CHAPTER 7

7.1

(a) Backward elimination: Drop x_3 (step 1); drop x_4 (step 2); next candidate x_2 for elimination can not be dropped. Model with x_1 and x_2.

(b) Forward selection: Enter x_4 (step1); enter x_1 (step 2); enter x_2 (step 3); next candidate x_3 for selection can not be entered. Model with x_1, x_2, and x_4.

(c) Stepwise Regression: Steps 1, 2 and 3 of forward selection; x_4 can be dropped from the model containing x_1, x_2, and x_4; no reason to add x_3 to the model with x_1 and x_2. Model with x_1 and x_2.

(d) Model with x_1 and x_2: C_p = 2.68, close to desired value 3. Full model: C_p = 5. Prefer model with x_1 and x_2.

(e) x_2 and x_4 are highly correlated.

(f) F = 68.6; p-value less than 0.001; reject $\beta_1 = \beta_3 = 0$.

7.2

(a) C_p: Model with x_1 and x_2 (C_p = 2.7)

R^2: Model with x_1 and x_2, or model with x_1 and x_4. Small gain by going to more complicated models.

(b) Backward elimination ($\alpha_{drop} = 0.1$): Model with x_1 and x_2.

Forward selection ($\alpha_{enter} = 0.1$): Model with x_1, x_2, and x_4.

Stepwise regression ($\alpha_{drop} = \alpha_{enter} = 0.1$): Model with x_1 and x_2.

7.3

Minitab Best Subset Regression results:

Response is Y_1

Vars	R-Sq	R-Sq(adj)	C-p	S	X1	X2	X3	X4
1	49.3	45.4	9.8	1470.5				X
1	34.0	29.0	16.1	1677.2			X	
2	63.3	57.2	6.1	1301.8			X	X
2	49.6	41.2	11.7	1526.3	X			X
3	66.8	57.8	6.6	1293.4	X		X	X
3	64.6	54.9	7.5	1335.8		X	X	X
4	75.6	65.9	5.0	1162.2	X	X	X	X

Response is Y_2

Vars	R-Sq	R-Sq(adj)	C-p	S	X 1	X 2	X 3	X 4
1	98.4	98.3	7.3	43.517		X		
1	97.8	97.6	14.6	51.392	X			
2	99.1	99.0	1.1	33.550	X	X		
2	98.5	98.2	8.5	44.288		X		X
3	99.1	98.9	3.0	34.965	X	X	X	
3	99.1	98.9	3.0	35.021	X	X		X
4	99.1	98.8	5.0	36.644	X	X	X	X

Response is Y_3

Vars	R-Sq	R-Sq(adj)	C-p	S	X 1	X 2	X 3	X 4
1	36.1	31.2	8.1	90.890			X	
1	5.6	0.0	17.2	110.45	X			
2	66.3	60.7	1.1	68.686	X		X	
2	65.1	59.3	1.4	69.938		X	X	
3	66.4	57.3	3.0	71.616	X		X	X
3	66.3	57.1	3.0	71.731	X	X	X	
4	66.5	53.1	5.0	75.051	X	X	X	X

Minitab Stepwise Regression results:

Response is Y_1

The regression equation is
Y1 = 7770 + 49.6 X3 + 45.1 X4

Predictor	Coef	SE Coef	T	P
Constant	7770	2349	3.31	0.006
X3	49.55	23.14	2.14	0.053
X4	45.07	14.56	3.10	0.009

S = 1302   R-Sq = 63.3%   R-Sq(adj) = 57.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	35115127	17557564	10.36	0.002
Residual Error	12	20335325	1694610		
Total	14	55450452			

Response is Y_2

The regression equation is
Y2 = - 67.4 + 5.66 X1 + 8.02 X2

Predictor	Coef	SE Coef	T	P
Constant	-67.40	41.20	-1.64	0.128

```
X1               5.662       1.802       3.14    0.009
X2               8.018       1.864       4.30    0.001

S = 33.55       R-Sq = 99.1%     R-Sq(adj) = 99.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2     1546691      773346    687.05    0.000
Residual Error    12       13507        1126
Total             14     1560198
```

Response is Y_3

```
The regression equation is
Y3 = 292 - 2.68 X1 + 5.94 X3

Predictor        Coef     SE Coef          T        P
Constant        292.4       122.2       2.39    0.034
X1            -2.6796      0.8168      -3.28    0.007
X3              5.943       1.278       4.65    0.001

S = 68.69       R-Sq = 66.3%     R-Sq(adj) = 60.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2      111462       55731     11.81    0.001
Residual Error    12       56613        4718
Total             14      168075
```

(a) For production overhead costs (y_1): x_3 and x_4 are important. For direct production costs (y_2): x_1 and x_2 are important. For marketing costs (y_3): x_1 and x_3 are important.
(b) For production overhead costs (y_1), the change in production from the last period (x_4) is the single most important variable. For direct production costs (y_2), the production quantity (x_2) is the single most important variable.

7.4

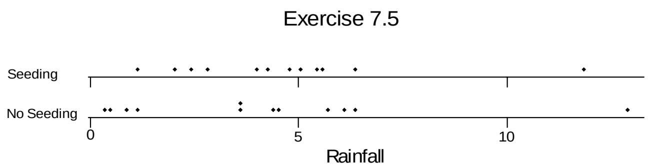
(a) False; different models may result if multicollinearity is present
(b) True
(c) False; can stay the same

7.5

Dot plots of rainfall for days with and without seeding are shown below. We see little difference between the two groups. The results of the two-sample t-test shown below indicate that the group difference is not significant.

```
Two-sample T for Rainfall
SA           N      Mean     StDev   SE Mean
0  (NO)     12      4.17      3.52       1.0
1  (YES)    12      4.63      2.78      0.80

Difference = mu (0) - mu (1)
Estimate for difference:  -0.46
95% CI for difference: (-3.16, 2.24)
T-Test of difference = 0 (vs not =):T-Value = -0.36 P-Value = 0.725 DF=20
```



The question now becomes whether the significance of the seeding action changes when other explanatory variables are included in the model. The results of the full model shown below are:
F = 1.77 for overall regression; p-value = 0.1647; the evidence for including any of the variables is quite weak;
t-values of the regression coefficients are small; their p-values are large, indicating that the variables are not important given that the other variables are in the model.
Seeding action is insignificant, indicating that it is difficult to justify cloud seeding.
Case diagnostics reveal that case 2 has a large studentized residual = -2.278, Cook's D = 4.748 and leverage = 0.865.

```
The regression equation is
y=Rainfall = 4.65 + 1.01 SA - 0.0321 Time - 0.911 SC + 0.006 EchoCov
           + 2.17 EchoMot + 1.84 PreWet

Predictor        Coef     SE Coef          T        P
Constant        4.654       3.337       1.39    0.181
SA              1.013       1.203       0.84    0.411
Time         -0.03212     0.02892      -1.11    0.282
SC            -0.9109      0.7512      -1.21    0.242
EchoCov        0.0057      0.1149       0.05    0.961
EchoMot         2.168       1.579       1.37    0.188
PreWet          1.844       2.758       0.67    0.513

S = 2.836       R-Sq = 38.5%     R-Sq(adj) = 16.8%
```

```
Analysis of Variance

Source            DF        SS        MS       F       P
Regression         6    85.584    14.264    1.77   0.165
Residual Error    17   136.751     8.044
Total             23   222.335
```

We also investigate the effects of interaction effects between the seeding action (SA) and the other explanatory variables. Using stepwise regression leads to a model with SA, the interaction between SA and SC, and time.

```
                      Parameter       Standard
Variable      DF       Estimate          Error    t Value    Pr > |t|

Intercept      1        6.27308        1.04889       5.98      <.0001
SA             1        7.81779        3.47088       2.25      0.0357
Time           1       -0.06076        0.02132      -2.85      0.0099
SA*SC          1       -2.18142        0.99308      -2.20      0.0400
```

The significant estimate of SA indicates that seeding action may be effective. However, the negative interaction SA*SC is difficult to explain; it indicates that the rainfall under cloud seeding decreases with increasing suitability. Also, there are two cases with relatively large Cook's distances (0.38 and 0.56). Omitting these two cases makes the effects of SA and SA*SC insignificant, leaving time (with a negative coefficient) as the only significant variable. In summary, this small data set is not particularly helpful in settling the issue whether cloud seeding is effective.

7.6 The Minitab Best Subset Regression procedure suggests a model with police expenditures (PE), the number of families per 1,000 earning below one half of the median income (IncInequ), the mean number of years of schooling x 10 of the population (Ed), and the number of males aged 14-24 per 1,000 of total state population (Age). Case #29 exhibits the largest leverage (0.471):

```
The regression equation is
Crime Rate = - 425 + 1.30 PE + 0.641 IncInequ + 1.66 Ed + 0.760 Age

Predictor        Coef     SE Coef         T        P
Constant      -424.92       85.85     -4.95    0.000
PE             1.2980      0.1438      9.03    0.000
IncInequ       0.6409      0.1527      4.20    0.000
Ed             1.6605      0.4580      3.63    0.001
Age            0.7602      0.3442      2.21    0.033

S = 22.15       R-Sq = 70.0%     R-Sq(adj) = 67.2%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	48196	12049	24.55	0.000
Residual Error	42	20614	491		
Total	46	68809			

7.7
$\hat{\mu} = -5.0359 + 0.0671\text{AirFlow} + 0.1295\text{CoolTemp}$; $R^2 = 0.909$; $C_p = 2.9$.
Last case (AirFlow = 70; CoolTemp = 20; StackLoss = 1.5) is an influential observation and should be scrutinized. Without this case:
$\hat{\mu} = -5.1076 + 0.0863\text{AirFlow} + 0.0803\text{CoolTemp}$; $R^2 = 0.946$

7.8
Stepwise regression ($\alpha_{\text{drop}} = \alpha_{\text{enter}} = 0.15$):

$\hat{\mu} = -62.60 + 7.427\%\text{ASurf} + 6.828\%\text{ABase} - 5.2685\text{Run}$;

$R^2 = 0.724$; $R^2_{\text{adj}} = 0.693$; $C_p = 1.3$.

Similar model: $\hat{\mu} = -23.00 + 5.975\%\text{ASurf} - 5.4058\text{Run}$;

$R^2 = 0.695$; $R^2_{\text{adj}} = 0.673$; $C_p = 1.9$.

Cases 13 and 15 with large Cook's influence. Second set of runs with considerably smaller change in rut depth.

7.9 Case 89 with age =197 should be omitted from the data set. The age of this child is very different from the ages of the other children. Results of the remaining n = 108 students are shown below:

Correlation among the variables:

	age	iq	math1	math2	read1
iq	-0.724				
math1	0.095	-0.024			
math2	-0.293	**0.542**	-0.418		
read1	-0.286	**0.474**	0.133	0.176	
read2	-0.071	-0.006	0.380	-0.357	0.314

Math problem solving and reading speed are positively correlated with IQ; IQ and age are correlated. Since we don't really know how students were selected into this study it is unclear what to make of this strong negative correlation between age and IQ.

Strongest results for Math2 (mathematics problem solving). No gender effect, rather weak age effect, but strong relationship with IQ.

```
The regression equation is
math2 = - 85.6 + 0.319 age + 0.623 iq + 0.33 gender

Predictor        Coef     SE Coef          T        P
Constant       -85.59       30.33      -2.82    0.006
age            0.3186      0.1804       1.77    0.080
iq             0.6230      0.1060       5.88    0.000
gender          0.327       2.575       0.13    0.899

S = 13.24       R-Sq = 31.4%     R-Sq(adj) = 29.4%
The regression equation is
math2 = - 85.3 + 0.317 age + 0.623 iq

Predictor        Coef     SE Coef          T        P
Constant       -85.28       30.08      -2.84    0.005
age            0.3173      0.1793       1.77    0.080
iq             0.6227      0.1055       5.90    0.000

S = 13.18       R-Sq = 31.4%     R-Sq(adj) = 30.1%


The regression equation is
math2 = - 34.0 + 0.488 iq

Predictor        Coef     SE Coef          T        P
Constant      -33.998       8.170      -4.16    0.000
iq            0.48754     0.07349       6.63    0.000

S = 13.31       R-Sq = 29.3%     R-Sq(adj) = 28.7%
```

Similar results for Read1 (reading speed). No gender effect, rather weak age effect, but strong relationship with IQ.

```
The regression equation is
read1 = - 14.2 + 0.0921 age + 0.241 iq + 1.19 gender

Predictor        Coef     SE Coef          T        P
Constant       -14.19       15.13      -0.94    0.351
age           0.09211     0.09001       1.02    0.309
iq            0.24059     0.05290       4.55    0.000
gender          1.193       1.285       0.93    0.355

S = 6.609       R-Sq = 23.8%     R-Sq(adj) = 21.6%


The regression equation is
read1 = - 13.0 + 0.0875 age + 0.240 iq

Predictor        Coef     SE Coef          T        P
Constant       -13.02       15.07      -0.86    0.390
age           0.08749     0.08981       0.97    0.332
iq            0.23953     0.05285       4.53    0.000

S = 6.604       R-Sq = 23.2%     R-Sq(adj) = 21.7%
```

```
The regression equation is
read1 = 1.12 + 0.202 iq

Predictor        Coef     SE Coef          T        P
Constant        1.118       4.052       0.28    0.783
iq            0.20226     0.03645       5.55    0.000

S = 6.603       R-Sq = 22.5%     R-Sq(adj) = 21.8%
```

7.10

The stepwise procedure in SAS (with Alpha-to-Enter = Alpha-to-Drop = 0.15) includes the proportion of males (%Male), the proportion of males older than 18 (%Male18), the proportion of the population older than 65 (%Pop65), the proportion of the rural (nonmetro) population (%nonMetro) and the proportion of households earning more than 100 thousand dollars %Inc100).

```
The regression equation is
% Votes for Bush = - 717 + 59.6 %Male - 44.3 %Male18 - 0.893 %Pop65
           + 0.149 %NonMetro - 2.04 %Incom100

Predictor        Coef     SE Coef          T        P
Constant       -717.4       156.0      -4.60    0.000
%Male           59.57       12.78       4.66    0.000
%Male18       -44.347       9.994      -4.44    0.000
%Pop65        -0.8928      0.5187      -1.72    0.092
%NonMetro     0.14864     0.04455       3.34    0.002
%Incom100     -2.0361      0.5481      -3.72    0.001

S = 5.531       R-Sq = 74.6%     R-Sq(adj) = 71.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5     4034.86      806.97     26.38    0.000
Residual Error    45     1376.56       30.59
Total             50     5411.42
```

States 2 (Alaska) and 9 (District of Columbia) have large Cook’s distance and leverage values. They have smaller population compared with other states. The proportion of votes for Bush was small (compared to other states) in the District of Columbia, and it was large (compared to other states) in Alaska.