

4. Multiple Regression in Practice

The preceding chapters have helped define the broad principles on which regression analysis is based. What features one should look for in the computer output and their interpretation has been discussed. Several details still need to be addressed. As in previous chapters, we would do so by beginning with an example.

Crime Rate in 47 states of the USA with 13 explanatory variables

The crime rate in 47 states in the U.S. was reported together with possible factors that might influence it. The factors recorded are as follows;

X1 = the number of males aged 14 - 24 per 1000 of total state population

X2 = binary variable distinguishing southern states (X2 =1) from the rest (X2=0).

X3 = the mean number of years of schooling x 10 of the population, 25 years old and over

X4 = police expenditure (in dollars) per person by state and local government in 1960

X5 = police expenditure (in dollars) per person by state and local government in 1959

X6 = labour force participation rate per 1000 civilian urban males in the age group 14-24

X7 = the number of males per 1000 females

X8 = state population size in hundred thousands

X9 = the number of non-whites per 1000

X10 = unemployment rate of urban males per 1000 in the age group 14-24

X11 = unemployment rate of urban males per 1000 in the age group 35-59.

X12 = the median value of family income or transferable goods and assets (unit 10 dollars)

X13 = the number of families per 1000 earning below one-half of the median income.

The data file is given on the pages that follow:

X1	X2	X3	X4	X5	X6	X7	X8
MalePop	Sth-Rest	YrsSchl	expend60	Expend59	Labrfrce	M/F	Popnsiz
151	1	91	58	56	510	950	33
143	0	113	103	95	583	1012	13
142	1	89	45	44	533	969	18
136	0	121	149	141	577	994	157
141	0	121	109	101	591	985	18
121	0	110	118	115	547	964	25
127	1	111	82	79	519	982	4
131	1	109	115	109	542	969	50
157	1	90	65	62	553	955	39
140	0	118	71	68	632	1029	7
124	0	105	121	116	580	966	101
134	0	108	75	71	595	972	47
128	0	113	67	60	624	972	28
135	0	117	62	61	595	986	22
152	1	87	57	53	530	986	30
142	1	88	81	77	497	956	33
143	0	110	66	63	537	977	10
135	1	104	123	115	537	978	31
130	0	116	128	128	536	934	51
125	0	108	113	105	567	985	78
126	0	108	74	67	602	984	34
157	1	89	47	44	512	962	22
132	0	96	87	83	564	953	43
131	0	116	78	73	574	1038	7
130	0	116	63	57	641	984	14
131	0	121	160	143	631	1071	3
135	0	109	69	71	540	965	6
152	0	112	82	76	571	1018	10
119	0	107	166	157	521	938	168
166	1	89	58	54	521	973	46
140	0	93	55	54	535	1045	6
125	0	109	90	81	586	964	97
147	1	104	63	64	560	972	23
126	0	118	97	97	542	990	18
123	0	102	97	87	526	948	113
150	0	100	109	98	531	964	9
177	1	87	58	56	638	974	24
133	0	104	51	47	599	1024	7
149	1	88	61	54	515	953	36
145	1	104	82	74	560	981	96
148	0	122	72	66	601	998	9
141	0	109	56	54	523	968	4
162	1	99	75	70	522	996	40
136	0	121	95	96	574	1012	29
139	1	88	46	41	480	968	19
126	0	104	106	97	599	989	40
130	0	121	90	91	623	1049	3

X9	X10	X11	X12	X13	Y
Non-White	Unempl4	Unemp35	Incm	Poverty	Crimrate
301	108	41	394	261	79.1
102	96	36	557	194	163.5
219	94	33	318	250	57.8
80	102	39	673	167	196.9
30	91	20	578	174	123.4
44	84	29	689	126	68.2
139	97	38	620	168	96.3
179	79	35	472	206	155.5
286	81	28	421	239	85.6
15	100	24	526	174	70.5
106	77	35	657	170	167.4
59	83	31	580	172	84.9
10	77	25	507	206	51.1
46	77	27	529	190	66.4
72	92	43	405	264	79.8
321	116	47	427	247	94.6
6	114	35	487	166	53.9
170	89	34	631	165	92.9
24	78	34	627	135	75
94	130	58	626	166	122.5
12	102	33	557	195	74.2
423	97	34	288	276	43.9
92	83	32	513	227	121.6
36	142	42	540	176	96.8
26	70	21	486	196	52.3
77	102	41	674	152	199.3
4	80	22	564	139	34.2
79	103	28	537	215	121.6
89	92	36	637	154	104.3
254	72	26	396	237	69.6
20	135	40	453	200	37.3
82	105	43	617	163	75.4
95	76	24	462	233	107.2
21	102	35	589	166	92.3
76	124	50	572	158	65.3
24	87	38	559	153	127.2
349	76	28	382	254	83.1
40	99	27	425	225	56.6
165	86	35	395	251	82.6
126	88	31	488	228	115.1
19	84	20	590	144	88
2	107	37	489	170	54.2
208	73	27	496	224	82.3
36	111	37	622	162	103
49	135	53	457	249	45.5
24	78	25	593	171	50.8
22	113	40	588	160	84.9

The results of the regression analysis using all 13 explanatory variables are given below:

The regression equation is

$$\begin{aligned} \text{CrmRate} = & - 692 + 1.04 \text{ PopMale} - 8.3 \text{ StateSth} + 1.80 \text{ School} + 1.61 \text{ Expens60} \\ & - 0.67 \text{ Expend59} - 0.041 \text{ LabrFrce} + 0.165 \text{ NmbrMen} - 0.041 \text{ Popn} \\ & + 0.0072 \text{ NonWhite} - 0.602 \text{ Unemplo} + 1.79 \text{ Unemp35+} + 0.137 \text{ Incom} \\ & + 0.793 \text{ Povrty} \end{aligned}$$

Predictor	Coef	StDev	T	P
Constant	-691.8	155.9	-4.44	0.000
PopMale	1.0398	0.4227	2.46	0.019
StateSth	-8.31	14.91	-0.56	0.581
School	1.8016	0.6497	2.77	0.009
Expens60	1.608	1.059	1.52	0.138
Expend59	-0.667	1.149	-0.58	0.565
LabrFrce	-0.0410	0.1535	-0.27	0.791
NmbrMen	0.1648	0.2099	0.78	0.438
Popn	-0.0413	0.1295	-0.32	0.752
NonWhite	0.00717	0.06387	0.11	0.911
Unemplo	-0.6017	0.4372	-1.38	0.178
Unemp35+	1.7923	0.8561	2.09	0.044
Incom	0.1374	0.1058	1.30	0.203
Povrty	0.7929	0.2351	3.37	0.002

S = 21.94 R-Sq = 76.9% R-Sq(adj) = 67.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	13	52930.6	4071.6	8.46	0.000
Residual Error	33	15878.7	481.2		
Total	46	68809.3			

Source	DF	Seq SS
PopMale	1	550.8
StateSth	1	153.7
School	1	9056.7
Expens60	1	30760.3
Expend59	1	1530.2
LabrFrce	1	611.3
NmbrMen	1	1110.0
Popn	1	426.5
NonWhite	1	142.0
Unemplo	1	70.7
Unemp35+	1	2696.6
Incom	1	347.5
Povrty	1	5474.2

Unusual Observations

Obs	PopMale	CrmRate	Fit	StDev Fit	Residual	St Resid
11	124	167.40	116.84	11.24	50.56	2.68R

R denotes an observation with a large standardized residual

These 13 explanatory variables account for almost 77% of the variation in crime rate in the 47 states of the United States for which data were available. The β coefficients of all the variables are relatively small, except for whether the state was in the South or elsewhere. Southern states have rates on average less by a factor of 8.3 compared to other states.

However, a discrepancy is noticeable on studying the regression equation with regard to expenditure on police services between the years 1959 and 1960. Why should police expenditure in one year be associated with increase in crime rate and decrease in the previous year? It does not make sense. Secondly, even though the F statistic is highly significant, and which provides evidence for the presence of a linear relationship between all 13 variables and the response variable, the β coefficients of both expenditures for 1959 and 1960 have non-significant t ratios. Non-significant t means there is no slope! In other words police expenditure has no effect whatsoever on crime rate!

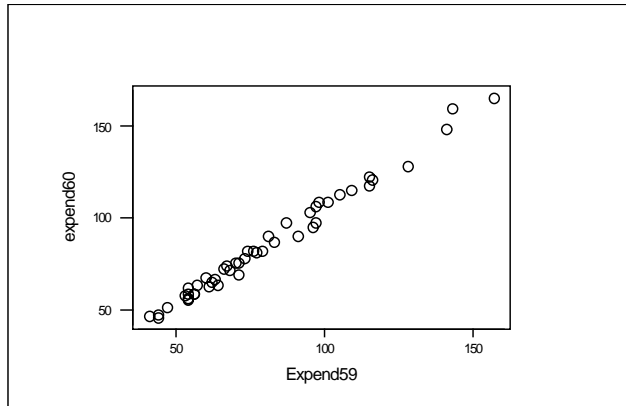
The discrepancy is explained by the next analysis which is a correlation matrix between all the 14 variables

Correlations (Pearson)

	PopMale	State-st	School	Expen60	Expen59	LabrForc	M/F	Popn
State-st	0.584							
School	-0.530	-0.703						
Expen60	-0.506	-0.373	0.483					
Expen59	-0.513	-0.376	0.499	0.994				
LabrForc	-0.161	-0.505	0.561	0.121	0.106			
m/f	-0.029	-0.315	0.437	0.034	0.023	0.514		
Popn	-0.281	-0.050	-0.017	0.526	0.514	-0.124	-0.411	
NonWhit	0.593	0.767	-0.665	-0.214	-0.219	-0.341	-0.327	0.095
Unemp14	-0.224	-0.172	0.018	-0.044	-0.052	-0.229	0.352	-0.038
Unemp35	-0.245	0.072	-0.216	0.185	0.169	-0.421	-0.019	0.270
Inc	-0.670	-0.637	0.736	0.787	0.794	0.295	0.180	0.308
Pvrty	0.639	0.737	-0.769	-0.631	-0.648	-0.270	-0.167	-0.126
Crime	-0.089	-0.091	0.323	0.688	0.667	0.189	0.214	0.337
	NonWhit	Unemp14	Unemp35	Inc.	Povrty			
Unemp14	-0.156							
Unemp35	0.081	0.746						
Inc	-0.590	0.045	0.092					
Pvrty	0.677	-0.064	0.016	-0.884				
Crime	0.033	-0.050	0.177	0.441	-0.179			

The correlation matrix above shows the correlation coefficient (Pearson's r) for each of the explanatory variable with every other variable including the response variable - crime rate. Some variables are more closely correlated with crime rate than others, but expenditure on Police Services in 1959 and in 1960 show the highest correlation coefficients for crime rates at 0.667 and 0.668 respectively. However, they are also closely correlated with each other at 0.994. When explanatory variables are closely correlated with each other this type of discrepancy arises. This is because β coefficients are partial correlation coefficients, and each β describes the relationship between the corresponding X and the response Y **taking into account the other X variables**. In other words, each of 13 β coefficients tells us about its respective variable's contribution after allowing for the contributions of other explanatory variables. Therefore, if a variable is highly correlated with another it will have no additional contribution to make over and above the contribution of the other. In this particular data set, X_4 and X_5 are closely correlated. Each tells only part of the story and we cannot reliably estimate their effects. Unreliable parameter estimate means large standard error and lower t statistic. How closely the two variables are correlated can be judged from Fig. 4.1

Fig. 4.1 Scatter plot of Expend 60 against expend 59



If we were to now perform a regression analysis excluding police expenditure in 1959 we get the following results:

Regression Analysis excluding X5.

$$\begin{aligned} \text{Crimrate} = & - 704 + 1.06 \text{ MalePop} - 7.9 \text{ Sth-Rest} + 1.72 \text{ YrsSchl} + 1.01 \text{ expend60} \\ & - 0.017 \text{ Labrfrce} + 0.163 \text{ m/f} - 0.039 \text{ Popnsiz} - 0.0001 \text{ Non-White} \\ & - 0.585 \text{ Unemp14} + 1.82 \text{ Unemp35} + 0.135 \text{ Incm} + 0.804 \text{ Pvrty} \end{aligned}$$

Predictor	Coef	Stdev	t-ratio	p
Constant	-704.1	152.9	-4.60	0.000
MalePop	1.0645	0.4165	2.56	0.015
Sth-Rest	-7.87	14.75	-0.53	0.597
YrsSchl	1.7216	0.6286	2.74	0.010
expend60	1.0097	0.2436	4.15	0.000
Labrfrce	-0.0172	0.1464	-0.12	0.907
m/f	0.1630	0.2079	0.78	0.438
Popnsiz	-0.0389	0.1282	-0.30	0.764
Non-White	-0.00013	0.06200	-0.00	0.998
Unemp14	-0.5848	0.4319	-1.35	0.185
Unemp35	1.8192	0.8465	2.15	0.039
Incm	0.1351	0.1047	1.29	0.206
Pvrty	0.8040	0.2320	3.47	0.001

s = 21.72 R-sq = 76.7% R-sq(adj) = 68.5%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	12	52768.2	4397.4	9.32	0.000
Error	34	16041.0	471.8		
Total	46	68809.3			

SOURCE	DF	SEQ SS
MalePop	1	550.8
Sth-Rest	1	153.7
YrsSchl	1	9056.7
expend60	1	30760.3
Labrfrce	1	1207.0
m/f	1	1381.5
Popnsiz	1	528.8
Non-White	1	72.7
Unemp14	1	155.1
Unemp35	1	2806.6
Incm	1	429.4
Pvrty	1	5665.5

Unusual Observations

Obs.	MalePop	Crimrate	Fit	Stdev.Fit	Residual	St.Resid
11	124	167.40	118.47	10.77	48.93	2.59R
19	130	75.00	113.76	11.09	-38.76	-2.08R

R denotes an obs. with a large st. resid.

In this analysis expenditure in 1960 acquires a significant t ratio proving that the discrepancy noted earlier was due to **multicollinearity** between the two X variables. The topic of multicollinearity is discussed further in Chapter 5.

Finding the most economical model

Regression Analysis with several explanatory variables tells us about the effect of each individual variable in the model taking into account the effects of other variables included in the model. The magnitude, significance, and interpretation of any one β coefficient depend on what other variables are included. In observational studies like the present one about crime rates in U.S. states it is always difficult to ensure that all the X variables of importance have been considered. The tendency is to throw the net wide, and include all possible influences on the outcome variable. However, when planning interventions, like measures to reduce crime, it becomes necessary to focus on the most significant variables in order to obtain the maximum benefit with the most economical use of resources. Secondly, if the regression equation is a model to represent the data, then it stands to reason that one would have a better grasp of the situation by focusing onto the main sources of influence on the response variable. By using just enough predictor variables to capture the main features of the model one is also able to improve prediction. These observations are illustrated by means of the next regression analysis, which takes into account only 5 out of the 13 explanatory variables.

Regression analysis with 5 explanatory variables

$$\text{CrmRate} = -524 + 1.02 \text{ PopMale} + 2.03 \text{ School} + 1.23 \text{ Expens60} + 0.914 \text{ Unemp35+} + 0.635 \text{ Povrty}$$

Predictor	Coef	StDev	T	P
Constant	-524.37	95.12	-5.51	0.000
PopMale	1.0198	0.3532	2.89	0.006
School	2.0308	0.4742	4.28	0.000
Expens60	1.2331	0.1416	8.71	0.000
Unemp35+	0.9136	0.4341	2.10	0.041
Povrty	0.6349	0.1468	4.32	0.000

S = 21.30 R-Sq = 73.0% R-Sq(adj) = 69.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	50206	10041	22.13	0.000
Residual Error	41	18604	454		
Total	46	68809			

Source	DF	Seq SS
PopMale	1	551
School	1	7260
Expens60	1	31739
Unemp35+	1	2174
Povrty	1	8483

Unusual Observations

Obs	PopMale	CrmRate	Fit	StDev Fit	Residual	St Resid
11	124	167.40	104.44	6.74	62.96	3.12R
19	130	75.00	118.39	6.22	-43.39	-2.13R
29	119	104.30	149.64	11.02	-45.34	-2.49R

R denotes an observation with a large standardized residual

Eight variables have been removed from the original 13 at a cost of reducing the R-sq value from 76.9% to 73%.

How does one determine the most economical model?

One method is by considering the analysis of variance table of the full model remaining after excluding those variables which are causes of any multicollinearity. In the present data set the variable X5 (Police Expenditure in 1959) was excluded leaving 12 explanatory variables. In the resulting regression analysis one examines the table of **t-ratio** and the corresponding **p** value of significance. Those variables which have a significant t-ratio are selected for inclusion into the sub-set. (Recall that a significant t-ratio for a β coefficient means that a slope is present). Next the Analysis of Variance table is examined in the column headed **SEQ SS**. The amount in each row of this column corresponding to the variable tells us the contribution made by the respective explanatory variable to the Regression Sum of Squares. One verifies that the explanatory variables selected to form the sub-set are indeed making a sizable contribution to the Regression Sum of Squares. (Recall that for any explanatory variable the larger the Regression Sum of Squares the smaller is the Residual Sum of Squares, since $TSS = ESS + RSS$. A small Residual Sum of Squares means that the data points are not so scattered, and that there is greater clustering about the line of least squares).

An alternative strategy is based on adding or dropping one variable at a time from a given model. The idea is to compare the current model with a new model obtained by adding or deleting an explanatory variable from the current model. Call the smaller model (i.e. with fewer variables) Model I and the bigger model as Model II. One can compute the *F* statistic (called partial *F*) by $RSS \text{ Model I} - RSS \text{ Model II} \div (RSS \text{ Model II} / \text{degree of freedom of Model II})$. If partial *F* value is bigger than that of *F* the smaller model is better.

Once the important β coefficients have been identified the next step is to determine their relative roles in explaining the variability in the outcome variable Y. The actual numeric value of the β coefficients is not a guide. For example, the β coefficient of StateSth is 7.2 and that of Popn is 0.324. This does not mean that StateSth is 21 times more important than Popn. This is because of the unit of measurement. Popn is measured in actual numbers and StateSth is categorical. The *t* statistic provides the true measure. The *t* statistic for StateSth is 0.39 and that for Popn is 2.10.

In the regression analysis with 5 explanatory variables the sub-set was chosen in the manner just described. For comparison the regression analysis resulting from the 7 remaining explanatory variables is now presented.

The regression equation is

$$\text{CrmRate} = - 615 + 7.2 \text{ StateSth} - 0.045 \text{ LabrFrce} + 0.630 \text{ NmbrMen} + 0.324 \text{ Popn} + 0.142 \text{ NonWhite} - 0.362 \text{ Unemplo} + 0.224 \text{ Incom}$$

Predictor	Coef	StDev	T	P
Constant	-614.9	188.8	-3.26	0.002
StateSth	7.16	18.45	0.39	0.700
LabrFrce	-0.0455	0.1863	-0.24	0.808
NmbrMen	0.6300	0.2534	2.49	0.017
Popn	0.3238	0.1545	2.10	0.043
NonWhite	0.14219	0.07508	1.89	0.066
Unemplo	-0.3620	0.3492	-1.04	0.306
Incom	0.22365	0.07084	3.16	0.003

S = 30.86 R-Sq = 46.0% R-Sq(adj) = 36.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	7	31673.2	4524.7	4.75	0.001
Residual Error	39	37136.1	952.2		

```

Total                46      68809.3

Source      DF      Seq SS
StateSth    1        565.3
LabrFrce    1       1891.3
NmbrMen     1       1292.6
Popn        1      15155.3
NonWhite    1         713.2
Unemplo     1       2564.5
Incom       1       9490.9

Unusual Observations
Obs  StateSth  CrmRate      Fit  StDev Fit  Residual  St Resid
 36     0.00   127.20   68.14   11.01     59.06     2.05R

R denotes an observation with a large standardized residual

```

The R-sq. value using these 7 explanatory variables is only 46%.

Stepwise Regression Analysis

Another method of selecting a sub-set from a list of explanatory variables is Stepwise Regression Analysis. Most computer software have a facility for carrying out **stepwise** forward selection and backward elimination of variables based on their overall contribution to the F statistic. Significance of F can be set by the individual or left to the default value of 4. Stepwise regression removes and adds variables to the regression model for the purpose of identifying a useful subset of the predictors. Stepwise first finds the explanatory variable with the highest correlation (R^2) to start with. It then tries each of the remaining explanatory variables until it finds the two with the highest R^2 . Then it tries all of them again until it finds the three variables with the highest R^2 , and so on. The overall R^2 gets larger as more variables are added.

In resorting to the **stepwise** procedure the following should be bone in mind:

- 1). By chance about 5% of sample relationships can be expected to be significant, and one can wrongly single them out as being truly significant.
- 2). When there are strong relationships between several explanatory variables, stepwise procedures tend to exclude one or more of these variables. One may then mistakenly conclude that the excluded variable is unimportant.
- 3). If two variables X_1 and X_2 are positively related but their effects on Y have opposite signs, or X_1 and X_2 are negatively related but their effects on Y have the same signs, stepwise procedures may exclude one or both. We then understate the importance of both variables.
- 4). Stepwise regression tends to choose only those observations which have no missing values for any of the variables. So the final model obtained by the Stepwise procedure will have fewer subjects if the data set has missing values. Hence the model selected by Stepwise may not fit a larger subsequent data set well.
- 5). The P values are to be discounted because they do not take into account the very many tests that have been carried out.
- 6). Different methods of Stepwise like 'forward selection' and 'backward elimination' are likely to produce different models. Also the same model seldom emerges when a second data set is analysed.

In conclusion, Stepwise regression is useful in the early exploratory phase of analysis, but not to be relied on for the confirmatory stage. Once the exploratory stage has identified the significant explanatory variables they can then be further assessed.

Stepwise Regression on the Crime data file gave the following output:
 [In MINITAB → Stat | Regression | Stepwise]

Stepwise Regression

F-to-Enter: 4.00 F-to-Remove: 4.00

Response is Crimrate on 13 predictors, with N = 47

Step	1	2	3	4	5
Constant	14.45	-94.47	-327.54	-424.92	-524.37
expend60	0.89	1.24	1.24	1.30	1.23
T-Ratio	6.35	7.58	8.41	9.03	8.71
Pvrty		0.41	0.75	0.64	0.63
T-Ratio		3.36	4.98	4.20	4.32
YrsSchl			1.58	1.66	2.03
T-Ratio			3.31	3.63	4.28
MalePop				0.76	1.02
T-Ratio				2.21	2.89
Unemp35					0.91
T-Ratio					2.10
S	28.4	25.6	23.1	22.2	21.3
R-Sq	47.28	58.03	66.56	70.04	72.96

It is of interest to note that:

- At each step as a further variable is added to the regression equation the β coefficients of X variables already in the equation change in value, and so also does the t ratio which is in fact equal to β coefficient \div its standard deviation.
- At each step as a further X variable is added, improvement takes place in R-sq value. Stepwise stops when no further improvement in R-sq can occur by the addition or deletion of a variable.

Another procedure is for obtaining the **best subset**, also provided by most software.

[In Minitab → Stat | Regression | Best subsets]

The **best** procedure gave the following output

Best Subsets Regression

Response is Crimrate

Vars	R-sq	Adj. R-sq	C-p	s	S e E L P N
1	47.3	46.1	32.4	28.393	M t Y x x a o o U U
1	44.5	43.2	36.4	29.144	a h r p p b p n n n
					l - s e e r n - e e P
					e R S n n f s W m m I v
					P e c d d r m i h p p n r
					o s h 6 5 c / z t 1 3 c t
					p t l 0 9 e f e e 4 5 m y

2	58.0	56.1	19.0	25.619		X				X
2	56.2	54.3	21.6	26.159	X	X				
3	66.6	64.2	8.8	23.131		X	X			X
3	64.6	62.2	11.6	23.788		X		X		X
4	70.0	67.2	5.8	22.154	X	X	X			X
4	68.3	65.2	8.4	22.802		X	X	X		X
5	73.0	69.7	3.7	21.301	X	X	X		X	X
5	72.4	69.1	4.4	21.516	X	X	X			X
6	74.8	71.0	3.1	20.827	X	X	X		X	X
6	73.9	70.0	4.3	21.189	X	X	X		X	X
7	75.4	71.0	4.2	20.843	X	X	X		X	X
7	75.3	70.8	4.4	20.885	X	X	X	X		X
8	76.4	71.4	4.8	20.680	X	X	X	X		X
8	75.8	70.7	5.6	20.928	X	X	X	X		X
9	76.6	70.9	6.4	20.847	X	X	X	X	X	X
9	76.6	70.9	6.5	20.859	X	X	X	X	X	X
10	76.8	70.3	8.2	21.066	X	X	X	X	X	X
10	76.7	70.2	8.3	21.110	X	X	X	X	X	X
11	76.9	69.6	10.1	21.323	X	X	X	X	X	X
11	76.8	69.6	10.1	21.336	X	X	X	X	X	X
12	76.9	68.8	12.0	21.615	X	X	X	X	X	X
12	76.9	68.7	12.1	21.634	X	X	X	X	X	X
13	76.9	67.8	14.0	21.936	X	X	X	X	X	X

Notice the variation in the R-sq value, which reaches its maximum at step 10. The crosses on the Right hand side of the output indicate which variables have been included.

So to summarise, the best set of X variables in a data set can be identified by:

1. The investigator's understanding of the topic being researched, the research question, and whether it is an observational or intervention study. In most intervention studies the investigator has already decided which explanatory variables to include. In observational studies the situation is fluid, and one is likely to throw the net wide and include as many variables as appear to have a likely bearing on the response variable. This was indeed the case with the "crime" data set.
2. A first round of selection of candidate variables by studying the results of analysing the full model after the exclusion of variables causing multicollinearity, if any.
3. Using the 'stepwise' or 'best' procedure. However, one must exercise caution when going by these automatic procedures. They do not consider the practical importance of any of the explanatory variables. It would be advisable to do subset selection by means of steps 1-3, and then verify by an automatic procedure.

Why should one look for a sub-set amongst explanatory variables?

The regression equation is often referred to as **the model**, because it helps us to understand a situation by identifying the main sources of influence on the response variable. In that case it is best to use just enough explanatory variables to capture the main features of the data set. This is the **principle of parsimony**. If two models explain the data about equally well, choose the simpler one.

In our discussions about whether a model was adequate or not we have so far referred to the value of R-sq. In the case of the data set on Crime in the United States we found that a 13 variables model gave R-sq. value of 76.9%; a 12 variables model gave the value of 76.7%; a 5 variables model gave the value as 73%; step-wise regression chose a model with 5 variables to give an R-sq. value of 72.96%, and so on. There are several other criteria for judging the best model. These are now described.

Criteria for judging the appropriateness of a model

1. R-sq. the coefficient of determination

The problem with R-sq. is that its value always increases as new variables are added to the model even though they are not contributing significant independent information. Hence by itself R-sq is not a good measure. But increase in R-sq. when another variable is added can provide useful insight into how much additional predictive information is gained by adding this variable. R-sq. is most useful when choosing between two models with the same number of explanatory variables.

2. Adjusted R-sq.

Although similar to R-sq. it takes into account the number of coefficients ' p ' in the model and the number of subjects ' n '. (R-sq. = Regression Sum of Squares ÷ Total Sum of Squares).
Adjusted R-sq = $1 - (n - 1 / n - p - 1) \times (1 - R\text{-sq})$. The formula involves ' p ', the number of explanatory variables in the model which makes comparison between Adjusted R-sq. with different ' p ' meaningful because Adjusted R-sq. is not automatically greater for larger ' p '.

R-sq. tends to overestimate the amount of variability accounted for, so that if one were to apply the regression equation derived from one sample to another independent sample one would almost always get a smaller R-sq value in the new sample compared to the original.

3. Standard error of the estimate

This is also called the standard deviation of Y about the regression line, and is denoted as s in the computer output. One could take the model with the smallest value of s as the best model.

4. C-p statistic.

In general, among candidate models we select the one with the smallest C-p value, and where the value of C-p is closest to p , the number of parameters in the model (i.e. intercept plus the number of explanatory variables in the model).

$C\text{-}p = \text{Residual Sum of Squares}_p \div \text{Mean Residual Sum of Squares}_m - (n - 2p)$, where
Residual Sum of Squares _{p} = RSS for the model with p parameters including the intercept and
Mean Residual Sum of Squares _{m} = Mean RSS with all the predictors.

C-p is commonly used to select a subset of explanatory variables after an initial regression analysis employing all possible explanatory variables. One then selects the smallest model that produces a C-p value near to p , the number of parameters. Small C-p means small variance in estimating the regression coefficients. In other words a precise model is achieved and adding more explanatory variables is unlikely to improve the precision any more. Simple models are always preferable because they are easy to interpret and would be less prone to multicollinearity.

In the example of the 'best' subset regression analysis provided step 6 with C-p value of 3.1 has been highlighted as giving the best subset. The corresponding R-sq. value is 74.8%, adjusted R-sq. value is 71%, and $s = 20.827$.

Caution needs to be exercised when resorting to automatic selection procedures like 'Stepwise' and 'Best'. These procedures are machine led and do not take into account the practical importance or biological plausibility of the predictors.

Finally, before accepting the results of a regression analysis it is advisable to always apply regression diagnostics. That is the subject of the next chapter. But before proceeding to the next topic let us summarise below what has been learnt so far about multiple linear regression.

Commentary on multiple linear regression

Multiple linear regression is an extension of bivariate regression in which several explanatory variables instead of one are combined to obtain a value on a response variable. The goal of regression is to arrive at a set of β values called regression coefficients for the explanatory variables. Using the β values one is able to calculate a predicted Y value for any given subject in the sample such that it would be as close as possible to the Y value obtained by measurement.

As a statistical tool regression can be used to answer a number of questions related to the data set. We next consider these questions:

1. Does the regression equation really provide better than chance prediction?

Multiple regression is used to find out how well a set of variables predict a particular outcome. Amongst a set of candidate variables which ones are best predictors of an outcome, and especially whether a particular predictor variable can still predict the outcome after controlling for the influence of another variable. The F ratio and its level of significance provide the answer. This may not necessarily be the case with step-wise regression because all the candidate explanatory variables do not enter the equation. Some textbooks provide tables that show how large R-sq. must be to be considered statistically significant for a given sample size, the number of candidate explanatory variables and the level specified for F to enter.

2. Out of a set of explanatory variables in a regression equation how can one select the more important ones?

When the explanatory variables are uncorrelated with each other, the assessment of the contribution that each makes is straightforward. Explanatory variables with bigger t-ratios are usually more influential than those with smaller ones.

If the explanatory variables are correlated with each other assessment of the importance of each of them becomes ambiguous.

3. Having obtained a regression equation how can one judge whether adding one or more explanatory variables would improve prediction of the response?

Improvement in R-sq. and particularly adjusted R-sq. provides a useful guide. The addition of new explanatory variables can affect the relative contributions of those variables already in the equation. In the selection of additional variables the research question and the theoretical rationale behind it must guide the researcher. To leave the selection entirely to the software would be a mistake.

4. How can one find out what influence a particular explanatory variable has in the context of another or a set of explanatory variables?

The investigator can enter the explanatory variables in sequence according to theoretical or logical considerations. For example, variables that are considered more influential are entered first, either one at a time or in blocks. At each step the computer output provides the relevant information.

5. How good is the regression equation at predicting the response in a new sample of subjects on whom only the data for explanatory variables are available?

The question of generalisability is at the heart of all research. The primary requirement is the accuracy of data. If accuracy has been ascertained, a procedure called cross validation may be employed. In this procedure a regression equation is developed from a portion of the sample, and then applied to the other portion. If it works then it is very likely that the regression equation will predict the response variable better than chance for new cases.

Another factor influencing generalisability is the size of the sample. Obviously with small samples similar results cannot be expected when applied widely.

One should always bear in mind that regression functions do not involve any assumptions of time, order or causal relations. Regression coefficients and quantities derived from them represent measures of associations, not measures of effects.

6. Does sample size have any bearing on the number of explanatory variables that can be included in developing the regression equation?

A simple rule of thumb is $N \geq 104 + \text{number of explanatory variables}$. For example, if there are 5 explanatory variables one should have at least 109 cases. This assumes a medium sized relationship between the response and explanatory variables, a significance level of 0.05 and a power of 80. A higher number of cases would be needed if the response variable is skewed, a small effect size is anticipated, or there is a high likelihood of measurement error.

7. How does multiple regression contribute to answering a research question?

- Most of the time investigators are looking for an input/output relationship. The explanatory variables comprising the input are being related to the output in a way as to enable them to measure the contribution of each of them. The relationship is often complicated by other factors that may be related to both the input and the output. Such confounding factors are often taken care of in the regression.
- The regression equation provides a way of predicting the outcome.
- It is possible to assess the simultaneous effects of a number of explanatory variables on the outcome.