Case Study 5: Hypothesis testing: Tax and Fuel consumption in the USA

Richard Wilkinson

The data

US Federal Highway Administration collected the following data in order to understand the effect of state gasoline tax on fuel consumption.

They collected information on the following quantities:

- TAX =Gasoline state tax rate, cents per gallon
- DLIC = Number of licenced drivers per 1000 people in the state
- INC = Per capita personal income for the year 2000 (in \$1000s) for the state
- ROAD = Miles of federal-aid highway (in 1000s) for the state
- FUEL = Gallons of gasoline sold for road use per capita
- State = State name

Lets start by downloading the data

```
filepath <- "https://www.maths.nottingham.ac.uk/personal/pmzrdw/FuelData.txt"

# Download the data from the internet
download.file(filepath, destfile = "FuelData.txt", method = "curl")
FuelData <- read.table(file='FuelData.txt', header=TRUE,sep="&")
FuelData[1:10,]
```

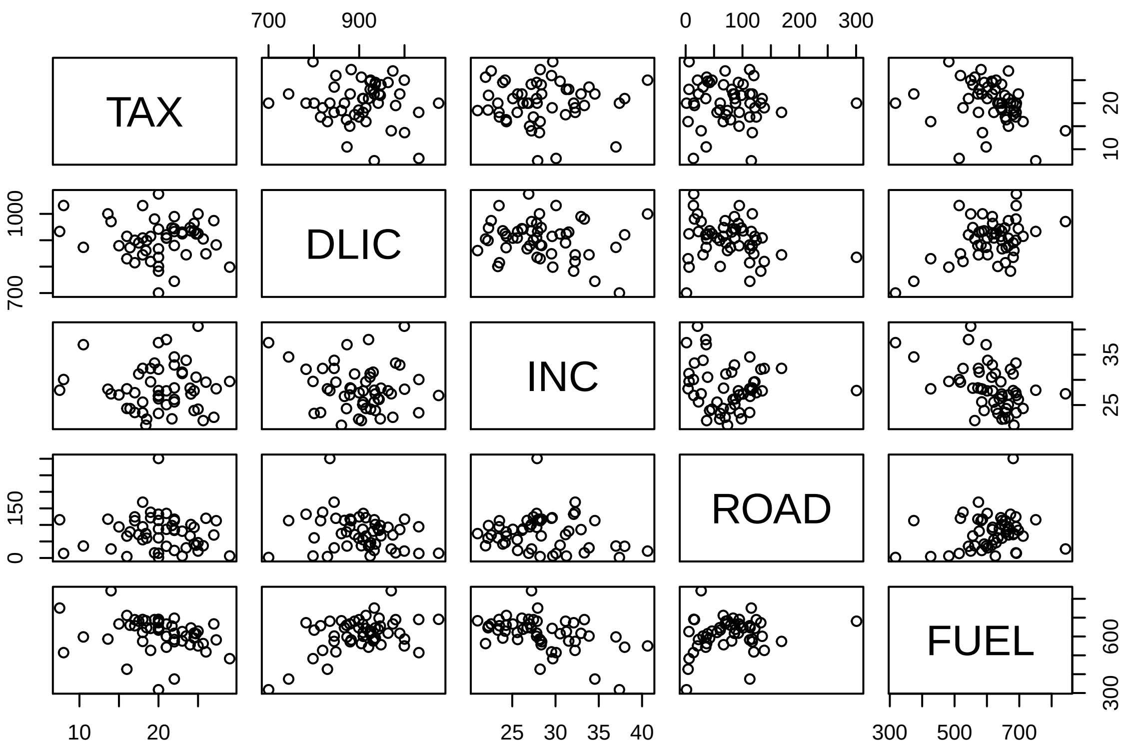
```
##             State  TAX    DLIC    INC    ROAD   FUEL
## 1  Alabama       18.0 1031.38 23.471  94.440 690.26
## 2  Alaska         8.0 1031.64 30.064  13.628 514.27
## 3  Arizona       18.0  908.59 25.578  55.245 621.47
## 4  Arkansas      21.7  946.57 22.257  98.132 655.29
## 5  California    18.0  844.70 32.275 168.771 573.91
## 6  Colorado      22.0  989.60 32.949  85.854 616.61
## 7  Connecticut   25.0  999.59 40.640  20.910 549.99
## 8  Delaware      23.0  924.34 31.255   5.814 626.02
## 9  Dist of Col   20.0  700.19 37.383   1.534 317.49
## 10 Florida       13.6 1000.12 28.145 117.299 586.34
```

```
str(FuelData) # look at the data structure
```

```
## 'data.frame':    51 obs. of  6 variables:
##  $ State: Factor w/ 51 levels "Alabama      ",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ TAX  : num  18 8 18 21.7 18 22 25 23 20 13.6 ...
##  $ DLIC : num  1031 1032 909 947 845 ...
##  $ INC  : num  23.5 30.1 25.6 22.3 32.3 ...
##  $ ROAD : num  94.4 13.6 55.2 98.1 168.8 ...
##  $ FUEL : num  690 514 621 655 574 ...
```

Lets first visualise the data with a scatter-plot matrix

```
plot(FuelData[,-1])  ## -1 so as not to plot the state name
```



or we can use the following, which draws a smooth on the data suggesting the general trend (useful when n is large).

```
require(car)
scatterplotMatrix(FuelData[,-1]) ## gives slightly more info
```

The plots give the impression that FUEL decreases on average with TAX, but it is hard to say anything for certain as there is a lot of variation. The impression is that FUEL is at best weakly related to the other variables.

However, the scatter-plot matrix just shows marginal relationships between pairs of variables (i.e. FUEL vs TAX ignores the information in DLIC, ROAD and INC). It doesn't help us to understand how fuel is related to all four predictors simultaneously.

Multiple linear regression

Consider the multiple linear regression model

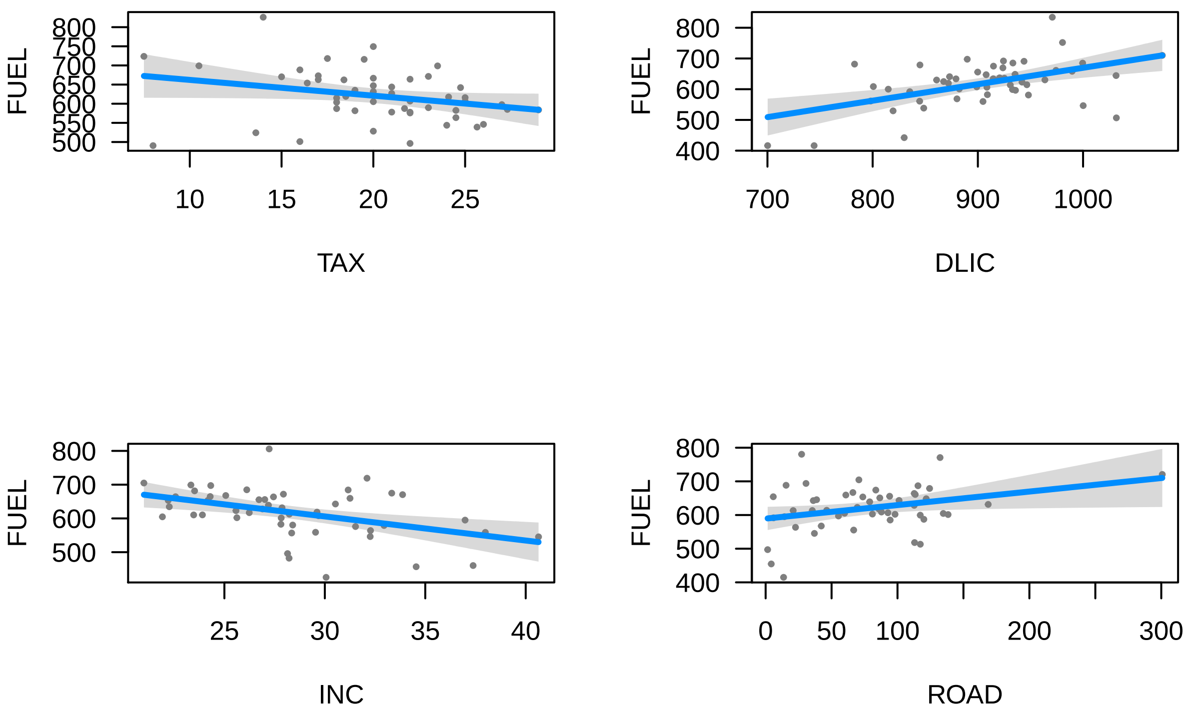
$$FUEL_i = \beta_0 + \beta_1 TAX_i + \beta_2 DLIC_i + \beta_3 INC_i + \beta_4 ROAD_i + \epsilon_i$$

```
fit <- lm(FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData  )
coef(fit)
```

```
## (Intercept)          TAX         DLIC          INC         ROAD
## 383.9544095  -4.1145064   0.5352993  -7.1373711   0.4015963
```

The coefficients tell us how important each variable is for predicting the fuel consumption, but its *always* nice to visualise things when possible. Visualising a model with 4 inputs isn't easy, but the R package visreg helps.

```r
library(visreg) # You will have to install the package first time you use it
# install.packages('visreg')
par(mfrow=c(2,2))
visreg(fit)
```



Testing

The key question we want to answer:

- Is TAX useful for predicting FUEL after including ROAD, INC, DLIC?

Test $H_0 : \beta_1 = 0$; vs $H_1 : \beta_1 \neq 0$;

In this case we can use a t-test or an F-test as there is just a single constraint, and $F_{1,n-p} = (t_{n-p})^2$. The test statistic is:

$$T = \frac{\widehat{\beta}_1}{\text{std.error}(\widehat{\beta}_1)} \sim t_{n-p} \quad \text{under } H_0.$$

i.e. we should reject H_0 at the $100\alpha\%$ level if

$$|T_{obs}| \geqslant t_{46}\left(1 - \alpha/2\right).$$

R automatically carries out this test when you run the summary command.

```r
summary(fit)
```

```
##
## Call:
## lm(formula = FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -179.900  -35.945    5.394   36.820  180.187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 383.9544   165.7540   2.316 0.025044 *
## TAX          -4.1145     2.1074  -1.952 0.056988 .
## DLIC          0.5353     0.1373   3.898 0.000313 ***
## INC          -7.1374     2.2054  -3.236 0.002247 **
## ROAD          0.4016     0.1873   2.144 0.037315 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.17 on 46 degrees of freedom
## Multiple R-squared:  0.4755, Adjusted R-squared:  0.4299
## F-statistic: 10.43 on 4 and 46 DF,  p-value: 4.271e-06
```

We can simply read off the t-statistic and the corresponding p-value. R also provides a visual indication of the significance, with '.' in this case, showing that the p-value is between 0.05 and 0.1, i.e., not enough evidence to reject H_0 at the 5% level.

- This table suggests that TAX is the only variable that does not contain much information about the response once the other variables have been considered.
- It is important to note that the t-tests on each β_j for including that parameter are not independent. For example if two of the input variables were not significant then leaving just one of them out of the model may cause the other one to become significant.
- Conversely, if only TAX and the intercept are included in the model, then TAX might well be significant.

This contains all of the information in a concise format. You should make sure you understand what every number in this output means, how to interpret it, and how to calculate it. If you wanted to do the analysis using the F-test, you could type

```r
fit2 <- lm(FUEL ~ DLIC+INC+ROAD, data=FuelData)
anova(fit, fit2)
```

Test for the existence of regression

We want to test whether the full model is a significant improvement over the null model, i.e., test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0;\ \beta_0 \text{ arbitrary}$$

vs

$$H_1 : \beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ \beta_4 \text{ arbitrary}$$

```r
fit0<-lm(FUEL~1, data=FuelData)
anova(fit0, fit) ## compares the full model with the null model
```

```
## Analysis of Variance Table
##
## Model 1: FUEL ~ 1
## Model 2: FUEL ~ TAX + DLIC + INC + ROAD
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     50 395700
## 2     46 207533  4    188167 10.427 4.271e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can read from the table that $F = 10.43$ and $p < 0.001$.

- This implies that the full model is a significant improvement on the null model, i.e. that at least one of the input variables is informative about the response variable.
- It does not imply that *all* of the input variables are informative though.

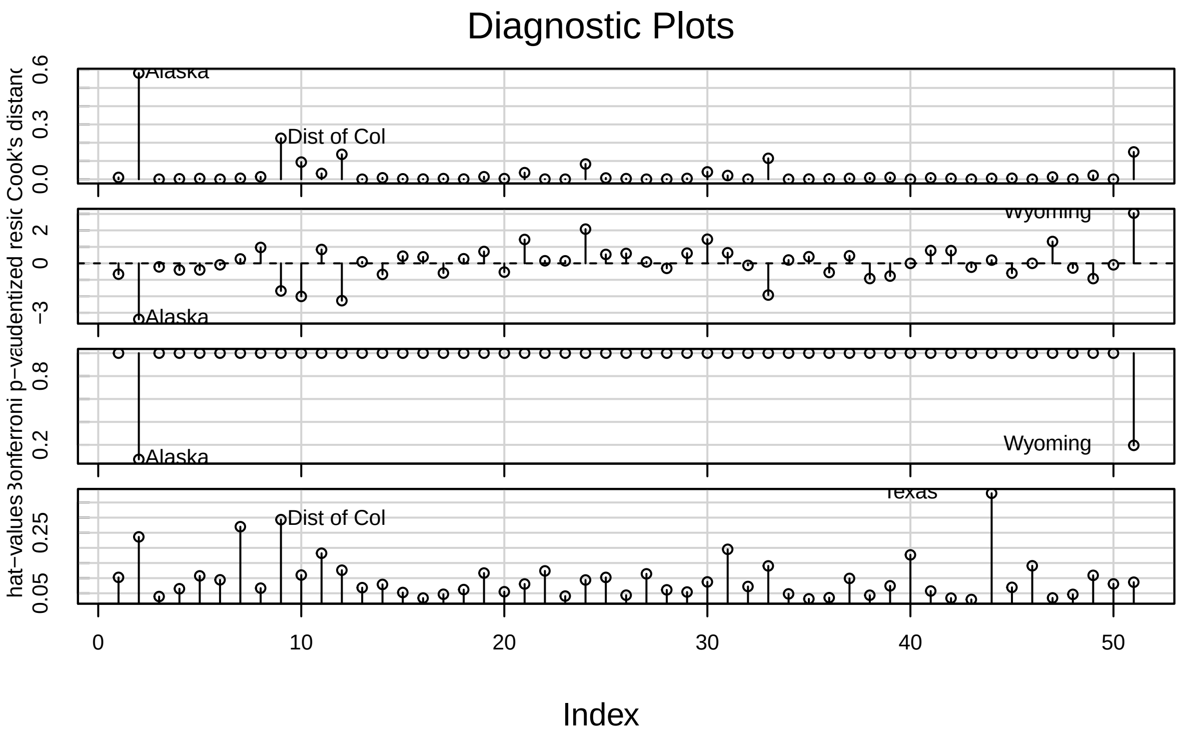
Since $F > F_{4,46}(0.99) = 3.76$ we have strong evidence that at least some of the explanatory variables are useful for prediction.

Detecting outliers

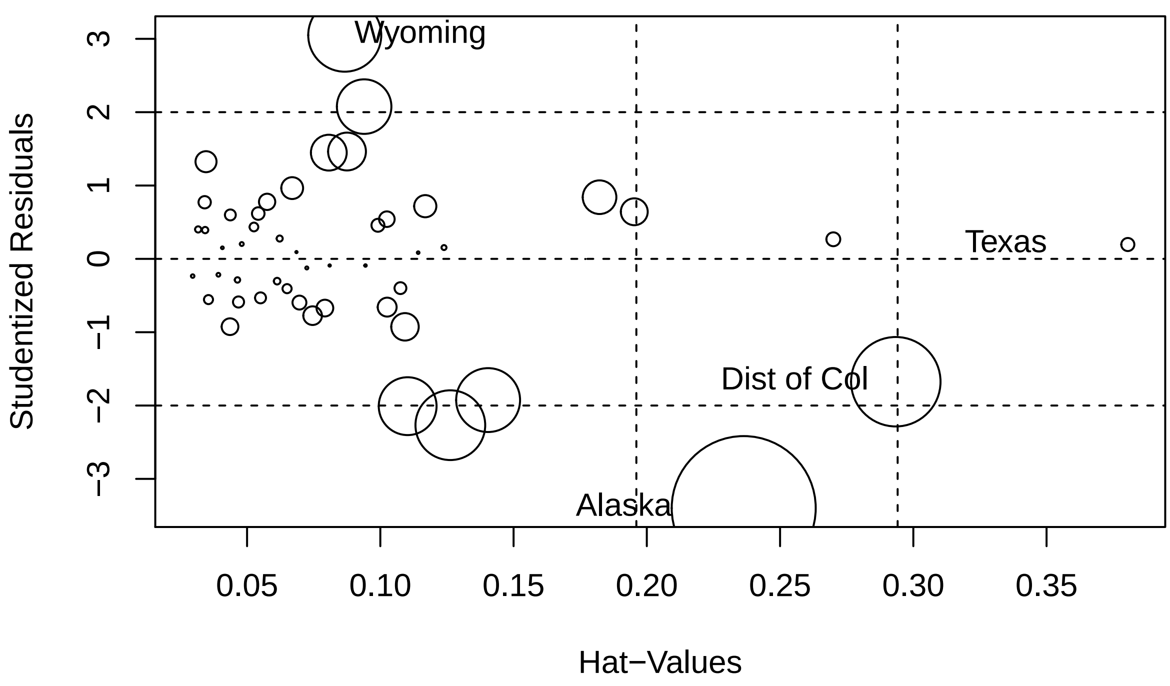
```r
require(car)
```

```
## Loading required package: car
```

```r
influenceIndexPlot(fit, id.n=2, labels=FuelData$State)
```




```r
influencePlot(fit, id.n=2, labels=FuelData$State)
```




```
##                  StudRes        Hat      CookD
## Alaska        -3.3982194 0.23636154 0.76257332
## Dist of Col   -1.6753422 0.29343215 0.47361957
## Texas          0.1954799 0.38048945 0.06923925
## Wyoming        3.0497586 0.08664797 0.38664771
```

These suggest that Alaska is by far the most influential point as it is a large outlier and has reasonably high leverage. It could be argued that it is an unusual state and should be left out of the analysis.

```
FuelData2 <- FuelData[-2,] # remove Alaska - alternatively use select command in dply package
fit3 <- lm(FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData2)
compareCoefs(fit, fit3)
```

```
##
## Call:
## 1: lm(formula = FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData)
## 2: lm(formula = FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData2)
##              Est. 1    SE 1  Est. 2    SE 2
## (Intercept) 383.954 165.754 354.986 149.741
## TAX          -4.115   2.107  -6.826   2.061
## DLIC          0.535   0.137   0.626   0.127
## INC          -7.137   2.205  -6.657   1.994
## ROAD          0.402   0.187   0.311   0.171
```

```
summary(fit3)
```

```
##
## Call:
## lm(formula = FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.115  -29.596    4.543   31.075  148.863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 354.9862   149.7406   2.371  0.02210 *
## TAX          -6.8258     2.0614  -3.311  0.00184 **
## DLIC          0.6256     0.1267   4.939 1.13e-05 ***
## INC          -6.6575     1.9941  -3.339  0.00170 **
## ROAD          0.3113     0.1710   1.821  0.07527 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.58 on 45 degrees of freedom
## Multiple R-squared:  0.5718, Adjusted R-squared:  0.5338
## F-statistic: 15.03 on 4 and 45 DF,  p-value: 7.127e-08
```

This has made us much more certain that TAX has an impact on FUEL usage. We can check if removing any of the other points with large Cook's distance has much effect, but only perhaps Dist. of Col. and Hawaii can be justified on the grounds of being unusual in some way.

```
FuelData3 <- FuelData[-c(2,9,12),]
fit4 <- lm(FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData3)
compareCoefs(fit, fit3, fit4, se = FALSE)
```

```
##
## Call:
```

```
## 1: lm(formula = FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData)
## 2: lm(formula = FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData2)
## 3: lm(formula = FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData3)
##                Est. 1   Est. 2   Est. 3
## (Intercept) 383.9544 354.9862 604.7502
## TAX          -4.1145  -6.8258  -8.3581
## DLIC          0.5353   0.6256   0.4021
## INC          -7.1374  -6.6575  -6.2424
## ROAD          0.4016   0.3113   0.0387
```

```r
summary(fit4)
```

```
##
## Call:
## lm(formula = FUEL ~ TAX + DLIC + INC + ROAD, data = FuelData3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -135.706  -29.059    0.283   26.639  133.670
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 604.75021  144.65077   4.181  0.00014 ***
## TAX          -8.35815    1.83258  -4.561  4.2e-05 ***
## DLIC          0.40207    0.12623   3.185  0.00269 **
## INC          -6.24245    1.76758  -3.532  0.00100 ***
## ROAD          0.03872    0.16348   0.237  0.81388
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.63 on 43 degrees of freedom
## Multiple R-squared:  0.5358, Adjusted R-squared:  0.4926
## F-statistic: 12.41 on 4 and 43 DF,  p-value: 8.545e-07
```

Again, removing these two additional states has strengthed the evidence against H_0. Finally, to be confident in our conclusion, we should check that there are no obvious violations of the modelling assumptions.