$\boldsymbol{y} = \mathbf{1}\beta_0 + \boldsymbol{x}\beta_1 + \boldsymbol{\epsilon}$, and obtain the vector of residuals $\boldsymbol{e}$. Suppose that the true model is the quadratic model

$$\boldsymbol{y} = \mathbf{1}\beta_0 + \boldsymbol{x}\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{\epsilon}$$

where $\boldsymbol{x}_2' = (x_1^2, x_2^2, \ldots, x_n^2)$.

Show that $E(\boldsymbol{e}) = \beta_2(I - H)\boldsymbol{x}_2$, where H is the hat matrix from the fitted linear model.

6.4. Consider the linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with X an $n \times (p+1)$ matrix with rank $(p+1)$, and $\boldsymbol{\epsilon}$ a vector of uncorrelated errors with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I$. Let $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the vector of least squares estimates.

a. Find the mean vector and the covariance matrix of $\hat{\boldsymbol{\mu}}$.

b. Show that $\frac{1}{n}\sum_{i=1}^{n} V(\hat{\mu}_i) = \frac{(p+1)}{n}\sigma^2$.

Hint: Find the trace of $V(\hat{\boldsymbol{\mu}})$; use the fact that trace of AB = trace of BA if the products are defined.

c. Let $H = (h_{ij})$ be any $n \times n$ symmetric idempotent matrix: $H' = H$ and $HH = H$. Show that the diagonal elements h_{ii} must lie between zero and one.
Hint: Consider $\boldsymbol{a}_i'H$, where $\boldsymbol{a}_i$ is a $n \times 1$ vector with all components 0 except for the ith element, which is 1.

d. Assume that the linear model includes a constant term. Then the diagonal elements h_{ii} of the hat matrix $H = X(X'X)^{-1}X'$ satisfy $h_{ii} \geq \frac{1}{n}$.

Hint: Parameterize the model by centering the regressor variables $(x_{ij} - \bar{x}_j)$, for $j = 1, 2, \ldots, p$.

e. Consider the linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the X matrix has **rank less than** $p+1$. Then $X'X\boldsymbol{\beta} = X'\boldsymbol{y}$ has infinitely many solutions for $\boldsymbol{\beta}$. Suppose that $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are two solutions and let $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\mu}} = \mathrm{X}\tilde{\boldsymbol{\beta}}$ be the corresponding fitted values. Show that $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}$. This shows that both solutions of the normal equations will produce the same fitted values and residuals.

6.5. a. Suppose I is the $r \times r$ identity matrix, $\boldsymbol{w}$ and $\boldsymbol{v}$ are $r \times 1$ column vectors, and α is a constant. Show by direct multiplication that

$$(I + \alpha\boldsymbol{v}\boldsymbol{w}')^{-1} = I - \left(\frac{\alpha}{1 + \alpha\boldsymbol{v}'\boldsymbol{w}}\right)\boldsymbol{v}\boldsymbol{w}'$$

b. Use the result in (a) to obtain an expression for $(A + \boldsymbol{w}\boldsymbol{w}')^{-1}$ in terms of A^{-1} and $\boldsymbol{w}$.

c. Suppose we use least squares to fit the model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

to data from n subjects. Data $(y_{n+1}, x_{n+1,1}, \ldots, x_{n+1,p})$ become available on one more case so that the model becomes

$$\begin{pmatrix} \boldsymbol{y} \\ \cdots\cdot \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} X \\ \cdots\cdot \\ \boldsymbol{w}' \end{pmatrix}\boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \cdots\cdot \\ \epsilon_{n+1} \end{pmatrix}$$

where $\boldsymbol{w}' = (1, x_{n+1,1}, \ldots, x_{n+1,p})$, or

$$\boldsymbol{y}_1 = X_1\boldsymbol{\beta} + \boldsymbol{\epsilon}_1$$

i. Find an expression for $(X_1'X_1)^{-1}$ in terms of $(X'X)^{-1}$ and $\boldsymbol{w}$.

ii. Find an expression for $\hat{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'\boldsymbol{y}_1$ in terms of $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$.

This provides a simple way of updating the least squares estimate as more data become available. It is used in deletion diagnostics.

6.6. Consider the multiple regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where X consists of the k columns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$. Prove that $\hat{\beta}_k$ can be obtained by the following three steps:

Step 1. Regress $\boldsymbol{y}$ on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k-1}$ and denote the vector of residuals by $\boldsymbol{r}$.

Step 2. Regress $\boldsymbol{x}_k$ on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k-1}$ and denote the vector of residuals by $\boldsymbol{u}$.

Step 3. Fit the model $\boldsymbol{r} = \beta_k\boldsymbol{u} + \boldsymbol{\epsilon}$. The resulting estimate $\hat{\beta}_k$ is identical to the estimate $\hat{\beta}_k$ in $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Hint: Use $\tilde{X}$ to denote the first $k-1$ columns of X. Then $\hat{\boldsymbol{\beta}}$ satisfies $X'X\hat{\boldsymbol{\beta}} = X'\boldsymbol{y}$, where

$$X'X = \begin{pmatrix} \tilde{X}'\tilde{X} & \tilde{X}'\boldsymbol{x}_k \\ \boldsymbol{x}_k'\tilde{X} & \boldsymbol{x}_k'\boldsymbol{x}_k \end{pmatrix} \text{ and } X'\boldsymbol{y} = \begin{pmatrix} \tilde{X}'\boldsymbol{y} \\ \boldsymbol{x}_k'\boldsymbol{y} \end{pmatrix}.$$

6.7. Explain why the following statements are true or false:

is related to the rate of exchange of gases in the lungs) were measured. The expired ventilation (y) and the oxygen uptake (x) are related nonlinearly.

Graph expired ventilation against oxygen uptake. Repeat the graph for various appropriate transformations, and develop a model that relates the transformed variables. Consider the Box–Cox family of transformations. Estimate the appropriate transformation.

x = Oxygen Uptake	y = Expired Ventilation
574	21.9
592	18.6
664	18.6
667	19.1
718	19.2
770	16.9
927	18.3
947	17.2
1,020	19.0
1,096	19.0
1,277	18.6
1,323	22.8
1,330	24.6
1,599	24.9
1,639	29.2
1,787	32.0
1,790	27.9
1,794	31.0
1,874	30.7
2,049	35.4
2,132	36.1
2,160	39.1
2,292	42.6
2,312	39.9
2,475	46.2
2,489	50.9
2,490	46.5
2,577	46.3
2,766	55.8
2,812	54.5
2,893	63.5
2,957	60.3

x = Oxygen Uptake	y = Expired Ventilation
3,052	64.8
3,151	69.2
3,161	74.7
3,266	72.9
3,386	80.4
3,452	83.0
3,521	86.0
3,543	88.9
3,676	96.8
3,741	89.1
3,844	100.9
3,878	103.0
4,002	113.4
4,114	111.4
4,152	119.9
4,252	127.2
4,290	126.4
4,331	135.5
4,332	138.9
4,390	143.7
4,393	144.8

6.16. The data are taken from Robertson, J. D., and Armitage, P. Comparison of two hypertensive agents. *Anaesthesia,* 14, 53–64, 1959. The data are given in the file **recovery**.

Hypertensive drugs are used routinely to lower a patient's blood pressure, and such drugs are administered continuously during surgery. Since surgery times vary, the total amount of the drug that is administered varies from case to case. Also, patients react differently to such drugs, and hence blood pressure during surgery varies across patients. The sooner blood pressure rises to normal levels, the better. The recovery time (i.e., the time it takes for a patient's systolic blood pressure to return to normal) is an important variable.

The following table lists, for a sample of 53 patients, the recovery time, the logarithm of the administered dose, and the average systolic blood pressure while the drug is being administered.

Discuss how recovery time is related to the dose and the blood pressure that is achieved during surgery. Fit appropriate regression models, check for model violations, and interpret the results. Explore the usefulness of transformations on the response.

x_1 = Log Dose	x_2 = Blood Pressure	y = Recovery Time
2.26	66	7
1.81	52	10
1.78	72	18
1.54	67	4
2.06	69	10
1.74	71	13
2.56	88	21
2.29	68	12
1.80	59	9
2.32	73	65
2.04	68	20
1.88	58	31
1.18	61	23
2.08	68	22
1.70	69	13
1.74	55	9
1.90	67	50
1.79	67	12
2.11	68	11
1.72	59	8
1.74	68	26
1.60	63	16
2.15	65	23
2.26	72	7
1.65	58	11
1.63	69	8
2.40	70	14
2.70	73	39
1.90	56	28
2.78	83	12
2.27	67	60
1.74	84	10

x_1 = Log Dose	x_2 = Blood Pressure	y = Recovery Time
2.62	68	60
1.80	64	22
1.81	60	21
1.58	62	14
2.41	76	4
1.65	60	27
2.24	60	26
1.70	59	28
2.45	84	15
1.72	66	8
2.37	68	46
2.23	65	24
1.92	69	12
1.99	72	25
1.99	63	45
2.35	56	72
1.80	70	25
2.36	69	28
1.59	60	10
2.10	51	25
1.80	61	44

6.17. The data are taken from Brown, B. M. and Maritz, J. S. Distribution-free methods in regression. *Australian Journal of Statistics,* 24, 318–331, 1982. The data are given in the file **rigidity**.

Measurements on 50 varieties of timber are made on their rigidity, elasticity, and air-dried density. The objective is to predict rigidity as a function of elasticity and air-dried density. Pay careful attention to the case diagnostics.

y = Rigidity	x_1 = Elasticity	x_2 = Density
1,000	99.0	25.3
1,112	173.0	28.2
1,033	188.0	28.6
1,087	133.0	29.1
1,069	146.0	30.7
925	91.0	31.4

Example: Power Plant Data Continued

We use preset significance levels α to enter $= 0.15$ and α to drop $= 0.15$. The procedure terminates with PT, S, D, NE, and CT. In fact, no variables were removed along the way. The model summary is identical to the one in Table 7.8.

This example shows these procedures at their best. All three algorithms lead to the same conclusion: a model that involves the five explanatory variables PT, S, D, NE, and CT. However, we have previously seen that several other quite reasonable models describe the data equally well but involve other variables.

All "automatic" algorithms should be used with caution. In situations in which there is an appreciable degree of multicollinearity among the explanatory variables, the three methods may lead to quite different final models. In such situations, it is preferable to examine all possible regressions because such an analysis can show that several different models perform quite similarly (in terms of R^2, s^2, C_p).

Most observational studies will have some degree of multicollinearity. Hence, one should be cautious with automatic model selection procedures.

EXERCISES

7.1. In an experiment involving one dependent variable (y) and four explanatory variables x_1, x_2, x_3, and x_4, all possible regressions are fit to a data set consisting of $n = 13$ cases. A constant term is routinely included in all models. The results are summarized as follows:

Regressors in Model	Residual Sum of Squares
None	4,073.6
x_1	1,898.5
x_2	1,359.5
x_3	2,909.1
x_4	1,325.8
x_1, x_2	86.9
x_1, x_3	1,840.6
x_1, x_4	112.1
x_2, x_3	623.1
x_2, x_4	1,303.3
x_3, x_4	263.6
x_1, x_2, x_3	72.2
x_1, x_2, x_4	72.0
x_1, x_3, x_4	76.2
x_2, x_3, x_4	110.7
x_1, x_2, x_3, x_4	71.8

a. What model will result from automatic backward elimination with significance level (α to drop) 0.05?

b. What model will result from automatic forward selection with a significance level (α to enter) 0.1?

c. What model will result from automatic stepwise regression with significance levels α to enter $= \alpha$ to drop $= 0.1$?

d. Compare the value of the C_p statistic for the model you found in (a) with that of the model that includes all four x's.

e. In the one-variable models, x_2 and x_4 seem to be important. However, the model with x_1 and x_2 and the model with x_1 and x_4 are the best in the two-variable group, and not the model with x_2 and x_4. Explain.

f. Consider the regression model with the variables x_1, x_2, x_3, and x_4. Test the hypothesis $\beta_1 = \beta_3 = 0$.

7.2. Consider the data given in the file **hald**. It contains the variables y, x_1, x_2, x_3, x_4.

a. For each of the following criteria, indicate which set of independent variables is best for predicting y.

i. R^2
ii. C_p

b. Using (i) backward elimination, (ii) forward selection, and (iii) stepwise regression, find the best sets of independent variables.

7.3. A company studies its marketing and production processes in order to better predict production overhead costs (y_1), direct production costs (y_2), and marketing costs (y_3). It selects as predictor variables direct labor input (x_1), production quantity (x_2), sales quantity (x_3), and the change in production from the last period (x_4). Data on these variables for the past 15 months are in the file **market**.

a. For each of the three response variables, select the best model(s) for prediction. Do these models change with the selected response variable?

b. Assess which of the input factors are most important for influencing (i) overhead costs and (ii) direct production costs.

7.4. Explain why the following statements are true or false.

a. All criteria for the selection of the best regression equation lead to the same set of regressor variables.

b. Addition of a variable to a regression equation does not decrease R^2.

c. Addition of a variable to a regression model always decreases the residual mean square.

7.5. The data are taken from Woodley, W. L., Biondini, R., and Berkeley, J. Rainfall results 1970–75: Florida Area Cumulus Experiment. *Science,* 195, 735–742; 1977. The data are given in the file **rainseeding**.

These particular data come from an experiment during the summer of 1975 that investigated the usefulness of silver iodide to increase rainfall. Experiments were carried out on 24 days that were judged suitable for seeding. Suitability was judged on the basis of a suitability criterion (SC) that had to be at least 1.5 (with larger values indicating better suitability). On each day, the decision to seed or not to seed was made at random ($A = 1$ if seeding occurred; A = 0 if no seeding). The response is the amount of rain (in cubic meters $\times 10^7$) that fell on the target area during a 6-hr period during that day. In addition, the data set includes the following covariates:

- *Time:* Number of days after the first day of the experiment (June 1, 1975)
- *Echo coverage:* The percentage cloud cover in the experimental area, determined from radar measurements
- *Echo motion:* An indicator whether the radar echo was moving (1) or stationary (2)
- *Prewetness:* The total rainfall in the target area 1 hr before seeding (in cubic meters $\times 10^7$)

Investigate appropriate models that relate the amount of rainfall to the explanatory variables. Use model selection procedures

Seeding Action	Time	Suitability Criterion	Echo Coverage	Echo Motion	Pre-wetness	$y =$ Rainfall
0	0	1.75	13.4	2	0.274	12.85
1	1	2.70	37.9	1	1.267	5.52
1	3	4.10	3.9	2	0.198	6.29
0	4	2.35	5.3	1	0.526	6.11
1	6	4.25	7.1	1	0.250	2.45
0	9	1.60	6.9	2	0.018	3.61
0	18	1.30	4.6	1	0.307	0.47
0	25	3.35	4.9	1	0.194	4.56
0	27	2.85	12.1	1	0.751	6.35
1	28	2.20	5.2	1	0.084	5.06
1	29	4.40	4.1	1	0.236	2.76
1	32	3.10	2.8	1	0.214	4.05
0	33	3.95	6.8	1	0.796	5.74
1	35	2.90	3.0	1	0.124	4.84
1	38	2.05	7.0	1	0.144	11.86
0	39	4.00	11.3	1	0.398	4.45
0	53	3.35	4.2	2	0.237	3.66
1	55	3.70	3.3	1	0.960	4.22
0	56	3.80	2.2	1	0.230	1.16
1	59	3.40	6.5	2	0.142	5.45
1	65	3.15	3.1	1	0.073	2.02
0	68	3.15	2.6	1	0.136	0.82
1	82	4.01	8.3	1	0.123	1.09
0	83	4.65	7.4	1	0.168	0.28

(all possible regressions, backward elimination and stepwise regression). Assess the effectiveness of cloud seeding, after having adjusted your analysis for important covariates. Check for unusual cases, and determine the sensitivity of your results to these cases.

7.6. The data are taken from Vandaele, W. Participation in illegitimate activities: Erlich revisited. In: *Deterrence and Incapacitation* (Blumstein, A., Cohen, J., and Nagin, D., Eds.,). Washington, DC: National Academy of Sciences, pp. 270–335, 1978. The data are given in the file **crimerate**.

Data on crime-related statistics for 47 U.S. states in 1960 are given. The data set includes

- Crime rate: Number of offenses known to police per 1,000,000 population
- Age: Age distribution—Number of males aged 14–24 per 1,000 of total state population
- *S*: Binary variable distinguishing southern states (1) from the rest of the states
- Ed: Mean number of years of schooling x 10 of the population, 25 years or older
- PE: Police expenditures—Per capita expenditure on police protection by state and local government in 1960
- PE-1: Police expenditures—Per capita expenditure on police protection by state and local government in 1959
- LF: Labor force participation rate per 1,000 civilian urban males in the age group 14–24
- *M*: The number of males per 1,000 females
- Pop: The state population size in 100,000
- NW: The number of nonwhites per 1,000
- UE1: Unemployment rate of urban males per 1,000 in the age group 14–24
- UE2: Unemployment rate of urban males per 1,000 in the age group 35–39
- Wealth: Median value of transferable goods and assets or family income (units 10 dollars)
- IncIneq: Income inequality—Number of families per 1,000 earning below one-half of the median income

Crime Rate	Age	S	Ed	PE	PE-1	LF	M	Pop	NW	UE1	UE2	Wealth	Inc Ineq
79.1	151	1	91	58	56	510	950	33	301	108	41	394	261
163.5	143	0	113	103	95	583	1,012	13	102	96	36	557	194
57.8	142	1	89	45	44	533	969	18	219	94	33	318	250
196.9	136	0	121	149	141	577	994	157	80	102	39	673	167
123.4	141	0	121	109	101	591	985	18	30	91	20	578	174
68.2	121	0	110	118	115	547	964	25	44	84	29	689	126
96.3	127	1	111	82	79	519	982	4	139	97	38	620	168
155.5	131	1	109	115	109	542	969	50	179	79	35	472	206
85.6	157	1	90	65	62	553	955	39	286	81	28	421	239
70.5	140	0	118	71	68	632	1,029	7	15	100	24	526	174
167.4	124	0	105	121	116	580	966	101	106	77	35	657	170
84.9	134	0	108	75	71	595	972	47	59	83	31	580	172
51.1	128	0	113	67	60	624	972	28	10	77	25	507	206
66.4	135	0	117	62	61	595	986	22	46	77	27	529	190
79.8	152	1	87	57	53	530	986	30	72	92	43	405	264
94.6	142	1	88	81	77	497	956	33	321	116	47	427	247
53.9	143	0	110	66	63	537	977	10	6	114	35	487	166
92.9	135	1	104	123	115	537	978	31	170	89	34	631	165
75.0	130	0	116	128	128	536	934	51	24	78	34	627	135
122.5	125	0	108	113	105	567	985	78	94	130	58	626	166
74.2	126	0	108	74	67	602	984	34	12	102	33	557	195
43.9	157	1	89	47	44	512	962	22	423	97	34	288	276
121.6	132	0	96	87	83	564	953	43	92	83	32	513	227
96.8	131	0	116	78	73	574	1,038	7	36	142	42	540	176
52.3	130	0	116	63	57	641	984	14	26	70	21	486	196
199.3	131	0	121	160	143	631	1,071	3	77	102	41	674	152
34.2	135	0	109	69	71	540	965	6	4	80	22	564	139
121.6	152	0	112	82	76	571	1,018	10	79	103	28	537	215
104.3	119	0	107	166	157	521	938	168	89	92	36	637	154
69.6	166	1	89	58	54	521	973	46	254	72	26	396	237
37.3	140	0	93	55	54	535	1,045	6	20	135	40	453	200
75.4	125	0	109	90	81	586	964	97	82	105	43	617	163
107.2	147	1	104	63	64	560	972	23	95	76	24	462	233
92.3	126	0	118	97	97	542	990	18	21	102	35	589	166
65.3	123	0	102	97	87	526	948	113	76	124	50	572	158
127.2	150	0	100	109	98	531	964	9	24	87	38	559	153
83.1	177	1	87	58	56	638	974	24	349	76	28	382	254
56.6	133	0	104	51	47	599	1,024	7	40	99	27	425	225
82.6	149	1	88	61	54	515	953	36	165	86	35	395	251
115.1	145	1	104	82	74	560	981	96	126	88	31	488	228
88.0	148	0	122	72	66	601	998	9	19	84	20	590	144
54.2	141	0	109	56	54	523	968	4	2	107	37	489	170
82.3	162	1	99	75	70	522	996	40	208	73	27	496	224
103.0	136	0	121	95	96	574	1,012	29	36	111	37	622	162
45.5	139	1	88	46	41	480	968	19	49	135	53	457	249
50.8	126	0	104	106	97	599	989	40	24	78	25	593	171
84.9	130	0	121	90	91	623	1,049	3	22	113	40	588	160