MULTIPLE REGRESSION

Testing for Associations in Chocolates

by Melissa Lavides & Erin Norton

There is a yummy variety of chocolates with different nutritional values based on their composition. Nutritional value can be associated with one or more variables such as protein, fat, carbohydrates, and sodium content. Additionally, perhaps the sizes, prices and unit prices of the chocolates are correlated with their nutritional value.

Linear models are a good way to represent data that has either one independent variable, or multiple independent variables measured against a dependent variable. A simple linear regression is a linear model which represents the slope and intercept of a line best fitting a dependent variable against a single independent variable, while multiple regressions aid us in visualizing those data sets with multiple independent variables (either in quantitative or factor form). We can use analyses of multiple regressions to determine the most optimal model for a particular set of variables.

The example used in the following pages refers to 16 different chocolate bar brands in terms of their energy level (the dependent variable, relying—at least in this case—on the other variables) in relation to their size, price, unit price, protein content, fat content, protein content, and sodium content (the independent variables).

We can use a **Multiple Regression Analysis** to identify any correlations between the nutritional values of different brands of chocolate and their respective compositions and prices. Correlations between nutritional values and compositions can even be used to classify the large variety of chocolates into distinct collections such as chocolates with "high nutritional value" and chocolates considered "diet."

THE DATA SET

- 16 brands of chocolates sold in Queensland, Australia in 2002
- **Energy** is a standard measurement of nutritional value and is the chosen **dependent variable** (Y) against which we test the composition for correlations
- **7 independent variables** (X_i) chosen from the standard composition of each brand of chocolate

Y	Energy	kilojoules per 100 grams of the chocolate
X1	Size	size of the chocolate in grams
X2	Price	price in dollars of the chocolate
X3	Unit.Price	unit price in dollars of the chocolate
X4	Protein	% of protein in the chocolate
X5	Fat	% of fat in the chocolate
X6	Carbo	% of carbohydrate in the chocolate
X7	Sodium	sodium content of the chocolate in milligrams

	Size	Price	Unit.Price	Energy	Protein	Fat	Carbo	Sodium
Dark.Bounty	50.0	0.88	1.76	1970	3.1	27.2	53.2	75
Bounty	50.0	0.88	1.76	2003	4.6	26.5	59.0	115
Milo.Bar	40.0	1.15	2.88	2057	9.9	23.0	60.9	116
Viking	80.0	1.54	1.93	1920	5.1	18.4	67.5	220
KitKat.White	45.0	1.15	2.56	2250	7.2	30.1	59.4	110
KitKat.Chunky	78.0	1.40	1.79	2186	7.0	28.4	59.7	93
Cherry.Ripe	55.0	1.28	2.33	1930	3.5	24.5	56.4	40
Snickers	60.0	0.97	1.62	1980	10.2	22.9	59.9	190
Mars	60.0	0.97	1.62	1890	4.7	19.5	67.9	160
Crunchie	50.0	1.28	2.56	2030	5.6	20.4	67.4	250
Tim.Tam	40.0	1.10	2.75	2180	5.5	26.8	67.3	160
Turkish.Delight	55.0	1.28	2.33	1623	2.2	9.2	73.3	90
Mars.Lite	44.5	0.97	2.18	1640	3.7	12.0	77.9	220
Dairy.Milk.King	75.0	1.58	2.11	2210	8.2	29.8	57.0	110
Maltesers	60.0	1.55	2.58	1980	8.5	20.6	63.3	130
MandMs	42.5	1.18	2.78	1970	5.0	20.0	69.0	148

USING THE R INTERPRETER

- Upload and save the **chocolates** data set into your working directory

```
CHOC=read.table("chocolates.txt",header=TRUE)
```

- Attach the data set into R interpreter

```
attach(CHOC)
CHOC
```

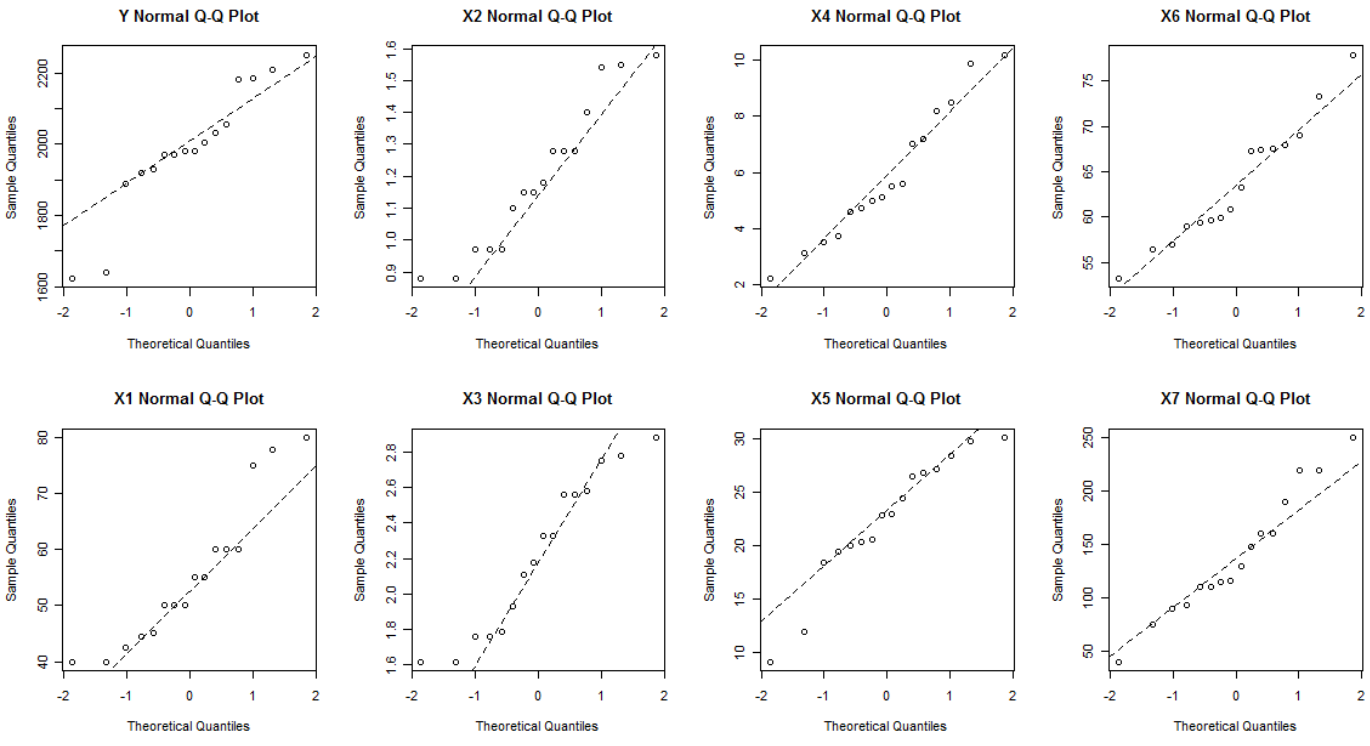
- Assign the dependent variable Y to "**Energy**" and the independent variables X_i to the rest

```
Y=Energy
X1=Size
X2=Price
X3=Unit.Price
X4=Protein
X5=Fat
X6=Carbo
X7=Sodium
```

TEST THE NORMALITY OF THE SAMPLE

- Before we perform statistical tests, use the **qqnorm()** or the **qqmath()** (run the lattice package) function to test the Normal Distribution of each variable in the chocolate data set. A normal distribution is an underlying assumption for creating a multiple regression, so this is a necessary step.
- We can use the **par(mfcol=c(,))** function to display all of the graphs onto one screen

```
par(mfcol=c(2,4))
qqnorm(Y,main="Y Normal Q-Q Plot")
qqline(Y,lty=2)
qqnorm(X1,main="X1 Normal Q-Q Plot")
qqline(X1,lty=2)
qqnorm(X2,main="X2 Normal Q-Q Plot")
qqline(X2,lty=2)
qqnorm(X3,main="X3 Normal Q-Q Plot")
qqline(X3,lty=2)
qqnorm(X4,main="X4 Normal Q-Q Plot")
qqline(X4,lty=2)
qqnorm(X5,main="X5 Normal Q-Q Plot")
qqline(X5,lty=2)
qqnorm(X6,main="X6 Normal Q-Q Plot")
qqline(X6,lty=2)
qqnorm(X7,main="X7 Normal Q-Q Plot")
qqline(X7,lty=2)
par(mfcol=c(1,1))
```

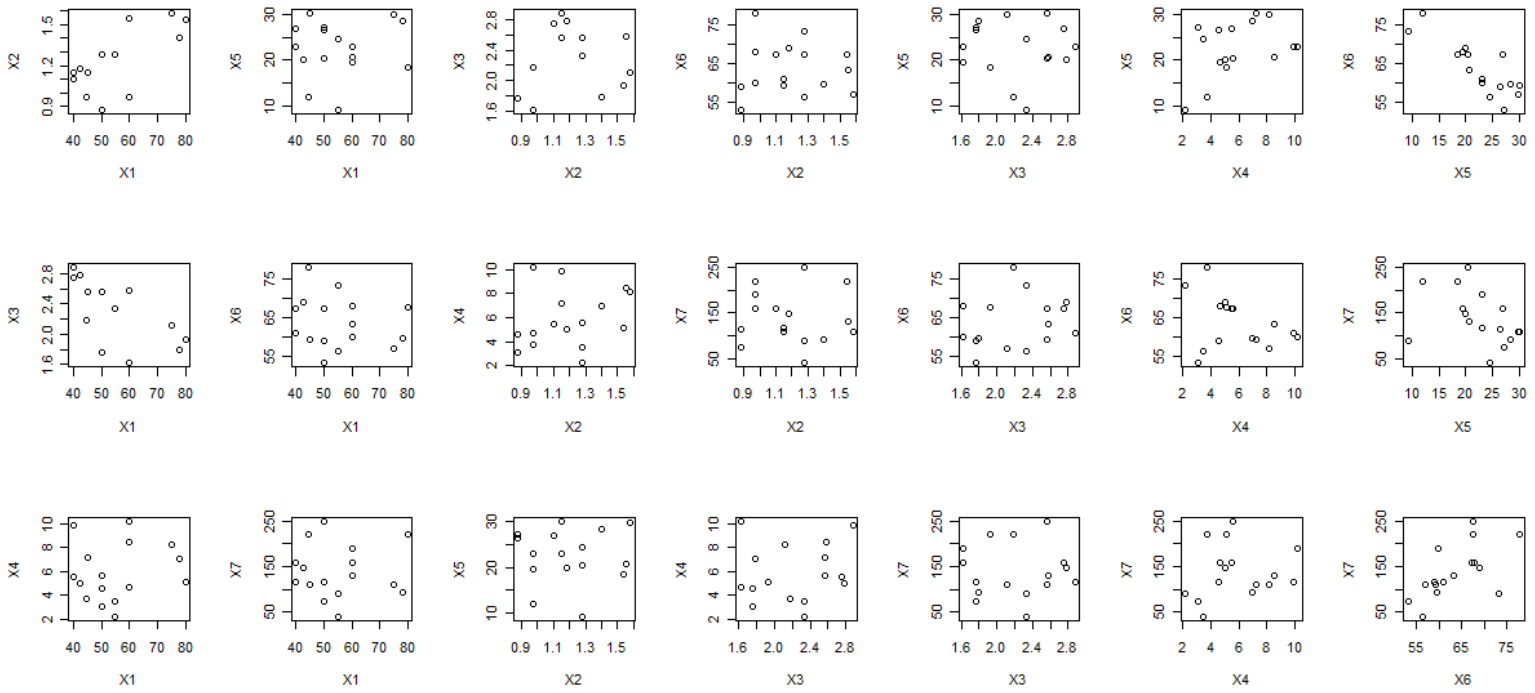


- Although the graphs show slight S-shape, the **qqline()** function shows that the data fits without striking evidence of non-normality.

TEST THE CORRELATIONS USING X vs Y GRAPHS

- It is also important to use a scatter plot to plot each independent variable against each other to determine whether there is correlation between any of them. Correlation between the independent variables themselves can skew the multiple regression and report false results.

```
par(mfcol=c(3,7))
plot(X1,X2)
plot(X1,X3)
plot(X1,X4)
plot(X1,X5)
plot(X1,X6)
plot(X1,X7)
plot(X2,X3)
plot(X2,X4)
plot(X2,X5)
plot(X2,X6)
plot(X2,X7)
plot(X3,X4)
plot(X3,X5)
plot(X3,X6)
plot(X3,X7)
plot(X4,X5)
plot(X4,X6)
plot(X4,X7)
plot(X5,X6)
plot(X5,X7)
plot(X6,X7)
par(mfcol=(1,1)))
```

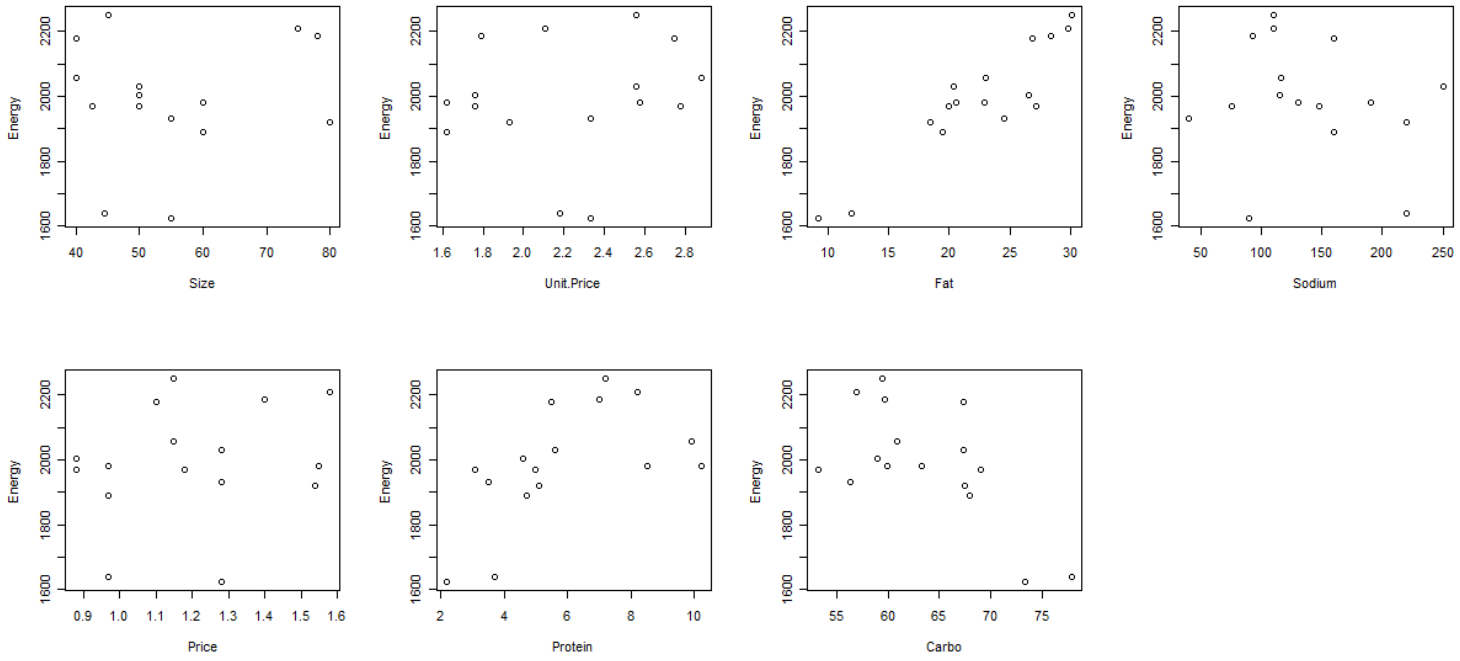


- We expect the independent variables to graph in random scatter. However, a linear set of points illustrates an association between the variables. We must remove these related variables because they will skew our regression tests.

- According to our graphs, we remove the interrelated variables **Size** (X1)**, Price** (X2) and **Unit.Price** (X3). This makes sense because biologically Energy (nutritional value) should not depend on these external variables (they are not parts of the chocolate's composition)

- Next, we plot each independent variable (X) against Energy (Y) for any visual correlations

```
par(mfcol=c(2,4))
plot(X1,Y,xlab="Size",ylab="Energy")
plot(X2,Y,xlab="Price",ylab="Energy")
plot(X3,Y,xlab="Unit.Price",ylab="Energy")
plot(X4,Y,xlab="Protein",ylab="Energy")
plot(X5,Y,xlab="Fat",ylab="Energy")
plot(X6,Y,xlab="Carbo",ylab="Energy")
plot(X7,Y,xlab="Sodium",ylab="Energy")
par(mfcol=(1,1)))
```



- The graphs further prove that Size, Price and Unit.Price are not correlated with Energy.
- In contrast, the **Protein** (X4) and **Fat** (X5) variables show positive correlation with Energy, and **Carbo** (X6) shows a negative correlation with Energy. These results make sense because biologically these variables are expected to supply Energy (and affect nutritional value).

SIMPLE LINEAR REGRESSION

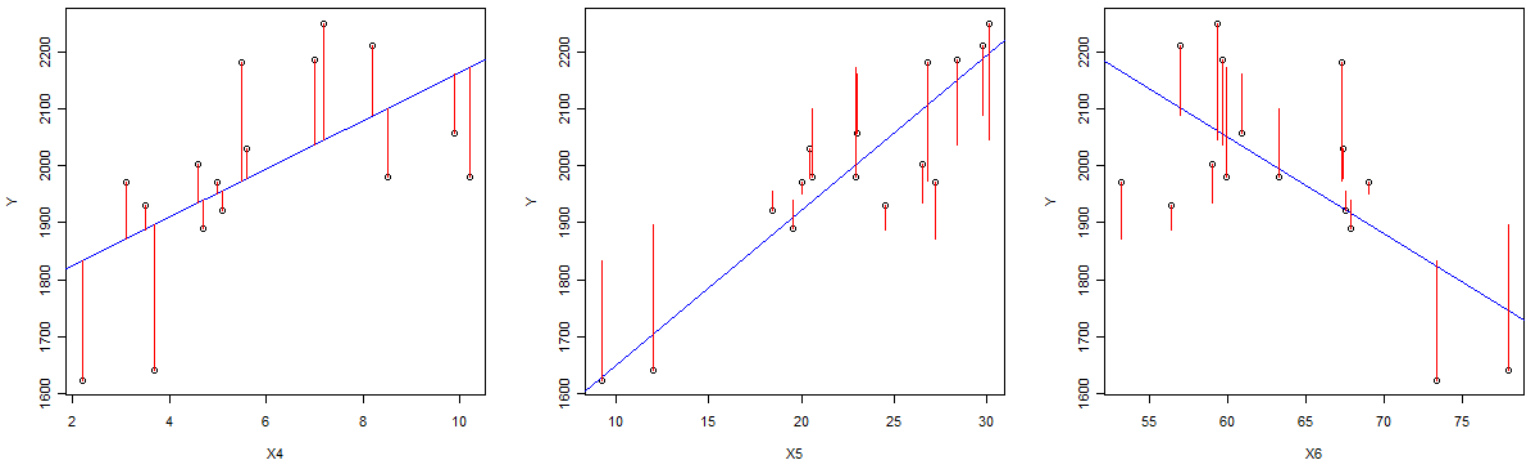
- Now that we have reduced our original model to a simpler equation, we can run simple linear regressions on each of the independent variables (X4, X5, X6 and X7) against the dependent variable in order to determine whether there is an association between each one individually against the dependent variable. This is important because once we begin the multiple regression, there is no way to determine the impact each independent variable has on its own on the dependent variable.

```
LM=lm(Y~X4)
LM
Yhat=fitted(LM)
e=residuals(LM)
RESULTS=data.frame(X4,Y,Yhat,e)
anova(LM)
```

VARIABLE	PROBABILITY (α = 0.05)	CONCLUSION	CORRELATED?
Protein (X4)	0.01989	p < α **reject H_0**	**YES**
Fat (X5)	6.818e-07	p < α **reject H_0**	**YES**
Carbo (X6)	0.007538	p < α **reject H_0**	**YES**
Sodium (X7)	0.5032	p > α **do not reject H_0**	**NO**

- Model: $\mathbf{Y = \alpha + \beta X + \varepsilon}$
  - $\boldsymbol{\alpha}$ = y intercept of regression line (translation)
  - $\boldsymbol{\beta}$ = slope of regression line (scaling coeff)
  - $\boldsymbol{\varepsilon}$ = error factor in prediction of Y given that it is a random variable distributed as $N(0, \sigma^2)$.
- Hypotheses:
  H_0: $\beta = 0$ the slope of regression equals zero, that is, X & Y are not related
  H_1: $\beta \neq 0$ the slope of regression does not equal zero, X & Y are related
- Using the Type I error set to α = 0.05 and comparing the probability value from ANOVA table, we conclude that **Protein, Fat** and **Carbo** but not Sodium are correlated with Energy.
- These results make sense because in biological comparison to the other variables, Sodium should not contribute to the Energy (nutritional value)
- **THEREFORE**: our reduced model is $\mathbf{Y = \alpha + \beta(X4+X5+X6) + \varepsilon}$
- Plot the regression line and points for X4, X5 and X6

```
par(mfcol=c(1,3))
plot(X4,Y)
abline(LM4,col="blue")
segments(X4,Yhat,X4,Y,col="red")
plot(X5,Y)
abline(LM5,col="blue")
segments(X5,Yhat,X5,Y,col="red")
plot(X6,Y)
abline(LM6,col="blue")
segments(X6,Yhat,X6,Y,col="red")
par(mfcol=c(1,1)))
```



LINEAR MODEL

- Now we construct the linear model using a multiple regression.
- The **null hypothesis** for this test is:
  H_0: all slope β's = 0
  This is to say that none of the regression lines for any of the dependent vs independent variable relationships has a slope other than zero. Thus none have a relationship.
- And the **alternative hypothesis** is:
  H_1: at least some slope β's ≠ 0
  This is to say that at least one of the regression line (representing a particular independent variable vs the dependent variable) has a slope other than zero, indicating that that/those particular comparison(s) have a relationship(s).
- We run the model by assigning the X and Y variables to the example variables, and telling R what we want to be included in our full LM (linear model). As discussed before, we are including Protein, Fat, Carbo, and Sodium, all in relation to Energy.

```
Y=Energy
X4=Protein
X5=Fat
X6=Carbo
X7=Sodium
LM=lm(Y~X4+X5+X6+X7)
LM
```

This LM command gives us a set of information, none of which is very useful to us:

Call:
lm(formula = Y ~ X4 + X5 + X6 + X7)

Coefficients:

(Intercept)	X4	X5	X6	X7
366.7358678495	17.6682265304	35.0077388533	11.5704086885	-0.0355955293

- Then we can use a summary function on the full linear model, which will most importantly give us the F statistic and p-value for the overall F-test on the linear model. This is provided at the very end of the test's output. This tests for differences in slopes of all the regression lines in the model, and we are able to use this p-value to reject or retain the null hypothesis. The summary also gives us p-values for each individual/partial t-test for each of the variables. These statistics are provided in the table, and can be used to reject hypotheses specifically designed to determine a relationship between that particular variable and the dependent variable.

```
summary(LM)
```

Call:
lm(formula = Y ~ X4 + X5 + X6 + X7)

Residuals:

Min	1Q	Median	3Q	Max
-105.70499263	-18.03618468	7.27507654	19.63443565	79.21752771

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	366.7358678495	353.8391196641	1.03645	0.322241

```
X4          17.6682265304   6.7532750361  2.61625   0.023989 *
X5          35.0077388533   4.3319486262  8.08129 5.9327e-06 ***
X6          11.5704086885   4.5432453565  2.54673   0.027153 *
X7          -0.0355955293   0.3306187071 -0.10766   0.916201
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.1448944 on 11 degrees of freedom
Multiple R-squared: 0.936823508,        Adjusted R-squared: 0.913850238
F-statistic: 40.7788494 on 4 and 11 DF,  p-value: 1.55635612e-06
```

The p-value we see is less than an alpha value of either 0.05 or 0.01, which is sufficient to reject the null hypothesis and thus conclude that at least one of the regression lines is not equal to zero, and therefore represents a relationship between a particular independent variable and the dependent variable. From the partial t-tests, we can see that the p-value for the association between X7 (sodium) and Y (energy) is very high, which leads us to conclude that there is no relationship between the two. The other independent variables have relatively low p-values, so we will definitely continue to look into their association with Energy.

- Next we can use an anova function on the linear model to further analyze the data:

```
anova(LM)
```

```
Analysis of Variance Table

Response: Y
          Df      Sum Sq      Mean Sq  F value     Pr(>F)
X4         1 156339.57930 156339.57930 57.49702 1.0831e-05 ***
X5         1 261830.57846 261830.57846 96.29346 8.9255e-07 ***
X6         1  25323.77157  25323.77157  9.31333   0.011013 *
X7         1     31.51807     31.51807  0.01159   0.916201
Residuals 11  29909.99009   2719.09001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

CREATING REDUCED LINEAR MODELS

- Next we want to use a marginal sums of squares test to create reduced linear models. Reduced models are variations of the full model, which includes all of the variables. The reduced forms of the model strive to be more fit for the model, by reducing the error term. If eliminating only one of the independent variables at a time causes the model to fit the data better, the marginal sums of squares test will tell us so. The key concept here is that this test runs marginally, meaning it only eliminates one X variable from the test at a time. R requires that we generate these reduced models in the following way:

```
RM1=lm(Y~X5+X6+X7)
RM2=lm(Y~X4+X6+X7)
RM3=lm(Y~X4+X5+X7)
RM4=lm(Y~X4+X5+X6)
```

Notice that each reduced model (RM) eliminates one X variable. In RM1, it is X4 that is eliminated, in RM2, it is X5 that is eliminated, and so on.

- We are then able to do an Anova on each reduced model. Notice that this is Anova, a command with a capital A, an ANOVA function dedicated to the marginal test. A summary of each reduced model can also be used, with the command **summary(RM1)**, etc. In the case of summary, the overall F-test will again provide, at the very end, the p-value to which the following hypotheses can be applied:
- The **null hypothesis** is:
  H_0: $\beta_i = 0$
  This is to say that the slope of the reduced model is zero, and that there is no relationship between the independent and dependent variables in the model
- And the **alternative hypothesis** is:
  H_1: $\beta_i \neq 0$
  This is to say that the slope of the reduced model is a value other than zero, implying that some relationship exists between the independent and dependent variables in the model.

*For Anova p-values (not the overall F-test p-values provided by the summary function), their meaning is relative to the probability that the variable of that row is associated with the dependent variable in the absence of the variable that has been eliminated from that particular reduced model. The results of each of our reduced Anovas are shown below.

Anova(RM1)

Anova Table (Type II tests)

Response: Y

	Sum Sq	Df	F value	Pr(>F)	
X5	199405.61074	1	49.31565	1.3899e-05	***
X6	10758.31127	1	2.66067	0.12881	
X7	3002.76531	1	0.74262	0.40571	
Residuals	48521.45797	12			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova(RM2)

Anova Table (Type II tests)

Response: Y

	Sum Sq	Df	F value	Pr(>F)	
X4	40440.73471	1	2.33890	0.152102	
X6	78972.61782	1	4.56739	0.053865	.
X7	4765.34189	1	0.27560	0.609160	
Residuals	207486.33400	12			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova(RM3)

Anova Table (Type II tests)

Response: Y

```
            Sum Sq Df  F value     Pr(>F)
X4        11734.24575  1  2.96160    0.11092
X5       238913.42832  1 60.29929 5.0934e-06 ***
X7         7719.75624  1  1.94839    0.18805
Residuals  47545.52350 12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Anova(RM4)
```

Anova Table (Type II tests)

```
Response: Y
            Sum Sq Df  F value     Pr(>F)
X4        21582.71512  1  8.64995  0.0123472 *
X5       182310.16773  1 73.06653 1.8962e-06 ***
X6        25323.77157  1 10.14930  0.0078364 **
Residuals  29941.50816 12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

AIC CRITERION FOR CREATING REDUCED MODELS

- Next we need to utilize a very important function in R for developing a reduced model based on a statistic other than p-values. There is a criterion called the AIC, or Akaike’s Information Criterion, which considers the sums of squares of different forms of reduced models to determine the ideal one. The reason it works this way is that when an independent variable is dropped from a linear model, such as in the instance of a reduced model, that variable’s sums of squares are transferred to the error term. This correlates to a good way to evaluate each reduced model.
- The AIC actually has a unique function called the automated backwards stepwise regression. This procedure begins with the full linear model (the dependent variable as well as all the independent variables) and eliminates one at a time, measuring the AIC value with each elimination. The eliminations continue for as long as at least one independent variable as a lower AIC value than the dependent variable (coded as <none>) in the data table. When the dependent variable has the lowest AIC, the eliminations stop and the remaining ones form the reduced model.

```
FM=lm(Y~X4+X5+X6+X7)
step(FM,direction="backward")
```

The program gives us the following:

```
Start:  AIC=130.53
Y ~ X4 + X5 + X6 + X7

      Df    Sum of Sq        RSS        AIC
- X7   1     31.51807  29941.50816 128.5505969
<none>                 29909.99009 130.5337456
- X6   1  17635.53340  47545.52350 135.9496672
- X4   1  18611.46788  48521.45797 136.2747630
- X5   1 177576.34391 207486.33400 159.5237125
```

The <none> variable is always the dependent variable, and we look for the AIC values that are lower than that one. We see that the one lowest than that AIC is that of variable 7. R removes it and generates a new AIC for the next step. Then the program starts over, beginning with the now-reduced model

```
Step:  AIC=128.55
Y ~ X4 + X5 + X6

      Df    Sum of Sq        RSS        AIC
<none>                  29941.50816 128.5505969
- X4    1  21582.71512  51524.22328 135.2354977
- X6    1  25323.77157  55265.27974 136.3569826
- X5    1 182310.16773 212251.67589 157.8870284

Call:
lm(formula = Y ~ X4 + X5 + X6)

Coefficients:
(Intercept)         X4         X5         X6
383.4489316  17.3716578  34.9262970  11.2862945
```

In this step, we see that the lowest AIC value is that of the dependent variable, which is the program's cue to stop. This is because the object is to have the dependent variable have the lowest AIC value of all of the variables in the model.

Since the program eliminated variables 4 and 6 before it concluded, the reduced linear model is displayed, and is Y ~ X4 + X5 + X6

This is the reduced linear model, with the X7 (Sodium) variable excluded. This result is consistent with our previous findings with the original simple regression analysis of Sodium against Energy, as well as with the multiple regression in which we began to suspect that Sodium is not associated with the dependent variable. We are then able to analyze this reduced model against the full model using an ANOVA or summary function.

FINAL ANOVA ANALYSIS

- The last step in our evaluation is to run an ANOVA on the reduced model versus the full model to decide whether or not the reduced model is truly the ideal option. The test we use is the general F test for model comparison.
- The **null hypothesis** is**:**
  H_0: coefficients in j but not included in k = 0
  In other words, the parsimonious model is preferred. The parsimonious (more simple) model is always preferred unless there is significant statistical proof otherwise. Thus the null hypothesis implies that the reduced model is more ideal.
- The **alternative hypothesis** is**:**
  H_1: at least some of these coefficients not 0
  This is to say that there is statistical reason to prefer the full model over the more parsimonious reduced model.

```
FM=lm(Y~X4+X5+X6+X7)
RM=lm(Y~X4+X5+X6)
anova(RM,FM)
```

Analysis of Variance Table

```
Model 1: Y ~ X4 + X5 + X6
Model 2: Y ~ X4 + X5 + X6 + X7
  Res.Df       RSS Df   Sum of Sq      F Pr(>F)
1     12 29941.50816
2     11 29909.99009  1 31.51807241 0.01159 0.9162
```

This table's most important feature is the p-value, 0.9162. Since this p-value is not less than alpha, we are not able to reject our null hypothesis. Thus we retain the null hypothesis, which implies that the parsimonious model is preferred. We conclude that the reduced model is preferred, and thus that the model including only variables X4, X5 and X6 (Protein, Fat, and Carbo) are meaning full in relation to the dependent variable of the model (Energy).

Based on a series of statistical analyses, we conclude that the optimal linear model for this dataset is the reduced model Y ~ X4 + X5 + X6.