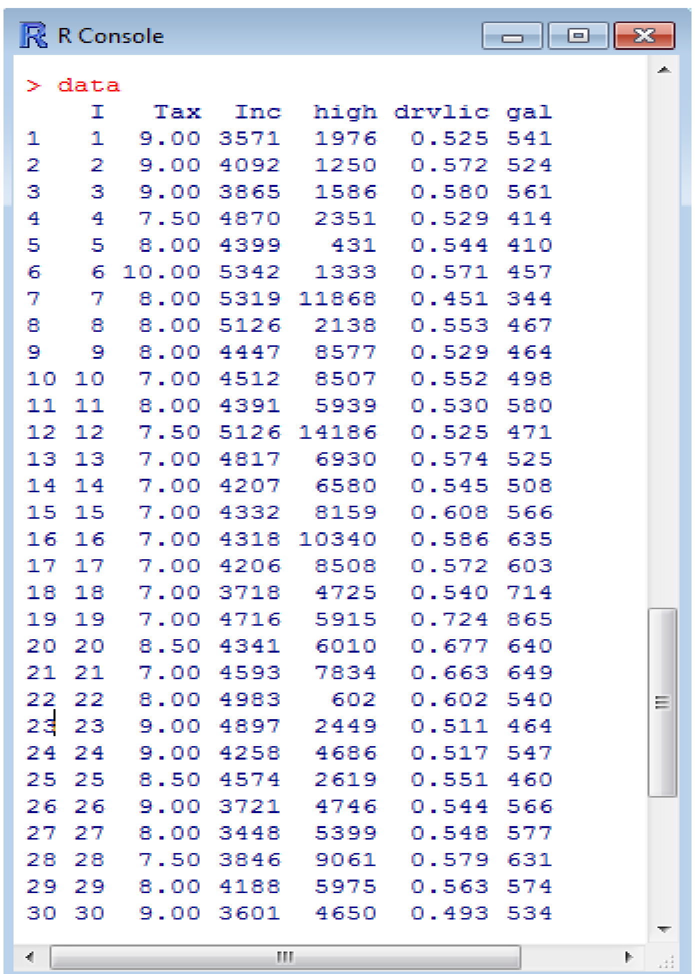
# General Linear Modeling

**Introduction:**

- General Linear Modeling is the new hot thing in statistics. The idea of this category of statistical tests is that it tries to fit the values of a dependent variable, Y, to an explanatory variable using a linear function.
- We found data that recorded the petroleum tax (cents per gallon), average income (dollars), paved highways (miles), Proportion of population with driver's licenses, and the gallons of petroleum consumed (millions). The gallons of petroleum consumed is the dependent variable because the explanatory variables affect this directly.

**Multiple Linear Regression:**

- Since our data has more than one explanatory variable we can compare them all at once using in R using multiple regression. The data looks like this:



- **Model**

$Y_i = \beta_1 X_{1\,i} + \beta_2 X_{2\,i} + \beta_3 X_{3\,i} + \alpha + \varepsilon_i$

Where:
$\alpha$ = is the intercept of the regression line
$\beta i$ = is the slope of the regression line for each X variable
$\varepsilon i$ = the residuals of the prediction of Y given they are normal and random

- **Overall F-test for Multiple Regression:**
  - $H_0$: all $\beta$s = 0 --> there is no relation between the variables
  - $H_1$: at least 1 $\beta$s ≠ 0

- **Partial T/F-tests of single Coefficients:**
  - Hypotheses:
    - $H_0$: A single $\beta$ = 0 --> There is no relation between the variables
    - $H_1$: A single $\beta$ ≠ 0 --> There is a relation between the variable
  - Test Statistic:
    - $t_j$: $b_j / sb_j$
    - $F_j$: $t_j^{\,2}$
  - Decision Rule:
    - If $P < \alpha$ reject the null hypothesis
    - If $P > \alpha$ fail to reject the null hypothesis

R Console

```
> Y=gal
> X1=Tax
> X2=Inc
> X3=high
> X4=drvlic
>
> LM=lm(Y~X1+X2+X3+X4)
> summary(LM)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)

Residuals:
    Min      1Q  Median      3Q     Max
-122.03  -45.57  -10.66   31.53  234.95

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.773e+02  1.855e+02   2.033 0.048207 *
X1          -3.479e+01  1.297e+01  -2.682 0.010332 *
X2          -6.659e-02  1.722e-02  -3.867 0.000368 ***
X3          -2.426e-03  3.389e-03  -0.716 0.477999
X4           1.336e+03  1.923e+02   6.950 1.52e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66.31 on 43 degrees of freedom
Multiple R-squared: 0.6787,     Adjusted R-squared: 0.6488
F-statistic: 22.71 on 4 and 43 DF,  p-value: 3.907e-10
```

## Conclusion:

- The overall p-value ($3.907 \times 10^{-10}$) is less than alpha of .05 (chosen a priori), so we reject the null hypothesis and conclude that at least one of the slopes is not equal to zero.
- Looking at the partial t-tests, we see that the p-value for the t-test regarding variable X3 fails to reject the null hypothesis. This means that this single β is equal to 0. It would follow that since this variable seems to not be relating to the dependent variable, we can take that variable out of the linear model in order to achieve a reduced, optimal model that will reduce the sums of square error and create a more parsimonious model.

- Here is the new reduced model without variable X3:

R Console

```
> RM1=lm(Y~X1+X2+X4)
> summary(RM1)

Call:
lm(formula = Y ~ X1 + X2 + X4)

Residuals:
    Min      1Q  Median      3Q     Max 
-110.10  -51.22  -12.89   24.49  238.77 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  307.32790  156.83067   1.960 0.056394 .  
X1           -29.48381   10.58358  -2.786 0.007848 ** 
X2            -0.06802    0.01701  -3.999 0.000240 ***
X4          1374.76841  183.66954   7.485 2.24e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 65.94 on 44 degrees of freedom
Multiple R-squared: 0.6749,     Adjusted R-squared: 0.6527 
F-statistic: 30.44 on 3 and 44 DF,  p-value: 8.235e-11 

> |
```

- Conclusion for RM1:
  - Now the p-value for the overall F-test is $8.235 \times 10^{-11}$. This is still less than alpha so the null hypothesis is still rejected.
  - Now looking at the partial t-tests, and none of the p-values are greater than alpha, so variables X1, X2, X3 seem all to be important and relate to the dependent variable.

**Comparing Models:**

- Now that there are two models, it is beneficial to compare them using the general F-test for model comparison.
- This test will determine whether the fewer variables of the reduced model are sufficient enough to explain the dependent variable, Y in the data set.

GLM F-test:

Assumptions:

- Y--->$Y_n$ is random and normal and specified before the test.
- $X_{k1}$--->$X_{k,n}$ are fixed variables and are matched to a specific Y

Model:

- $Yi = \beta_0 + \Sigma\beta_j X_i + \varepsilon i$
- $Yi = \beta_0 + \Sigma\beta_k X_i + \varepsilon i$

Where:

$\beta_0$ = the y intercept of the regression line

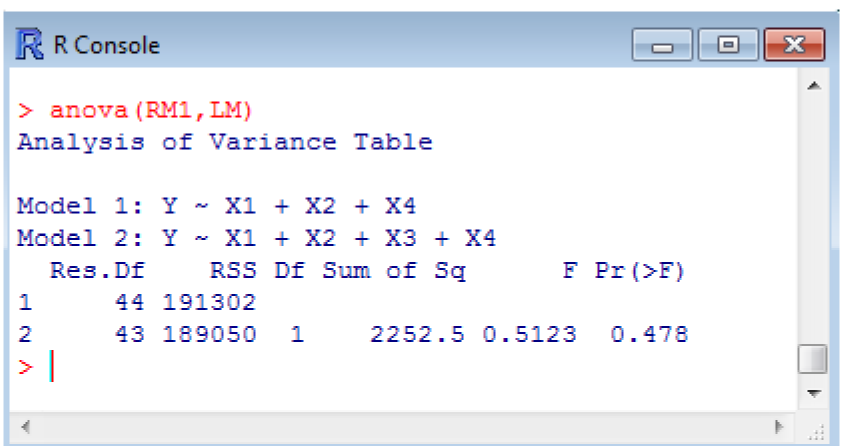
$\beta_j$ = are the slopes for the full model

$\varepsilon_i$ = the error factor of Y and a random normal variable

Hypothesis:

- $H_0$: the slope coefficients in J but not included in K are = 0.
- $H_1$: at least one of these coefficients is not 0.

Example:




```
> anova(RM1,LM)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 191302
2     43 189050  1    2252.5 0.5123  0.478
>
```

Conclusion:

- Since the p-value is .478, you fail to reject the null hypothesis and conclude that the reduced model is sufficient enough to describe the dependent variable, Y.
- So now I am sure that I have a good reduced model with RM1.

**Choosing an optimal linear model**

- Creating an optimal linear model is sometimes very difficult. There are ways to help you to reduce a full model into an optimal linear model based on parsimony.

Akaike's information criterion

- This is a criterion that reports the parsimony of the model. This can be utilized using the extractAIC() function in R.
- A good way to go about doing this is utilizing the automated tests in R.

Drop 1:

This is an automated test that drops one variable at a time from a specified full model and calculates the AIC for each independent varaible. Here is a the Drop one applied to my example

```
R Console
> drop1(LM)
Single term deletions

Model:
Y ~ X1 + X2 + X3 + X4
       Df Sum of Sq    RSS    AIC
<none>              189050 407.37
X1      1     31632 220682 412.80
X2      1     65729 254779 419.69
X3      1      2252 191302 405.94
X4      1    212355 401405 441.51
>
```

Conclusion:

- Since variable X3 has the lowest AIC, it would follow to possibly eliminate that variable from the full model and only have variables X1, X2, and X4 in the reduced model…the same conclusion that was taken from the summary() function shown above.

Backwards Stepwise Regression:

- This is another automated way of utilizing AIC to create an optimal linear model. This removes the variable with the smallest AIC for each round until a reduced model comes about that has the lowest AIC.

R Console

```
> step(LM,direction="backward")
Start:  AIC=407.37
Y ~ X1 + X2 + X3 + X4

       Df Sum of Sq    RSS    AIC
- X3    1      2252 191302 405.94
<none>              189050 407.37
- X1    1     31632 220682 412.80
- X2    1     65729 254779 419.69
- X4    1    212355 401405 441.51

Step:  AIC=405.94
Y ~ X1 + X2 + X4

       Df Sum of Sq    RSS    AIC
<none>              191302 405.94
- X1    1     33742 225044 411.74
- X2    1     69532 260834 418.82
- X4    1    243586 434889 443.36

Call:
lm(formula = Y ~ X1 + X2 + X4)

Coefficients:
(Intercept)           X1           X2           X4
  307.32790    -29.48381     -0.06802   1374.76841

> 
```

- **Conclusion:**
    - At the start of this test, the full model is Y ~ X1 +X2+X3+X4. The test drops the variable with the lowest AIC (X3) and compares that reduced model to the full model.
    - Now Y ~ X1 +X2+X4 is the full model and it drops the variable with the lowest AIC and compares it with the AIC of the full model. Since the full model has the lowest AIC, the test stops there. Conclude that Y ~ X1+X2+X4 is a possible optimal linear model.

- **Overall Conclusion:**

    - According to the tests we performed, X3 (the miles of highway) had no bearing on the dependent variable as shown above. This variable does not contribute to the gallons of gas consumed for a couple of reasons. First off, cars will get better gas mileage on a highway. So countries with more highways will consume less gas. Also, people may not even use the highway is the system if the country has a successful public transportation system. This would lessen the amount of people driving and lower the gas consumption of the country, devaluing the highways relation to the gas consumption.