Variable Selection Methods

SESSION 14

Structure

14.1 INTRODUCTION

In Lab Sessions 11, 12 and 13, you have learnt how to fit linear regression models for one and more than one predictors and the related inferential analysis. In real life, we may come across a long list of regressors available in the data which may be related to the response variable. In this situation, it is difficult to choose some important regressors, which can really contribute to the model. The basic screening may be done on the basis of prior experience, but some standard methods known as **variable selection methods** are also used to decide the appropriate regressors to be entered in the model.

Prerequisite

- Lab Session 12 of MSTL-002 (Industrial Statistics Lab).
- Unit 12 of MSTE-002 (Industrial Statistics-II).

You have learnt in Unit 12 of MSTE-002 (Industrial Statistics-II) that three methods are commonly used for variable selection: (i) forward selection, (ii) backward elimination and (iii) stepwise regression. These methods of variable selection are available in most statistical software packages. In this lab session, our aim is to train you in obtaining the best regression model using these three methods of variable selection in Excel 2007.

Objectives

After performing the activities of this session, you should be able to:

- prepare the spreadsheet for the variable selection methods with MS Excel 2007;
- test the significance of the regression parameters;
- choose the appropriate best fitted regressors for the regression model; and
- interpret the results of the forward selection, backward elimination and stepwise methods.

14.2 PROBLEM DESCRIPTION

For applying the variable selection methods, we consider the data of a juice manufacturing company. In Lab Session 12, we fitted a multiple regression model for the monthly sales as a linear function of advertisement cost and price. Here

we have also considered two other variables: temperature and number of stores. The data are recorded in Table 1.

Table 1: Data of a juice manufacturing company

S. No.	Monthly Sales	Advertisement Cost	Price	Temperature	Number of Stores
1	212000	3240	85	24	14
2	230000	3550	94	28	15
3	273000	3890	77	33	16
4	255000	3720	84	31	16
5	285000	4030	75	35	17
6	262000	3790	81	32	16
7	273000	3950	76	34	16
8	215000	3290	89	24	15
9	185000	2920	79	21	13
10	268000	3850	78	33	16
11	285000	4030	75	35	17
12	230000	3550	94	28	15
13	212000	3240	85	24	14
14	309000	4200	68	39	17
15	228000	3470	94	27	15
16	215000	3290	89	24	15
17	324000	4340	64	42	18
18	210000	3330	94	24	14
19	262000	3790	81	32	16
20	329000	4400	62	44	18
21	313000	4240	67	40	17
22	243000	3660	88	30	16
23	255000	3720	84	31	16
24	293000	4080	73	36	17
25	216000	3380	94	25	15
26	195000	2990	83	22	14
27	208000	3140	82	24	14
28	268000	3850	78	33	16
29	173000	2860	78	20	13
30	235000	3600	92	29	15
31	216000	3380	94	25	15
32	300000	4130	71	37	17
33	235000	3600	92	29	15
34	210000	3330	94	24	14
35	273000	3890	77	33	16
36	293000	4080	73	36	17
37	243000	3660	88	30	16
38	273000	3950	76	34	16
39	228000	3470	94	27	15
40	201000	3060	83	23	14

We can also decide on different values for α_{IN} and α_{OUT} as these show the level of significance for entering and eliminating a regressor, respectively. For example, we can use $\alpha_{IN} = 0.05$ and $\alpha_{OUT} = 0.03$.

We now build appropriate linear regression models for the monthly sales using (i) forward selection, (ii) backward elimination and (iii) stepwise approaches to determine the most appropriate regressors at 5% level of significance, i.e., $\alpha_{IN} = \alpha_{OUT} = 0.05$.

14.3 FORWARD SELECTION METHOD

You have already studied the forward selection method in Unit 12 of MSTE-002. Therefore, before describing how to apply the forward selection procedure in Excel, we provide a brief overview of the steps involved in this method. First, we assume a model without any regressor. Then we enter one regressor at each step on the basis of partial F-test into the model. We stop when no more regressors can be significantly entered in the model. In this way, we get a **final model**. This process consists of the following steps:

If we add an important regressor (X) to the model, it increases the regression sum of square and decreases the residual sum of squares, which result in increasing the F-value. On the other hand, adding an unimportant (unnecessary) regressor may decrease the regression sum of squares and increase the residual sum of squares, which results in decreasing the F value.

Step 1: Before starting the procedure for forward selection, we decide the level of significance to enter the regressor in the model, i.e., **Alfa-to-In** (α_{IN}) to compare the p-values calculated at each stage of selection.

Step 2: If Y is a dependent variable, we start by assuming a model without any regressor, i.e.,

$$Y = B_0 + e \qquad \ldots(1)$$

where B_0 is the intercept and e is a normally distributed random error component with mean zero and variance σ^2.

Step 3: We perform the regression analysis for all possible one-regressor models and decide the entering variable in the model. Then we identify the regressor having the largest absolute t value of the coefficient (or partial F value) and also calculate its p-value. If its p-value is significant, the corresponding regressor will enter in the model. We can also find the first variable to be entered in the model by calculating the correlation coefficient (r) for each pair of the dependent and independent variables (X, Y). We check the significance of the model for the pair having the highest correlation coefficient.

Step 4: We now perform the regression analysis for all possible two-regressor models containing one regressor determined from Step 3 and the second from the remaining regressors. We compute the p-value for the coefficients of the remaining regressors and include the most significant regressor in the model. In this way, we get 2 regressors in the model.

For example, if we add the regressor x_q in the model, which already contains the regressor x_p, the extra sum of squares measures the increase in the regression sum of squares due to the addition of the one more regressor. For example, the extra sum of squares due to x_q is

$$SS_{Reg}\left(b_q / b_p\right) = SS_{Reg}(b_p, b_q) - SS_{Reg}\left(b_p\right)$$

We can also define partial F-statistic to check the effect of x_q after eliminating the effect of x_p as follows:

$$F\left(b_q / b_p\right) = \frac{SS_{Reg}\left(b_q / b_p\right)}{SS_{Res}} \sim F_{(\alpha),\,[1,(n-p)]}$$

Step 5: Similarly, we carry out the regression analysis for all possible three-regressor models containing two regressors determined from Steps 3 and 4 and one from the remaining regressors. We select the most significant regressor from the remaining regressors.

For example, if we add one more regressor x_r in the model which already contains the regressors x_p and x_q then the extra sum of squares due to x_r is

$$SS_{Reg}\left(b_r / b_p, b_q\right) = SS_{Reg}\left(b_p, b_q, b_r\right) - SS_{Reg}\left(b_p, b_q\right)$$

and $$F\left(b_r / b_p, b_q\right) = \frac{SS_{Reg}\left(b_r / b_p, b_q\right)/2}{SS_{Res}} \sim F_{(\alpha),[2,(n-p)]}$$

Step 6: We repeat the procedure till all regressors are included in the model or no additional regressor gives a significant p-value. In the same manner, we can compute the extra sum of squares for every addition of a regressor.

Steps in Excel

Before starting the procedure, the first thing we need to do is to set a significance level for deciding a regressor to be entered into the model. We call this significance level **Alpha-to-In** and denote it as α_{IN}. In this session, we are considering $\alpha_{IN} = 0.05$. If the p-value is less than or equal to α_{IN}, we consider the corresponding regressor as significant, otherwise it is insignificant.

Note that the p-value associated with partial F-statistic is the same as the p-value associated with the t-statistic of the corresponding regressor. For the data given in Sec. 14.2, we denote the monthly sales, advertisement cost, price, temperature and number of stores by Y, X_1, X_2, X_3 and X_4, respectively.

The steps involved in the forward selection method to select the appropriate regressors in the model using Excel 2007 are given below:

This partial F test is equivalent to the t-test because it involves the testing of only the single regressor. We know that for single variable

$$F_{cal} = t^2_{cal}$$

Step 1: We enter the given data in Excel 2007 spreadsheet as shown in Fig. 14.1.

	A	B	C	D	E	F
1	**S.No.**	**Y**	**X1**	**X2**	**X3**	**X4**
2	1	212000	3240	85	24	14
3	2	230000	3550	94	28	15
4	3	273000	3890	77	33	16
5	4	255000	3720	84	31	16
6	5	285000	4030	75	35	17
7	6	262000	3790	81	32	16
8	7	273000	3950	76	34	16
9	8	215000	3290	89	24	15
10	9	185000	2920	79	21	13
11	10	268000	3850	78	33	16
12	11	285000	4030	75	35	17
13	12	230000	3550	94	28	15
14	13	212000	3240	85	24	14
15	14	309000	4200	68	39	17
16	15	228000	3470	94	27	15
17	16	215000	3290	89	24	15
18	17	324000	4340	64	42	18
19	18	210000	3330	94	24	14
20	19	262000	3790	81	32	16
21	20	329000	4400	62	44	18

Forward Selection

Fig. 14.1: Partial screenshot of the spreadsheet for the given data.

Step 2: To select the first regressor in the model

We perform the regression analysis for each regressor by considering only one regressor at a time as discussed in Step 4 of Sec. 11.4 in Lab Session 11. In variable selection methods, we only need the p-value related to the regression coefficients. So we specify only the following options as shown in Fig. 14.2:

- ✓ ***Input Y Range***.
- ✓ ***Input X Range***.
- ✓ Tick the ***Labels***.
- ✓ Specify the ***Output Range***.

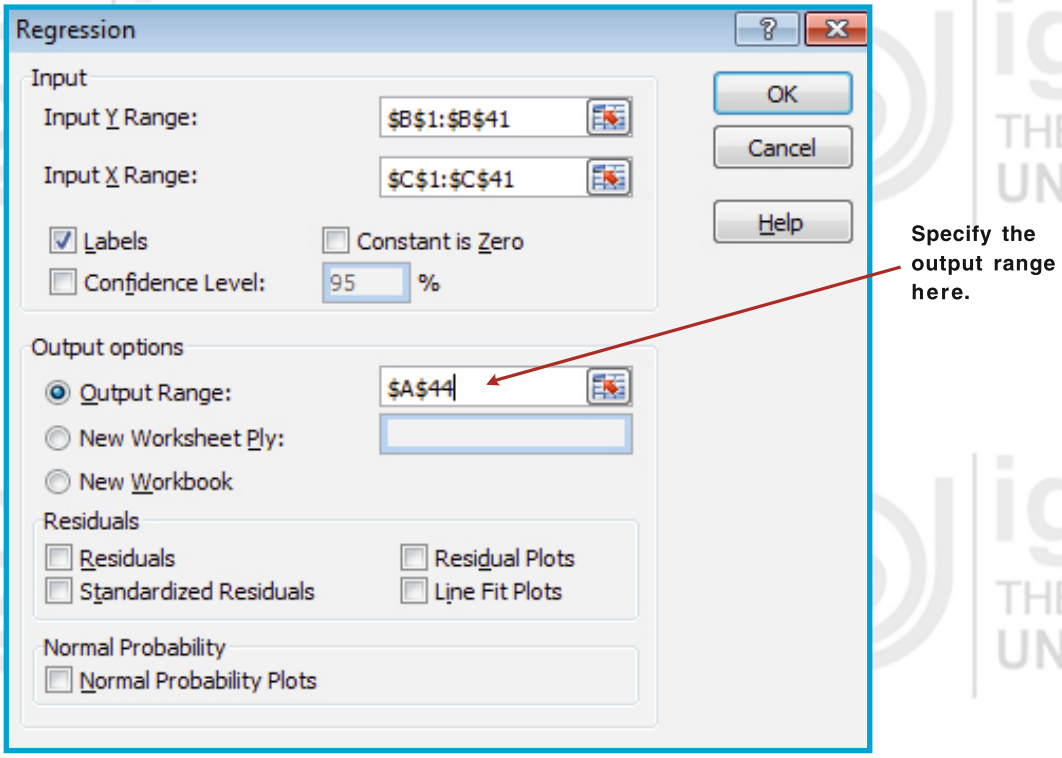

Fig. 14.2

We obtain the results after clicking ***OK***. We perform the regression analysis for X_1, X_2, X_3 and X_4 one at a time as given below:

For X_1: We consider **Y range** as Cells B1:B41, **X range** as Cells C1:C41 and **output range** as Cell A44 as shown in Fig. 14.2. The result is shown in Fig. 14.3.

	A	B	C	D	E	F	G
53	ANOVA						
54		*df*	*SS*	*MS*	*F*	*Significance F*	
55	Regression	1.0000	60035464734.5936	60035464734.5936	1783.2638	0.0000	
56	Residual	38.0000	1279310265.4064	33666059.6160			
57	Total	39.0000	61314775000.0000				
58							
59		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
60	Intercept	-109571.1015	8524.6940	-12.8534	0.0000	-126828.4421	-92313.7609
61	X1	98.0940	2.3229	42.2287	0.0000	93.3915	102.7966
62							

Fig. 14.3

For X_2 : We consider **Y range** as Cells B1:B41, **X range** as Cells D1:D41 and **output range** as Cell K44. The result is shown in Fig. 14.4.

	K	L	M	N	O	P	Q
53	ANOVA						
54		*df*	*SS*	*MS*	*F*	*Significance F*	
55	Regression	1.0000	34697762892.9791	34697762892.9791	49.5366	0.0000	
56	Residual	38.0000	26617012107.0209	700447687.0269			
57	Total	39.0000	61314775000.0000				
58							
59		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
60	Intercept	516401.6045	38317.8738	13.4768	0.0000	438831.1251	593972.0840
61	X2	-3264.2509	463.7892	-7.0382	0.0000	-4203.1431	-2325.3587
62							

Fig. 14.4

For X_3: We consider **Y range** as Cells B1:B41, **X range** as Cells E1:E41 and **output range** as Cell U44. The result is shown in Fig. 14.5.

	U	V	W	X	Y	Z	AA
53	ANOVA						
54		*df*	*SS*	*MS*	*F*	*Significance F*	
55	Regression	1.0000	60594399255.6855	60594399255.6855	3196.3697	0.0000	
56	Residual	38.0000	720375744.3145	18957256.4293			
57	Total	39.0000	61314775000.0000				
58							
59		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
60	Intercept	50197.4317	3571.4011	14.0554	0.0000	42967.5082	57427.3551
61	X3	6593.2635	116.6197	56.5364	0.0000	6357.1793	6829.3478
62							

Fig. 14.5

For X_4: We consider **Y range** as Cells B1:B41, **X range** as Cells F1:F41 and **output range** as Cell AE44. The result is given in Fig. 14.6.

	AE	AF	AG	AH	AI	AJ	AK
53	ANOVA						
54		*df*	*SS*	*MS*	*F*	*Significance F*	
55	Regression	1.0000	57283024294.0702	57283024294.0702	539.9032	0.0000	
56	Residual	38.0000	4031750705.9298	106098702.7876			
57	Total	39.0000	61314775000.0000				
58							
59		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
60	Intercept	-223668.8181	20378.3862	-10.9758	0.0000	-264922.7039	-182414.9322
61	X4	30402.1783	1308.4188	23.2358	0.0000	27753.4230	33050.9336
62							

Fig. 14.6

- The first regressor to be entered in the model is temperature (X_3).
- The lowest p-value of t test = 0.0000.
- From Fig. 14.5, the model is:

  $\hat{Y} = 50197.432 + 6593.264X_3$
- F test p-value = 0.0000.

Interpretation

We now compare Figs. 14.3 to 14.6 and note from Fig. 14.5 that the highest absolute value of t-statistic is 56.5364 given in Cell X61 (Fig. 14.5), which is associated with regressor X_3. The p-value of t-statistic associated with the X_3 is 0.0000 given in Cell Y61, which is less than $\alpha_{IN} = 0.05$. We can conclude that the regressor X_3 is contributing significantly to the model. We can include the regressor X_3 (temperature) into the model.

Step 3: To select the second regressor in the model

The variable X_3 is already entered in the model as shown in Step 2. We now repeat Step 2 for all possible 2 regressors in the models that include X_3 as one of the regressors, i.e., (X_3, X_1),

(X_3, X_2) and (X_3, X_4) to find the second regressor to be entered in the model. As we have already explained in Lab Session 12, we need to put independent variables in the adjacent columns in Excel 2007 to run regression analysis for more than one independent variable. So we copy and paste the values of (X_3, X_1), (X_3, X_2) and (X_3, X_4) in Cells H1:I41, K1:L41, and N1:O41, respectively, as shown in Fig. 14.7.

	H	I	J	K	L	M	N	O
1	**X3**	**X1**		**X3**	**X2**		**X3**	**X4**
2	24	3240		24	85		24	14
3	28	3550		28	94		28	15
4	33	3890		33	77		33	16
5	31	3720		31	84		31	16
6	35	4030		35	75		35	17
7	32	3790		32	81		32	16
8	34	3950		34	76		34	16
9	24	3290		24	89		24	15
10	21	2920		21	79		21	13
11	33	3850		33	78		33	16
12	35	4030		35	75		35	17

Forward Selection

Fig. 14.7

To find the second variable to be entered in the model, we perform the regression analysis for these combinations of the two regressors in the model one at a time as discussed in Step 2.

For X_3 and X_1: We consider **Y range** as Cells B1:B41, **X range** as Cells H1:I41 and **output range** as Cell A64. The result is shown in Fig. 14.8.

	A	B	C	D	E	F	G
73	ANOVA						
74		*df*	*SS*	*MS*	*F*	*Significance F*	
75	Regression	2.0000	60773651603.4472	30386825801.7236	2077.7378	0.0000	
76	Residual	37.0000	541123396.5528	14624956.6636			
77	Total	39.0000	61314775000.0000				
78							
79		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
80	Intercept	-4221.1064	15857.3263	-0.2662	0.7916	-36351.1012	27908.8883
81	X3	4436.6174	624.4765	7.1045	0.0000	3171.3078	5701.9270
82	X1	32.6780	9.3341	3.5009	0.0012	13.7654	51.5906
83							

Fig. 14.8

For X_3 and X_2: We consider **Y range** as Cells B1:B41, **X range** as Cells K1:L41 and **output range** as Cell K64. The result is shown in Fig. 14.9.

	K	L	M	N	O	P	Q
73	ANOVA						
74		*df*	*SS*	*MS*	*F*	*Significance F*	
75	Regression	2.0000	60598074331.3543	30299037165.6772	1564.2017	0.0000	
76	Residual	37.0000	716700668.6457	19370288.3418			
77	Total	39.0000	61314775000.0000				
78							
79		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
80	Intercept	56137.9637	14108.0130	3.9792	0.0003	27552.4142	84723.5132
81	X3	6534.7594	178.7085	36.5666	0.0000	6172.6616	6896.8572
82	X2	-50.9283	116.9213	-0.4356	0.6657	-287.8334	185.9769
83							

Fig. 14.9

For X_3 and X_4 : We consider **Y range** as Cells B1:B41 **X range** as Cells N1:O41 and **output range** as Cell U64. The result is shown in Fig. 14.10.

	U	V	W	X	Y	Z	AA
73	ANOVA						
74		*df*	*SS*	*MS*	*F*	*Significance F*	
75	Regression	2.0000	60722895691.0313	30361447845.5156	1897.9774	0.0000	
76	Residual	37.0000	591879308.9688	15996738.0802			
77	Total	39.0000	61314775000.0000				
78							
79		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
80	Intercept	2036.6194	17306.5584	0.1177	0.9070	-33029.7985	37103.0373
81	X3	5562.0325	379.2959	14.6641	0.0000	4793.5060	6330.5589
82	X4	5098.1839	1798.8116	2.8342	0.0074	1453.4455	8742.9223
83							

Fig. 14.10

Interpretation

We now study Figs. 14.8, 14.9 and 14.10 and notice that the highest absolute value of t-statistic is 3.5009 given in Cell D82 (Fig. 14.8), which is associated with regressor X_1. The p-value of t-statistic associated with the regressor X_1 is 0.0012 given in Cell E82, which is less than $\alpha_{IN} = 0.05$. Thus, the regressor X_1 is contributing significantly to the model. We now conclude that the regressor X_1 (advertising cost) enters in the model as a second regressor.

- The second regressor to be entered in the model is the advertising cost (X_1).
- The lowest p-value of t test = 0.0012.
- From Fig. 14.8, the model is:

  $\hat{y} = -4221.106 + 4436.617X_3 + 32.678X_1$
- F test p-value = 0.0000.

Step 4: To select the third regressor in the model

After we select two regressors in the model, we consider all possible models with 3 regressors, keeping in mind that X_3 and X_1 are already entered in the model. The possible combinations of the regressors will be (X_3, X_1, X_2) and (X_3, X_1, X_4). For performing the regression analysis in Excel, we keep the values of (X_3, X_1, X_2) and (X_3, X_1, X_4) in adjacent columns, i.e., Cells Q1:S41 and U1:W41, respectively, as shown in Fig. 14.11.

	Q	R	S	T	U	V	W
1	**X3**	**X1**	**X2**		**X3**	**X1**	**X4**
2	24	3240	85		24	3240	14
3	28	3550	94		28	3550	15
4	33	3890	77		33	3890	16
5	31	3720	84		31	3720	16
6	35	4030	75		35	4030	17
7	32	3790	81		32	3790	16
8	34	3950	76		34	3950	16
9	24	3290	89		24	3290	15
10	21	2920	79		21	2920	13
11	33	3850	78		33	3850	16
12	35	4030	75		35	4030	17
13	28	3550	94		28	3550	15
14	24	3240	85		24	3240	14
15	39	4200	68		39	4200	17
16	27	3470	94		27	3470	15
17	24	3290	89		24	3290	15
18	42	4340	64		42	4340	18
19	24	3330	94		24	3330	14
20	32	3790	81		32	3790	16
21	44	4400	62		44	4400	18
22	40	4240	67		40	4240	17
23	30	3660	88		30	3660	16
24	31	3720	84		31	3720	16
25	36	4080	73		36	4080	17
26	25	3380	94		25	3380	15

Forward Selection

Fig. 14.11

To discover the third regressor to be entered in the model, we execute the regression analysis for combinations of the three regressors in the model one at a time as discussed in Steps 2 and 3.

For X_3, X_1 and X_2: We consider **Y range** as Cells B1:B41, **X range** as Cells Q1:S41 and **output range** as Cell A85. The result is shown in Fig. 14.12.

	A	B	C	D	E	F	G
94	ANOVA						
95		*df*	*SS*	*MS*	*F*	*Significance F*	
96	Regression	3.0000	61009129636.2336	20336376545.4112	2395.2909	0.0000	
97	Residual	36.0000	305645363.7664	8490148.9935			
98	Total	39.0000	61314775000.0000				
99							
100		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
101	Intercept	2150.5738	12142.4667	0.1771	0.8604	-22475.4899	26776.6376
102	X3	1445.8446	740.8706	1.9515	0.0588	-56.7106	2948.3998
103	X1	68.2135	9.8034	6.9581	0.0000	48.3312	88.0958
104	X2	-561.9502	106.7039	-5.2664	0.0000	-778.3557	-345.5446
105							

Fig. 14.12

For X_3, X_1 and X_4: We consider **Y range** as Cells B1:B41, **X range** as Cells U1:W41 and **output range** as Cell K85. The result is shown in Fig. 14.13.

	K	L	M	N	O	P	Q
94	ANOVA						
95		*df*	*SS*	*MS*	*F*	*Significance F*	
96	Regression	3.0000	60809972541.5603	20269990847.1868	1445.5549	0.0000	
97	Residual	36.0000	504802458.4397	14022290.5122			
98	Total	39.0000	61314775000.0000				
99							
100		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
101	Intercept	-20674.8643	18590.6180	-1.1121	0.2735	-58378.3847	17028.6562
102	X3	4304.9125	616.9260	6.9780	0.0000	3053.7285	5556.0965
103	X1	25.4067	10.1954	2.4920	0.0174	4.7294	46.0839
104	X4	3023.5790	1878.6777	1.6094	0.1163	-786.5559	6833.7138
105							

Fig. 14.13

Interpretation

It is the same as for Step 3. We find the regressor having the highest absolute value of t-statistic. It is 5.2664 given in Cell D104 corresponding to the regressor X_2 (Fig. 14.12) and its p-value is 0.0000 given in Cell E104, which is less than $\alpha_{IN} = 0.05$. Hence, the regressor X_2 is contributing significantly to the model. We conclude that the regressor X_2 (price) enters in the model as a third regressor.

- The third regressor to be entered in the model is price (X_2).
- The lowest p-value of t test = 0.0000.
- From Fig.14.12, the Model is

  $$\hat{Y} = 2150.574 + 1445.845X_3 + 68.213X_1 - 561.95X_2$$

- F test p-value = 0.0000.

Step 5: To select the fourth regressor in the model

We repeat the previous steps by considering the combination of 4 regressors, i.e., (X_3, X_1, X_2, X_4). Here we have kept these regressors in Cells Y1:AB41 as shown in Fig. 14.14 to keep them in adjacent columns.

	Y	Z	AA	AB
1	**X3**	**X1**	**X2**	**X4**
2	24	3240	85	14
3	28	3550	94	15
4	33	3890	77	16
5	31	3720	84	16
6	35	4030	75	17
7	32	3790	81	16
8	34	3950	76	16
9	24	3290	89	15
10	21	2920	79	13
11	33	3850	78	16
12	35	4030	75	17
13	28	3550	94	15
14	24	3240	85	14
15	39	4200	68	17
16	27	3470	94	15
17	24	3290	89	15
18	42	4340	64	18
19	24	3330	94	14
20	32	3790	81	16
21	44	4400	62	18
22	40	4240	67	17
23	30	3660	88	16
24	31	3720	84	16
25	36	4080	73	17
26	25	3380	94	15

Forward Selection

Fig. 14.14

For X_3, X_1, X_2 and X_4 : We consider **Y range** as Cells B1:B41, **X range** as Cells Y1:AB41 and **output range** as Cell A107. The result is shown in Fig. 14.15.

	A	B	C	D	E	F	G
16	ANOVA						
17		*df*	*SS*	*MS*	*F*	*Significance F*	
18	Regression	4.0000	61059504597.6716	15264876149.4179	2092.9597	0.0000	
19	Residual	35.0000	255270402.3284	7293440.0665			
20	Total	39.0000	61314775000.0000				
21							
22		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
23	Intercept	-17069.7198	13421.7442	-1.2718	0.2118	-44317.3089	10177.8693
24	X3	1195.1257	693.2697	1.7239	0.0936	-212.2867	2602.5381
25	X1	60.7615	9.5184	6.3836	0.0000	41.4381	80.0850
26	X2	-579.8462	99.1325	-5.8492	0.0000	-781.0958	-378.5966
27	X4	3569.2517	1358.1134	2.6281	0.0127	812.1350	6326.3684
28							

Fig. 14.15

- The fourth regressor to be entered in the model is the number of stores (X_4).
- The lowest p-value of t test = 0.0127.
- From Fig. 14.15, the model is:

  $\hat{y} = -17069.720 + 1195.126X_3 + 60.762X_1 - 579.846X_2 + 3569.252X_4$
- F test p-value = 0.0000.

Interpretation

Fig. 14.15 reveals that the absolute value of t-statistic corresponding to the regressor X_4 is 2.6281 (given in Cell D127) and its p-value is 0.0127 (given in Cell E127), which is less than α_{IN} = 0.05. So we can keep the last regressor X_4 (number of stores) in the model.

In this way, the forward selection method is used to build an appropriate model. The final model is

$$Y = -17069.7198 + 1195.1257X_3 + 60.7615X_1 - 579.8462X_2 + 3569.2517X_4$$

We can also note from Fig. 14.15 that the p-value corresponding to F-statistic is 0.0000 given in Cell F118, which is less than 0.05. We can conclude that the overall model is also significant at 5% level of significance.

14.4 BACKWARD ELIMINATION METHOD

In Sec. 14.3, you have learnt how to choose the important regressors which can really contribute in the model using the forward selection method in Excel 2007. In this section, you will learn the backward elimination method of variable selection. As the name itself suggests, this method goes in the reverse direction of the forward selection method.

As you have already studied the backward elimination method in Unit 12 of MSTE-002, we provide only a brief overview of the steps involved. In this method, we assume a model with all regressors and eliminate (drop) one regressor from the model at each step on the basis of p-values of the t-statistic(s) of the regression coefficients. We stop when no more regressors can be significantly eliminated from the model. In this way, we get a **final model**. This process consists of the following steps:

Step 1: Before starting the procedure for the backward elimination, we decide the level of significance to eliminate the regressor from the model, which is called **Alfa-to-Out** (α_{OUT}) to compare the p-values calculated at each stage of the elimination.

Step 2: Let $X_1, X_2, ..., X_p$ be the p regressors. Here, we start by considering all regressors in the model:

$$\hat{Y} = \hat{B}_0 + \hat{B}_1 X_1 + \hat{B}_2 X_2 + ... + \hat{B}_p X_p \qquad ...(2)$$

Step 3: We perform the regression analysis by considering all regressors in the model. To decide the variable that has to be eliminated from the model, we identify the regressor having the lowest absolute t value of the regression coefficient (or partial F value) and compare its p-value with α_{OUT}. If p-value is insignificant, the corresponding regressor will be eliminated from the model.

Step 4: We now leave the regressor eliminated in Step 3 and carry out the regression analysis for the remaining $(p - 1)$ regressors in the model. We eliminate the most insignificant regressor from the model. We now have $(p - 2)$ regressors in the model.

Step 5: Similarly, we carry out the regression analysis for the remaining $(p - 2)$ regressors in the model excluding the two regressors eliminated in Steps 3 and 4. Then we eliminate the most insignificant regressor from the remaining regressors.

Step 6: We repeat the procedure till no regressor gives insignificant p-value.

Steps in Excel

For eliminating a regressor from the model, the significance level or **Alpha-to-Out** denoted by α_{OUT} is given as 0.05. If the p-value is less than α_{OUT}, we consider the corresponding regressor as significant and retain it. Otherwise, we consider it as insignificant and eliminate that regressor.

Step 1: We enter the given data in a new Excel sheet and name it "**Backward Elimination**" as shown in Fig. 14.16.

	A	B	C	D	E	F
1	**S.No.**	**Y**	**X1**	**X2**	**X3**	**X4**
2	1	212000	3240	85	24	14
3	2	230000	3550	94	28	15
4	3	273000	3890	77	33	16
5	4	255000	3720	84	31	16
6	5	285000	4030	75	35	17
7	6	262000	3790	81	32	16
8	7	273000	3950	76	34	16
9	8	215000	3290	89	24	15
10	9	185000	2920	79	21	13
11	10	268000	3850	78	33	16
12	11	285000	4030	75	35	17
13	12	230000	3550	94	28	15
14	13	212000	3240	85	24	14
15	14	309000	4200	68	39	17
16	15	228000	3470	94	27	15
17	16	215000	3290	89	24	15
18	17	324000	4340	64	42	18
19	18	210000	3330	94	24	14
20	19	262000	3790	81	32	16
21	20	329000	4400	62	44	18

Backward Selection

Fig. 14.16: Partial screenshot of the spreadsheet for the given data.

Step 2: To eliminate the first regressor from the model

We perform the regression analysis by considering all the regressors, i.e., X_1, X_2, X_3 and X_4 in the model as discussed in Sec. 14.3. Here we need only the p-value related to regression coefficients. So we specify only the following options as shown in Fig. 14.17:

- ✓ ***Input Y Range*** as Cells B1:B41.
- ✓ ***Input X Range*** as Cells C1:F41.
- ✓ Tick the ***Labels***.
- ✓ ***Output Range*** as Cell A44.

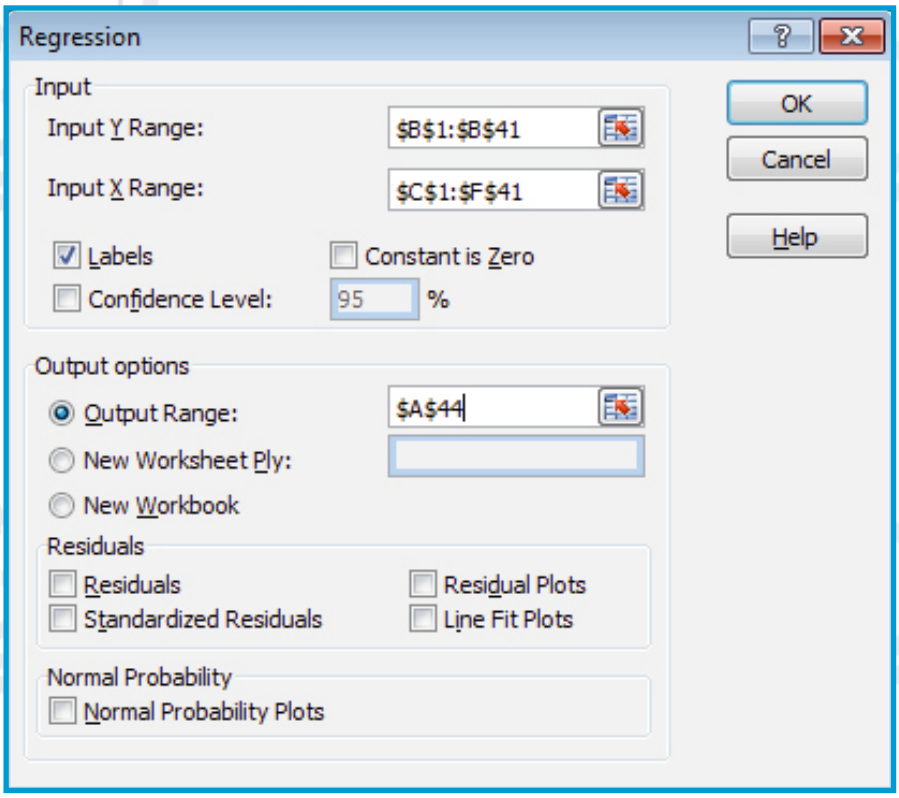

Fig. 14.17

After clicking on ***OK***, we obtain the output shown in Fig. 14.18.

	A	B	C	D	E	F	G
53	ANOVA						
54		*df*	*SS*	*MS*	*F*	*Significance F*	
55	Regression	4.0000	61059504597.6716	15264876149.4179	2092.9597	0.0000	
56	Residual	35.0000	255270402.3284	7293440.0665			
57	Total	39.0000	61314775000.0000				
58							
59		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
60	Intercept	-17069.7198	13421.7442	-1.2718	0.2118	-44317.3089	10177.8693
61	X1	60.7615	9.5184	6.3836	0.0000	41.4381	80.0850
62	X2	-579.8462	99.1325	-5.8492	0.0000	-781.0958	-378.5966
63	X3	1195.1257	693.2697	1.7239	0.0936	-212.2867	2602.5381
64	X4	3569.2517	1358.1134	2.6281	0.0127	812.1350	6326.3684
65							

Fig. 14.18

Interpretation

From Fig. 14.18, the lowest absolute value of t-statistic is 1.7239 given in Cell D63, which is associated with the regressor X_3. The p-value of t-statistic associated with the regressor X_3 is 0.0936 given in Cell E63, which is more than $\alpha_{OUT} = 0.05$. We can conclude that the regressor X_3 is insignificant and it is not contributing to the model. So we can eliminate X_3 (temperature) from the model.

Step 3: To eliminate the second regressor from the model

In Step 2, the regressor X_3 has been eliminated from the model. We now repeat the Step 2 for 3 regressors in the models that exclude X_3, i.e., (X_1, X_2, X_4) to find the second regressor to be eliminated from the model. We have explained in Lab

- We start with the model considering all regressors (Fig. 14.18):

  $\hat{y} = -17069.720 + 60.762X_1 - 579.846X_2 + 1195.1257X_3 + 3569.252X_4$
- The first regressor to be eliminated from the model is the temperature (X_3).
- The highest p-value of t test = 0.0936.
- F test p-value = 0.0000.
- After eliminating X_3, the model is (Fig. 14.21)

  $\hat{y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$

Session 12 that we need to put independent variables in the adjacent columns in Excel 2007 to run the regression analysis in case of more than one independent variables. So we have placed the values of (X_1, X_2, X_4) in Cells H1:J41 as shown in Fig. 14.19.

	H	I	J
1	**X1**	**X2**	**X4**
2	3240	85	14
3	3550	94	15
4	3890	77	16
5	3720	84	16
6	4030	75	17
7	3790	81	16
8	3950	76	16
9	3290	89	15
10	2920	79	13
11	3850	78	16
12	4030	75	17
13	3550	94	15
14	3240	85	14
15	4200	68	17
16	3470	94	15
17	3290	89	15
18	4340	64	18
19	3330	94	14
20	3790	81	16
21	4400	62	18
22	4240	67	17
23	3660	88	16
24	3720	84	16
25	4080	73	17
26	3380	94	15

Backward Selection

Fig. 14.19

To discover the second variable which can be eliminated from the model, we perform the regression analysis for the three regressors in the model as discussed in Step 2 by specifying the following:

- ✓ ***Input Y range*** as Cells B1:B41.
- ✓ ***Input X range*** as Cells H1:J41.
- ✓ Tick the ***Labels***.
- ✓ ***Output Range*** as Cell A67 as shown in Fig. 14.20.

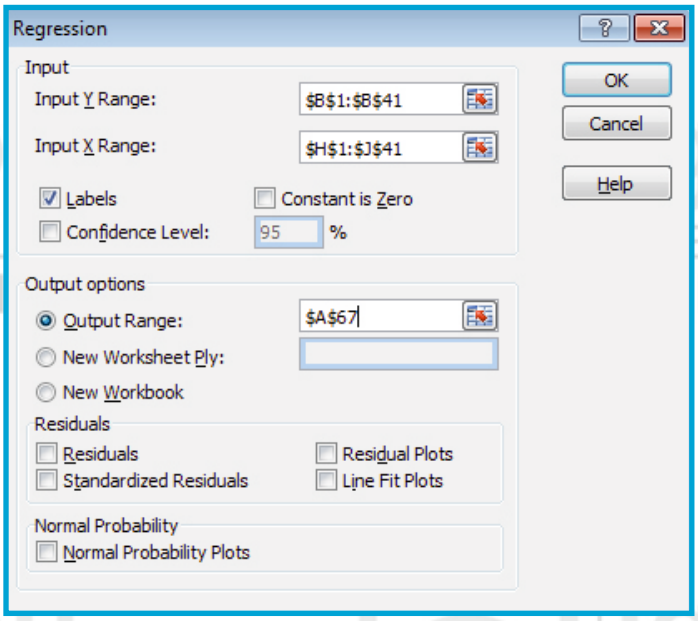

Fig. 14.20

After clicking on ***OK***, we obtain the output shown in Fig. 14.21.

	A	B	C	D	E	F	G
76	ANOVA						
77		*df*	*SS*	*MS*	*F*	*Significance F*	
78	Regression	3.0000	61037829797.5672	20345943265.8557	2644.7613	0.0000	
79	Residual	36.0000	276945202.4328	7692922.2898			
80	Total	39.0000	61314775000.0000				
81							
82		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
83	Intercept	-28820.1147	11874.5560	-2.4270	0.0204	-52902.8303	-4737.3992
84	X1	75.4046	4.4114	17.0931	0.0000	66.4578	84.3513
85	X2	-710.9027	65.3406	-10.8800	0.0000	-843.4195	-578.3859
86	X4	3891.4261	1381.5422	2.8167	0.0078	1089.5287	6693.3236
87							

Fig. 14.21

Interpretation

Fig. 14.21 reveals that the lowest absolute value of t-statistic is 2.8167 given in Cell D86, which is associated with the regressor X_4. The p-value of t-statistic associated with X_4 is 0.0078 given in Cell E86, which is less than $\alpha_{OUT} = 0.05$. So the regressor X_4 is contributing significantly to the model and we cannot eliminate X_4 from the model. Since no regressor is showing insignificant contribution, we terminate the process at this step. We have three regressors, i.e., advertising cost (X_1), price (X_2) and number of stores (X_4) in the model. Hence the resulting model is

$$\hat{Y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$$

Notice from Fig. 14.21 that the value of F-statistic is 2644.7613 given in Cell E78 and its p-value is 0.0000 given in Cell F78, which is less than 0.05. We conclude that the overall model is also significant.

- No more regressor is insignificant.
- The highest p-value of t test = 0.0078.
- From Fig. 14.21, the model is:

  $\hat{y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$
- F test p-value = 0.0000.

14.5 STEPWISE REGRESSION METHOD

You have learnt in Sec. 14.3 that the forward selection method begins without considering any regressor in the model and then we include the most significant variable one at a time, at each step. You have also learnt in Sec.14.4 that backward elimination method is just the reverse of the forward selection method. This procedure starts with all regressors in the model. Then we drop the most insignificant regressor variable one at a time, at each step. In this section, you will learn how to apply stepwise regression in Excel 2007. The stepwise regression method is a combination of both forward selection and backward elimination methods.

In this method, we use hypothesis testing approach as discussed in Secs. 14.3 and 14.4 to check the significance of the regression coefficients at each stage for possible selection and elimination.

In stepwise regression method, we start with forward selection method and also check for the possible elimination of the insignificant variable at each step. You have studied the stepwise regression method in Unit 12 of MSTE-002. We provide only a brief overview of the steps involved as follows:

Step 1: Before starting the procedure for stepwise regression, we decide the levels of significance to enter and eliminate the regressor, i.e., α_{IN} and α_{OUT}, respectively, to compare the p-values.

Step 2: We start by assuming a model with no regressor in it, i.e., we take

$$Y = B_0 + e \qquad ...(3)$$

Step 3: We perform the regression analysis for all possible one-regressor models and decide the variable to be entered in the model. We identify the regressor having the largest absolute t value of the regression coefficient (or partial F value) and also calculate its p-value. If the p-value is significant, the corresponding regressor enters in the model.

Step 4: We perform the regression analysis for all possible two-regressor models containing one regressor obtained from Step 3 and the second from the remaining regressors. We compute the p-value for the regression coefficients of the remaining regressors and include the most significant regressor in the model. Thus, we obtain 2 regressors in the model.

Step 5: Note that before checking the possibility for addition of the next regressor, we also check the significance of the regressor, which was included in the model at Step 3. If both regressors are still significant, we go to Step 7. Otherwise, we go to Step 6.

Step 6: We eliminate the insignificant regressor if its p-value is more than α_{OUT}. We repeat Step 4 with two regressors in the models, taking one regressor which was included at Step 4 and the other regressor from the remaining regressors excluding the regressor which we eliminated in this step.

Step 7: Similarly, we carry out the regression analysis for all possible three-regressor models containing two regressors obtained from Steps 5 or 6 and the third regressor from the remaining regressors. We select the most significant regressor from the remaining regressors. At every step, we check for the possiblility of elimination of the insignificant regressor.

Step 8: We repeat the procedure till all regressors are included in the model or no regressor gives significant p-value to enter into the model and all regressors in the model are significant.

Steps in Excel

We first set a significance level for deciding a regressor to be entered into the model, i.e., α_{IN} (**Alpha-to-In**) and significance level for deciding a regressor to be eliminated from the model, i.e., α_{OUT} (**Alpha-to-Out**). It is given that $\alpha_{IN} = \alpha_{OUT} = 0.05$. If the p-value is less than or equal to α_{IN} or α_{OUT}, we consider the corresponding regressor as significant, otherwise insignificant.

Step 1: We enter the data in an Excel Sheet and name it "**Stepwise Regression**" as shown in Fig. 14.22.

	A	B	C	D	E	F
1	S.No.	Y	X1	X2	X3	X4
2	1	212000	3240	85	24	14
3	2	230000	3550	94	28	15
4	3	273000	3890	77	33	16
5	4	255000	3720	84	31	16
6	5	285000	4030	75	35	17
7	6	262000	3790	81	32	16
8	7	273000	3950	76	34	16
9	8	215000	3290	89	24	15
10	9	185000	2920	79	21	13
11	10	268000	3850	78	33	16
12	11	285000	4030	75	35	17
13	12	230000	3550	94	28	15
14	13	212000	3240	85	24	14
15	14	309000	4200	68	39	17
16	15	228000	3470	94	27	15
17	16	215000	3290	89	24	15
18	17	324000	4340	64	42	18
19	18	210000	3330	94	24	14
20	19	262000	3790	81	32	16
21	20	329000	4400	62	44	18

Stepwise Regression

Fig. 14.22: Partial screenshot of the spreadsheet for the given data.

Step 2: To select the first regressor in the model

In stepwise regression method, the selection of the first regressor variable is done in the same way as discussed in Step 2 of Sec. 14.3. From Fig. 14.5, the highest absolute value of t-statistic was 56.5364 given in Cell X61, which is associated with regressor X_3. Since the p-value of t-statistic was 0.0000 (given in Cell Y61) and it is less than $\alpha_{IN} = 0.05$, the regressor X_3 is contributing significantly to the model. So we include X_3 (temperature) in the model.

- The first regressor to be entered in the model is temperature (X_3).
- The lowest p-value of t test = 0.0000.
- From Fig. 14.5, the model is:

  $\hat{Y} = 50197.432 + 6593.264X_3$
- F test p-value = 0.0000.

Step 3: To select the second regressor in the model and check for possible elimination

This step is also the same as Step 3 of Sec.14.3. So we follow Step 3 of Sec. 14.3 where the highest absolute value of t-statistic was 3.5009 given in Cell D82 and was associated with the regressor X_1. The p-value of t-statistic associated with the X_1 was 0.0012 given in Cell E82 (Fig. 14.8), which is less than $\alpha_{IN} = 0.05$. So we include X_1 (advertising cost) in the model since the regressor X_1 is contributing significantly to the model as shown in Fig. 14.8.

The only difference is that now we look again at the regressor which was included in the model at Step 2 for the possible elimination. From Fig. 14.8 of Sec. 14.3, the p-value corresponding to X_3 is 0.0000 given in Cell E81. It is still significant at 5% level of significance and so we retain the regressor X_3.

At this step, we have two regressors X_3 and X_1 in the model.

- The second regressor to be entered in the model is the advertising cost (X_1).
- The lowest p-value of t test = 0.0012.
- From Fig. 14.8, the model is:

  $\hat{y} = -4221.106 + 4436.617X_3 + 32.678X_1$
- F test p-value = 0.0000.
- Since all p-values of t tests < 0.05, no regressor will be eliminated at this step.

- The third regressor to be entered in the model is price (X_2).
- The lowest p-value of t test = 0.0000.
- From Fig.14.12, the Model is

$$\hat{Y} = 2150.574 + 1445.845X_3 + 68.213X_1 - 561.95X_2$$

- F test p-value = 0.0000.
- From Fig. 14.12, since p-value = 0.0936 > 0.05, the regressor to be eliminated from the model is the temperature (X_3).

Step 4: To select the third regressor in the model and check for possible elimination

The selection procedure of third regressor for this data set is the same as Step 4 of Sec. 14.3. You have noted there that we entered the regressor X_2 in the model as it had the highest absolute value of t-statistic, i.e., 5.2664 and the corresponding p-value is 0.0000 given in Cells D104 and E104, respectively (Fig. 14.12). It was contributing significantly to the model. So we include X_2 (price) in the model. To check for possible elimination, we go back to Fig. 14.12 of Sec. 14.3.

Notice from Fig. 14.12 that the p-value corresponding to X_3 is 0.0588 given in Cell E102. It shows insignificant contribution to the model after including the regressor X_2 at 5% level of significance. So we eliminate X_3 from the model. At this step we again have two regressors in the model, i.e., X_1 and X_2.

Step 5: To select the third regressor in the model and check for possible elimination

We now consider all possible models with 3 regressors by keeping in mind that X_1 and X_2 are already there in the model and we have eliminated X_3 from the model. So the possible combination is (X_1, X_2, X_4). For performing the regression analysis in Excel, we keep the values of (X_1, X_2, X_4) in adjacent columns, i.e., Cells Y1:AA41 as shown in Fig. 14.23.

	Y	Z	AA	AB
1	**X1**	**X2**	**X4**	
2	3240	85	14	
3	3550	94	15	
4	3890	77	16	
5	3720	84	16	
6	4030	75	17	
7	3790	81	16	
8	3950	76	16	
9	3290	89	15	
10	2920	79	13	
11	3850	78	16	
12	4030	75	17	
13	3550	94	15	
14	3240	85	14	
15	4200	68	17	
16	3470	94	15	
17	3290	89	15	
18	4340	64	18	
19	3330	94	14	
20	3790	81	16	
21	4400	62	18	

Stepwise Regression

Fig. 14.23

To discover the third regressor to be entered in the model, we perform the regression analysis for this combination of three regressors in the model.

For X_1, X_2 and X_4: We consider **Y range** as Cells B1:B41, **X range** as Cells Y1:AA41 and **output range** as Cell A107. The output is shown in Fig. 14.24.

	A	B	C	D	E	F	G
116	ANOVA						
117		*df*	*SS*	*MS*	*F*	*Significance F*	
118	Regression	3	61037829797.5672	20345943265.8557	2644.7613	0.0000	
119	Residual	36	276945202.4328	7692922.2898			
120	Total	39	61314775000				
121							
122		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
123	Intercept	-28820.1147	11874.5560	-2.4270	0.0204	-52902.8303	-4737.3992
124	X1	75.4046	4.4114	17.0931	0.0000	66.4578	84.3513
125	X2	-710.9027	65.3406	-10.8800	0.0000	-843.4195	-578.3859
126	X4	3891.4261	1381.5422	2.8167	0.0078	1089.5287	6693.3236
127							

Fig. 14.24

Note that the regressor having the highest absolute value of t-statistic, i.e., 2.8167 given in Cell D126 corresponds to X_4 (see Fig. 14.24). Its p-value is 0.0078 given in Cell E126, which is less than $\alpha_{IN} = 0.05$. Therefore, the regressor X_4 is contributing significantly to the model. We can conclude that X_4 (number of stores) enters in the model as a third regressor. We now check for possible elimination. Fig. 14.24 reveals that the p-value corresponding to both the previously added regressors X_1 and X_2 is 0.0000 given in Cells E124 and E125, respectively, which is less than α_{OUT}. Both regressors X_1 and X_2 are still contributing significantly to the model. So we will not eliminate them. In this way we have three regressors X_1, X_2 and X_4 in the model.

Step 6: To select the fourth regressor in the model and check for possible elimination

We now consider the combination of 4 regressors, i.e., (X_1, X_2, X_4, X_3). We ***Copy*** and ***Paste*** the values of these regressors in Cells AC1:AF41 to keep them in the adjacent columns as shown in Fig. 14.25.

	AC	AD	AE	AF
1	**X1**	**X2**	**X4**	**X3**
2	3240	85	14	24
3	3550	94	15	28
4	3890	77	16	33
5	3720	84	16	31
6	4030	75	17	35
7	3790	81	16	32
8	3950	76	16	34
9	3290	89	15	24
10	2920	79	13	21
11	3850	78	16	33
12	4030	75	17	35
13	3550	94	15	28
14	3240	85	14	24
15	4200	68	17	39
16	3470	94	15	27
17	3290	89	15	24
18	4340	64	18	42
19	3330	94	14	24
20	3790	81	16	32
21	4400	62	18	44

Stepwise Regression

Fig. 14.25

We perform the regression analysis as discussed in Step 5 of Sec. 14.3 and Step 2 of Sec. 14.4.

We consider **Y range** as Cells B1:B41, **X range** as Cells AC1:AF41 and output **range** as Cell A129. The output is shown in Fig. 14.26.

- The third regressor to be entered in the model is number of stores (X_4).
- The lowest p-value of t test = 0.0078.
- From Fig. 14.24, the model is (Fig. 14.24)

  $\hat{y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$
- F test p-value = 0.0000.
- Since all p-values of t tests < 0.05, no regressor will be eliminated at this step.

	A	B	C	D	E	F	G
138	ANOVA						
139		*df*	*SS*	*MS*	*F*	*Significance F*	
140	Regression	4.0000	61059504597.6716	15264876149.4179	2092.9597	0.0000	
141	Residual	35.0000	255270402.3284	7293440.0665			
142	Total	39.0000	61314775000.0000				
143							
144		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
145	Intercept	-17069.7198	13421.7442	-1.2718	0.2118	-44317.3089	10177.8693
146	X1	60.7615	9.5184	6.3836	0.0000	41.4381	80.0850
147	X2	-579.8462	99.1325	-5.8492	0.0000	-781.0958	-378.5966
148	X4	3569.2517	1358.1134	2.6281	0.0127	812.1350	6326.3684
149	X3	1195.1257	693.2697	1.7239	0.0936	-212.2867	2602.5381
150							

Fig. 14.26

Note from Fig. 14.26 that the absolute value of t-statistic corresponding to the regressor X_3 is 1.7239 given in Cell D149. Its p-value is 0.0936 given in Cell E149, which is greater than $\alpha_{IN} = 0.05$. Therefore, we cannot enter the last regressor X_3 (temperature) in the model as it is not contributing significantly. From Fig. 14.26, you can see that the p-values corresponding to the other regressors, i.e., X_1, X_2 and X_4 are less than $\alpha_{OUT} = 0.05$. So we can conclude that the regressors X_1, X_2 and X_4 contribute significantly to the model and we do not eliminate X_1, X_2 and X_4. In this way, we have obtained the best fitted regression model with regressors X_1, X_2 and X_4 using stepwise regression method. Hence from Fig. 14.24, the resulting model is

- No more regressor is significant.
- From Fig. 14.24, the model is:

  $\hat{y} = -28820.1147 + 75.4046\,X_1 - 710.9027\,X_2 + 3891.4261\,X_4$
- F test p-value = 0.0000.

$$\hat{Y} = -28820.1147 + 75.4046X_1 - 710.9027X_2 + 3891.4261X_4$$

Notice from Fig.14.24 that the value of F-statistic is 2644.7613 given in Cell E118 and its p-value is 0.0000 given in Cell F118, which is less than 0.05. We conclude that the overall model is also significant.

You should now apply the above methods on other problems for practice.

Activity

Work out the following exercises with the help of MS Excel 2007 and interpret the results:

A1) Example 2 given in Unit 12 of MSTE-002.

A2) Exercise E2 given in Unit 12 of MSTE-002.

Match the results with the manual calculation done in Unit 12 of MSTE-002.

Continuous Assessment 14

Consider the data given in Continuous Assesment 12 of Lab Session 12 to build an appropriate linear regression model for electricity consumption of a household. Consider two more regressors, i.e., temperature and number of persons living in the house as given in Table 2.

Build a regression model by selecting appropriate regressors in the model using (i) forward selection, (ii) backward elimination and (iii) stepwise regression methods.

Table 2: Electricity consumption data

S. No.	Unit	Area	AC	Temperature	No. of Persons
1	1060	1316	5	31	2
2	1150	1420	7	36	4
3	1365	1556	12	42	5
4	1275	1488	9	32	5
5	1425	1612	13	40	3
6	1310	1516	10	37	5
7	1365	1556	12	36	2
8	1075	1352	6	34	4
9	925	1168	4	32	3
10	1340	1540	11	37	5
11	1425	1612	13	40	5
12	1150	1420	8	32	4
13	1060	1316	5	33	4
14	1545	1680	15	41	6
15	1140	1388	7	32	4
16	1075	1352	6	33	2
17	1620	1736	16	33	7
18	1050	1296	5	34	4
19	1310	1516	10	35	5
20	1645	1760	16	44	7
21	1565	1696	15	42	6
22	1215	1464	9	39	5
23	1275	1488	10	40	5
24	1465	1632	13	42	3
25	1080	1356	7	44	4
26	975	1196	4	32	3
27	1040	1256	5	33	4
28	1340	1540	11	37	5
29	865	1144	4	31	3
30	1175	1440	8	38	4
31	1080	1356	7	37	4
32	1500	1652	15	43	6
33	1175	1440	9	36	4
34	1050	1296	5	31	2
35	1365	1580	12	37	5
36	1465	1632	15	43	6
37	1215	1464	9	33	5
38	1365	1580	12	34	5
39	1140	1388	7	35	4
40	1005	1224	4	36	3

Home Work: Do It Yourself

1) Follow the steps explained in Secs. 14.3 to 14.5 to comprehend the variable selection methods in regression analysis for the data of Table 1. Take the screenshots and keep them in your record book.

2) Develop the spreadsheets for the exercise "Continuous Assessment 14" as explained in this lab session. Take screenshots of the final spreadsheets.

3) **Do not forget** to keep the screenshots in your record book as these will contribute to your continuous assessment in the Laboratory.