Exploring and modeling data on American fuel consumption

The fuel dataset

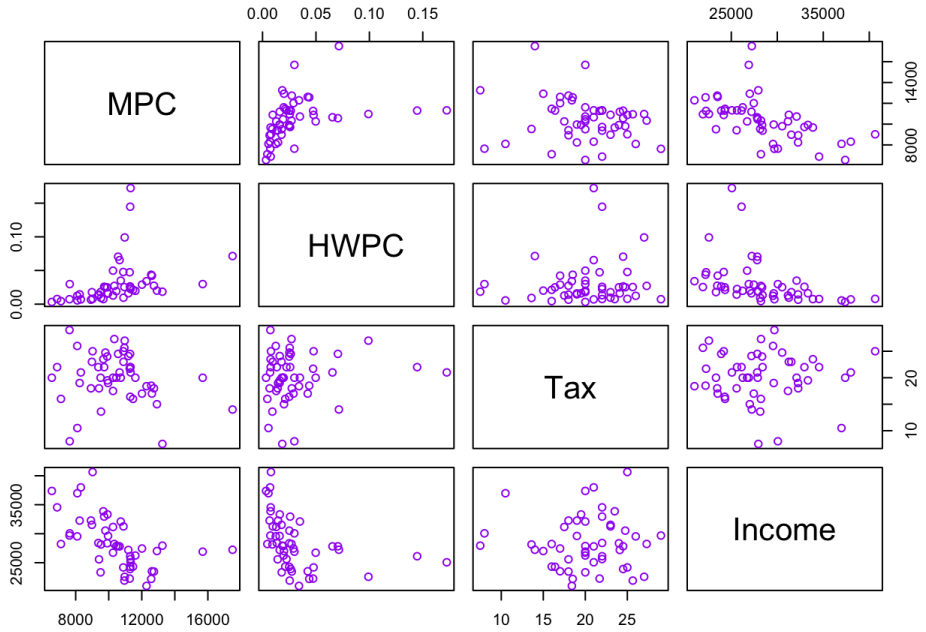
```
## [1] "Drivers" "FuelC"   "Income"  "Miles"   "MPC"     "Pop"     "Tax"
```

The data set “fuel2001” is about fuel consumption in the US, broken down by state, including DC. It contains information about the number of licensed drivers (‘Drivers’), the amount of gasoline sold for road use (‘FuelC’, in thousands of gallons), the number of miles of federal-aid highway, per-capita income as of the year 2000, estimated numer of miles driven per-capita, etc.

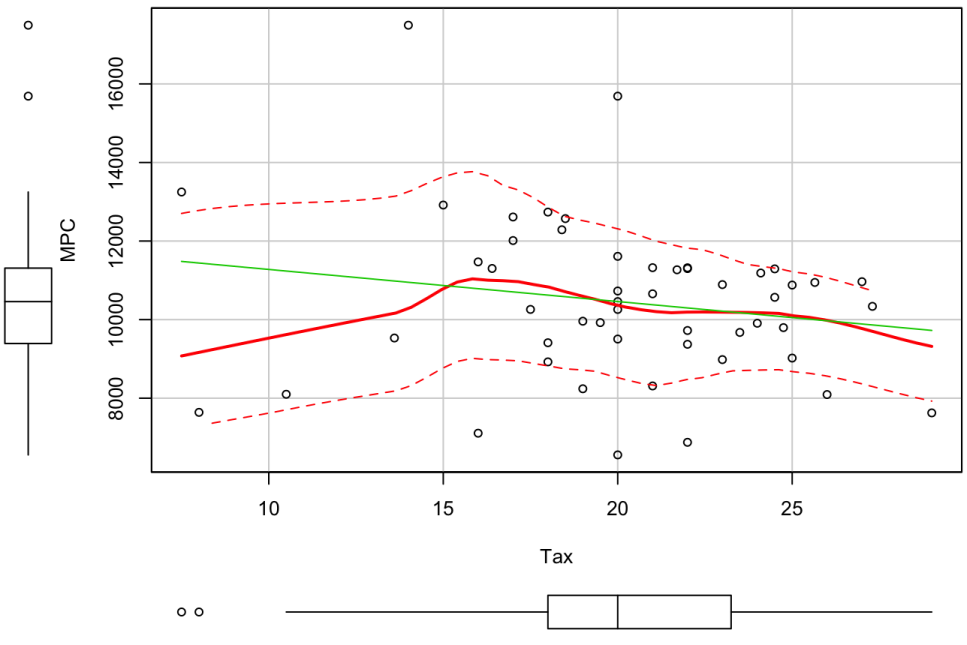
Fancy graphs

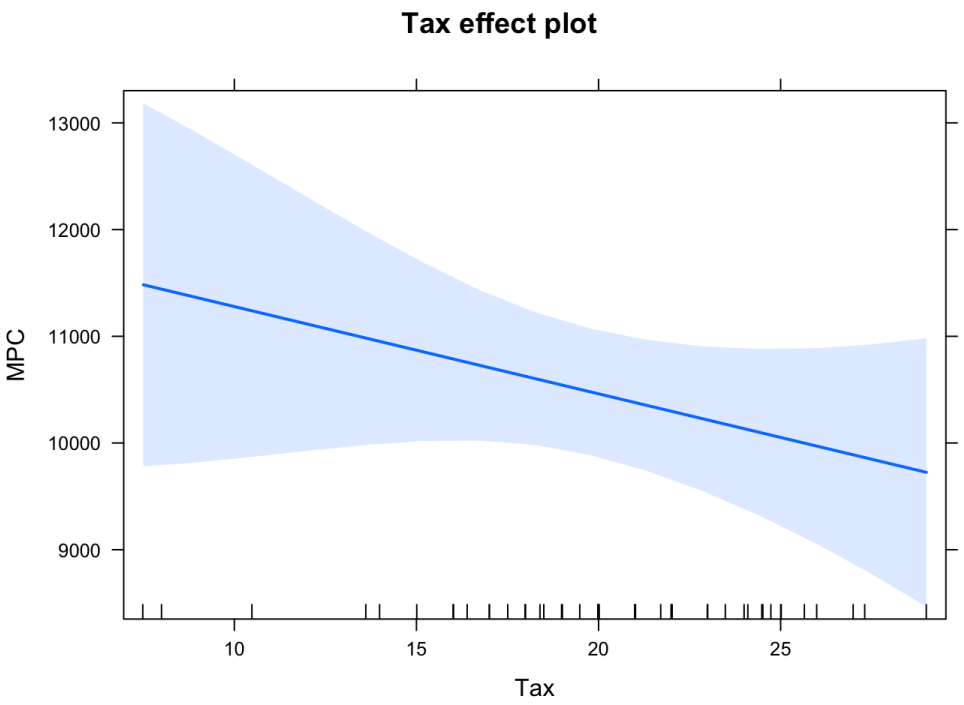
Here’s a scatterplot matrix of the data, with Highway Miles transformed to Highway Miles Per Capita:

```
fuel2001$HWPC <- fuel2001$Miles/fuel2001$Pop
# ggpairs(fuel2001,columns=c('MPC','Tax','HWPC','Income'),aes(color=I('purple')))
pairs(MPC~HWPC+Tax+Income,fuel2001,col='purple')
```



The following fancy graphs might make it look like there’s a relationship between the gas tax and the number of miles driven per capita:





But a quick statistical test casts doubt on this relationship:

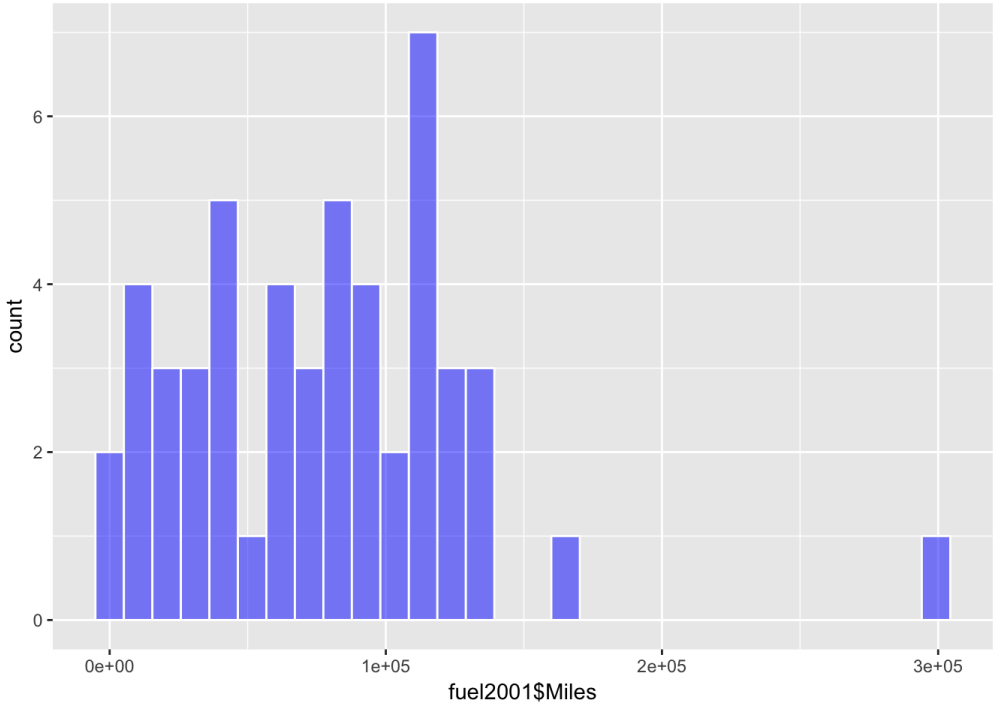
```
##
## Call:
## lm(formula = MPC ~ Tax, data = fuel2001)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3905.0 -1121.9   267.4  1041.4  6543.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12096.06    1301.83   9.292 2.18e-12 ***
## Tax           -81.76      63.04  -1.297    0.201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2026 on 49 degrees of freedom
## Multiple R-squared:  0.03319,    Adjusted R-squared:  0.01346
## F-statistic: 1.682 on 1 and 49 DF,  p-value: 0.2007
```

The p-value for the test is relatively high (0.2), suggesting that a similarly strong apparent relationship between variables like “MPC”" and “Tax”" would occur just by chance about 20% of the time assuming there is in fact no relationship between the two variables.

Exploring the dataset

Let’s back up and get acquainted with the data.

First, to get a handle on some of the main variables, a few plots:



Notice that the number of miles of federal-aid highway is very skewed among states: most states have a moderate length of federal-aid highway, with just one having quite a large amount. Which is the outlying state?

```
which.max(fuel2001$Miles)  # returns an index
```

```
## [1] 44
```

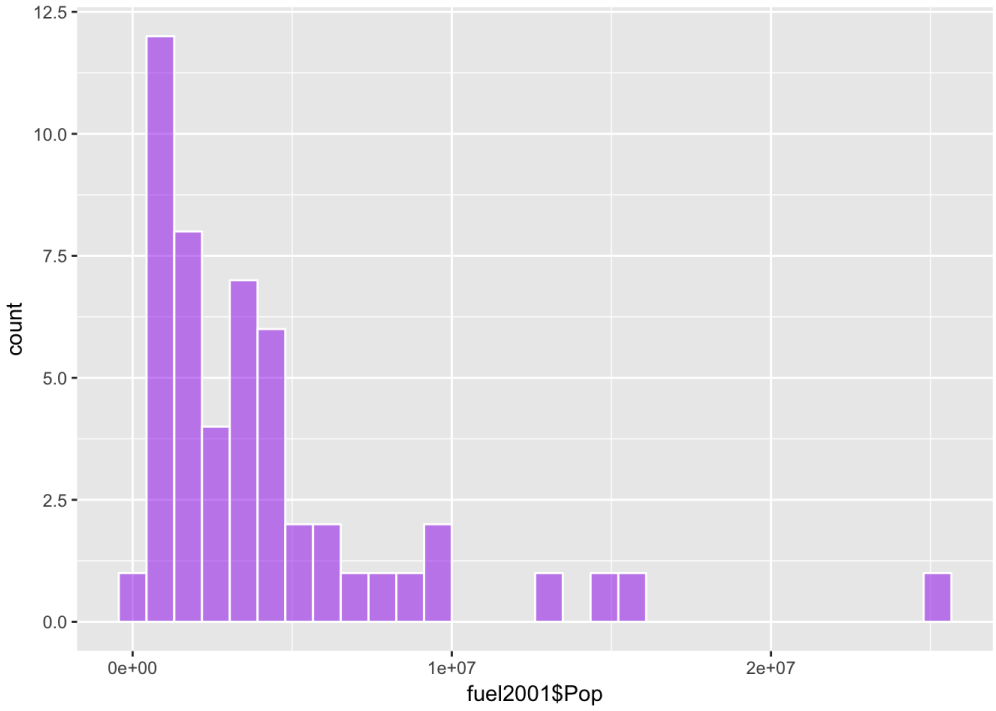
```
rownames(fuel2001)[which.max(fuel2001$Miles)]
```

```
## [1] "TX"
```

The outlier is Texas. Now let's take a look at the populations:

```
ggplot(data=fuel2001,aes(fuel2001$Pop))+geom_histogram(col=I('white'),fill=I('purple'),alpha=.5)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Most states have populations below 10 million; which state is most populous?

```
## [1] "CA"
```

What are some of the other most populous states?

```
order(fuel2001$Pop)
```

```
##  [1] 51  2  9 46 35 42  8 27 40 12 30 13 20 28 32 49 29 45 17  4 25 16  7
## [24] 37 38 41 18  6 19  1 24  3 21 50 26 43 48 15 22 47 11 34 31 23 36 14
## [47] 39 10 33 44  5
```

```
head(rownames(fuel2001)[order(fuel2001$Pop, decreasing = TRUE)], 10)
```

```
##  [1] "CA" "TX" "NY" "FL" "PA" "IL" "OH" "MI" "NJ" "NC"
```

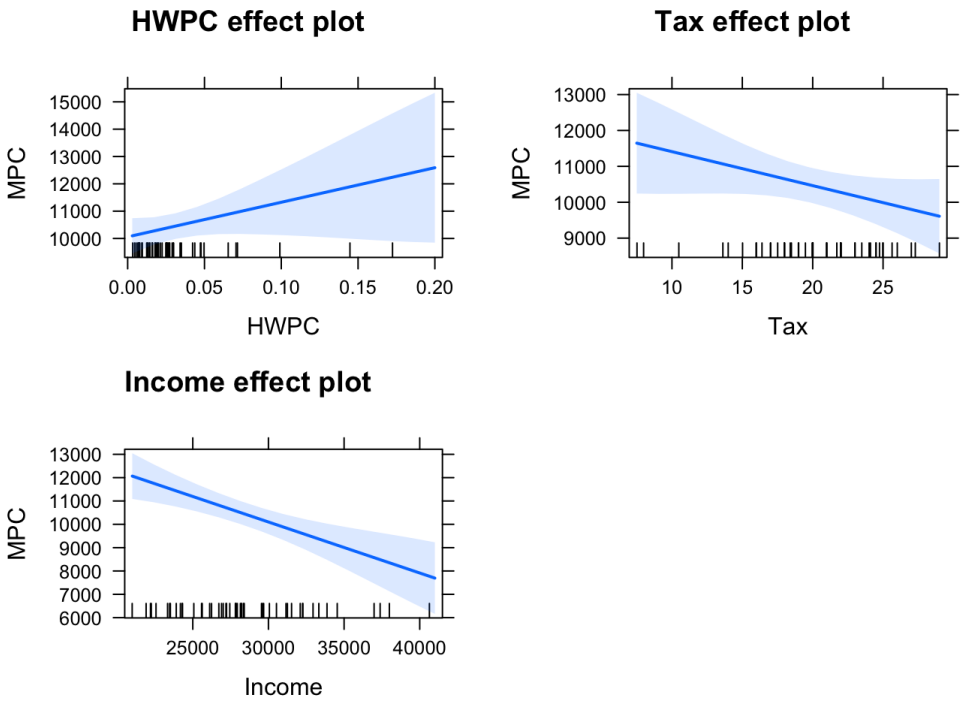
Modeling Miles Driven per Capita

Now let's try to predict the number of miles driven per capita from some of the other variables. Let's try using income, gas tax, and number of miles of federal highway per capita as predictors.

```
mMHTI <- lm(MPC ~ HWPC + Tax + Income, fuel2001)
summary(mMHTI)
```

```
##
## Call:
## lm(formula = MPC ~ HWPC + Tax + Income, data = fuel2001)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3585.3  -946.2  -115.7   663.4  5691.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.818e+04  2.004e+03   9.070 6.79e-12 ***
## HWPC         1.265e+04  7.935e+03   1.595 0.117476
## Tax         -9.472e+01  5.198e+01  -1.822 0.074795 .
## Income      -2.186e-01  5.767e-02  -3.790 0.000428 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1658 on 47 degrees of freedom
## Multiple R-squared:  0.3792, Adjusted R-squared:  0.3396
## F-statistic: 9.569 on 3 and 47 DF,  p-value: 4.814e-05
```

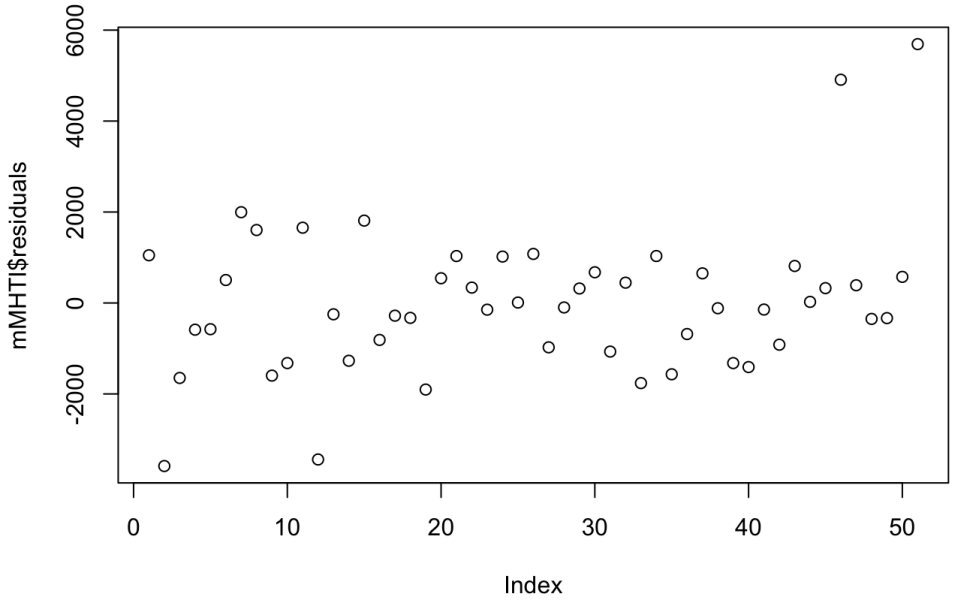
```
plot(allEffects(mMHTI))
```



These plots show the relationships among each of the predictors and the response variable (miles driven per capita). Note that the fact that these are “effect” plots does not mean that any of these variables actually “affects” the number of miles driven per capita: for example, this model is not good evidence that high gas taxes cause people to drive less.

It is a good idea to look at the discrepancies between the modeled values and the actual values to see whether there are any patterns:

```
plot(mMHTI$residuals)
```



There is no clear pattern in the residuals, though there are a couple of interesting outliers:

```
tail(sort(mMHTI$residuals))
```

```
##       DE       GA       IN       CT       VT       WY
## 1604.168 1655.326 1810.716 1995.909 4908.587 5691.663
```

Wyoming and Vermont have a surprisingly high number of miles driven per capita given their respective per-capita income, federal highway miles per capita, and gas tax rates.

Now we can try a simple prediction exercise: we can predict Georgia’s miles per capita using the same predictors, but leaving Georgia’s original data out of the model.

Here’s Georgia’s data:

```
fuel2001["GA",]
```

```
##    Drivers   FuelC Income  Miles     MPC     Pop Tax       HWPC
## GA 5833802 4693703  27940 115534 13248.6 6250708 7.5 0.01848335
```

And here’s a new model with a dataset that leaves out Georgia’s data:

```
fuel2001GA <- fuel2001[rownames(fuel2001) !='GA',]
model.MHTI.noGA <- lm(MPC~HWPC+Tax+Income,fuel2001GA)
```

```
summary(model.MHTI.noGA)
```

```
##
## Call:
## lm(formula = MPC ~ HWPC + Tax + Income, data = fuel2001GA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3295.7  -903.5  -124.4   679.3  5877.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.759e+04  2.069e+03   8.504 5.41e-11 ***
## HWPC         1.281e+04  7.918e+03   1.618 0.112594
## Tax         -7.024e+01  5.642e+01  -1.245 0.219473
## Income      -2.169e-01  5.756e-02  -3.769 0.000466 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1654 on 46 degrees of freedom
## Multiple R-squared:  0.371,  Adjusted R-squared:   0.33
## F-statistic: 9.043 on 3 and 46 DF,  p-value: 8.107e-05
```

We can use R's 'predict' command to get a prediction for Georgia's number of miles driven per capita.

```
GApredict <- predict(model.MHTI.noGA,newdata=fuel2001['GA',],interval='prediction')[1]
GAactual <- fuel2001$MPC[rownames(fuel2001)=='GA']
GAactual
```

```
## [1] 13248.6
```

Here are the two values compared side-by-side:

```
com <- matrix(c(GAactual,GApredict),ncol=2)
colnames(com) <- c("Actual","Predicted")
com
```

```
##       Actual Predicted
## [1,] 13248.6  11241.43
```

We can also get a prediction interval, constructed so that it has a 95% chance of containing Georgia's true MPC.

```
p <- matrix(predict(model.MHTI.noGA,newdata=fuel2001['GA',],interval='prediction')[2:3],ncol=2)
colnames(p) <- c("Lower bound", "Upper bound")
p
```

```
##      Lower bound Upper bound
## [1,]    7575.883    14906.97
```

Notice that Gerogia's true value of 13,248.6 miles driven per capita is in fact within this interval. But notice also that this prediction interval is fairly wide (7,575 to 14,906 miles driven per capita), as would be expected with such a small data set - even counting Washington DC, 51 data points is not many.