2. Multiple Regression

In multiple regression several variables are used as predictors of the response variable.

For each of n cases we have the observed values of response variable (Y) and the predictors $X_1, X_2, \ldots, X_p$.

The data forms an array of dimension n x (p+1)

Case	**Y**	**X_1**	**X_2**	**…**	**X_p**
1	y_1	x_{11}	x_{12}	…	x_{1p}
2	y_2	x_{21}	x_{22}	…	x_{2p}
.	.	.	.	…	.
.	.	.	.	…	.
.	.	.	.	…	.
n	y_n	x_{n1}	x_{n2}	…	x_{np}

x_{ij} refers to value of the i^{th} case of the j^{th} predictor variable (i = 1 to n, j = 1 to p).

In multiple regression an equation expresses the response as a linear function of the predictor variables. This equation is estimated from the data.

The model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + e_i, \quad e_i \sim NID(0, \sigma^2)$$

Fuel Data Set

A study carried out in 48 states of America in 1974. The following variables were collected.

TAX = Tax on motor fuel (cents per gallon)
INC = Income per capita ($1,000)
ROAD = Length of primary roads in the state (thousands of miles)
DLIC = Proportion of population licensed to drive (%)
FUEL = Fuel consumption per capita (gallons per person)

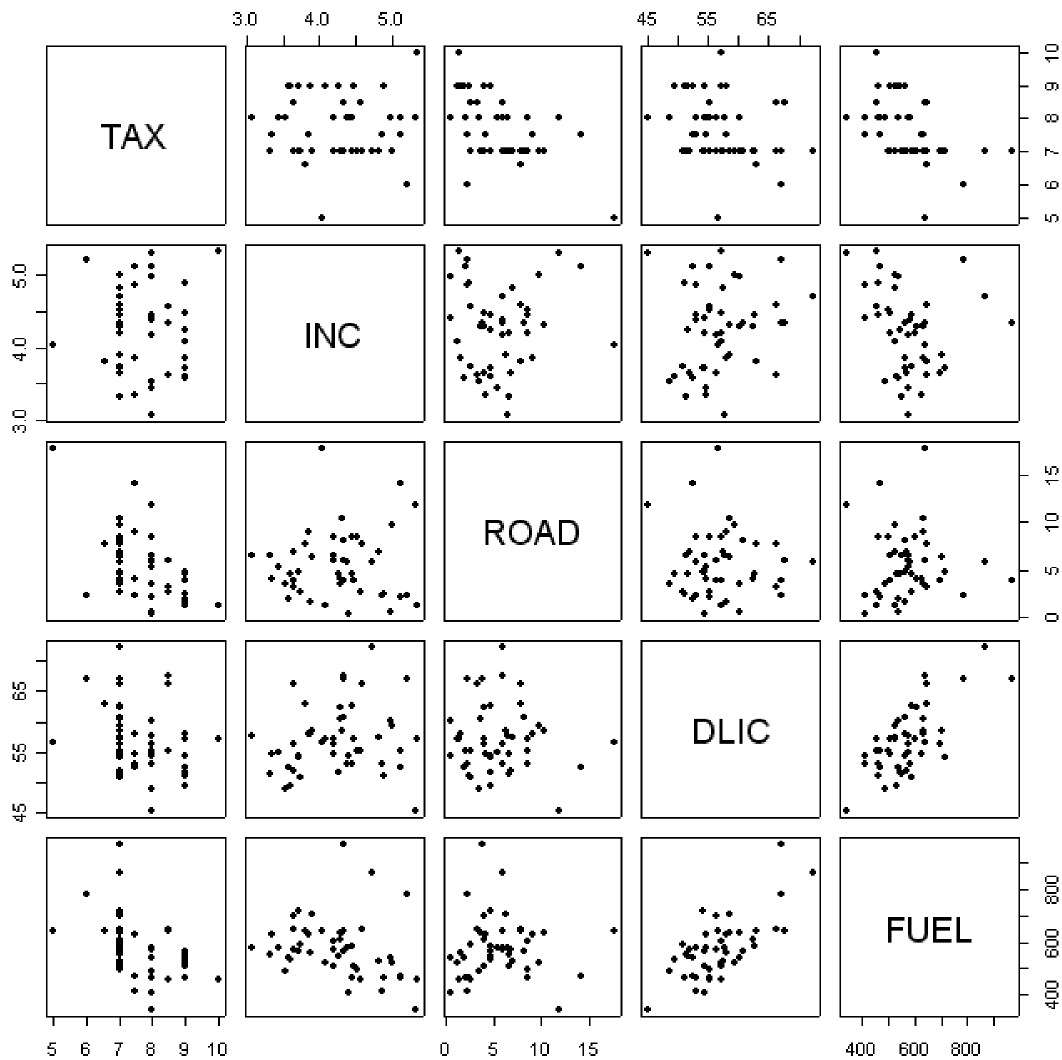
These data will be used to model fuel consumption as a function of the other variables.

	TAX	INC	ROAD	DLIC	FUEL
ME	9.0	3.571	1.976	52.5	541
NH	9.0	4.092	1.250	57.2	524
VT	9.0	3.865	1.586	58.0	561
MA	7.5	4.870	2.351	52.9	414
RI	8.0	4.399	0.431	54.4	410
CN	10.0	5.342	1.333	57.1	457
NY	8.0	5.319	11.868	45.1	344
NJ	8.0	5.126	2.138	55.3	467
PA	8.0	4.447	8.577	52.9	464
OH	7.0	4.512	8.507	55.2	498
IN	8.0	4.391	5.939	53.0	580
IL	7.5	5.126	14.186	52.5	471
MI	7.0	4.817	6.930	57.4	525
WI	7.0	4.207	6.580	54.5	508
MN	7.0	4.332	8.159	60.8	566
IA	7.0	4.318	10.340	58.6	635
MO	7.0	4.206	8.508	57.2	603
ND	7.0	3.718	4.725	54.0	714
SD	7.0	4.716	5.915	72.4	865
NE	8.5	4.341	6.010	67.7	640
KS	7.0	4.593	7.834	66.3	649
DE	8.0	4.983	0.602	60.2	540
MD	9.0	4.897	2.449	51.1	464
VA	9.0	4.258	4.686	51.7	547
WV	8.5	4.574	2.619	55.1	460
NC	9.0	3.721	4.746	54.4	566
SC	8.0	3.448	5.399	54.8	577
GA	7.5	3.846	9.061	57.9	631
FL	8.0	4.188	5.975	56.3	574
KY	9.0	3.601	4.650	49.3	534
TN	7.0	3.640	6.905	51.8	571
AL	7.0	3.333	6.594	51.3	554
MS	8.0	3.063	6.524	57.8	577
AR	7.5	3.357	4.121	54.7	628
LA	8.0	3.528	3.495	48.7	487
OK	6.6	3.802	7.834	62.9	644
TX	5.0	4.045	17.782	56.6	640
MT	7.0	3.897	6.385	58.6	704
ID	8.5	3.635	3.274	66.3	648
WY	7.0	4.345	3.905	67.2	968
CO	7.0	4.449	4.639	62.6	587
NM	7.0	3.656	3.985	56.3	699
AZ	7.0	4.300	3.635	60.3	632
UT	7.0	3.745	2.611	50.8	591
NV	6.0	5.215	2.302	67.2	782
WN	9.0	4.476	3.942	57.1	510
OR	7.0	4.296	4.083	62.3	610
CA	7.0	5.002	9.794	59.3	524

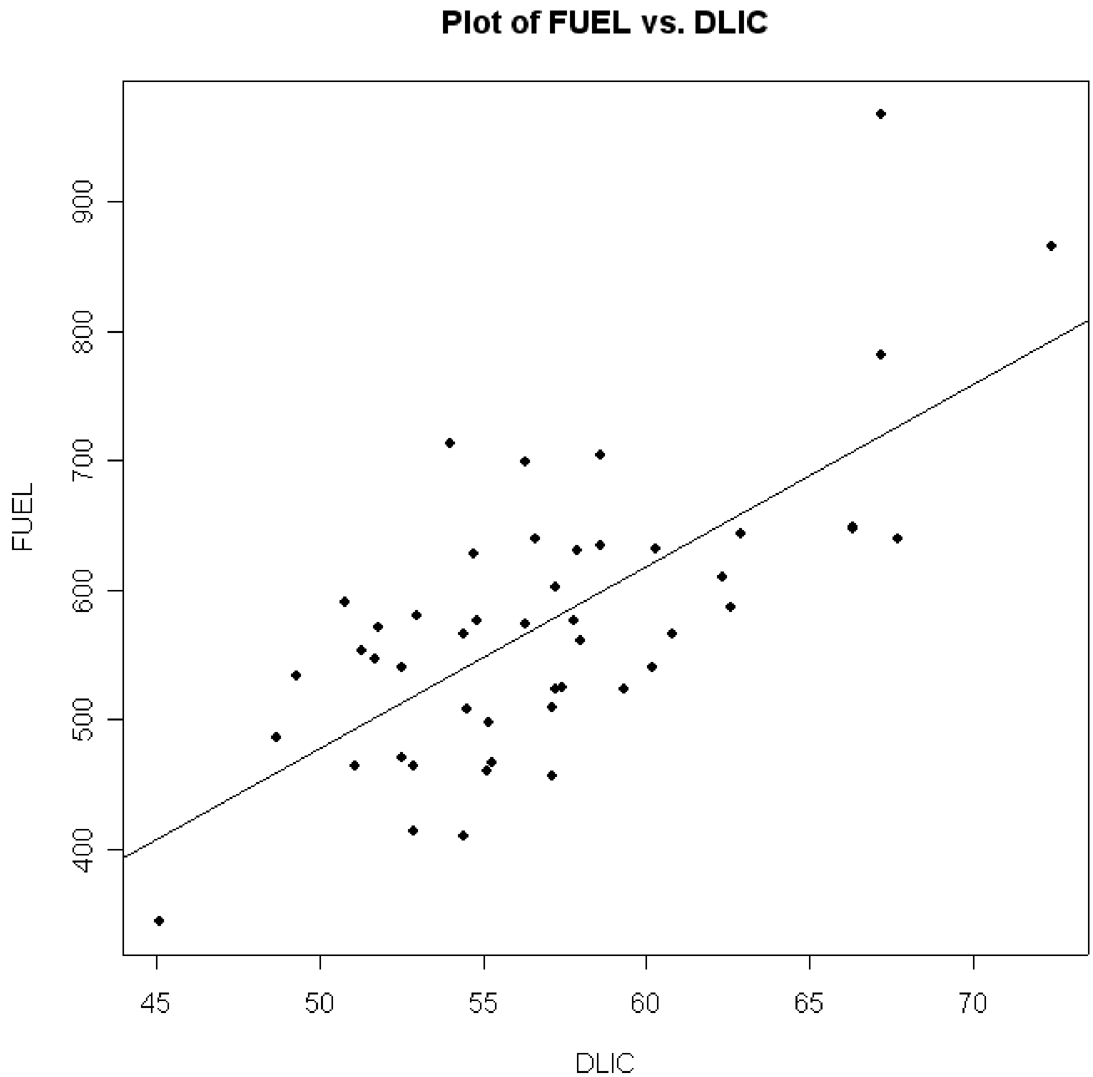
Matrix of Correlation coefficients:

	TAX	INC	ROAD	DLIC	FUEL
TAX	1.00000000	0.01266516	-0.52213014	-0.2880372	-0.45128028
INC	0.01266516	1.00000000	0.05016279	0.1570701	-0.24486207
ROAD	-0.52213014	0.05016279	1.00000000	-0.0641295	0.01904194
DLIC	-0.28803717	0.15707008	-0.06412950	1.0000000	0.69896542
FUEL	-0.45128028	-0.24486207	0.01904194	0.6989654	1.00000000

Matrix of Scatter-plots:



Model of FUEL on DLIC

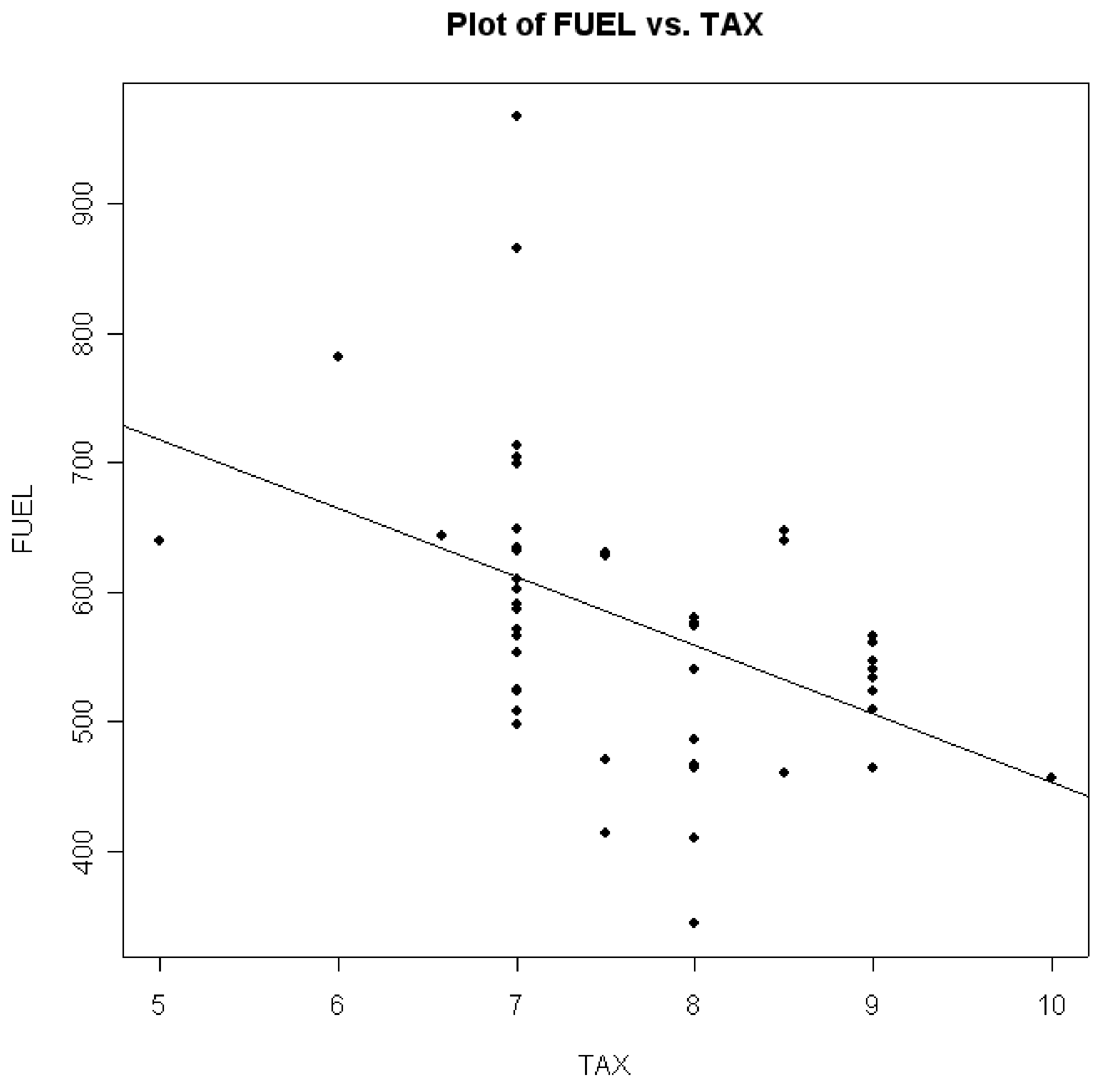


FUEL = $\hat{\beta}_0 + \hat{\beta}_1$ DLIC = –227.21 + 14.40 DLIC

Each percent increase in DLIC (percentage of the population with driver's licenses) corresponds to an estimated 14.40-gallon **increase** per capita in FUEL consumption

$R^2 = (0.6990)^2 = 0.4886$, so 48.9% of the variability in FUEL is explained by DLIC

Model of FUEL on TAX



FUEL = $\hat{\beta}_0$ + $\hat{\beta}_1$ TAX = –948.091 – 53.11 TAX

Each cent increase in TAX (tax in cents per gallon) corresponds to an estimated 53.11-gallon **decrease** per capita in FUEL consumption

$R^2 = (-0.4513)^2 = 0.2037$, so 20.4% of the variability in FUEL is explained by TAX

In the above models, we interpreted the predictor variable without regard to the other.

Model of FUEL on DLIC and TAX

We now fit a model with both predictors.

FUEL = $\hat{\beta}_0$ + $\hat{\beta}_1$ TAX + $\hat{\beta}_2$ DLIC = 108.97 – 32.08 TAX + 12.51 DLIC

In this section, we study two main questions:

Q1 In the model of FUEL on TAX and DLIC, how do we interpret $\hat{\beta}_1$ and $\hat{\beta}_2$?

A1 $\hat{\beta}_1$ = –32.08 measures the effect of TAX on FUEL, **adjusted for DLIC**

The effect of increasing TAX by one cent per gallon, **with DLIC held constant**, is to decrease FUEL consumption by 32.08 gallons per capita

$\hat{\beta}_2$ = 12.51 measures the effect of DLIC on FUEL, **adjusted for TAX**

The effect of one percentage increase in DLIC, **with TAX held constant**, is to increase FUEL consumption by 12.51 gallons per capita.

In multiple regression, $R^2 = \frac{SSreg}{SYY}$

Let R^2(TAX,DLIC) be the multiple-R^2 in the model of FUEL on TAX and DLIC.

Then R^2(TAX) = r^2(FUEL, TAX) = $(-0.4513)^2$ = 20.4%
and R^2(DLIC) = r^2(FUEL, DLIC) = $(0.6990)^2$ = 48.9%

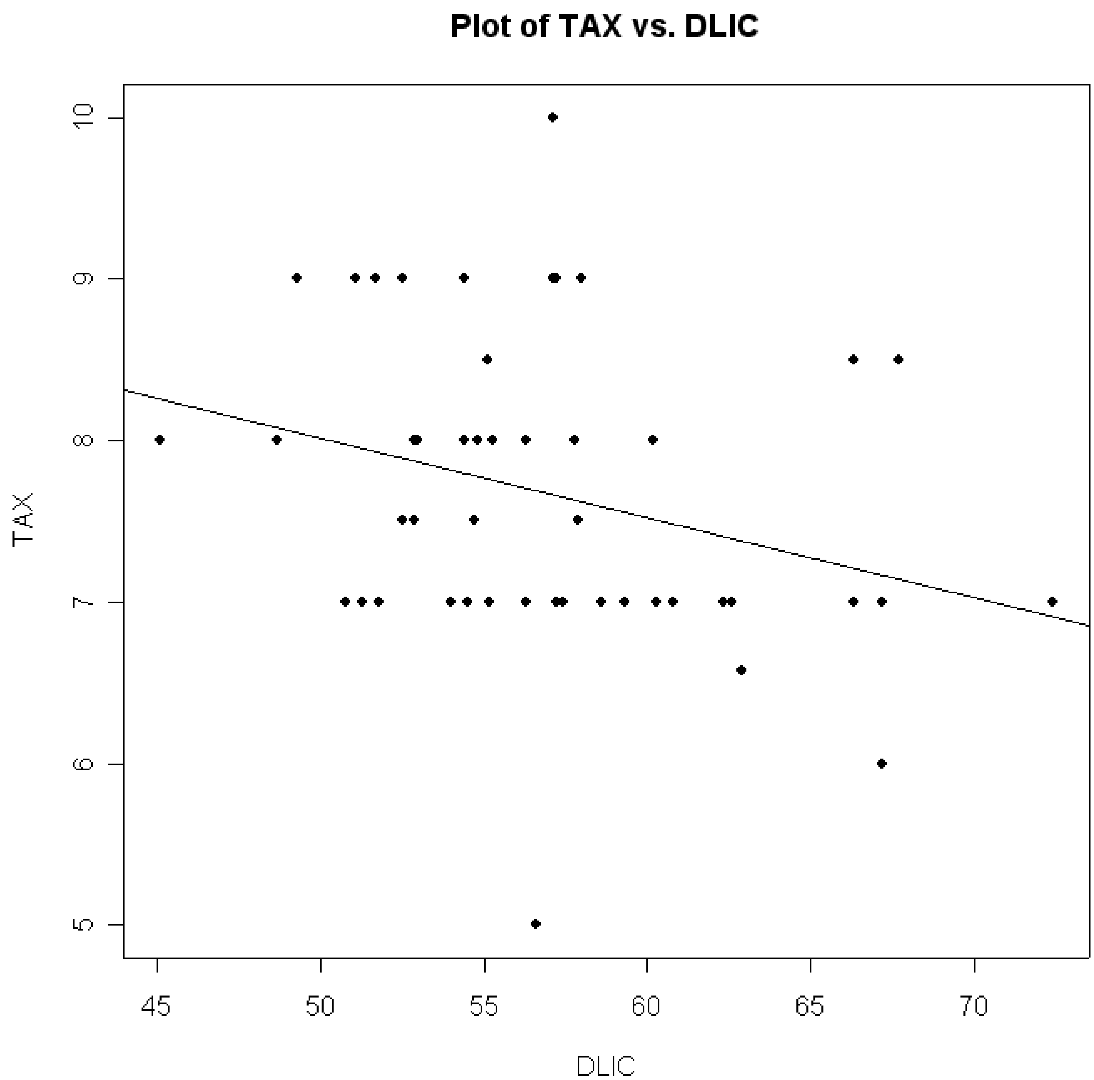
Q2 How does R^2(TAX, DLIC) relate to R^2(TAX) and R^2(DLIC)?

*SS*reg (TAX, DLIC) ≥ max(*SS*reg (TAX), *SS*reg (DLIC))
Including TAX and DLIC must be at least as good as including either separately.

R^2(TAX, DLC) ≥ max(R^2(TAX) , R^2(DLIC)) = max(20.4%, 48.9%) = 48.9%

Q When is R^2(TAX, DLIC) **equal to** R^2(TAX) + R^2(DLIC) = 69.3% here ?

A Only if TAX and DLIC are **completely unrelated** and measure completely different things, i.e. if **r(TAX, DLIC) = 0**



Here r(TAX, DLIC) = –0.2880, so TAX and DLIC are related

Q Can R^2(TAX, DLIC) be **less than** R^2(TAX) + R^2(DLIC) ?

A Yes, if TAX and DLIC are **related** to each other and are both explaining the same variability

Q Can R^2(TAX, DLIC) be **greater than** R^2(TAX) + R^2(DLIC) ?

A Yes, if TAX and DLIC **interact** so that knowing both gives more information than knowing just one of them

Added variable plot for TAX after DLIC

We will show that $\hat{\beta}_1$ in the larger model (FUEL = $\hat{\beta}_0$ + $\hat{\beta}_1$ TAX + $\hat{\beta}_2$ DLIC) shows the effect of TAX on FUEL, **adjusted for DLIC.**

The **residuals** from the model of FUEL on DLIC represent **that part of FUEL which is not explained by DLIC**.

Now model TAX on DLIC as follows: TAX = 10.48 – 0.0494 DLIC

The **residuals** from this model represent **that part of TAX which is not explained by DLIC**.

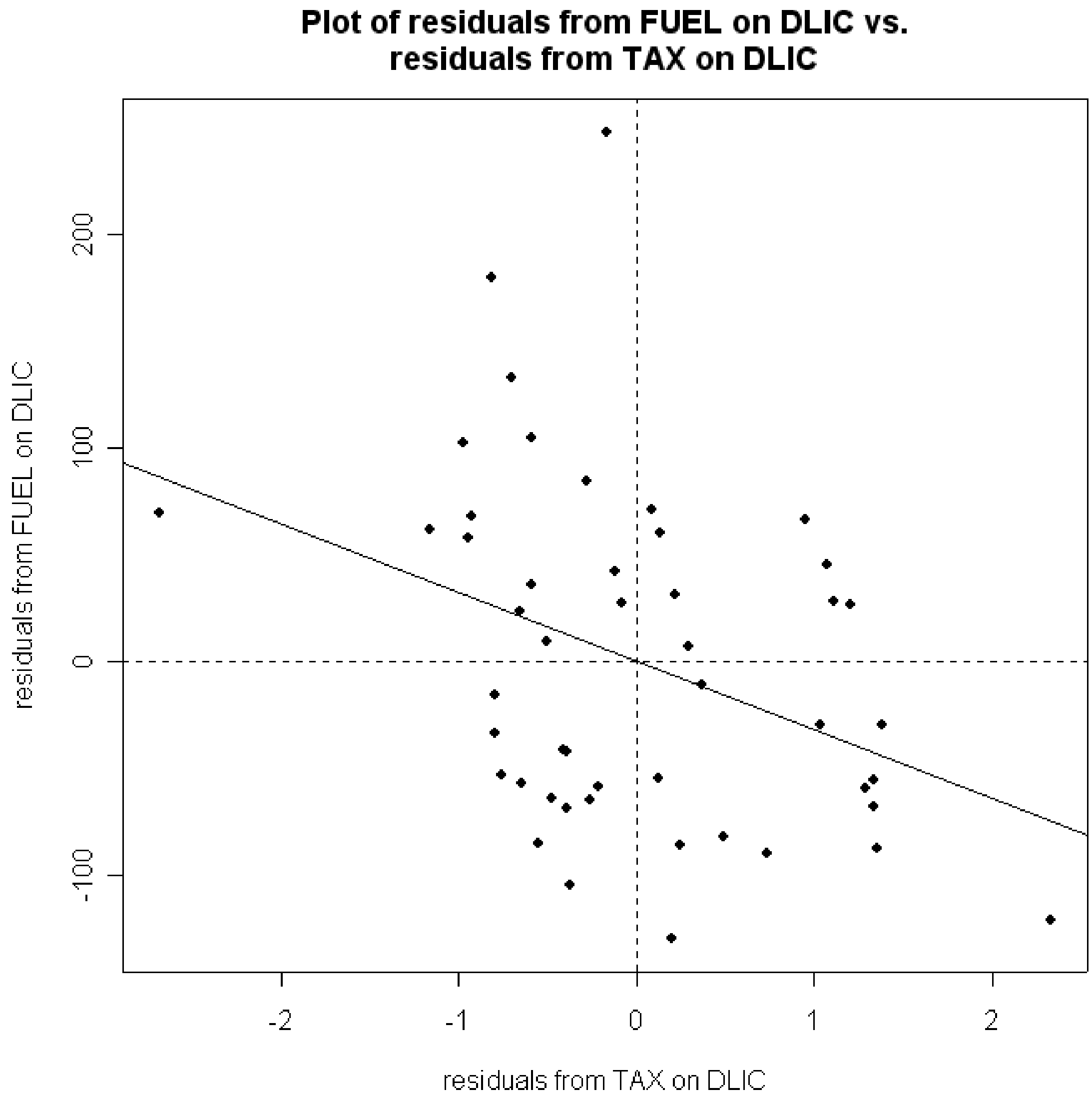
Consider the scatter-plot of the **residuals** from the model of FUEL on DLIC (on the Y axis) against the **residuals** from the model of TAX on DLIC (on the X axis). This called the **added variable plot for TAX after DLIC.** The **added variable** is TAX.

Added variable plots are used to display the relationship between a response variable (FUEL) and a predictor variable(s) (TAX) adjusted for the relationship between the predictor variable(s) and another predictor variable (DLIC).

These plots are interpreted the same as scatter-plots.

If there is a strong relationship the added variable helps to explain previously unexplained variability (in the smaller model).

If there is no relationship the added variables does not explain any of the previously unexplained variability.



Fit a regression line to this scatter-plot:

FUEL which is not explained by DLIC $= \hat{\beta}_0 + \hat{\beta}_1$ TAX which is not explained by DLIC.

Because both sets of residuals sum to 0, this regression line will pass though (0, 0), i.e. the intercept is 0.

```
  (Intercept) resid(fuel.cons3.lm)
2.513722e-15        -3.207532e+01
```

The **slope** of this fitted regression line is $\hat{\beta}_1 = -32.08$

Thus $\hat{\beta}_1 = -32.08$ in this model can be interpreted as measuring the effect of TAX on FUEL **adjusted for DLIC**.

Note that $\hat{\beta}_1$ = –32.08 is also the coefficient of TAX in the model of FUEL on TAX and DLIC (FUEL = $\hat{\beta}_0$ + $\hat{\beta}_1$ TAX + $\hat{\beta}_2$ DLIC = 108.97 – 32.08 TAX + 12.51 DLIC)

Similarly, $\hat{\beta}_2$ = 12.51 can be interpreted as measuring the effect of DLIC on FUEL **adjusted for TAX**.

In multiple regression, the $\hat{\beta}_j\,'s$ are adjusted for all other predictors in the model.

Partial correlation coefficients

The **partial correlation coefficient between FUEL and TAX, adjusted for DLIC** is denoted by

r(FUEL,TAX | DLIC)

and is defined as the ordinary correlation coefficient between the points in the added variable plot i.e.

r(FUEL,TAX | DLIC)

= r(residuals from FUEL on DLIC, residuals from TAX on DLIC)

Here r(FUEL,TAX | DLIC) = -0.3650 (from R).

This partial correlation coefficient can also be calculated directly as follows

$$\text{r(FUEL,TAX}\,|\,\text{DLIC)} = \frac{\text{r(FUEL,TAX)} - \text{r(FUEL,DLIC)r(TAX,DLIC)}}{\sqrt{1 - \text{r}^2\text{(FUEL,DLIC)}}\sqrt{1 - \text{r}^2\text{(TAX,DLIC)}}} \qquad (*)$$

$$= \frac{-0.4513 - (0.6990)(-0.2880)}{\sqrt{1 - 0.6990^2}\sqrt{1 - 0.2880^2}} = -0.3650$$

In general, $\text{r}(X_1,X_2\,|\,X_3) = \dfrac{\text{r}(X_1,X_2) - \text{r}(X_1,X_3)\text{r}(X_2,X_3)}{\sqrt{1 - \text{r}^2(X_1,X_3)}\sqrt{1 - \text{r}^2(X_2,X_3)}}$.

Here r(FUEL,TAX) = -0.4513. This is the correlation coefficient between the points in the scatter-plot of FUEL and TAX.

Note that |r(FUEL,TAX | DLIC)| = 0.3650 < |r(FUEL,TAX)| = 0.4513, so that

r^2(FUEL,TAX | DLIC) is **less than** r^2(FUEL,TAX),

i.e. the relationship in the added variable plot is **weaker** than the relationship in the scatter-plot. In this case, it may be shown that

R^2(TAX,DLIC) is **less than** R^2(TAX) + R^2(DLIC)

In this case, TAX and DLIC are **related** to each other and are **both explaining some of the same variability**.

From R, R^2(TAX,DLIC) = 55.7% < R^2(TAX) + R^2(DLIC) = 69.3%.

If r^2(FUEL,TAX | DLIC) is **greater than** r^2(FUEL,TAX),

i.e. the relationship in the added variable plot is **stronger** than the relationship in the scatter-plot, then it **may** happen that

R^2(TAX,DLIC) is **greater than** R^2(TAX) + R^2(DLIC)

In that case, TAX and DLIC **interact** so that knowing both gives much more information than knowing just one of them

It may be shown that R^2(TAX,DLIC) is related to R^2(DLIC) and r(FUEL,TAX | DLIC) as follows:

$$\mathbf{R^2(TAX,DLIC) =}$$

$$\mathbf{R^2(DLIC) + r^2(FUEL,TAX \mid DLIC) - R^2(DLIC)\, r^2(FUEL,TAX \mid DLIC)} \qquad (**)$$

$$= (0.6990)^2 + (0.3650)^2 - (0.6990)^2(0.3650)^2 = 55.7\%$$

From (**), it follows that if

$$\mathbf{r^2(FUEL,TAX \mid DLIC) - R^2(DLIC)\, r^2(FUEL,TAX \mid DLIC) > r^2(FUEL,TAX)},$$

then

$$R^2(TAX,DLIC) > R^2(DLIC) + r^2(FUEL,TAX) = R^2(DLIC) + R^2(TAX)$$

Similarly, if

$$\mathbf{r^2(FUEL,TAX \mid DLIC) < r^2(FUEL,TAX)},$$

then

$$R^2(TAX,DLIC) < R^2(DLIC) + r^2(FUEL,TAX) = R^2(DLIC) + R^2(TAX)$$

From (*) and (**), it follows that if **r(TAX,DLIC) =0**, then

$$R^2(TAX,DLIC) = R^2(DLIC) + R^2(TAX)$$

Model of FUEL on TAX, DLIC, INC and ROAD

FUEL = β_0 + β_1 TAX + β_2 DLIC + β_3 INC + β_4 ROAD + e_i

Fitted model: FUEL = 377.29 −34.79 TAX+ 13.37 DLIC − 66.59 INC −2.43 ROAD

$\hat{\beta}_1$ **= −**34.79 in this model measures the effect of TAX on FUEL, **adjusted for DLIC, INC and ROAD**

The effect of increasing TAX by one cent per gallon, **with DLIC, INC and ROAD held constant**, is to decrease FUEL consumption by 34.79 gallons per capita.

R^2(TAX,DLIC,INC,ROAD) = 0.6787 = 68%
So 68% of the variability in FUEL is explained by TAX, DLIC, INC and ROAD.

The **fitted value** $\hat{y}_i$ for Maine (ME) with TAX = 9, DLIC = 52.5, INC = 3.571 and ROAD = 1.976 is $\hat{y}_i$ = 377.29 −34.79(9)+ 13.37(52.5)− 66.59(3.571) −2.43(1.976) = 523.51

The actual value of FUEL for Maine is 541, so the residual is $\hat{e}_i = y_i - \hat{y}_i$ = 17.49.

Thus the model **underestimates** the FUEL consumption in Maine by 17.49 gallons per capita.

The residual sum of squares is $RSS = \sum_{i=1}^{n} \hat{e}_i^2$ and the regression sum of squares is

$SSreg = SYY - RSS$

Regression Models in Matrix Notation

The response variable values can be expressed as an nx1 vector, Y.

The errors can be expressed as an nx1 vector, e.

The regression coefficients can be expressed as a (p+1)x1 vector.

The data values can be expressed as a nx(p+1) matrix.

The regression model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + e_i$, $e_i \sim NID(0,\sigma^2)$ can then be expressed used matrix notation as follows:

$$Y = X\beta + e$$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix}$$

With $e \sim N(0, \sigma^2 I_n)$, I_n is the nxn identity matrix: $I_n = \begin{bmatrix} 1 & 0 & 0 & . & . & 0 \\ 0 & 1 & 0 & . & . & 0 \\ . & . & 1 & . & . & . \\ . & . & . & 1 & . & . \\ 0 & 0 & . & . & . & 1 \end{bmatrix}$

Analysis of Variance

From R, the analysis of variance table is:

```
Analysis of Variance Table

Response: FUEL
          Df Sum Sq Mean Sq F value    Pr(>F)    
TAX        1 119823  119823 27.2541 4.901e-06 ***
DLIC       1 207709  207709 47.2441 1.963e-08 ***
INC        1  69532   69532 15.8152 0.0002632 ***
ROAD       1   2252    2252  0.5123 0.4779989    
Residuals 43 189050    4397                      
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1
```

We can test the hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
against the alternative H_1: $\beta_1, \beta_2, \beta_3, \beta_4$ not all zero

If H_0 is true, then **none** of variables TAX, DLIC, INC, ROAD should be included in the model.

If H_1 is true, then **at least one** of variables TAX, DLIC, INC, ROAD should be included in the model

If H_0 is true, it may be shown that

$$F = \frac{SSreg / p}{RSS/(n-p')} \sim F(p, n-p'), \text{ where } p' = p+1$$

If the observed value of F is **too big**, we reject H_0 and accept H_1.

From R:
```
F-statistic: 22.71 on 4 and 43 DF,  p-value: 3.907e-10
```

Conclude that **at least one** of TAX, DLIC, INC and ROAD should be included in the model.

Partial F-test of the hypothesis H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$

If H_0 is true, then, **given that DLIC, INC and ROAD are already in the model**, the variable TAX **should not** be included in the model.

If H_1 is true, then, **given that DLIC, INC and ROAD are already in the model**, the variable TAX **should** be included in the model.

Let *RSS*(TAX,DLIC,INC,ROAD) = residual sum of squares in the model of FUEL on TAX, DLIC, INC and ROAD.

Let *RSS*(DLIC,INC,ROAD) = residual sum of squares in the model of FUEL on DLIC, INC and ROAD.

The **Extra Sum of Squares due to TAX, given DLIC, INC and ROAD** is denoted by ***SSreg*(TAX | DLIC,INC,ROAD)**

= *RSS*(DLIC,INC,ROAD) − *RSS*(TAX,DLIC,INC,ROAD) = 31,632 (from R).

If **H_0 is true**, then $F = \dfrac{SSreg(\text{TAX | DLIC,INC,ROAD})/1}{RSS(\text{TAX,DLIC,INC,ROAD}/(n-p')} \sim F(1, n-p')$

Here $F = \dfrac{32,632/1}{189,050/43} = 7.19$, while F(0.01;1,43) = 7.26, so the p-value is approximately 0.01.

We could also use a t-test. Estimates of coefficients from R:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  377.291    185.541   2.033 0.048207 *
TAX          -34.790     12.970  -2.682 0.010332 *
DLIC          13.364      1.923   6.950 1.52e-08 ***
INC          -66.589     17.222  -3.867 0.000368 ***
ROAD          -2.426      3.389  -0.716 0.477999
```

Remember the equivalence of the t-test and the F-test: $(-2.682)^2 = 7.19$

Partial F-test of the hypothesis H_0: $\beta_1 = \beta_4 = 0$ against H_1: β_1, β_4 not both 0

If H_0 is true, then, **given that DLIC and INC are already in the model**, the variables TAX and ROAD **should not** be included the model

If H_1 is true, then, **given that DLIC and INC are already in the model**,
at least one of the variables TAX and ROAD **should** be included in the model

The **Extra Sum of Squares due to TAX and ROAD given DLIC and INC** is denoted by ***SSreg*(TAX,ROAD | DLIC,INC)**

= *RSS*(DLIC,INC) − *RSS*(TAX,DLIC,INC,ROAD) = 35,994.4 (from R)

If **H_0 is true**, then $F = \dfrac{SSreg(\text{TAX,ROAD | DLIC,INC})/2}{RSS(\text{TAX,DLIC,INC,ROAD}/(n - p')} \sim F(2, n - p')$

Here $F = \dfrac{35{,}994.4/2}{189{,}050/43} = 4.09,$ with a p-value of 0.02 (from R).

Sequential Analysis of Variance Tables

Consider fitting the model FUEL = $\beta_0 + \beta_1$ DLIC + β_2TAX + β_3INC +β_4 ROAD + e.
The ANOVA table from R is:

```
Response: FUEL
          Df Sum Sq Mean Sq F value    Pr(>F)    
DLIC       1 287448  287448 65.3809 3.584e-10 ***
TAX        1  40084   40084  9.1173 0.0042477 ** 
INC        1  69532   69532 15.8152 0.0002632 ***
ROAD       1   2252    2252  0.5123 0.4779989    
Residuals 43 189050    4397
```

SS_{reg}(DLIC) = 287,448
SS_{reg}(TAX|DLIC) = 40,084
SS_{reg}(INC|DLIC, TAX) = 69,532
SS_{reg}(ROAD|DLIC, TAX, INC) = 2,252
RSS = 189,050

Consider fitting the model in different order:
FUEL = $\beta_0 + \beta_1$ ROAD + β_2INC + β_3DLIC +β_4 TAX + e.
The ANOVA table from R is:

```
Response: FUEL
          Df Sum Sq Mean Sq F value   Pr(>F)    
ROAD       1    213     213  0.0485 0.826693    
INC        1  35642   35642  8.1070 0.006735 ** 
DLIC       1 331829  331829 75.4755 5.15e-11 ***
TAX        1  31632   31632  7.1948 0.010332 *  
Residuals 43 189050    4397
```

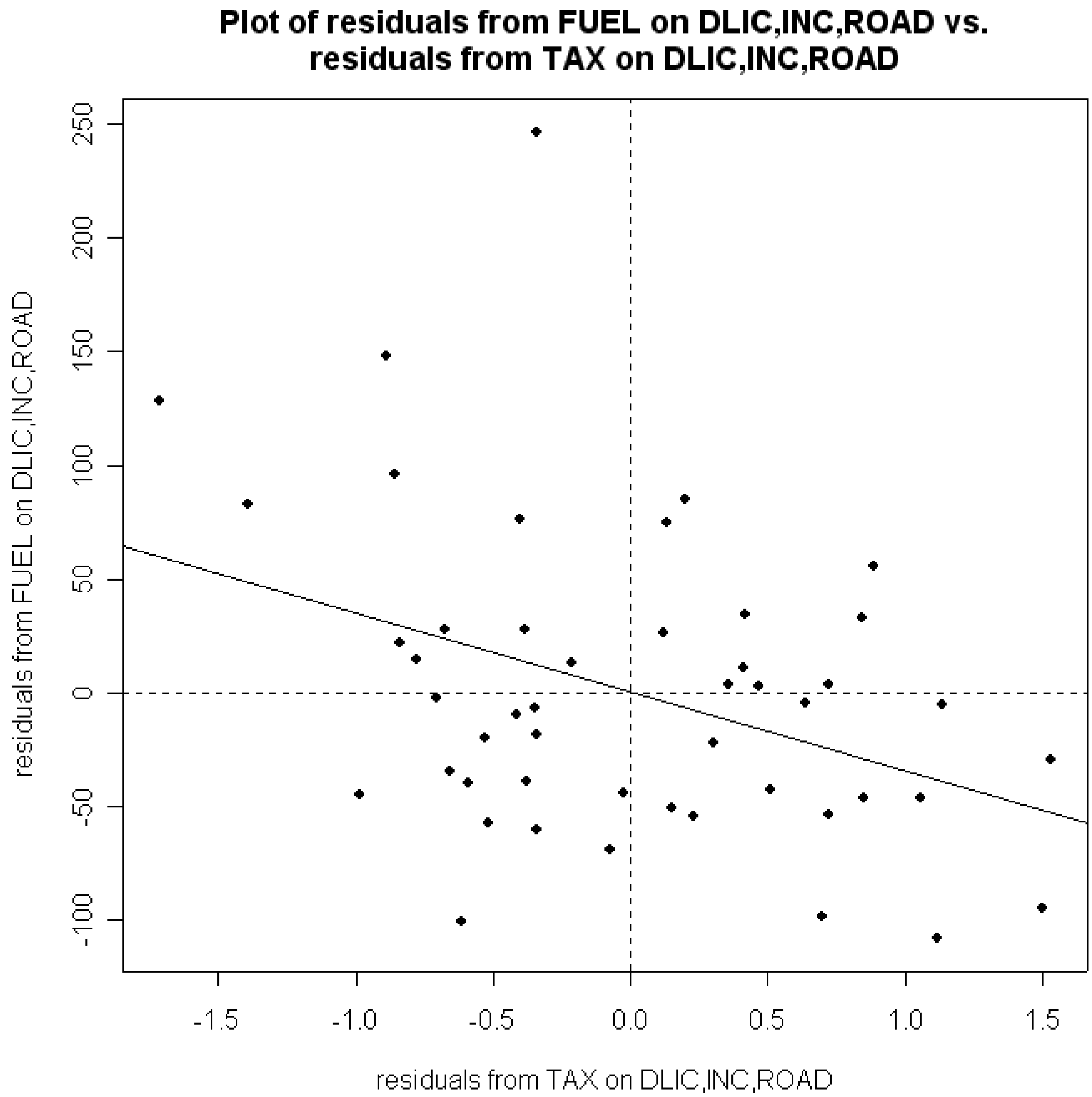
SS_{reg}(ROAD) = 213
SS_{reg}(INC|ROAD) = 35,642
SS_{reg}(DLIC|ROAD, INC) = 331,829
SS_{reg}(TAX|ROAD, INC, DLIC) = 31,632
RSS = 189,050

SYY, RSS and the parameter estimates will be the same for both models but the SS_{reg} for the individual predictors is dependent on the order they were (listed) added to the model.

This will always occur unless the correlation coefficients between each pair of predictors are all 0.

Added variable plot for TAX after DLIC, INC and ROAD

Plot the residuals from the model of FUEL on DLIC, INC and ROAD (on the Y- axis) against the residuals form the model of TAX on DLIC, INC and ROAD (on the X-axis). This is the added variable plot of TAX after DLIC, INC and ROAD. TAX is the added variable.



Fit a regression line to the points of this scatter-plot. The slope of the fitted line is $\hat{\beta}_1$ = −34.79 (from R).
This is the same of the coefficient of TAX in the model of FUEL on TAX, DLIC, INC and ROAD:

$$FUEL = 377.29 - 34.79\ TAX + 13.37\ DLIC - 66.59\ INC - 2.43\ ROAD$$

Thus $\hat{\beta}_1 = -34.79$ in this model can be interpreted as measuring the effect of TAX on FUEL, **adjusted for DLIC, INC and ROAD**.

Drawing Conclusions

Interpretation of regression coefficients

In the fuel consumption data, the fitted model of FUEL on TAX, DLIC, INC and ROAD was

$$\text{FUEL} = 377.29 - 34.79\ \text{TAX} + 13.36\ \text{DLIC} - 66.59\ \text{INC} - 2.43\ \text{ROAD}$$

Increasing the TAX rate by one cent per gallon should decrease consumption, **all other factors being held constant**, by 34.79 gallons per capita.

This assumes that we <u>can</u> change TAX by one unit and without affecting the other predictors.

In this example, the values of TAX, DLIC, INC and ROAD were **observed** in the 48 states. We cannot **assign** values to these variables, as we could in a **Designed Experiment**. In such an experiment, we would allocate the TAX value, the DLIC proportion, the INC and ROAD to each state (clearly not a viable experiment). We could then increase one predictor by one unit, while holding the values of the other predictors constant:

$\mathbf{X_1 = 9}$, $X_2 = 10$, $X_3 = 20$, $X_4 = 30$, $Y = ?$
$\mathbf{X_1 = 10}$, $X_2 = 10$, $X_3 = 20$, $X_4 = 30$, $Y = ?$
$\mathbf{X_1 = 9}$, $X_2 = 15$, $X_3 = 25$, $X_4 = 35$, $Y = ?$
$\mathbf{X_1 = 10}$, $X_2 = 15$, $X_3 = 25$, $X_4 = 35$, $Y = ?$

Consequently, whether fuel consumption would be changed by increasing taxes **cannot** be directly assessed here.

Instead, from the correlation coefficient r(FUEL, TAX) = −0.4513 and the scatter-plot of FUEL vs. TAX we can make the following more conservative statement:

"States with higher tax rates are **observed** to have lower fuel consumption".

To draw conclusions about the effects of changing tax rates, the rates must in fact be changed and the results observed.

Signs of parameter estimates

In previous examples, we have seen that the **value** of a parameter estimate can **change** depending on the other predictors in the model. Thus the value of $\hat{\beta}_1$ in the model

$$Y = \beta_0 + \beta_1 X_1 \qquad \text{(Model 1)}$$

is not, in general, the same as the value of $\hat{\beta}_1$ in the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad \text{(Model 2)}$$

This is due to **correlations** between the predictor variables X_1 and X_2.

It may be shown that if the predictors X_1 and X_2 are **uncorrelated**,
i.e. if

$$\mathbf{r_{12} = 0},$$

where r_{12} is the correlation coefficient between X_1 and X_2, then the value of $\hat{\beta}_1$ is the **same** in model 1 and 2. In this case, the predictors X_1 and X_2 are said to be **orthogonal**.

Thus it is possible that in model 1 $\hat{\beta}_1$ is **positive**, while in model 2, $\hat{\beta}_1$ is **negative**.

Thus the effect of X_1 on Y could be **positive**, while the effect of X_1 on Y, **adjusted for X_2**, is **negative**.

In general, it is possible that the **sign** of a parameter estimate can **change** depending on the other predictors in the model.

This makes interpretation of the parameter estimates more difficult.

Sometimes this problem can be removed by **redefining the predictors into new linear combinations** that are easier to interpret.
Thus instead of the predictors X_1, X_2 and X_3, we could use the predictors X_1, $X_2 – X_1$ and $X_3 – X_1$.

Since the second set of predictors can be obtained from the first set via a **linear transformation**, the two sets of predictor variables carry exactly the **same information** about the response variable Y. This is illustrated in the next example.

Berkeley Girls Data Set (Illustrating how parameters change)

A longitudinal study carried out on a sample of girls from birth to age 18 in Berkeley, California.

ID = unique identifier for each girl
WT2 = Weight (kg) at age 2
WT9 = Weight (kg) at age 9
WT18 = Weight (kg) at age 18
SOMA = Somatotype – a 7-point scale (from 1 = slender to 7 = fat) measuring fatness at age 18.

ID	WT2	WT9	WT18	SOMA
331	12.6	33.0	71.2	6.0
334	12.0	34.2	58.2	5.0
335	10.9	28.1	56.0	5.0
351	12.7	27.5	64.5	4.0
352	11.3	23.9	53.0	5.0
353	11.8	32.2	52.4	4.0
354	15.4	29.4	56.8	4.5
355	10.9	22.0	49.2	4.0
356	13.2	28.8	55.6	4.5
357	14.3	38.8	77.8	6.5
358	11.1	36.0	69.6	5.5
359	13.6	31.3	56.2	3.5
361	13.5	33.3	64.9	5.0
362	16.3	36.2	59.3	4.5
364	10.2	23.4	49.8	4.0
365	12.6	33.8	62.6	5.0
366	12.9	34.5	66.6	5.0
367	13.3	34.4	65.3	5.0
368	13.4	38.2	65.9	5.5
369	12.7	31.7	59.0	5.5
370	12.2	26.6	47.4	4.0
371	15.4	34.2	60.4	4.0
372	12.7	27.7	56.3	3.0
373	13.2	28.5	61.7	4.5
374	12.4	30.5	52.4	5.0
376	13.4	39.0	58.4	6.5
377	10.6	25.0	52.8	5.0
380	12.7	29.8	67.4	5.0
382	11.8	27.0	56.3	4.5
383	13.3	41.4	82.8	7.0
384	13.2	41.6	68.1	5.5
385	15.9	42.4	63.1	5.5

We wish to model somatotype (SOMA) by weights at age 2, 9 and 18 (WT2, WT9, WT18) for n= 32 girls.

The correlation coefficient between SOMA and WT2 is 0.1234. Thus the value of $\hat{\beta}_1$ in the model

SOMA = $\beta_0 + \beta_1$ WT2 (Model 1)

is **positive.** The effect of WT2 on SOMA is **positive**.

However, the value of $\hat{\beta}_1$ in the model

SOMA = $\beta_0 + \beta_1$ WT2 + β_2 WT9 + β_3 WT18 (Model 2)

is $\hat{\beta}_1 = -0.22$. Thus heavier girls at age 2 tend to be thinner at age 18.
The effect of WT2 on SOMA, **adjusted for WT9 and WT18**, is **negative**.
This may be due to correlations between the predictors.

Consider the following set of predictors:

WT2 = Weight at age 2.
DW9 = WT9 – WT2 = difference in weight from age 2 to 9.
DW18 = WT18– WT9 = difference in weight gain from age 9 to 18.

The value of $\hat{\beta}_1$ in the model

SOMA = $\beta_0 + \beta_1$ WT2 + β_2 DW9 + β_3 DW18 (Model 3)

is $\hat{\beta}_1 = -0.08$. Thus the effect of WT2 on SOMA, adjusted for DW9 and DW18, is still (barely) **negative**.

However, in model 2, the effect of WT2, adjusted for WT9 and WT18, appears **substantial** ($\hat{\beta}_1 = -0.22$ with t = –2.47), while in model 3, the effect of WT2, adjusted for DW9 and DW18, appears to be **negligible** ($\hat{\beta}_1 = -0.08$ with t = –1.05).

The value of $R^2 = 0.610$ is the same in models 2 and 3.
Since the three variables WT2, DW9 and DW18 can be obtained from WT2, WT9 and WT18 via a linear transformation, the two sets of variables carry exactly the same information concerning SOMA.
Consequently, the two models explain the same percentage of the variation in SOMA.

However, the estimated coefficient of WT2 in the two models depends on which set of variables is used. **Thus the interpretation of the effect of a variable depends not only on the other variables in a model, but also upon which linear transformation of those variables is used.**

Collinearity

Consider adding a **fourth** predictor variable

QUAD = WT2 –2WT9 + WT18 (= (WT18 – WT9) – (WT9 – WT2) = DW18 – DW9)

to the model of SOMA on WT2, DW9 and DW18:

SOMA = β_0 + β_1**QUAD** + β_2 WT2 + β_3 DW9 + β_4 DW18 (Model 4)

Since QUAD = DW18 – DW9 is a an **exact linear combination** of variables DW9 and DW18 that are already in the model, we say the four predictors WT2, DW9 and DW18 and QUAD are **linearly dependent**, since one can be determined exactly from the others.

To estimate the coefficient of QUAD in this model, consider drawing the **added variable plot for QUAD, after WT2, DW9 and DW18**.

Remember added variable plots are used to display the relationship between a response variable (SOMA) and a predictor variable (QUAD) adjusted for the relationship between that predictor variable and other predictor variable(s) (WT2, DW9, DW18).
These plots are interpreted the same as scatter-plots. If there is a strong relationship the added variable helps to explain previously unexplained variability (in the smaller model). If there is no relationship the added variables does not explain any of the previously unexplained variability.

We plot the residuals from a model of SOMA on WT2, DW9 and DW18 (on the Y axis) against the residuals from a model of QUAD on WT2, DW9 and DW18 (on the X axis).

However, since QUAD can be written as an **exact linear combination** of the predictors DW9 and DW18

QUAD = 0(WT2) + (-1)DW9 + (1)(DW18)

the fitted model of QUAD on WT2, DW9 and DW18 is

QUAD = 0 + (0)WT2 + (–1)DW9 + (1) DW18

There will be no errors in this model - **all exactly zero**. The residuals will all (approximately) equal zero.

Remember the formula for the slope in simple linear regression, $\hat{\beta}_1 = \frac{SXY}{SXX}$ and $SXX = \sum (x_i - \bar{x})^2$. As the x_i's here are all zero, SXX is also 0.

Thus the estimated slope coefficient for QUAD in the added variable plot for QUAD, after WT2, WT9 and DW9, is **not defined** and so the estimated coefficient for QUAD in model 4 is **not defined**. (demonstrated in practical 2)

In general, it is a fatal flaw to include a predictor which is an **exact linear combination** of other predictors in the model. The problem of **collinearity** arises if there is an exact (or approximate) linear relationship between the predictors in a linear model.

What is collinearity?

Two predictors X_1 and X_2 are **exactly collinear** if there is a linear equation such as

$$c_1X_1 + c_2X_2 = c_0$$

for some constants c_0, c_1, and c_2 that is true for all cases in the data.

Approximate collinearity is obtained if the linear equation holds approximately for the observed data.

The **degree of collinearity** between X_1 and X_2 is measured by the square of the sample correlation, r_{12}^2.

Exact collinearity corresponds to $r_{12}^2 = 1$; **noncollinearity** correponds to $r_{12}^2 = 0$.
As r_{12}^2 approaches1, **approximate collinearity** becomes stronger.

Usually, the adjective *approximate* is dropped and we would say that X_1 and X_2 are **collinear** if r_{12}^2 is **large**.

Why is collinearity a problem?

Consider a regression with two predictors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

Let $SX_jX_j = \sum_{j=1}^{n}(X_j - \bar{X}_j)^2$.

Then it may be shown that $\text{var}(\hat{\beta}_j) = \sigma^2\left(\frac{1}{1-r_{12}^2}\right)\left(\frac{1}{SX_jX_j}\right)$

The variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ are minimized if $r_{12}^2 = 0$, while as r_{12}^2 nears 1, these variances are greatly inflated.
For example, if $r_{12}^2 = 0.95$, but *SXjXj* stays the same, the variance of $\hat{\beta}_1$ is 20 times as large as if $r_{12}^2 = 0$.
Thus the use of **collinear** predictors can lead to **unacceptably variable** estimated coefficients compared to regression analyses with no collinearity.

If $r_{12}^2 = 1$ (i.e**. exact collinearity)**, it may be shown that the estimates of the coefficients β_1 and β_2 in are **not defined**.

The definition of collinearity extends naturally to p>2 predictors. A set of predictors $X_1, X_2, \ldots X_p$ are **collinear** is, for constants $c_0, c_1 \ldots c_p$,

$$c_1X_1 + c_2X_2 + \ldots + c_pX_p = c_0$$

holds (approximately). Thus at least one of the X_k can be (approximately) determined from the others. It can be shown that

$$\text{var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1-R_j^2} \right) \left(\frac{1}{SX_jX_j} \right) \quad j = 1, 2, \ldots, p,$$

where R_j^2 is the multiple R^2 for the regression of X_j on the other X's.

Detecting collinearity

The quantity $VIF_j = \dfrac{1}{1-R_J^2}$ is called the jth **Variance Inflation Factor**.

Assuming that the X_j's could have been sampled to make $R_j^2 = 0$ while keeping *SXjXj* constant, the VIF represents the increase in variance due to the correlation between the predictors and, hence, collinearity.

If the **maximum** VIF_j **exceeds 10**, or equivalently, if the **maximum** R_j^2 **exceeds 0.90**, this is taken as an indication of **serious collinearity** between the predictors.

The multiple regression model

$$Y = \beta_0 + \beta_1X_1 + \beta_2 X_2 + \ldots\ldots + \beta_p X_p + e$$

may be expressed in matrix terms as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

The least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

provided the matrix $\mathbf{X}^T\mathbf{X}$ has an inverse.

If, however, the predictors X_1, X_2, ... X_p are **linearly dependent**, i.e. **exactly collinear**, then the matrix $\mathbf{X}^T\mathbf{X}$ cannot be inverted and so the least squares estimate $\hat{\beta}$ is **not defined**.

Example:

Consider a multiple regression with 3 predictors:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i$, $e_i \sim NID(0,\sigma^2)$

Dropping the subscripts and the error specification:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

To assess the degree of multicollinearity we need to consider whether any of the variables are linear combinations of the others:

$X_1 = c_0 + c_2X_2 + c_3X_3$?
$X_2 = c_0 + c_1X_1 + c_3X_3$?
$X_3 = c_0 + c_1X_1 + c_2X_2$?

Fit $X_1 \sim X_2 + X_3$	Calculate R_1^2	Calculate VIF = $1/(1 - R_1^2)$
Fit $X_2 \sim X_1 + X_3$	Calculate R_2^2	Calculate VIF = $1/(1 – R_2^2)$
Fit $X_3 \sim X_1 + X_2$	Calculate R_3^2	Calculate VIF = $1/(1 – R_3^2)$

Calculate max(VIF_j).

If max(VIF) > 10 collinearity is a serious problem.

Equivalently,

If max(R_j^2) > 0.90 collinearity is a serious problem.

Can we Identify Collinearity from the Correlation Matrix?

	Y	X_1	X_2	X_3
Y	1			
X_1		1	r_{12}	r_{13}
X_2			1	r_{23}
X_3				1

Model: $Y \sim X_1 + X_2 + X_3$

Suppose $r_{12} = 0.96$ and $r_{13} = 0.46$.
Then if R_1^2 is the multiple R^2 in the model $X_1 \sim X_2 + X_3$

Then because $R_1^2 > r_{12}^2 = (0.96)^2$ and $R_1^2 > r_{13}^2 = (0.46)^2$

Thus $R_1^2 > (0.96)^2 = 0.9216$ and so collinearity is a problem.

Suppose all r_{ij}^2 are "small", say all less than 0.7.

Does this mean that all $R_j^2 < 0.90$? (collinearity not a problem)

No!

It could be that $R_1^2 > r_{12}^2 + r_{13}^2$ (if X_2 and X_3 interact)

So we could have $R_1^2 > 0.90$ (collinearity a problem) even if r_{12}^2 and r_{13}^2 are small.

Previous Exam Question 2

A hospital administrator wished to study the relationship between patient satisfaction and patient's age, severity of illness and anxiety level. The administrator randomly selected 23 patients and collected data on each of these variables. A model of the following form was fitted to these data:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i, \quad e_i \sim \text{NID}(0, \sigma^2),$$

where
Y = Satisfy = index of patient satisfaction,
X_1 = Age = patient's age in years,
X_2 = Severity = index of severity of illness,
X_3 = Anxiety = index of anxiety level.

Excerpts from the R output for this model are shown on the next page.

(a) Identify the estimate of β_3 and interpret it.

(b) Interpret the value of R-Squared.

(c) Test the hypothesis H_0: $\beta_3 = 0$ against H_1: $\beta_3 \neq 0$.
Explain the practical implications of your conclusion.

(d) Test the hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = 0$ against H_1: β_1 , β_2, β_3 not all 0.
Explain the practical implications of your conclusion.

(e) Test the hypothesis H_0: $\beta_2 = \beta_3 = 0$ against H_1: β_2 , β_3 not both 0.
Explain the practical implications of your conclusion.

(f) Comment on the values of the correlation coefficients between the variables in the model. Calculate the Variance Inflation Factor for each of the variables Age, Severity and Anxiety in the model. Interpret these factors. What implications do the values of these factors have with regard to multicollinearity in the model?

Note: In each the tests mentioned above, quote the value of the test statistic and the associated *p*-value.

R output for Question 2

```
> summary(patients1.lm)

Call:
lm(formula = Satisfy ~ Age + Severity + Anxiety)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 162.8759    25.7757   6.319 4.59e-06 ***
Age          -1.2103     0.3015  -4.015  0.00074 ***
Severity     -0.6659     0.8210  -0.811  0.42736
Anxiety      -8.6130    12.2413  -0.704  0.49021
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 10.29 on 19 degrees of freedom
Multiple R-Squared: 0.6727,     Adjusted R-squared: 0.621
F-statistic: 13.01 on 3 and 19 DF,  p-value: 7.482e-05

> patients2.lm <- lm(Satisfy ~ Age)
> anova(patients2.lm,patients1.lm)
Analysis of Variance Table

Model 1: Satisfy ~ Age
Model 2: Satisfy ~ Age + Severity + Anxiety
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     21 2466.8
2     19 2011.6  2     455.2 2.1497  0.144

> cor(patients.df)
            Satisfy        Age   Severity    Anxiety
Satisfy   1.0000000 -0.7736828 -0.5874444 -0.6023105
Age      -0.7736828  1.0000000  0.4666091  0.4976945
Severity -0.5874444  0.4666091  1.0000000  0.7945275
Anxiety  -0.6023105  0.4976945  0.7945275  1.0000000

> summary(lm(Age ~ Severity + Anxiety))$r.squared
[1] 0.2614395
> summary(lm(Severity ~ Age + Anxiety))$r.squared
[1] 0.6380081
> summary(lm(Anxiety ~ Age + Severity))$r.squared
[1] 0.6518792
```

Practical (Assignment) 2

Instructions for this practical

- Open the template “Surname Forename Chpt x” from Canvas (in “Practicals”).
- Complete the grid on the first page.
- Save this file (as a Word document) using your own surname, forename and the appropriate chapter number.

- Practice Question:
- Type the commands one by one into R.
- Compare the results in the R text output and graphics with the corresponding results and figures in your notes.
- Use appropriate R output to answer the questions, adapting the R code if necessary.

- Exam Question:
- Adapt the relevant R code you used for the practice question to answer the questions.
- Copy and paste the relevant R text output and graphics into your Word document to support your answers. Change the text font to “Courier New” to align columns.

- Restrict your Word document to a **maximum of 2 pages** (re-sizing graphics and deleting irrelevant R output will help).
- Submit this Word document **via Canvas** by **5.00pm** ____________________ (**STRICT** deadline)
- Note that submitting the practical is a declaration that the practical is your own work. Plagiarism/copying will not be tolerated.

Practice Question (not to be submitted)

Data was collected on the following variables for 48 contiguous states in the USA and stored in the **fuel.txt** dataset:

X_1 = TAX = motor fuel tax rate in cents per gallon
X_2 = DLIC = percent of population with driver's licenses
X_3 = INC = per capita income in thousands of dollars
X_4 = ROAD = thousands of miles of federal-aided primary highways
Y = FUEL = motor fuel consumption in gallons per person

A model of the following form was fitted to this data:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + e_i, \quad e_i \sim IN(0,\sigma^2)$$

(a) Identify the estimate of β_1 and interpret it.

(b) Interpret the value of R-Squared.

(c) Test the hypothesis $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$.
Explain the practical implications of your conclusion.

(d) Test the hypothesis $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
against $H_1: \beta_1, \beta_2, \beta_3, \beta_4$ not all 0.
Explain the practical implications of your conclusion.

(e) Test the hypothesis $H_0: \beta_1 = \beta_4 = 0$ against $H_1: \beta_1, \beta_4$ not both 0.
Explain the practical implications of your conclusion.

(f) Explain how you would construct an added variable plot for TAX after DLIC, INC and ROAD. Explain its connection with the coefficient β_1 in the above model.

Exam Question (Winter 2020-21, Question 2) (to be submitted)
Data were collected on a random sample of adults who were undergoing a physical examination. The data are stored in **BMI.txt (on Canvas & P: drive)**.

Fit a model of the following form to these data:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + e_i, \quad e_i \sim NID(0, \sigma^2),$$

where

Y = BMI = Body Mass Index (kg/m^2),
X_1 = Waist = waist circumference (cm),
X_2 = Leg = upper leg length (cm),
X_3 = Elbow = elbow breadth (cm),
X_4 = Wrist = wrist breadth (cm),
X_5 = Arm = arm circumference (cm).

(a) Interpret the estimate of β_3. (5 marks)

(b) Interpret and comment on the value of *R*-squared. (7 marks)

(c) Test the hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ against H_1: $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ not all 0. Quote the value of the test statistic and the associated p-value. Explain the practical implications of your conclusion. (7 marks)

(d) Can Elbow and Wrist be excluded from the current model? Specify an appropriate hypothesis to test this. Quote the value of the test statistic and the associated p-value. Explain the practical implication of your conclusion. (9 marks)

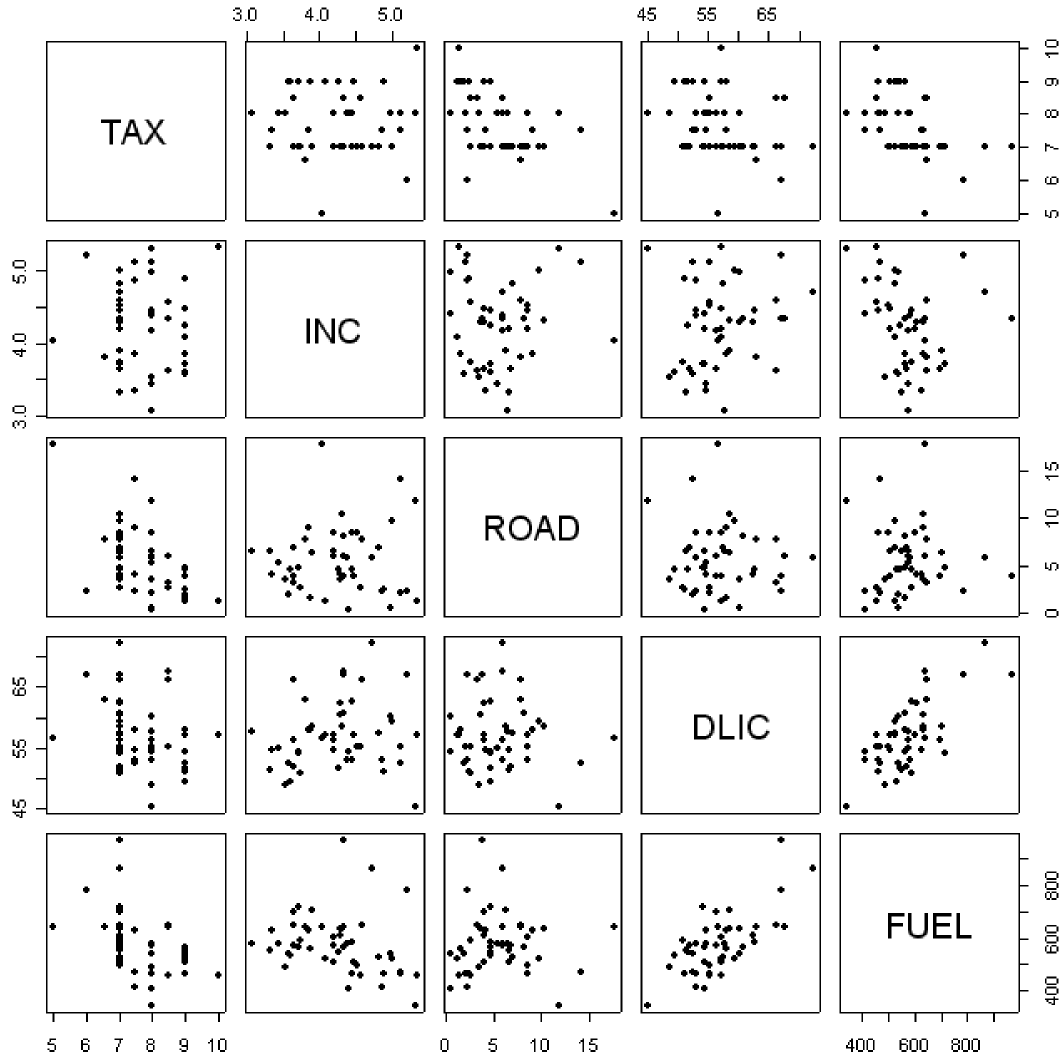
(e) Is there evidence of collinearity in the current model? What recommendation(s), if any, would you make? (10 marks)

(f) Test the hypothesis H_0: $\beta_3 = 0$, assuming all predictor variables in the current model are uncorrelated. (12 marks)

```
> # R code and output for Chapter 2

> # read fuel consumption data
> fuel.cons.df <- read.table("P:\\ST2053\\fuel.txt",
header=T)

> fuel.cons.df
     TAX   INC   ROAD DLIC FUEL
ME  9.00 3.571  1.976 52.5  541
NH  9.00 4.092  1.250 57.2  524
VT  9.00 3.865  1.586 58.0  561
MA  7.50 4.870  2.351 52.9  414
RI  8.00 4.399  0.431 54.4  410
CN 10.00 5.342  1.333 57.1  457
NY  8.00 5.319 11.868 45.1  344
NJ  8.00 5.126  2.138 55.3  467
PA  8.00 4.447  8.577 52.9  464
OH  7.00 4.512  8.507 55.2  498
IN  8.00 4.391  5.939 53.0  580
IL  7.50 5.126 14.186 52.5  471
MI  7.00 4.817  6.930 57.4  525
WI  7.00 4.207  6.580 54.5  508
MN  7.00 4.332  8.159 60.8  566
IA  7.00 4.318 10.340 58.6  635
MO  7.00 4.206  8.508 57.2  603
ND  7.00 3.718  4.725 54.0  714
SD  7.00 4.716  5.915 72.4  865
NE  8.50 4.341  6.010 67.7  640
KS  7.00 4.593  7.834 66.3  649
DE  8.00 4.983  0.602 60.2  540
MD  9.00 4.897  2.449 51.1  464
VA  9.00 4.258  4.686 51.7  547
WV  8.50 4.574  2.619 55.1  460
NC  9.00 3.721  4.746 54.4  566
SC  8.00 3.448  5.399 54.8  577
GA  7.50 3.846  9.061 57.9  631
FL  8.00 4.188  5.975 56.3  574
KY  9.00 3.601  4.650 49.3  534
TN  7.00 3.640  6.905 51.8  571
AL  7.00 3.333  6.594 51.3  554
MS  8.00 3.063  6.524 57.8  577
AR  7.50 3.357  4.121 54.7  628
LA  8.00 3.528  3.495 48.7  487
OK  6.58 3.802  7.834 62.9  644
TX  5.00 4.045 17.782 56.6  640
MT  7.00 3.897  6.385 58.6  704
ID  8.50 3.635  3.274 66.3  648
```

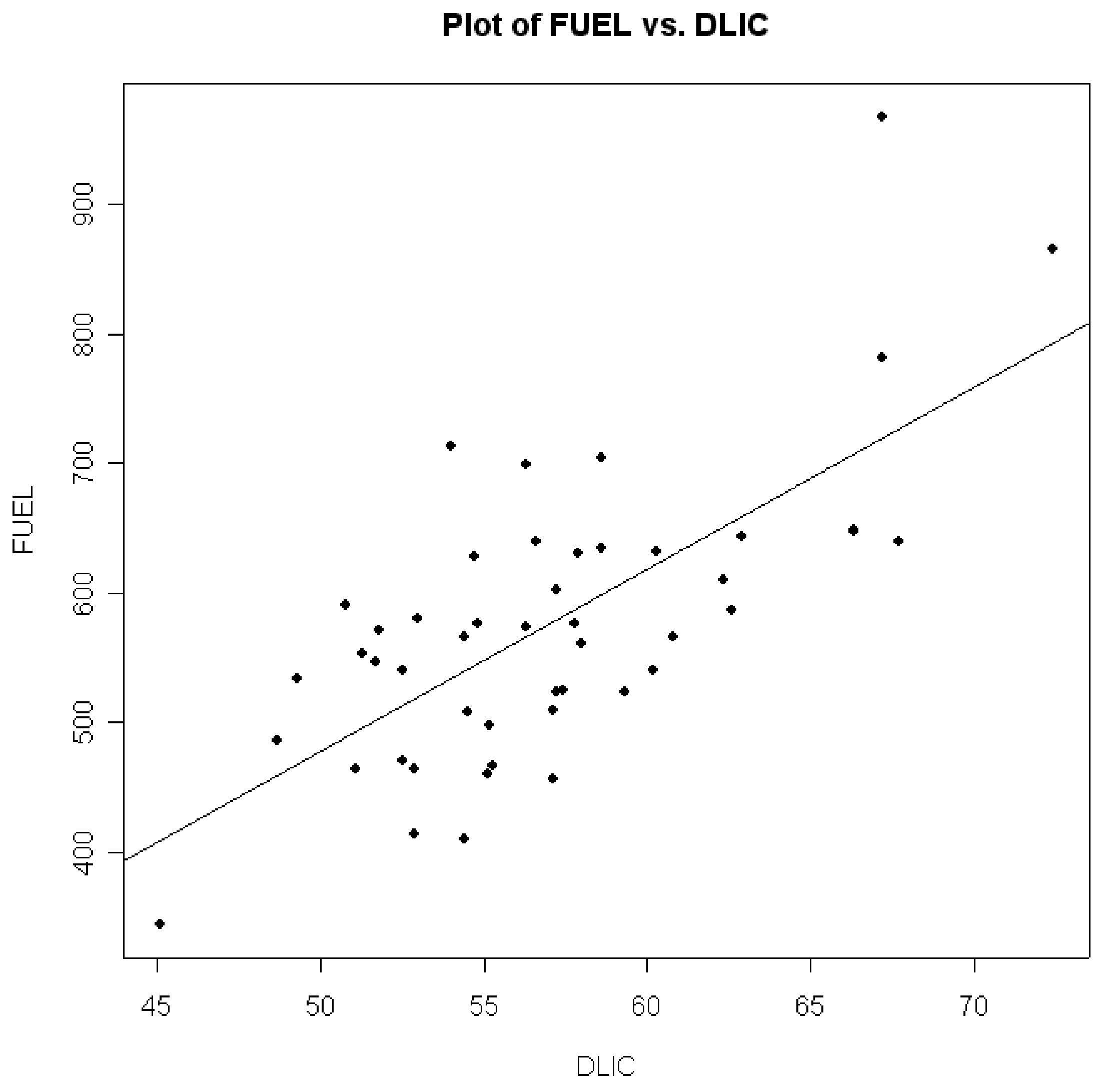
```
WY   7.00 4.345   3.905 67.2   968
CO   7.00 4.449   4.639 62.6   587
NM   7.00 3.656   3.985 56.3   699
AZ   7.00 4.300   3.635 60.3   632
UT   7.00 3.745   2.611 50.8   591
NV   6.00 5.215   2.302 67.2   782
WN   9.00 4.476   3.942 57.1   510
OR   7.00 4.296   4.083 62.3   610
CA   7.00 5.002   9.794 59.3   524
>
> # attach fuel.cons.df
> attach(fuel.cons.df)

> # summary statistics for fuel data
> # means, quartiles, maxima and minima
> summary(fuel.cons.df)

      TAX              INC             ROAD              DLIC            FUEL
 Min.   : 5.000   Min.   :3.063   Min.   : 0.431   Min.   :45.10   Min.   :344.0
 1st Qu.: 7.000   1st Qu.:3.739   1st Qu.: 3.110   1st Qu.:52.98   1st Qu.:509.5
 Median : 7.500   Median :4.298   Median : 4.736   Median :56.45   Median :568.5
 Mean   : 7.668   Mean   :4.242   Mean   : 5.565   Mean   :57.03   Mean   :576.8
 3rd Qu.: 8.125   3rd Qu.:4.579   3rd Qu.: 7.156   3rd Qu.:59.52   3rd Qu.:632.8
 Max.   :10.000   Max.   :5.342   Max.   :17.782   Max.   :72.40   Max.   :968.0

> # calculate variances and standard deviations
> variances <- vector()
> for ( i in 1:5) { variances [i] <- var(fuel.cons.df[,i])}
> names(variances) <- names(fuel.cons.df)
> variances
         TAX          INC         ROAD         DLIC         FUEL
9.039631e-01 3.290442e-01 1.219062e+01 3.076950e+01 1.251844e+04

> standevs <- variances^0.5
> standevs
        TAX         INC        ROAD        DLIC        FUEL
  0.9507698   0.5736238   3.4915072   5.5470265 111.8858156
```

```
> # scatter-plot matrix for fuel data
> pairs(fuel.cons.df,pch=16)
```

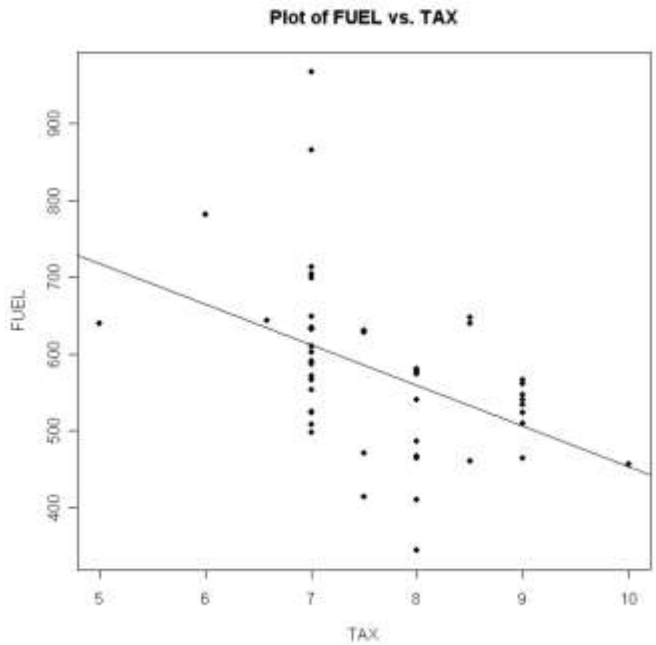



```
> cor(fuel.cons.df)
            TAX         INC        ROAD       DLIC        FUEL
TAX   1.00000000  0.01266516 -0.52213014 -0.2880372 -0.45128028
INC   0.01266516  1.00000000  0.05016279  0.1570701 -0.24486207
ROAD -0.52213014  0.05016279  1.00000000 -0.0641295  0.01904194
DLIC -0.28803717  0.15707008 -0.06412950  1.0000000  0.69896542
FUEL -0.45128028 -0.24486207  0.01904194  0.6989654  1.00000000
```

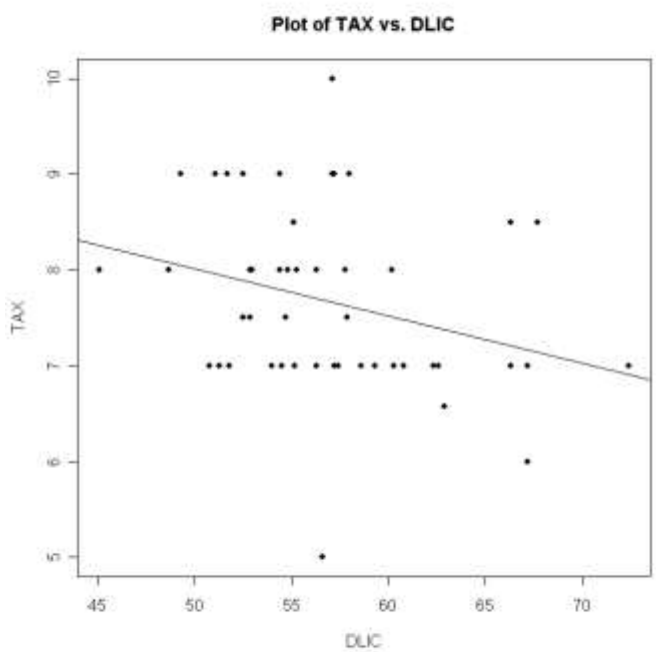
```
> # plot of FUEL vs. DLIC
> plot(DLIC,FUEL,main="Plot of FUEL vs. DLIC",pch=16)

> abline(lm(FUEL ~ DLIC))
```



```
> # plot of FUEL vs. TAX
> plot(TAX,FUEL,main="Plot of FUEL vs. TAX",pch=16)
> abline(lm(FUEL ~ TAX))
```

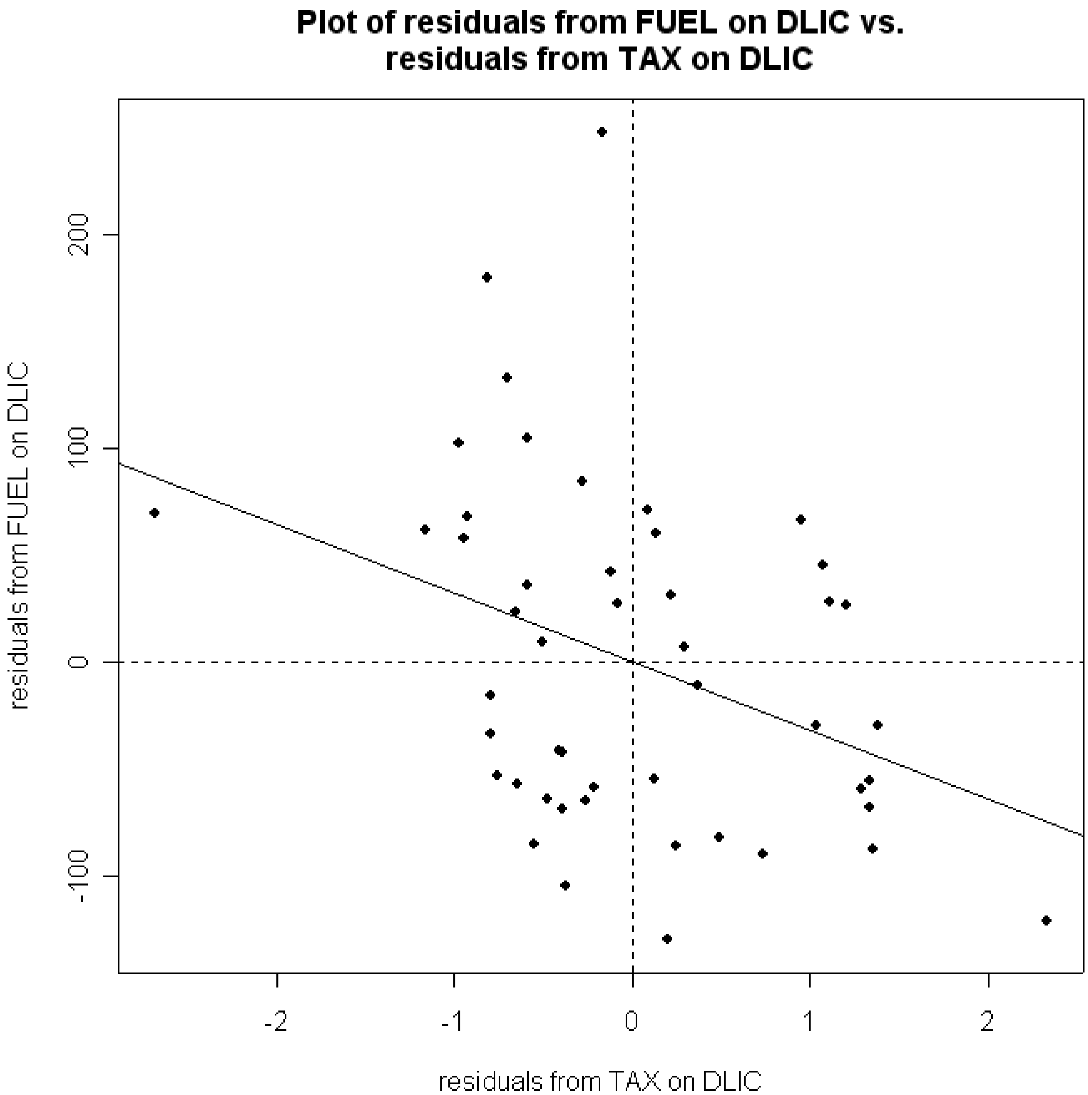



```
> # plot of TAX vs. DLIC
> plot(DLIC,TAX,main="Plot of TAX vs. DLIC",pch=16)
> abline(lm(TAX ~ DLIC))
```



```
> # regression line on FUEL on DLIC
> # regression coefficients and R-squared
> fuel.cons1.lm <- lm(FUEL ~ DLIC)
> summary(fuel.cons1.lm)

Call:
lm(formula = FUEL ~ DLIC)

Residuals:
    Min      1Q  Median      3Q     Max 
-129.64  -60.53  -13.03   58.57  247.90 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -227.309    121.862  -1.865   0.0685 .  
DLIC          14.098      2.127   6.629 3.29e-08 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1 

Residual standard error: 80.88 on 46 degrees of freedom
Multiple R-Squared: 0.4886,     Adjusted R-squared: 0.4774 
F-statistic: 43.94 on 1 and 46 DF,  p-value: 3.290e-08 

> # regression line on FUEL on TAX
> # regression coefficients and R-squared
> fuel.cons2.lm <- lm(FUEL ~ TAX)
> summary(fuel.cons2.lm)

Call:
lm(formula = FUEL ~ TAX)

Residuals:
     Min       1Q   Median       3Q      Max 
-215.157  -72.269    6.744   41.284  355.736 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   984.01     119.62   8.226 1.38e-10 ***
TAX           -53.11      15.48  -3.430  0.00128 ** 
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1 

Residual standard error: 100.9 on 46 degrees of freedom
Multiple R-Squared: 0.2037,     Adjusted R-squared: 0.1863 
F-statistic: 11.76 on 1 and 46 DF,  p-value: 0.001285 
```

```
> # regression line on TAX on DLIC
> # regression coefficients and R-squared
> fuel.cons3.lm <- lm(TAX ~ DLIC)
> summary(fuel.cons3.lm)

Call:
lm(formula = TAX ~ DLIC)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6897 -0.6058 -0.1886  0.5501  2.3350 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.48407    1.38663   7.561 1.32e-09 ***
DLIC        -0.04937    0.02420  -2.040   0.0471 *  
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 0.9203 on 46 degrees of freedom
Multiple R-Squared: 0.08297,    Adjusted R-squared: 0.06303 
F-statistic: 4.162 on 1 and 46 DF,  p-value: 0.04711 

> # added variable plot for TAX
> # Plot of residuals from FUEL on DLIC vs.
> # residuals from TAX on DLIC
> plot(resid(fuel.cons3.lm),resid(fuel.cons1.lm),
main="Plot of residuals from FUEL on DLIC vs. residuals
from TAX on DLIC", xlab="residuals from TAX on DLIC",
ylab="residuals from FUEL on DLIC",pch=16)
> abline(h=0,lty=2)
> abline(v=0,lty=2)
> # regression line of "residuals from FUEL on DLIC" on
> # residuals from TAX on DLIC
> fuel.cons4.lm <-
lm(resid(fuel.cons1.lm)~resid(fuel.cons3.lm))
> abline(fuel.cons4.lm)
```




```
> coef(fuel.cons4.lm)
        (Intercept) resid(fuel.cons3.lm)
      -2.052926e-15        -3.207532e+01
> # note intercept coefficient in above =0;
> # note slope coefficient in above = -32.07532
> # this equals the coefficient of TAX in the model
> # containing both TAX and DLIC as predictors below
```

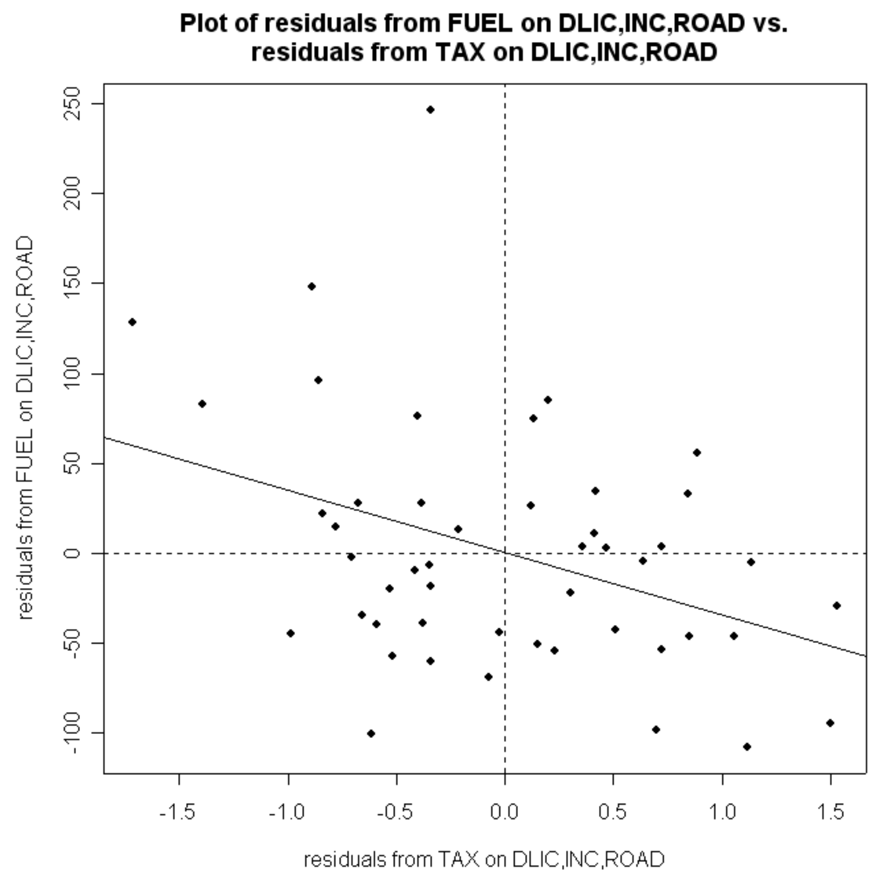
```
> # regression equation of FUEL on TAX and DLIC
> fuel.cons5.lm <- lm(FUEL~TAX+DLIC)
> summary(fuel.cons5.lm)

Call:
lm(formula = FUEL ~ TAX + DLIC)

Residuals:
     Min       1Q   Median       3Q      Max
-123.177  -60.172   -2.908   45.032  242.558

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  108.971    171.786   0.634   0.5291
TAX          -32.075     12.197  -2.630   0.0117 *
DLIC          12.515      2.091   5.986 3.27e-07 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 76.13 on 45 degrees of freedom
Multiple R-Squared: 0.5567,     Adjusted R-squared: 0.537
F-statistic: 28.25 on 2 and 45 DF,  p-value: 1.125e-08

> # coefficient of TAX in above = -32.075
> # Multiple R-squared in above model is 0.5567

> # correlation coefficient between FUEL and TAX,
> cor(FUEL,TAX)
[1] -0.4512803
> # cor(FUEL,TAX) = -0.4512803

> # partial correlation coefficient between FUEL and TAX,
> # adjusted for DLIC
> cor(resid(fuel.cons1.lm),resid(fuel.cons3.lm))
[1] -0.3649755
> # cor(FUEL,TAX|DLIC) = -0.3649755
```

```
> # regression equation of FUEL on TAX,DLIC,INC,ROAD
> # (includes overall F-Value, R-squared,
> # residual mean square )
> # Note : residual mean square = residual standard error^2
> fuel.cons6.lm <- lm(FUEL~TAX+DLIC+INC+ROAD)

> summary(fuel.cons6.lm)

Call:
lm(formula = FUEL ~ TAX + DLIC + INC + ROAD)

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)  377.291    185.541   2.033 0.048207 *
TAX          -34.790     12.970  -2.682 0.010332 *
DLIC          13.364      1.923   6.950 1.52e-08 ***
INC          -66.589     17.222  -3.867 0.000368 ***
ROAD          -2.426      3.389  -0.716 0.477999
---
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
```

```
> # X-matrix and Y-variable in above model
> model.matrix(fuel.cons6.lm)
   (Intercept)   TAX DLIC   INC   ROAD
ME           1  9.00 52.5 3.571  1.976
NH           1  9.00 57.2 4.092  1.250
............
OR           1  7.00 62.3 4.296  4.083
CA           1  7.00 59.3 5.002  9.794
attr(,"assign")
[1] 0 1 2 3 4

> FUEL
 [1] 541 524 561 414 410 457 344 467 464 498 580 471 525
508 566 635 603 714 865
[20] 640 649 540 464 547 460 566 577 631 574 534 571 554
577 628 487 644 640 704
[39] 648 968 587 699 632 591 782 510 610 524

> # ANOVA table
> # Note : for overall F-Value see summary(fuel.cons6.lm)
> anova(fuel.cons6.lm)
Analysis of Variance Table

Response: FUEL
          Df Sum Sq Mean Sq F value    Pr(>F)
TAX        1 119823  119823 27.2541 4.901e-06 ***
DLIC       1 207709  207709 47.2441 1.963e-08 ***
INC        1  69532   69532 15.8152 0.0002632 ***
ROAD       1   2252    2252  0.5123 0.4779989
Residuals 43 189050    4397
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

> # partial F-Test for TAX|ROAD,INC,DLIC
> fuel.cons7.lm <- lm(FUEL~ROAD+INC+DLIC)
> anova(fuel.cons7.lm,fuel.cons6.lm)
Analysis of Variance Table

Model 1: FUEL ~ ROAD + INC + DLIC
Model 2: FUEL ~ TAX + DLIC + INC + ROAD
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     44 220682
2     43 189050  1     31632 7.1948 0.01033 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1
```

```
> # sequential analysis of variance tables
> fuel.cons8.lm <- lm(FUEL~DLIC+TAX+INC+ROAD)

> anova(fuel.cons8.lm)
Analysis of Variance Table

Response: FUEL
          Df Sum Sq Mean Sq F value    Pr(>F)
DLIC       1 287448  287448 65.3809 3.584e-10 ***
TAX        1  40084   40084  9.1173 0.0042477 **
INC        1  69532   69532 15.8152 0.0002632 ***
ROAD       1   2252    2252  0.5123 0.4779989
Residuals 43 189050    4397
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1
> fuel.cons8.lm <- lm(FUEL~ROAD+INC+DLIC+TAX)

> anova(fuel.cons8.lm)
Analysis of Variance Table

Response: FUEL
          Df Sum Sq Mean Sq F value   Pr(>F)
ROAD       1    213     213  0.0485 0.826693
INC        1  35642   35642  8.1070 0.006735 **
DLIC       1 331829  331829 75.4755 5.15e-11 ***
TAX        1  31632   31632  7.1948 0.010332 *
Residuals 43 189050    4397
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

> # added variable plot for TAX after DLIC, INC and ROAD
> fuel.cons9.lm <- lm(TAX~DLIC+INC+ROAD)
> fuel.cons10.lm <-lm(FUEL~DLIC+INC+ROAD)

> # plot of residuals from FUEL on DLIC,INC,ROAD vs.
> # residuals from TAX on DLIC,INC,ROAD
> plot(resid(fuel.cons9.lm),resid(fuel.cons10.lm),
main="Plot of residuals from FUEL on DLIC,INC,ROAD vs.
residuals from TAX on DLIC,INC,ROAD", xlab="residuals from
TAX on DLIC,INC,ROAD", ylab="residuals from FUEL on
DLIC,INC,ROAD",pch=16)

> abline(h=0,lty=2)
> abline(v=0,lty=2)
```

```
> fuel.cons11.lm <-
lm(resid(fuel.cons10.lm)~resid(fuel.cons9.lm))

> abline(fuel.cons11.lm)
> coef(fuel.cons11.lm)
         (Intercept) resid(fuel.cons9.lm)
      -3.236995e-16        -3.479015e+01
> # note coefficient = -34.79015 = coefficient of TAX in
full model
```




```
> # quit R
> q("yes")
```