

3. Multiple Regression Analysis

The concepts and principles developed in dealing with simple linear regression (i.e. one explanatory variable) may be extended to deal with several explanatory variables.

We begin with an example of two explanatory variables, both of which are continuous. The regression equation in such a case becomes:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2$$

It is customary to replace α with β_0 , and so all future regression equations would be written as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots \dots \beta_n x_n$$

Regression with two explanatory variables

During certain surgical operations the surgeon may wish to lower the blood pressure of the patient by administering a drug. After the surgery is over the return to normal of the blood pressure depends on the dose of the drug administered, and the average systolic blood pressure reached during surgery.

A surgeon wishes to study the relationship between the dose of a new drug, the average systolic blood pressure during the operation and time for the blood pressure to come back to normal after administration of the drug has ceased.

The results obtained are in the following table:
(Data from *Anaesthesia* 1959;14:53-64).

Case	LogDose	B.P.Surg	Recov time
1	5.2	66	7
2	4.17	52	10
3	4.1	72	18
4	3.55	67	4
5	4.74	69	10
6	4.01	71	13
7	5.89	88	21
8	5.27	68	12
9	4.14	59	9
10	5.34	73	65
11	4.7	68	20
12	4.33	58	31
13	2.72	61	23
14	4.79	68	22
15	3.91	69	13
16	4.01	55	9
17	4.37	67	12
18	4.12	67	12

Case	LogDose	B.P.Surg	Recov time
19	4.86	68	11
20	3.96	59	8
21	4.01	68	26
22	3.68	63	16
23	4.95	65	23
24	5.2	72	7
25	3.8	58	11
26	3.75	69	8
27	5.53	70	14
28	6.22	73	39
29	4.37	56	28
30	6.4	83	12
31	5.23	67	60
32	4.01	84	10
33	6.03	68	60
34	4.14	64	22
35	4.17	60	21
36	3.64	62	14
37	5.55	76	4
38	3.8	60	27
39	5.16	60	26
40	3.91	59	28
41	5.64	84	15
42	3.96	66	8
43	5.46	68	46
44	5.13	65	24
45	4.42	69	12
46	4.58	72	25
47	4.58	63	45
48	5.41	56	72
49	4.14	70	25
50	5.43	69	28
51	3.66	60	10
52	4.84	51	25
53	4.14	61	44

We next perform scatter plots of the response variable 'recovery time' against the two explanatory variables viz. 'Log dose' and 'B.P. during surgery'

Fig. 3.1 Scatter plot of Recovery time against 'Log.dose'

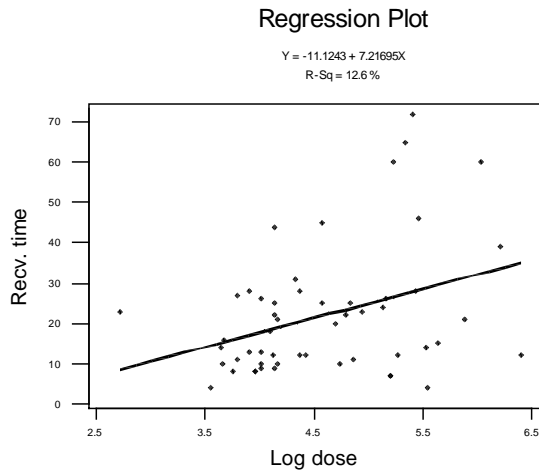
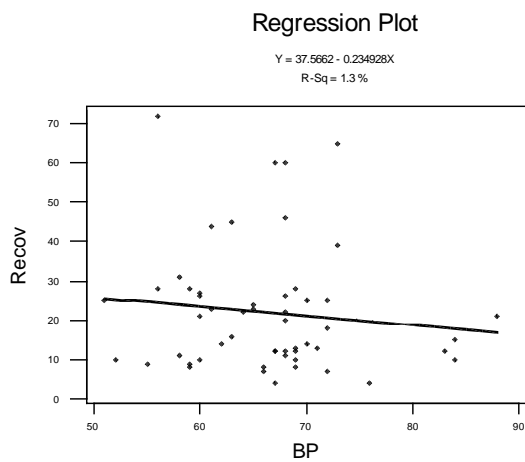


Fig. 3.2 Scatter plot of 'Recovery time' against 'B.P. during surgery'



The two scatter plots with their regression lines inserted show linear relationships. The higher the value of 'Log dose' the longer is the time to recovery. And the lower the blood pressure recorded during surgery the longer is the time for the blood pressure to reach normal. The question is "What is the effect of the two explanatory variables acting jointly on recovery time?" To answer the question we perform a multiple liner regression analysis, and the outcome is shown next:

[In MINITAB → Stat → Regression. In Response Box “Recov time” In Predictors Box “LogDose”; “B.P.Surg”]

Multiple Regression Analysis

The regression equation is
 Recv. time = 22.3 + 10.6 Log dose - 0.740 Surg.B.P.

Predictor	Coef	StDev	T	P
constant	22.27	17.55	1.27	0.210
Log dose	10.640	2.856	3.73	0.000
Surg.B.P	-0.7401	0.2893	-2.56	0.014

S = 14.24 R-Sq = 22.8% R-Sq(adj) = 19.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	2988.2	1494.1	7.36	0.002
Residual Error	50	10144.8	202.9		
Total	52	13133.0			

Source	DF	Seq SS
Log dose	1	1660.6
Surg.B.P	1	1327.6

Unusual Observations

Obs	Log dose	Recv. ti	Fit	StDev Fit	Residual	St Resid
7	5.89	21.00	19.81	5.92	1.19	0.09 X
10	5.34	65.00	25.06	2.88	39.94	2.86R
31	5.23	60.00	28.33	2.63	31.67	2.26R
32	4.01	10.00	2.77	6.37	7.23	0.57 X
47	5.41	72.00	38.39	4.99	33.61	2.52R

R denotes an observation with a large standardized residual
 X denotes an observation whose X value gives it large influence.

Reading the output

The plot of log dose against recovery time and similarly that of blood pressure achieved during surgery against recovery time show that there could be a linear relationship. The regression equation comprising both the explanatory variables is

$$\text{Recovery Time} = 22.3 + 10.6 \log\text{dose} - 0.740 \text{ Surg.B.P.}$$

The **F ratio** (in the Analysis of Variance Table) is 7.36 and significant at p = .002. This provides evidence of existence of a linear relationship between the response (Recovery time) and the two explanatory variables (logdose and BP).

We may wish to know whether there is a relationship between the response variable and the BP achieved during surgery over and above the dose of the hypotensive (i.e. the variable “logdose”). We can test for this by looking at the **t-ratio** in the second paragraph of the output. The T ratio for “Surg.B.P.” at -2.56 is a significant value p = 0.014. (The t-ratio is computed from the value of $\beta_2 \div$ standard deviation of β_2).

In a regression analysis with more than one explanatory variable the coefficient of determination R^2 is defined as:

Sum of squares explained by the model (i.e. Regression sum of Squares) ÷ Total Sum of Squares.

In the case of one explanatory variable the coefficient of determination is simply the square of the coefficient of correlation viz. r^2 .

These two values viz. **F ratio** and **t-ratio** tell us respectively whether there is a linear relationship between the response and explanatory variables taken together, and whether any given explanatory variable has an influence on the response variable over and above that of the other explanatory variables.

The output obtained when there are two explanatory variables differs slightly from the one in the case of a single explanatory variable by having one additional table under Analysis of Variance. This is the table of Sequential Sum of Squares (**SEQ SS** in the output). It provides a measure of the contribution made by each explanatory variable to the Regression Sum of Squares. In the present example Log dose contributes 1660.6 and Surg.B.P. adds further 1327.6 making a total of 2988.2 of the Regression SS.

In the analysis 5 observations have been flagged as being unusual (observations no. 7, 10, 31, 32, and 48). What to do about such unusual observations is discussed later. At the moment suffice it to say that one should carefully check these data points.

A question commonly asked is whether one explanatory variable is more important than the other. The effect of any given explanatory variable depends on which other variables have been included in the regression model. The question cannot be answered by simply looking at the respective values of the β coefficients, because the value of the β coefficient depends on the unit of the explanatory variable. In our example, dose of the drug is as logarithm of the dose used and BP is measured in mm. of Hg. There can be no comparison between such disparate quantities. Instead we look at the t-ratios of the two variables; that for logdose is 3.73 which is higher than -2.56 for Surg.B.P. So the effect of the dose is greater than the blood pressure achieved during surgery on recovery time.

Summary of Findings.

Figure 3.1 tells us about the effect of the dose of the hypotensive drug used on recovery time. R-sq = 12.6% indicating that about 13% of the variation in recovery time can be accounted for by the dose of the drug.

Figure 3.2 tells us about the effect of blood pressure achieved during surgery and time to recovery to the original level. The greater the drop in the blood pressure the longer it would take for it to return to its original level. The R-sq. value is 1.3%.

The Multiple Regression Analysis describes the effect of the two explanatory variables acting jointly on the recovery time. R-sq improves to 22.8% indicating that even though the dose of the drug is an important factor in lowering the blood pressure, the lower the blood pressure achieved during surgery the longer it takes for it to recover to normal value.

An interesting message here is about the individual variability of subjects in responding to the hypotensive drug. For the same dose those subjects experiencing a greater fall in their blood pressure would take longer for it to return to normal.

Second Example of Two Explanatory Variables

How does multiple regression analysis handle data when the explanatory variables are discrete numerical? The following example demonstrates this.

The Toxic Shock Syndrome

(Suttorp N, Galanos C, Neuhof H. *Am.J.Physiol.* 1987;253:C384-C390.)

Certain bacteria carry endotoxins, which cause shock when the bacteria invade the blood stream. The authors hypothesised that one way in which these bacteria cause shock is by acting on the endothelial cells lining the blood vessels and making them release prostacyclin, which dilates blood vessels resulting in fall in blood pressure. Prostacyclin is made from arachidonic acid.

To test the hypothesis, endothelial cells were grown in culture, then exposed to endotoxin and prostacyclin production was measured. Four different levels of endotoxin were used: 0, 10, 50 and 100. After exposure to endotoxin for several hours prostacyclin production was evaluated by stimulating the cells with three levels of the prostacyclin precursor (arachidonic acid): 10, 25 and 50 μM .

Since arachidonic acid is a precursor of prostacyclin, the amount of prostacyclin produced will depend on the amount of arachidonic acid present. The question is: Did production of prostacyclin also depend on the level of endotoxin?

The data are in the table below :

Prosta (P)	Arachi (A)	Endoto (E)
19.2	10	0
10.8	10	0
33.6	10	0
11.9	10	10
15.9	10	10
33.3	10	10
81.1	10	50
36.7	10	50
58	10	50
60.8	10	100
50.6	10	100
69.4	10	100
30.8	25	0
27.6	25	0
13.2	25	0
38.8	25	10
37	25	10
38.3	25	10
65.2	25	50
66.4	25	50
63.2	25	50
49.9	25	100
89.5	25	100
60.5	25	100
102.9	50	0

Prosta (P)	Arachi (A)	Endoto (E)
57.1	50	0
76.7	50	0
70.5	50	10
66.4	50	10
76.3	50	10
83.1	50	50
61.7	50	50
101.5	50	50
86.2	50	100
115.9	50	100
102.1	50	100

The output from regression analysis is given below:

[In MINITAB → Stat → Regression. In Response Box “Prosta(P)”; in Predictors Box “Arach (A) , “Endotox”]

Regression Analysis

The regression equation is

$$\text{Prosta(P)} = 10.9 + 1.11 \text{ Arach(A)} + 0.372 \text{ Endotox}$$

Predictor	Coef	StDev	T	P
Constant	10.854	5.708	1.90	0.066
Arach(A)	1.1140	0.1550	7.19	0.000
Endotox	0.37159	0.06496	5.72	0.000

S = 15.35 R-Sq = 71.9% R-Sq(adj) = 70.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	19866.0	9933.0	42.18	0.000
Residual Error	33	7770.6	235.5		
Total	35	27636.5			

Source	DF	Seq SS
Arach(A)	1	12160.9
Endotox	1	7705.0

Unusual Observations

Obs	Arach(A)	Prosta(P)	Fit	StDev Fit	Residual	St Resid
7	10.0	81.10	40.57	3.88	40.53	2.73R
25	50.0	102.90	66.55	4.96	36.35	2.50R

R denotes an observation with a large standardized residual

Conclusions from the output

The regression equation to describe the data is as follows:

$$\text{Prosta} = 10.9 + 1.11 \mu\text{M Arachi.} + 0.372 \text{ Endo.}$$

This analysis reveals that prostacyclin production depends on the concentration of endotoxin. The t-ratio for endotoxin is highly significant. The conclusion to be drawn is that endotoxin increases the production of prostacyclin. Since prostacyclin production requires the presence of a key enzyme, the researchers concluded that endotoxin stimulates endothelial cells to increase the activity of this enzyme.

Example of explanatory variables which are categorical

How does multiple regression analysis handle explanatory variables which are categorical? Let us assume that there are two explanatory variables x_1 and x_2 , and that x_2 is categorical, and that we have a regression equation

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

The problem of categorical variables is solved by having two values for the category and assigning value ‘1’ to one group and ‘0’ to the other. So for the category where $x_2=1$, we have

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

and for the category where $x_2 = 0$ we have

$$Y = \beta_0 + \beta_1 X_1$$

The explanatory variable with values equal to ‘1’ and ‘0’ is referred to as **Indicator variable**. In the case of several categories, most computer software allow the creation of ‘dummy’ indicator variables.

The method of handling categorical explanatory variables is illustrated by means of an example below:

Leucine Metabolism in Human Newborns

(Denne SC, Kalhan SC. Am.J.Physiol.1987;253:E608-E615)

Infants grow faster than adults, and therefore produce more skeletal muscle. It may be hypothesized that newborns have a different protein metabolism than adults. To test the hypothesis leucine turnover as a marker for protein metabolism was measured in newborns and adults. Leucine is an important amino acid which is known to regulate protein metabolism by stimulating protein synthesis and retarding protein degradation. Because metabolism depends on body size, the researchers analyzed leucine metabolism as a function of body weight in newborns and adults. The data are given below:

Log leu.	Log wt.	Nbrn (0) Adlt (1)
2.54	0.44	0
2.57	0.42	0
2.59	0.46	0
2.73	0.46	0
2.74	0.51	0
2.81	0.52	0
2.81	0.57	0
2.86	0.58	0
2.93	0.61	0
3.61	1.73	1

Log leu.	Log wt.	Nbrn (0) Adlt (1)
3.61	1.77	1
3.64	1.81	1
3.65	1.7	1
3.66	1.76	1
3.75	1.75	1
3.75	1.87	1
3.84	1.92	1
3.88	1.97	1
3.89	1.92	1
3.93	1.86	1

In this data file Log leucine is the numerical variable and the subject is categorised as Nbrn/Adlt coded as 0/1. The regression equation may be written as

$$\text{Log wt} = \beta_0 + \beta_1 \times \text{Log leu.} + \beta_2 \times \text{Nbrn/Adlt}$$

In the case of the newborns, this works out as

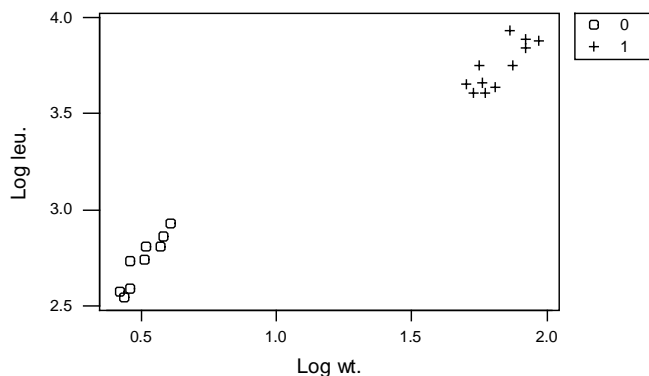
$$\text{Log wt} = \beta_0 + \beta_1 \times \text{Log leu.} + 0,$$

and for adults it is

$$\begin{aligned} \text{Log wt} &= \beta_0 + \beta_1 \times \text{Log leu.} + \beta_2 \times 1 \\ &= (\beta_0 + \beta_2) + \beta_1 \times \text{Log leu.} \end{aligned}$$

The scatter plot is provided in Fig. 3.3

Fig. 3.3 Scatter plot of log leucine against log wt.



A linear relationship is shown between log leucine and log wt. for both newborn and adult; adult (=1) is at the top R-hand corner and infants (=0) at the bottom L-hand corner.

Regression lines if fitted would represent the respective regression equations, each with a slope equal to β_1 , and the distance between the two slopes equal to β_2 , which in this case is 0.761.

The outcome of the regression analysis is given below:

[In MINITAB → Stat → Regression. In Response Box "Log leu." In Predictors Box "Log wt; Nbrn Adlt]

The regression equation is

$$\text{Log leu.} = 2.05 + 1.35 \text{ Log wt.} - 0.761 \text{ Nbrn(0)a} + \text{adl(1)}$$

Predictor	Coef	StDev	T	P
Constant	2.0459	0.1077	18.99	0.000
Log wt.	1.3495	0.2070	6.52	0.000
Nbrn(0)a	-0.7605	0.2743	-2.77	0.013

S = 0.07070 R-Sq = 98.4% R-Sq(adj) = 98.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	5.3145	2.6573	531.62	0.000
Residual Error	17	0.0850	0.0050		
Total	19	5.3995			

Source	DF	Seq SS
Log wt.	1	5.2761
Nbrn(0)a	1	0.0384

Unusual Observations

Obs	Log wt.	Log leu.	Fit	StDev Fit	Residual	St Resid
20	1.86	3.9300	3.7954	0.0226	0.1346	2.01R

R denotes an observation with a large standardized residual

Models.

In both the examples above, the relationship between explanatory variables and outcome is described by a mathematical model viz. the regression equation which relates the explanatory variables denoted by Xs with the outcome denoted by Y. The most commonly used models are known as ‘linear models’. All that it means is that the X variables combine in a linear fashion to predict Y. Thus an equation of the form $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$ summarises in mathematical terms that X_1 and X_2 are predictors of Y where α , β_1 , and β_2 are constants (numbers). Regression is the method of estimating the parameters and so these are referred to as regression parameters. Linear models are appropriate when the outcome variable is normally distributed. However, the method can be generalised to include situations where the outcome variable is discrete numeric.

When we have taken a sample and estimated the parameters of the model we can relate them to the population from which the sample was taken. It is, however, important to bear in mind that models can only be an approximation to reality.

The usefulness of scatter plots

In all the three examples described so far scatter plots were obtained prior to data analysis in order to display the relationship between response and explanatory variables. This is a useful first step, and often reveals associations between variables which would be otherwise missed if one were to plunge into regression analysis straight away. This is illustrated in the next example.

Water hardness and Mortality Rates

In an investigation of environmental causes of disease data were obtained on annual mortality rates per 100 000 for males, averaged over the years 1958 – 1964, and the calcium

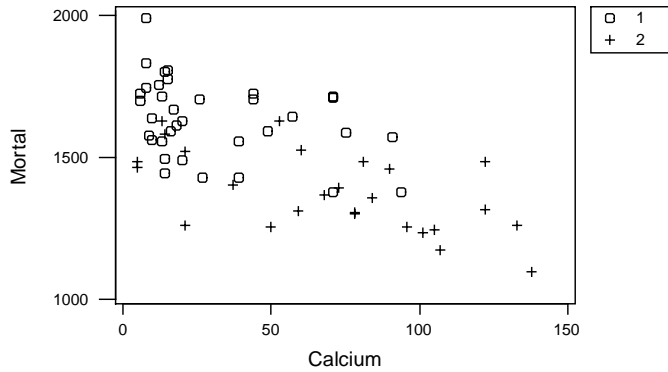
concentration (in parts per million) in the water supply for 61 large towns in England and Wales. The data are shown in the table below.

Mortal	Calcium	N/S
1247	105	2
1668	17	1
1466	5	2
1800	14	1
1609	18	1
1558	10	1
1807	15	1
1299	78	2
1637	10	1
1359	84	2
1392	73	2
1755	12	1
1307	78	2
1254	96	2
1491	20	1
1555	39	1
1428	39	1
1318	122	2
1260	21	2
1723	44	1
1379	94	1
1742	8	1
1574	9	1
1569	91	1
1096	138	2
1591	16	1
1402	37	2
1772	15	1
1828	8	1
1704	26	1
1702	44	1
1309	59	2
1259	133	2
1427	27	1
1724	6	1
1175	107	2
1486	5	2
1456	90	2
1696	6	1
1236	101	2
1711	13	1
1444	14	1
1591	49	1
1987	8	1
1495	14	1
1369	68	2
1257	50	2
1587	75	1
1713	71	1
1557	13	1
1640	57	1
1709	71	1
1625	20	1
1527	60	2
1627	53	2
1486	122	2
1485	81	2
1378	71	1
1519	21	2
1581	14	2
1625	13	2

Indicator variables need to be created in place of column 3 (N/S) such that in column 4 (not shown above) North = 1 South = 0.

We now proceed with scatter plot and regression analysis in the usual manner.

Fig. 3.4 Scatter plot of mortality rates by hardness of water



There appears to be a linear relationship on the scatter plot, and the regression analysis gives the following result:

[In MINITAB → Stat → Regression. In Response Box “Mortal”. In Predictors Box “Calcium; N/S]

The regression equation is
 Mortal = 1872 - 2.03 Calcium - 177 N/S

Predictor	Coef	StDev	T	P
Constant	1872.15	47.92	39.07	0.000
Calcium	-2.0341	0.4829	-4.21	0.000
N/S	-176.71	36.89	-4.79	0.000

S = 122.1 R-Sq = 59.1% R-Sq(adj) = 57.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1248318	624159	41.86	0.000
Residual Error	58	864856	14911		
Total	60	2113174			

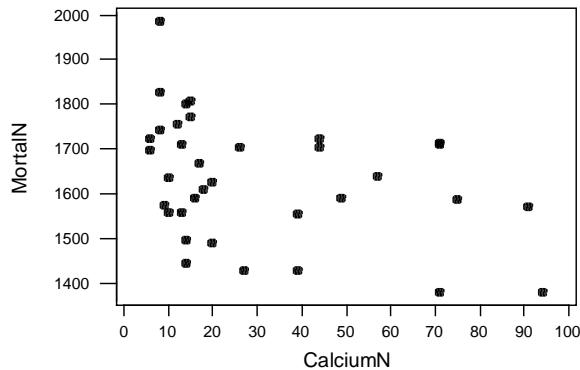
Source	DF	Seq SS
Calcium	1	906185
N/S	1	342132

Unusual Observations

Obs	Calcium	Mortal	Fit	StDev Fit	Residual	St Resid
44	8	1987.0	1679.2	23.3	307.8	2.57R

R denotes an observation with a large standardized residual

A closer look at Fig. 3.4 makes one wonder whether this is the full story. Towns in the North are all clustered in the top right hand corner, and towns in the South show a wide scatter. We can look at each group separately by ‘unstacking’ the data. (All computer software provide facility for doing so).

Fig. 3.5 Scatter plot of mortality rates by hardness of water in towns in the North

The result of regression analysis is as below:

(In MINITAB → Stat → Regression. In Response Box "MortalN". In Predictors Box "CalciumN")

The regression equation is
 $MortalN = 1692 - 1.93 \text{ CalciumN}$

Predictor	Coef	StDev	T	P
Constant	1692.31	33.78	50.09	0.000
CalciumN	-1.9313	0.8479	-2.28	0.029

S = 129.2 R-Sq = 13.6% R-Sq(adj) = 11.0%

Analysis of Variance

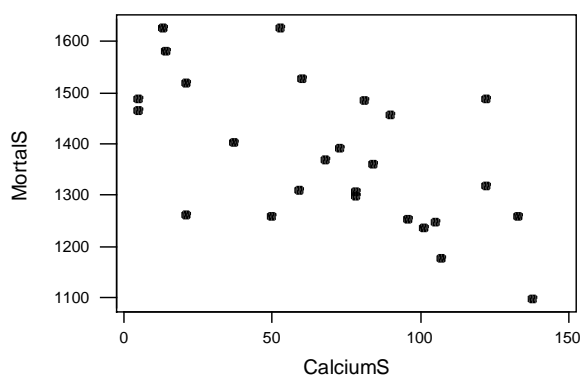
Source	DF	SS	MS	F	P
Regression	1	86621	86621	5.19	0.029
Residual Error	33	550937	16695		
Total	34	637558			

Unusual Observations

Obs	CalciumN	MortalN	Fit	StDev Fit	Residual	St Resid
12	94.0	1379.0	1510.8	58.2	-131.8	-1.14 X
15	91.0	1569.0	1516.6	55.8	52.4	0.45 X
27	8.0	1987.0	1676.9	28.9	310.1	2.46R

R denotes an observation with a large standardized residual
 X denotes an observation whose X value gives it large influence.

The scatter plot for towns of the South and results of the regression analysis next follow:

Fig. 3.6 Scatter plot of mortality rates by hardness of water for towns in the South
Regression Analysis

(In MINITAB → Stat → Regression. In Response Box “MortalS”. In Predictors Box “CalciumS”.)

The regression equation is
 Mortals = 1523 - 2.09 CalciumS

Predictor	Coef	StDev	T	P
Constant	1522.82	45.43	33.52	0.000
CalciumS	-2.0927	0.5664	-3.69	0.001

S = 114.3 R-Sq = 36.3% R-Sq(adj) = 33.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	178352	178352	13.65	0.001
Residual Error	24	313534	13064		
Total	25	491886			

Unusual Observations

Obs	CalciumS	MortalS	Fit	StDev Fit	Residual	St Resid
9	21	1260.0	1478.9	35.6	-218.9	-2.01R
22	122	1486.0	1267.5	37.1	218.5	2.02R

R denotes an observation with a large standardized residual

The first regression analysis gives us a measure of the trend viz.
 (Mortal = 1872 – 2.03 Calcium – 177 N/S) The ‘t’ ratios and corresponding significant P values give the evidence of a slope. The R-sq. value tells us that 59% of the variation in mortality can be accounted for by the hardness of water.

Regression analysis for towns in the North gives the equation
 MortalN = 1692 – 193CalciumN. The t ratio is still significant indicating a slope and so also is the F statistic indicating a linear relationship between mortality and hardness of water. But the R-sq value is now 13.6%.

Regression analysis for towns in the South gives a different picture. The ‘t’ ratio and F statistic are still significant, but the R-sq value is 36.3%.

The overall conclusion is that even though there is evidence of a relationship between hardness of water and mortality rates, the Southern towns have generally lower mortality rates and water hardness is much more compared to Northern towns. There may be other factors in addition to water hardness, which require further investigation.