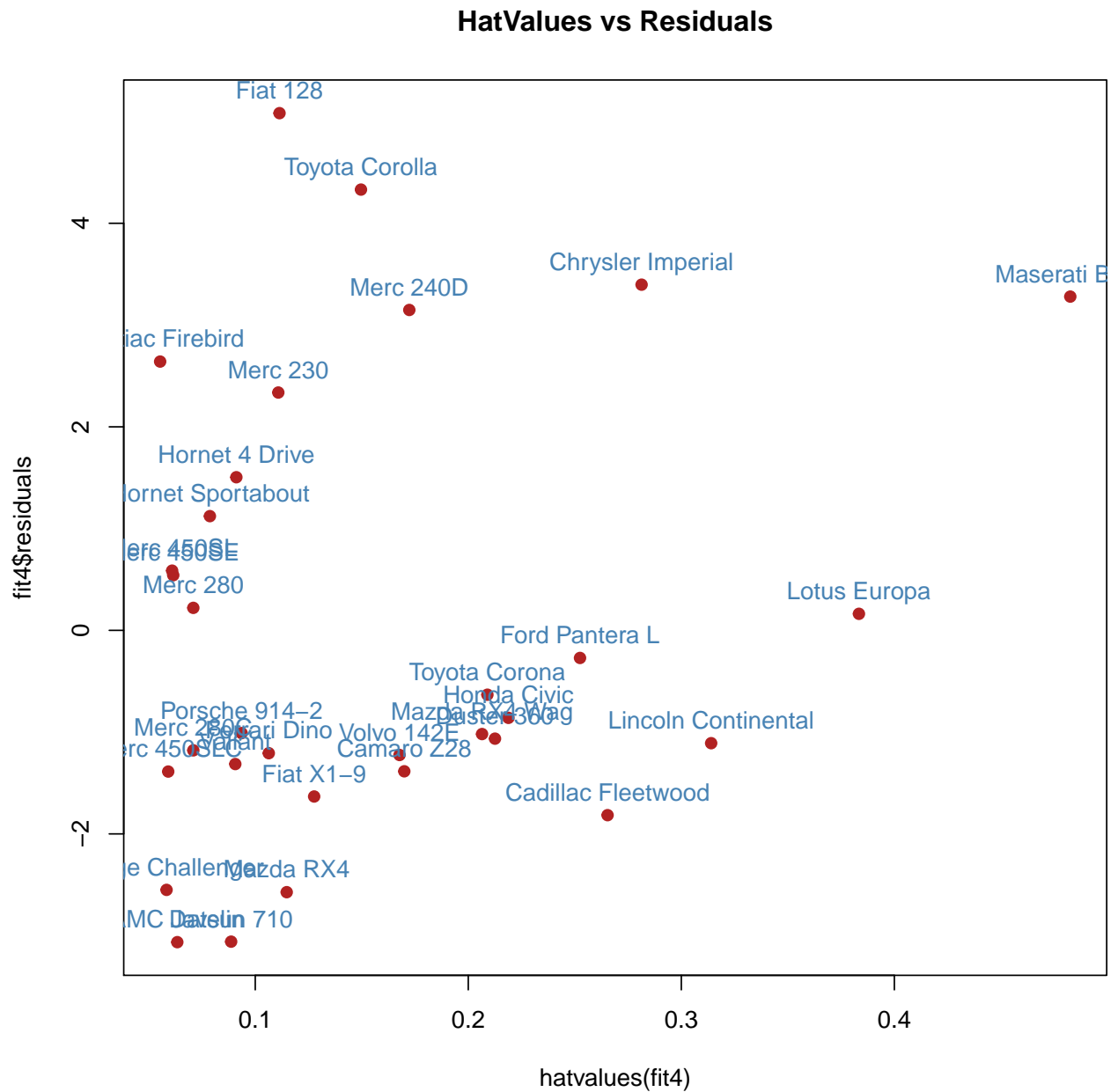
Weight vs Transmission Type



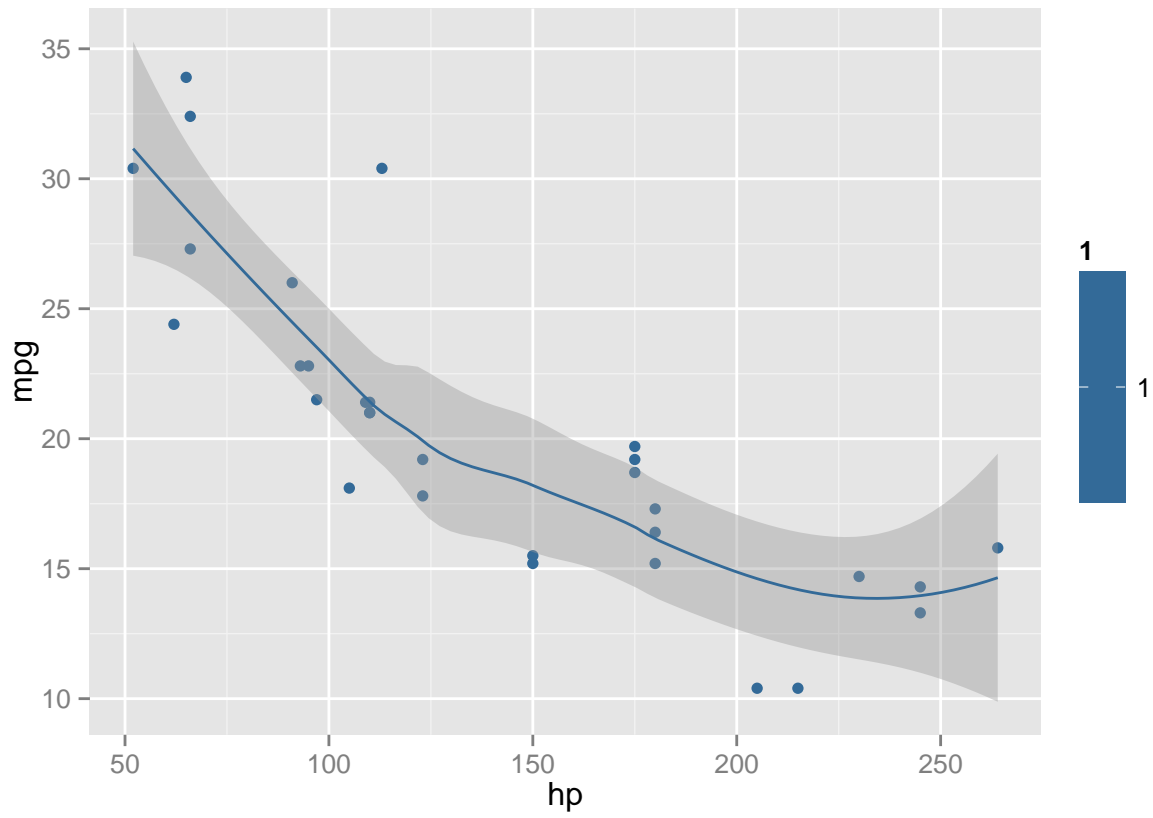
Horsepower vs Trans. Type



Checking for High Influence Points



Quadratic Response of MPG with Horsepower



Comparing models with and without Horsepwer Squared Term

