D/Solutions to exercises

Chapter 1: The forecasting perspective

1.1 Look for pragmatic applications in the real world. Note that there are no fixed answers in this problem.

- **(a)** Dow theory: There is an element of belief that past patterns will continue into the future. So first, look for the patterns (support and resistance levels) and then project them ahead for the market and individual stocks. This is a quantitative time series method.
- **(b)** Random walk theory: This is quantitative, and involves a time series rather than an explanatory approach. However, the forecasts are very simple because of the lack of any meaningful information. The best prediction of tomorrow's closing price is today's closing price. In other words, if we look at first differences of closing prices (i.e., today's closing price minus yesterday's closing price) there will be no pattern to discover.
- **(c)** Prices and earnings: Here instead of dealing with only one time series (i.e., the stock price series) we look at the relation between stock price and earnings per share to see if there is a relationship—maybe with a lag, maybe not. Therefore this is an explanatory approach to forecasting and would typically involve regression analysis.

1.2 *Step 1: Problem definition* This would involve understanding the nature of the individual product lines to be forecast. For example, are they high-demand products or specialty biscuits produced for individual clients? It is also important to learn who requires the forecasts and how they will be used. Are the forecasts to be used in scheduling production, or in inventory management, or for budgetary planning? Will the forecasts be studied by senior management, or by the production manager, or someone else? Have there been stock shortages so that demand has gone unsatisfied in the recent past? If so, would it be better to try to forecast demand rather than sales so that we can try to prevent this happening again in the future? The forecaster will also need to learn whether the company requires one-off forecasts or whether the company is planning on introducing a new forecasting system. If the latter, are they intending it to be managed by their own employees and, if so, what software facilities do they have available and what forecasting expertise do they have in-house?

Step 2: Gathering information It will be necessary to collect historical data on each of the product lines we wish to forecast. The company may be interested in forecasting each of the product lines for individual selling points. If so, it is important to check that there are sufficient data to allow reasonable forecasts to be obtained. For each variable the company wishes to forecast, at least a few years of data will be needed.

There may be other variables which impact the biscuit sales, such as economic fluctuations, advertising campaigns, introduction of new product lines by a competitor, advertising campaigns of competitors, production difficulties. This information is best obtained by key personnel within the company. It will be necessary to conduct a range of discussions with relevant people to try to build an understanding of the market forces.

If there are any relevant explanatory variables, these will need to be collected.

Step 3: Preliminary (exploratory) analysis Each series of interest should be graphed and its features studied. Try to identify consistent patterns such as trend and seasonality. Check for outliers. Can they be explained? Do any of the explanatory variables appear to be strongly related to biscuit sales?

Step 4: Choosing and fitting models A range of models will be fitted. These models will be chosen on the basis of the analysis in Step 3.

Step 5: Using and evaluating a forecasting model Forecasts of each product line will be made using the best forecasting model identified in Step 4. These forecasts will be compared with expert in-house opinion and monitored over the period for which forecasts have been made.

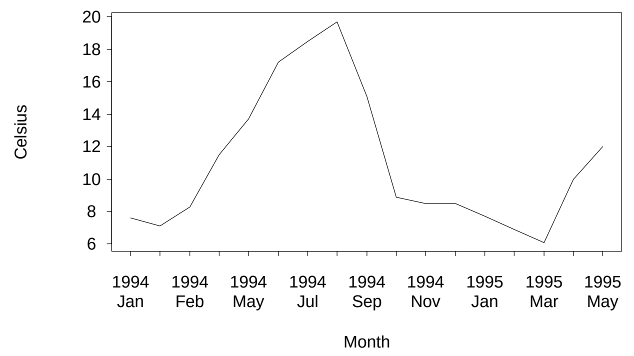
There will be work to be done in explaining how the forecasting models work to company personnel. There may even be substantial resistance to the introduction of a mathematical approach to forecasting. Some people may feel threatened. A period of education will probably be necessary.

A review of the forecasting models should be planned.

Chapter 2: Basic forecasting tools

2.1 **(a)** One simple answer: choose the mean temperature in June 1994 as the forecast for June 1995. That is, 17.2 °C.

(b) The time plot below shows clear seasonality with average temperature higher in summer.



Exercise 2.1(b): *Time plot of average monthly temperature in Paris (January 1994–May 1995).*

2.2 **(a)** Rapidly increasing trend, little or no seasonality.

(b) Seasonal pattern of period 24 (low when asleep); occasional peaks due to strenuous activity.

(c) Seasonal pattern of period 7 with peaks at weekends; possibly also peaks during holiday periods such as Easter or Christmas.

(d) Strong seasonality with a weekly pattern (low on weekends) and a yearly pattern. Peaks in either summer (air-conditioning) or winter (heating) or both depending on climate. Probably increasing trend with variation increasing with trend.

2.3 **(a)** Smooth series with several large jumps or direction changes; very large range of values; logs help stabilize variance.

(b) Downward trend (or early level shift); cycles of about 15 days; outlier at day 8; no transformation necessary.

(c) Cycles of about 9–10 years; large range and little variation at low points indicating transformation will help; logs help stabilize variance.

(d) No clear trend; seasonality of period 12; high in July; no transformation necessary.

(e) Initial trend; level shift end of 1982; seasonal period 4 (high in Q2 and Q3, low in Q1); no transformation necessary.

2.4 1-B, 2-A, 3-D, 4-C. The easiest approach to this question is to first identify D. Because it has a peak at lag 12, the time series must have a pattern of period 12. Therefore it is likely to be monthly. The slow decay in plot D shows the series has trend. The only series with both trend and seasonality of period 12 is Series 3. Next consider plot C which has a peak at lag 10. Obviously this cannot reflect a seasonal pattern since the only series remaining which is seasonal is series 2 and that has period 12. Series 4 is strongly cyclic with period approximately 10 and series 1 has no seasonal or strong cyclic patterns. Therefore C must correspond to series 4. Plot A shows a peak at lag 12 indicating seasonality of period 12. Therefore, it must correspond with series 2. That leaves plot B aligned with series 1.

2.5 **(a)**

	X	Y
Mean	52.99	43.70
Median	52.60	44.42
MAD	3.11	2.47
MSE	15.94	8.02
St.dev.	4.14	2.94

(b) Mean and median give a measure of center; MAD, MSE and St.dev. are measures of spread.

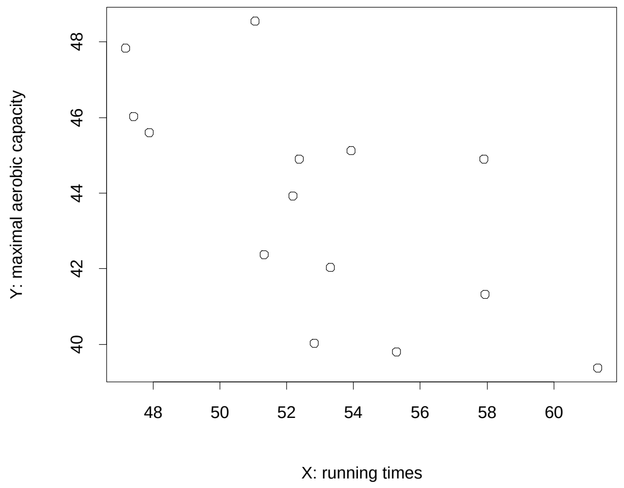
(c) $r = -0.660$. See plot on next page.

(d) It is inappropriate to compute autocorrelations since there is no time component to these data. The data are from 14 different runners. (Autocorrelation would be appropriate if they were data from the same runner at 14 different times.)

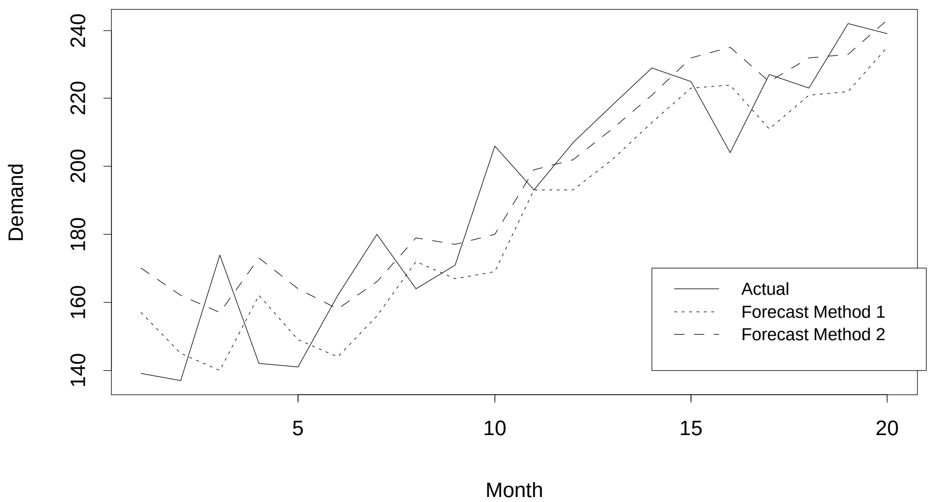
2.6 **(a)** See plot on following page.

(b) and (c)

Notation: Error 1 = (actual demand) − (method 1 forecast)
Error 2 = (actual demand) − (method 2 forecast)



Exercise 2.5(c): *Plot of running times versus maximal aerobic capacity.*



Exercise 2.6(a): *Time plots of data and forecasts.*

Period	Actual	Method 1	Error 1	Method 2	Error 2
1	139	157	−18	170	−31
2	137	145	−8	162	−25
3	174	140	34	157	17
4	142	162	−20	173	−31
5	141	149	−8	164	−23
6	162	144	18	158	4
7	180	156	24	166	14
8	164	172	−8	179	−15
9	171	167	4	177	−6
10	206	169	37	180	26
11	193	193	0	199	−6
12	207	193	14	202	5
13	218	202	16	211	7
14	229	213	16	221	8
15	225	223	2	232	−7
16	204	224	−20	235	−31
17	227	211	16	225	2
18	223	221	2	232	−9
19	242	222	20	233	9
20	239	235	4	243	−4
Analysis of errors		ME	6.25		−4.80
(periods 1–20)		MAE	14.45		14.29
		MSE	307.25		294.00
		MPE	2.55		−3.61
		MAPE	7.87		8.24
		Theil's U	0.94		0.85

On MAE and MSE, Method 2 is better than Method 1. On MAPE, Method 1 is better than Method 2. Note that this is different from the conclusion drawn in Section 4/2/3 where these two methods are compared. The difference is that we have used a different time period over which to compare the results. Holt's method (Method 2) performs quite poorly at the start of the series. In Chapter 4, this period is excluded from the analysis of errors.

2.7 **(a)** Changes: −0.25, −0.26, 0.13, . . . , −0.09, −0.77. There are 78 observations in the DOWJONES.DAT file. Therefore there are 77 changes.

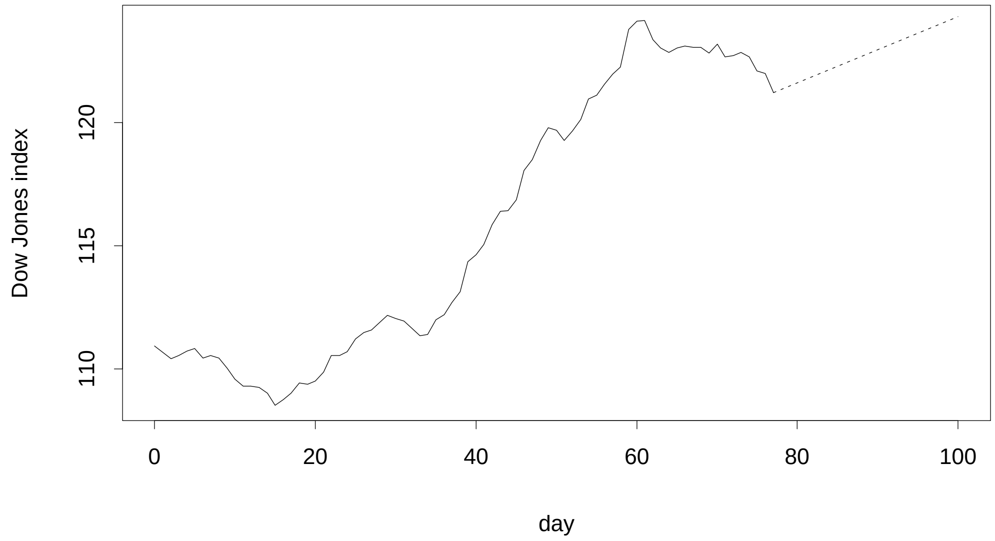
(b) Average change: 0.1336. So the next 20 changes are each forecast to be 0.1336.

(c) The last value of the series is 121.23. So the next 20 are forecast to be:

$$\begin{aligned}
\hat{X}_{79} &= 121.23 + 0.1336 = 121.36 \\
\hat{X}_{80} &= 121.36 + 0.1336 = 121.50 \\
\hat{X}_{81} &= 121.50 + 0.1336 = 121.63 \qquad \text{etc.}
\end{aligned}$$

In general, $\hat{X}_{79+h} = 121.23 + h(0.1336)$.

(d) See the plot below.



Exercise 2.7(d): *Plot of Dow Jones index (DOWJONES.DAT)*

(e) The average change is $c = \frac{1}{n-1}\sum_{t=2}^{n}(X_t - X_{t-1})$ and the forecasts are $\hat{X}_{n+h} = X_n + hc$. Therefore,

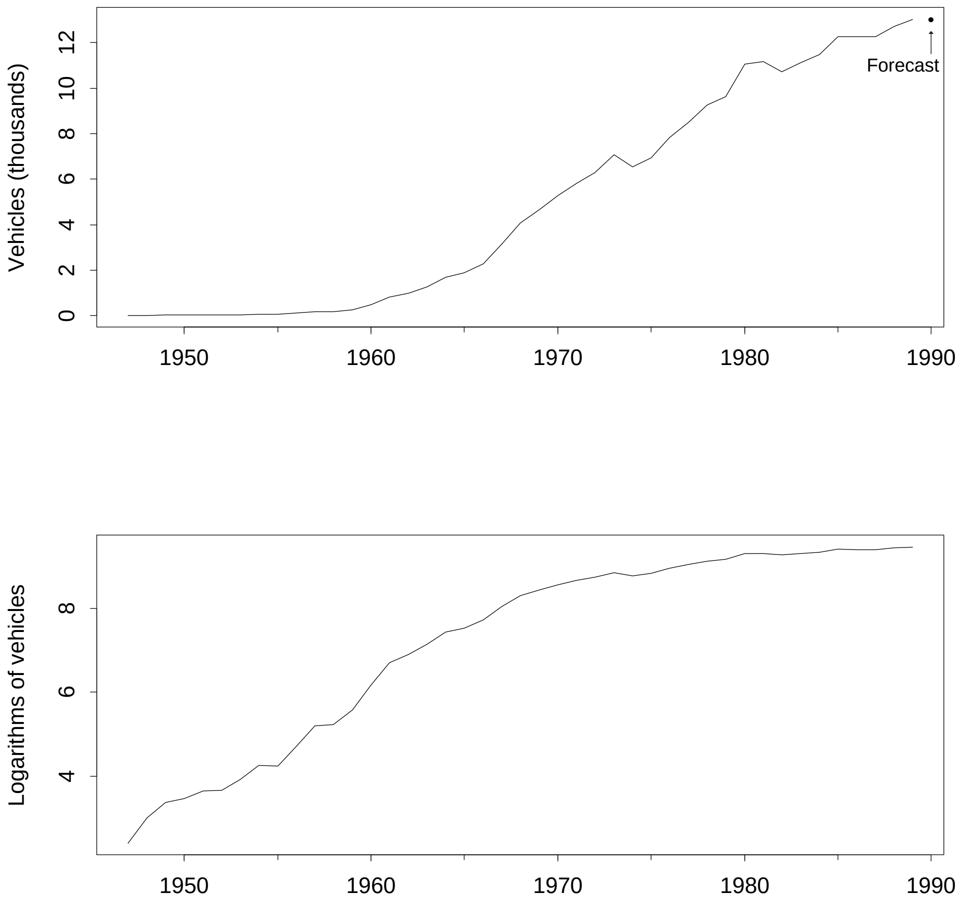
$$\begin{aligned} \hat{X}_{n+h} &= X_n + h\frac{1}{n-1}\sum_{t=2}^{n}(X_t - X_{t-1}) \\ &= X_n + \frac{h}{n-1}(X_n - X_1). \end{aligned}$$

This is a straight line with slope equal to $(X_n - X_1)/(n-1)$. When $h = 0$, $\hat{X}_{n+h} = X_n$ and when $h = -(n-1)$, $\hat{X}_{n+h} = X_1$. Therefore, the line is drawn between the first and last observations.

2.8 **(a)** See the plot on the next page. The variation when the production is low is much less than the variation in the series when the production is high. This indicates a transformation is required.

(b) See the plot on the next page.

(c) See the table on page 84.



Exercise 2.8 (a) and (b): *Time plots of Japanese automobile production and the logarithms of Japanese automobile production.*

Year	Data	Log	Forecast	Error	Error^2	|Error/Log|
1947	11	2.40				
1948	20	3.00	2.40	0.598	0.357	0.1996
1949	29	3.37	3.00	0.372	0.138	0.1103
1950	32	3.47	3.37	0.098	0.010	0.0284
1951	38	3.64	3.47	0.172	0.030	0.0472
1952	39	3.66	3.64	0.026	0.001	0.0071
1953	50	3.91	3.66	0.249	0.062	0.0635
1954	70	4.25	3.91	0.337	0.113	0.0792
1955	69	4.23	4.25	−0.014	0.000	0.0034
1956	111	4.71	4.23	0.475	0.226	0.1009
1957	182	5.20	4.71	0.495	0.245	0.0950
1958	188	5.24	5.20	0.032	0.001	0.0062
1959	263	5.57	5.24	0.336	0.113	0.0602
1960	482	6.18	5.57	0.606	0.367	0.0981
1961	814	6.70	6.18	0.524	0.275	0.0782
1962	991	6.90	6.70	0.197	0.039	0.0285
1963	1284	7.16	6.90	0.259	0.067	0.0362
1964	1702	7.44	7.16	0.282	0.079	0.0379
1965	1876	7.54	7.44	0.097	0.009	0.0129
1966	2286	7.73	7.54	0.198	0.039	0.0256
1967	3146	8.05	7.73	0.319	0.102	0.0396
1968	4086	8.32	8.05	0.261	0.068	0.0314
1969	4675	8.45	8.32	0.135	0.018	0.0159
1970	5289	8.57	8.45	0.123	0.015	0.0144
1971	5811	8.67	8.57	0.094	0.009	0.0109
1972	6294	8.75	8.67	0.080	0.006	0.0091
1973	7083	8.87	8.75	0.118	0.014	0.0133
1974	6552	8.79	8.87	−0.078	0.006	0.0089
1975	6942	8.85	8.79	0.058	0.003	0.0065
1976	7842	8.97	8.85	0.122	0.015	0.0136
1977	8514	9.05	8.97	0.082	0.007	0.0091
1978	9269	9.13	9.05	0.085	0.007	0.0093
1979	9636	9.17	9.13	0.039	0.002	0.0042
1980	11043	9.31	9.17	0.136	0.019	0.0146
1981	11180	9.32	9.31	0.012	0.000	0.0013
1982	10732	9.28	9.32	−0.041	0.002	0.0044
1983	11112	9.32	9.28	0.035	0.001	0.0037
1984	11465	9.35	9.32	0.031	0.001	0.0033
1985	12271	9.41	9.35	0.068	0.005	0.0072
1986	12260	9.41	9.41	−0.001	0.000	0.0001
1987	12249	9.41	9.41	−0.001	0.000	0.0001
1988	12700	9.45	9.41	0.036	0.001	0.0038
1989	13026	9.47	9.45	0.025	0.001	0.0027
1990			9.47			

Exercise 2.8 (c) and (d).

(d) MSE=0.059 (average of column headed Error^2)
MAPE=3.21% (average of values in last column multiplied by 100).

(e) See graph. Forecast is $e^{9.47} = 13026$.

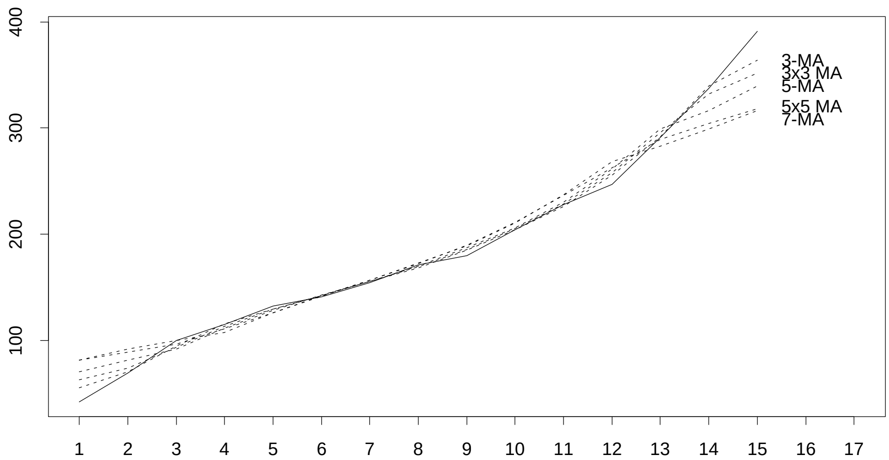
(f) There are a large number of possible methods. One method, which is discussed in Chapter 5, is to consider only data after 1970 and use a straight line fitted through the original data (i.e. without taking logarithms).

(g) The data for 1974 is lower than would be expected. If this information could be included in the forecasts, the MSE and MAPE would both be smaller because the forecast error in 1974 would be smaller.

Chapter 3: Time series decomposition

3.1

Y	3-MA	5-MA	7-MA	3×3-MA	5×5-MA
42	55.50	70.33	81.50	62.92	81.14
69	70.33	81.50	91.60	73.50	88.71
100	94.67	91.60	99.83	93.56	96.65
115	115.67	111.40	107.57	113.22	111.10
132	129.33	128.40	126.00	129.11	125.92
141	142.33	142.60	141.86	142.33	141.60
154	155.33	155.60	156.71	155.33	156.80
171	168.33	170.00	172.86	169.56	172.32
180	185.00	187.40	189.29	185.78	189.80
204	204.00	206.00	210.71	205.11	210.96
228	226.33	230.00	236.86	228.56	236.72
247	255.33	261.40	268.29	257.78	262.54
291	291.67	298.80	283.00	295.56	289.27
337	339.67	316.50	298.80	331.78	304.09
391	364.00	339.67	316.50	351.83	318.32



Exercise 3.1: *Smoothers fitted to the shipments data.*

The graph on the previous page shows the five smoothers. Because moving average smoothers are "flat" at the ends, the best smoother in this case is the one with the smallest number of terms, namely the 3-MA.

3.2

$$\begin{aligned}\hat{T}_t &= \tfrac{1}{3}\Big[\tfrac{1}{5}(Y_{t-3}+Y_{t-2}+Y_{t-1}+Y_t+Y_{t+1}) \\ &\quad + \tfrac{1}{5}(Y_{t-2}+Y_{t-1}+Y_t+Y_{t+1}+Y_{t+2}) \\ &\quad + \tfrac{1}{5}(Y_{t-1}+Y_t+Y_{t+1}+Y_{t+2}+Y_{t+3})\Big] \\ &= \tfrac{1}{15}Y_{t-3}+\tfrac{2}{15}Y_{t-2}+\tfrac{1}{5}Y_{t-1}+\tfrac{1}{5}Y_t+\tfrac{1}{5}Y_{t+1}+\tfrac{2}{15}Y_{t+2}+\tfrac{1}{15}Y_{t+3}.\end{aligned}$$

3.3 **(a)** The 4 MA is designed to eliminate seasonal variation because each quarter receives equal weight. The 2 MA is designed to center the estimated trend at the data points. The combination 2 × 4 MA also gives equal weight to each quarter.

(b) $\hat{T}_t = \frac{1}{8}Y_{t-2} + \frac{1}{4}Y_{t-1} + \frac{1}{4}Y_t + \frac{1}{4}Y_{t+1} + \frac{1}{8}Y_{t+2}$.

3.4 **(a)** Use 2×4 MA to get trend. If the end-points are ignored, we obtain the following results.

Data:

	Y1	Y2	Y3	Y4
Q_1	99	120	139	160
Q_2	88	108	127	148
Q_3	93	111	131	150
Q_4	111	130	152	170

Trend:

	Y1	Y2	Y3	Y4
Q_1		110.250	129.250	150.125
Q_2		114.875	134.500	154.750
Q_3	100.375	119.635	138.875	
Q_4	105.500	124.375	145.125	

Data − trend:

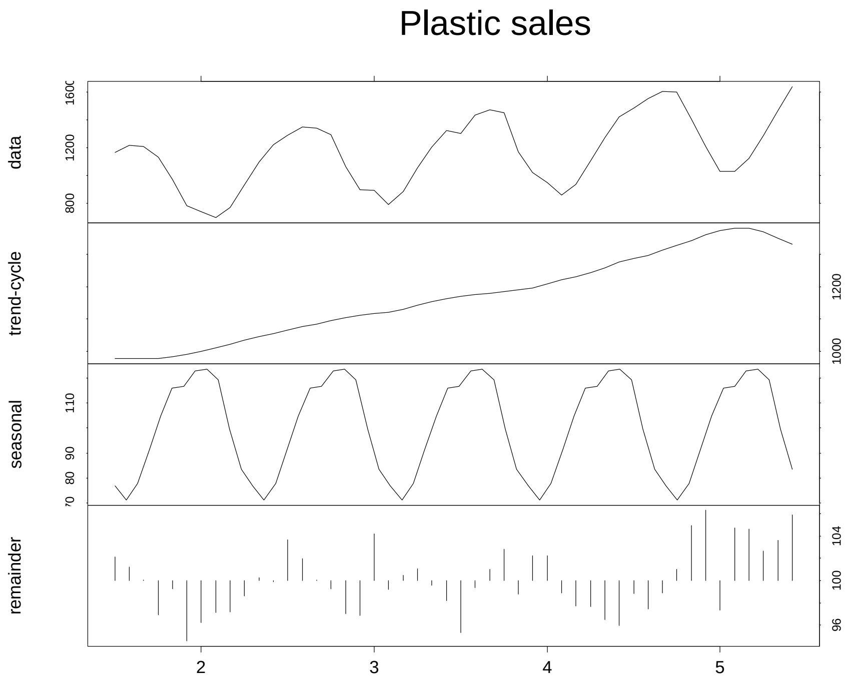
	Y1	Y2	Y3	Y4	Ave
Q_1		9.750	9.750	9.875	9.792
Q_2		–6.875	–7.500	–6.500	–7.042
Q_3	–7.375	–8.625	–8.875		–8.292
Q_4	5.500	5.625	6.875		6.000

(b) Hence, the seasonal indices are:
$\hat{S}_1 = 9.8$, $\hat{S}_2 = -7.0$, $\hat{S}_3 = -8.3$ and $\hat{S}_4 = 6.0$.
The seasonal component consists of replications of these indices.

(c) End points ignored. Other approaches are possible.

3.5 **(a)** See the top plot on the next page. There is clear trend which appears close to linear, and strong seasonality with a peak in August–October and a trough in January–March.

(b) Calculations are given at the bottom of the next page. The decomposition plot is shown at the top of the next page.



Year	J	F	M	A	M	J	J	A	S	O	N	D
Data												
1	742	697	776	898	1030	1107	1165	1216	1208	1131	971	783
2	741	700	774	932	1099	1223	1290	1349	1341	1296	1066	901
3	896	793	885	1055	1204	1326	1303	1436	1473	1453	1170	1023
4	951	861	938	1109	1274	1422	1486	1555	1604	1600	1403	1209
5	1030	1032	1126	1285	1468	1637	1611	1608	1528	1420	1119	1013
2×12 MA Trend												
1							977.0	977.0	977.1	978.4	982.7	990.4
2	1000.5	1011.2	1022.3	1034.7	1045.5	1054.4	1065.8	1076.1	1084.6	1094.4	1103.9	1112.5
3	1117.4	1121.5	1130.7	1142.7	1153.6	1163.0	1170.4	1175.5	1180.5	1185.0	1190.2	1197.1
4	1208.7	1221.3	1231.7	1243.3	1259.1	1276.6	1287.6	1298.0	1313.0	1328.2	1343.6	1360.6
5	1374.8	1382.2	1381.2	1370.6	1351.2	1331.2						
Ratios												
1							119.2	124.5	123.6	115.6	98.8	79.1
2	74.1	69.2	75.7	90.1	105.1	116.0	121.0	125.4	123.6	118.4	96.6	81.0
3	80.2	70.7	78.3	92.3	104.4	114.0	111.3	122.2	124.8	122.6	98.3	85.5
4	78.7	70.5	76.2	89.2	101.2	111.4	115.4	119.8	122.2	120.5	104.4	88.9
5	74.9	74.7	81.5	93.8	108.6	123.0						
Seasonal indices												
Ave	77.0	71.3	77.9	91.3	104.8	116.1	116.8	122.9	123.6	119.3	99.5	83.6

Exercise 3.5(a) and (b): *Multiplicative classical decomposition of plastic sales data.*

(c) The trend does appear almost linear except for a slight drop at the end. The seasonal pattern is as expected. Note that it does not make much difference whether these data are analyzed using a multiplicative decomposition or an additive decomposition.

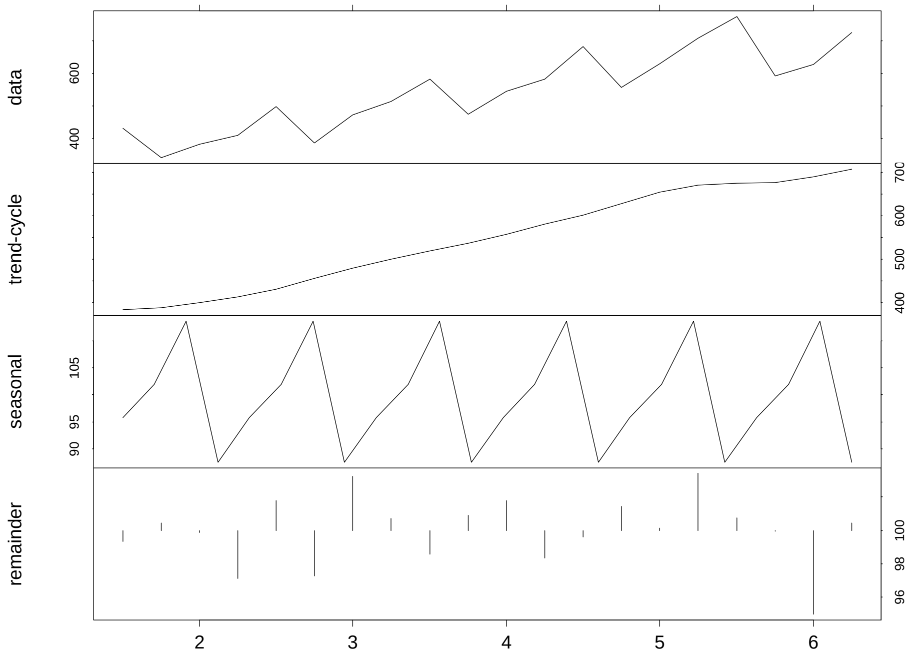
3.6

Period t	Trend T_t	Seasonal S_t	Forecast $\hat{Y}_t = T_t S_t/100$
61	1433.96	76.96	1103.6
62	1442.81	71.27	1028.3
63	1451.66	77.91	1131.0
64	1460.51	91.34	1334.0
65	1469.36	104.83	1540.3
66	1478.21	116.09	1716.1
67	1487.06	116.76	1736.3
68	1495.91	122.94	1839.1
69	1504.76	123.55	1859.1
70	1513.61	119.28	1805.4
71	1522.46	99.53	1515.3
72	1531.31	83.59	1280.0

3.7 **(a)** See the top of the figure on the previous page.

(b) The calculations are given below.

Year	Q1	Q2	Q3	Q4
Data				
1	362	385	432	341
2	382	409	498	387
3	473	513	582	474
4	544	582	681	557
5	628	707	773	592
6	627	725	854	661
4×2 MA				
1			382.5	388.0
2	399.3	413.3	430.4	454.8
3	478.3	499.6	519.4	536.9
4	557.9	580.6	601.5	627.6
5	654.8	670.6	674.9	677.0
6	689.4	708.1		
Ratios				
1			112.9	87.9
2	95.7	99.0	115.7	85.1
3	98.9	102.7	112.1	88.3
4	97.5	100.2	113.2	88.7
5	95.9	105.4	114.5	87.4
6	91.0	102.4		
Seasonal indices				
Ave	95.8	101.9	113.7	87.5



Exercise 3.7: *Decomposition plot for exports from French company.*

(c) Multiplicative decomposition seems appropriate here because the variance is increasing with the level of the series. The most interesting feature of the decomposition is that the trend has levelled off in the last year or so. Any forecast method should take this change in the trend into account.

3.8 **(a)** The top plot shows the original data followed by trend-cycle, seasonal and irregular components. The bottom plot shows the seasonal sub-series.

(b) The trend-cycle is almost linear and the small seasonal component is very small compared to the trend-cycle. The seasonal pattern is difficult to see in time plot of original data. Values are high in March, September and December and low in January and August. For the last six years, the December peak and March peak have been almost constant. Before that, the December peak was growing and the March peak was dropping. There are several possible outliers in 1991.

(c) The recession is seen by several negative outliers in the irregular component. This is also apparent in the data time plot. Note: the recession could be made part of the trend-cycle component by reducing the span of the loess smoother.

3.9 **(a) and (b)** Calculations are given below. Note that the seasonal indices are computed by averaging the de-trended values within each half-year.

Data	2×2 MA Trend	Detrended Data	Seasonal Component	Seasonal Adjusted Data
1.09			0.017	1.073
1.07	1.0825	-0.0125	-0.014	1.084
1.10	1.0825	0.0175	0.017	1.083
1.06	1.0750	-0.0150	-0.014	1.074
1.08	1.0625	0.0175	0.017	1.063
1.03	1.0450	-0.0150	-0.014	1.044
1.04	1.0300	0.0100	0.017	1.023
1.01	1.0225	-0.0125	-0.014	1.024
1.03	1.0075	0.0225	0.017	1.013
0.96			-0.014	0.974

(c) With more data, we could take moving averages of the detrended values for each half-year rather than a simple average. This would result in a seasonal component which changed over time.

Chapter 4: Exponential smoothing methods

4.1

		Period	Data	MA(3)		SES($\alpha = 0.7$)	
		t	Y_t	$\hat{Y}_t$	E_t	$\hat{Y}_t$	E_t
1974	1	1	5.4				
	2	2	5.3			5.40	-0.10
	3	3	5.3			5.33	-0.03
	4	4	5.6	5.33	0.27	5.31	0.29
1975	1	5	6.9	5.40	1.50	5.51	1.39
	2	6	7.2	5.93	1.27	6.48	0.72
	3	7	7.2	6.57	0.63	6.99	0.21
	4	8		7.10		7.14	
Accuracy statistics from period 4 through 7							
ME					0.92		0.65
MAE					0.92		0.65
MAPE					13.22		9.56
MSE					1.08		0.64
Theil's U					1.40		1.14

Theil's U statistic suggests that the naïve (or last value) method is better than either of these. If SES is used with an optimal value of α chosen, then $\alpha = 1$ is selected. This is equivalent to the naïve method. Note different packages may give slightly different results for SES depending on how they initialize the method. Some packages will also allow $\alpha > 1$.

4.2 **(a)** Forecasts for May 1992

Method	MA(3)	MA(5)	MA(7)	MA(9)	MA(11)
Forecast	24.0	48.6	55.6	51.7	53.1
MSE	1484.3	1031.2	757.5	860.8	1313.8

(b) Forecasts for May 1992

Method	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$
Forecast	45.5	41.9	33.1	29.1	28.7
MSE	1421.35	1211.80	1193.98	1225.40	1298.49

(c) Of these forecasting methods, the best MA(k) method has $k = 7$ and the best SES method has $\alpha = 0.5$. However, it should be noted that the MSE values for the MA methods are taken over different periods. For example, the MSE for the MA(7) method is computed only over 9 observations because it is not possible to compute an MA(7) forecast for the first seven observations. So the MSE values are not strictly comparable for the MA forecasts. It would be better to use a holdout sample but there are too few data.

4.3 Optimizing α for SES over the period 3 through 10:

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
MAPE	65.60	53.46	44.43	37.60	32.32	28.16	24.82	22.08	19.80	17.86
MSE	79.34	47.24	29.95	20.10	14.17	10.41	7.91	6.17	4.92	4.00

The optimal value is $\alpha = 1$.

With Holt's method, any combination of α and β will give MAPE=0. This is so because the differences between successive values of (4.13) are always going to be zero with this errorless series. Using $\alpha = 1$ for SES and $\alpha = 0.5$ and $\beta = 0.5$ for Holt's method gives the following results.

Data	SES		Holt's	
Y_t	$\hat{Y}_t$	E_t	$\hat{Y}_t$	E_t
2				
4	2	2	4	0
6	4	2	6	0
8	6	2	8	0
10	8	2	10	0
12	10	2	12	0
14	12	2	14	0
16	14	2	16	0
18	16	2	18	0
20	18	2	20	0

(a) Clearly Holt's method is better as it allows for the trend in the data.

(b) For SES, $\alpha = 1$. Because of the trend, the forecasts will always lag behind the actual values so that the forecast errors will always be at least 2. Choosing $\alpha = 1$ makes the forecast errors as small as possible for SES.

(c) See above.

4.4 **(a) (b) and (c)** See the table on the following page.

(d) There's not much to choose between these methods. They are both bad! Look at Theil's U values for instance. The last value method over the same period (13–28) gives MSE=6.0, MAPE=2.05 and Theil's U=1.0.

Period t	Data Y_t	Forecast MA(12)	Errors E_t	Forecast MA(6)	Errors E_t
1	108				
2	108				
3	110				
4	106				
5	108				
6	108				
7	105			108.00	-3.00
8	100			107.50	-7.50
9	97			106.17	-9.17
10	95			104.00	-9.00
11	95			102.17	-7.17
12	92			100.00	-8.00
13	95	102.67	-7.67	97.33	-2.33
14	95	101.58	-6.58	95.67	-0.67
15	98	100.50	-2.50	94.83	3.17
16	97	99.50	-2.50	95.00	2.00
17	101	98.75	2.25	95.33	5.67
18	104	98.17	5.83	96.33	7.67
19	101	97.83	3.17	98.33	2.67
20	99	97.50	1.50	99.33	-0.33
21	95	97.42	-2.42	100.00	-5.00
22	95	97.25	-2.25	99.50	-4.50
23	96	97.25	-1.25	99.17	-3.17
24	96	97.33	-1.33	98.33	-2.33
25	97	97.67	-0.67	97.00	0.00
26	98	97.83	0.17	96.33	1.67
27	94	98.08	-4.08	96.17	-2.17
28	92	97.75	-5.75	96.00	-4.00
29		97.33		95.50	
Accuracy criteria: periods 13–28					
ME			-1.51		-0.10
MAE			3.12		2.96
MSE			14.40		12.64
MAPE			3.23		3.03
Theil's U			1.58		1.45

Calculations for Exercise 4.4

4.5 (a) (b) and (c)

	Paperbacks		Hardcovers	
	SES	Holt	SES	Holt
Smoothing parameters	$\alpha = 0.213$	$\alpha = 0.335$	$\alpha = 0.347$	$\beta = 0.437$
		$\beta = 0.453$		$\beta = 0.157$
Forecast Day 31	210.15	224.24	240.38	250.73
Forecast Day 32	210.15	231.79	240.38	254.63
Forecast Day 33	210.15	239.33	240.38	258.53
Forecast Day 34	210.15	246.88	240.38	262.43
MAE	29.6	33.9	27.3	28.6
MSE	1252.2	1701.7	1060.6	1273.0
MAPE	17.1	18.4	13.5	14.3
Theil's U	0.68	0.92	0.81	0.92

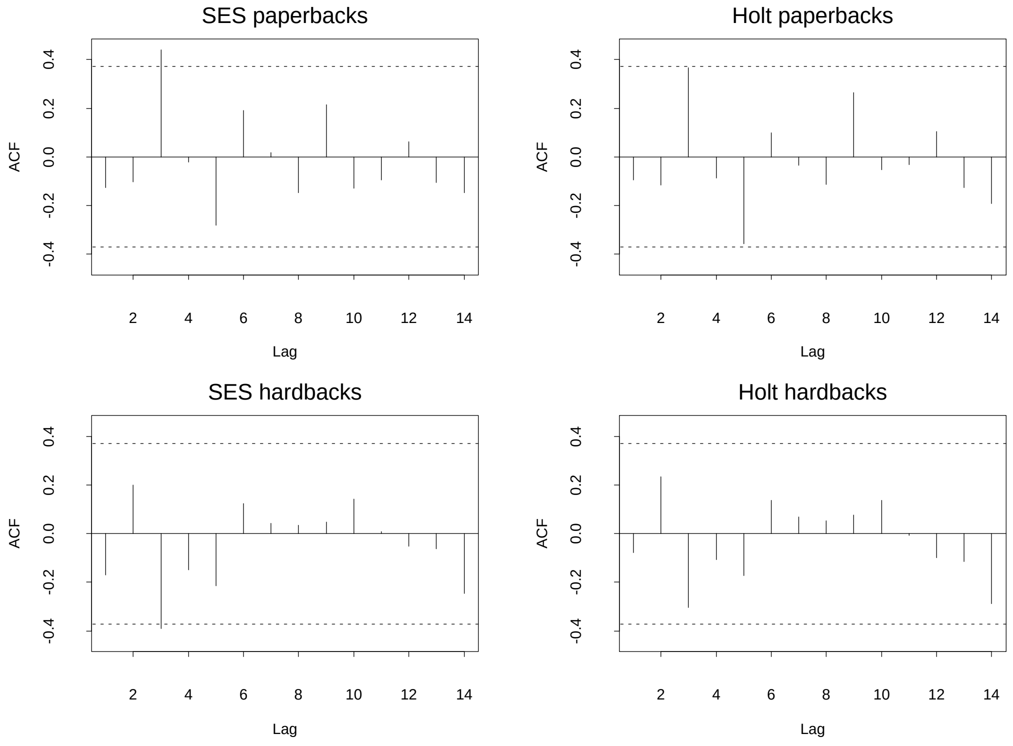
For both series, SES forecasting is performing better than Holt's method.

(d) SES forecasts are "flat" and Holt's forecasts show a linear trend. Both series show an upward linear trend and we would expect the forecasts to reflect that trend. Perhaps an out-of-sample analysis would give a better indication of the merits of the two methods.

(e) The autocorrelation functions of the forecast errors in each case are plotted on the next page. In each case, there is no noticeable pattern. Only a few spikes are just outside the critical bounds which is expected.

4.6 Here is a complete run for one set of values ($\beta = 0.1$ and $\alpha_1 = 0.1$). Note that in this program we have chosen to make the first three values of α be equal to the starting value. This is not crucial, but it does make a difference.

t	Y_t	F_t	E_t	A_t	M_t	α_t
1	200.0	200.00	0.00	0.00	0.00	0.100
2	135.0	200.00	-65.00	-6.50	6.50	0.100
3	195.0	193.50	1.50	-5.70	6.00	0.100
4	197.5	193.65	3.85	-4.74	5.79	0.950
5	310.0	197.31	112.69	7.00	16.48	0.820
6	175.0	289.74	-114.74	-5.18	26.30	0.425
7	155.0	241.00	-86.00	-13.26	32.27	0.197
8	130.0	224.08	-94.08	-21.34	38.45	0.411
9	220.0	185.43	34.57	-15.75	38.06	0.555
10	277.5	204.62	72.88	-6.89	41.55	0.414
11	235.0	234.77	0.23	-6.17	37.41	0.166
12		234.81				0.165



Exercise 4.5 (e): *Autocorrelation functions of forecast errors.*

For other combinations of values for β and starting values for α, here is what the final α value is:

	$\beta = 0.1$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.7$
$\alpha = 0.1$	0.165	0.327	0.732	0.143
$\alpha = 0.2$	0.058	0.454	0.783	0.133
$\alpha = 0.3$	0.140	0.618	0.797	0.133

The time series is not very long and therefore the results are somewhat fickle. In any event, it is clear that the β value and the starting values for α have a profound effect on the final value of α.

4.7 Holt-Winters' method is best because the data are seasonal. The variation increases with the level, so we use Holt-Winters' multiplicative method. The optimal smoothing parameters (giving smallest MSE) are $a = 0.479$, $b = 0.00$ and $c = 1.00$. These give the following forecasts (read left to right):

309.1	312.1	315.2	318.3	321.3	324.4
327.5	330.5	333.6	336.7	339.7	342.8
345.9	348.9	352.0	355.0	358.1	361.2
364.2	367.3	370.4	373.4	376.5	379.6

4.8 First choose any values for the three parameters. Here we have used $\alpha = \beta = \gamma = 0.1$. Different values will choose different initial values. Our program uses the method described in the textbook and gave the following results:

ME	MAE	MSE	MAPE	r_1	Theil's U
-240.5	240.5	62469.9	37.5	0.70	2.6

Now compare with the optimal values: $\alpha = 0.917$, $\beta = 0.234$ and $\gamma = 0.000$. Using the same initialization, we obtain the results in Table 4-11, namely

ME	MAE	MSE	MAPE	r_1	Theil's U
-9.46	24.00	824.75	3.75	0.17	0.29

Chapter 5: Simple regression

5.1 **(a)** -0.7, almost 1, 0.2

(b) False. The correlation is negative. So below-average values of one are associated with above-average values of the other variable.

(c) Wages have been increasing over time due to inflation. At the same time, population has been increasing and consequently, new houses need to be built. So, because they are both increasing with time, they are positively correlated.

(d) There are many factors affecting unemployment and it is simplistic to draw a causal connection with inflation on the basis of correlation. As in the previous question, both vary with time and the correlation could be induced by their time trends. Or they could both be related to some third variable such as business confidence or government spending.

(e) The older people in the survey had much less opportunity for education than the younger people. This negative correlation is caused by the increase in education levels over time.

5.2 **(a)** $\bar{X} = 5$, $\bar{Y} = 25$, $\sum(X_i - \bar{X})^2 = 20$, $\sum(X_i - \bar{X})(Y_i - \bar{Y}) = 78$. So $b = 78/20 = 3.9$ and $a = 25 - 3.9(5) = 5.5$. Hence, the regression line is $\hat{Y} = 5.5 + 3.9X$.

(b) $\sum(\hat{Y}_i - \bar{Y})^2 = 304.20$, $\sum(Y_i - \hat{Y}_i)^2 = 25.80$, $\sigma_e = \sqrt{25.80/(5-2)} = 2.933$. So

$$F = \frac{304.20/(2-1)}{25.80/(5-2)} = 35.4.$$

This has $(2-1) = 1$ df for the numerator and $(5-2) = 3$ df for the denominator. From Table C in Appendix III, the P-value is slightly smaller than 0.010. (Using a computer, it is 0.0095.) Standard errors:

$$\begin{aligned} \text{s.e.}(a) &= (2.93)\sqrt{\tfrac{1}{5} + \tfrac{25}{20}} = 3.53 \\ \text{s.e.}(b) &= (2.93)\sqrt{\tfrac{1}{20}} = 0.656. \end{aligned}$$

On 3 df, $t^* = 3.18$ for a 95% confidence interval. Hence 95% intervals are

$$\begin{aligned} \alpha: &\quad 5.500 \pm 3.18(3.53) = [-5.7, 16.7] \\ \beta: &\quad 3.900 \pm 3.18(0.656) = [1.8, 6.0] \end{aligned}$$

Output from Minitab for Exercise **5.2**:

```
Regression Analysis

The regression equation is
Y = 5.50 + 3.90X

Predictor       Coef       StDev          T        P
Constant       5.500       3.531       1.56    0.217
X             3.9000      0.6557       5.95    0.010

S = 2.933       R-Sq = 92.2%     R-Sq(adj) = 89.6%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    1      304.20      304.20     35.37    0.010
Error         3       25.80        8.60
Total         4      330.00
```

(c) $R^2 = 0.922$, $r_{XY} = r_{Y\hat{Y}} = 0.960$.

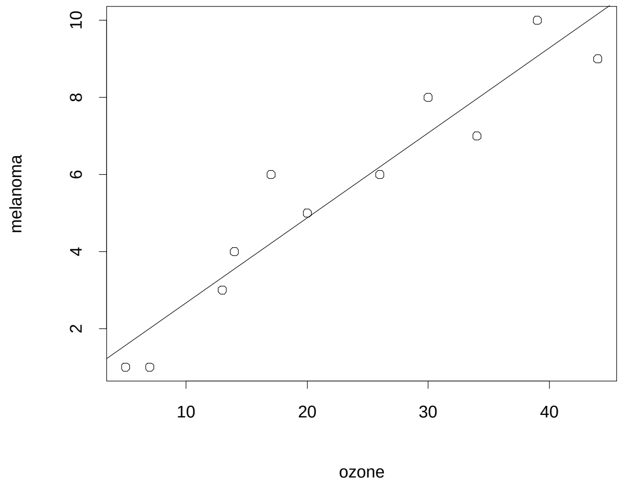
(d) The line through the middle of the graph is the line of best fit. The 95% prediction interval shown is the interval which would contain the Y value with probability 0.95 if the X value was 17. The 80% prediction interval shown is the interval which would contain the Y value with probability 0.80 if the X value was 26. The dotted line at the boundary of the light shaded region gives the ends of all the 95% prediction intervals. The dotted line at the boundary of the dark shaded region gives the ends of all the 80% prediction intervals.

5.3 **(a)** See the plot on the next page and the Minitab output on page 101. The straight line is $\hat{Y} = 0.46 + 0.22X$.

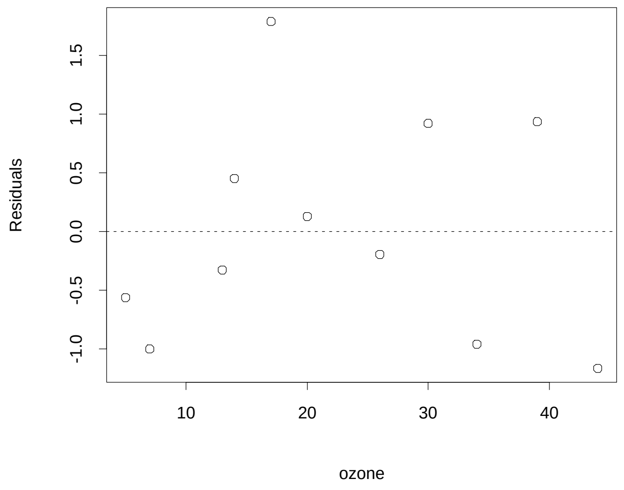
(b) See the plot on the next page. The residuals may show a slight curvature (Λ shaped). However, the curvature is not strong and the fitted model appears reasonable.

(c) $R^2 = 90.2\%$. Therefore, 90.2% of the variation in melanoma rates is explained by the linear regression.

(d) From the Minitab output:
Prediction: 9.286. Prediction interval: (6.749, 11.823)



Exercise 5.3(a): *Scatterplot of melanoma rate against ozone depletion.*



Exercise 5.3(b): *Scatterplot of residuals from the linear regression.*

Output using Minitab for Exercise **5.3**:

```
MTB > Regress 'Melanoma' 1 'Ozone';
SUBC> predict 40.

The regression equation is
Melanoma = 0.460 + 0.221 Ozone

Predictor       Coef       StDev          T        P
Constant      0.4598      0.6258       0.73    0.481
Ozone        0.22065     0.02426       9.09    0.000

S = 0.9947      R-Sq = 90.2%     R-Sq(adj) = 89.1%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    1      81.822      81.822     82.70    0.000
Error         9       8.905       0.989
Total        10      90.727

     Fit  StDev Fit       95.0% CI            95.0% PI
   9.286      0.517   (   8.116,  10.456)  (   6.749,  11.823)
```

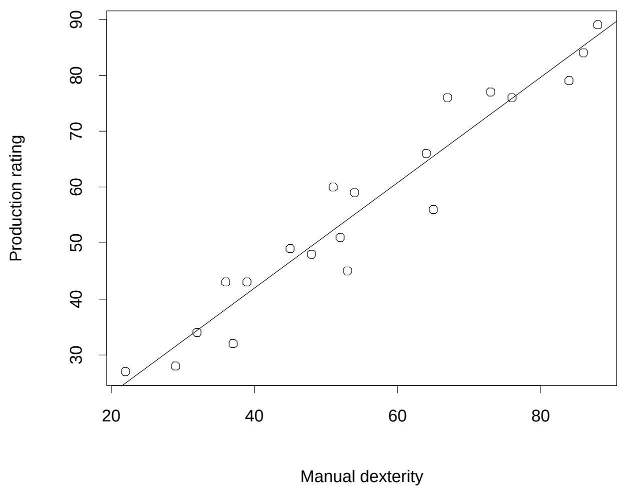
Note that it is the prediction interval (PI) we want here. Minitab also gives the confidence interval (CI) for the line at this point, something we have not covered in the book.

(e) This analysis has assumed that the susceptibility to melanoma among people living in the various locations is constant. This is unlikely to be true due to the diversity of racial mix and climate over the locations. Apart from ozone depletion, melanoma will be affected by skin type, climate, culture (e.g. is sun-baking encouraged?), diet, etc.

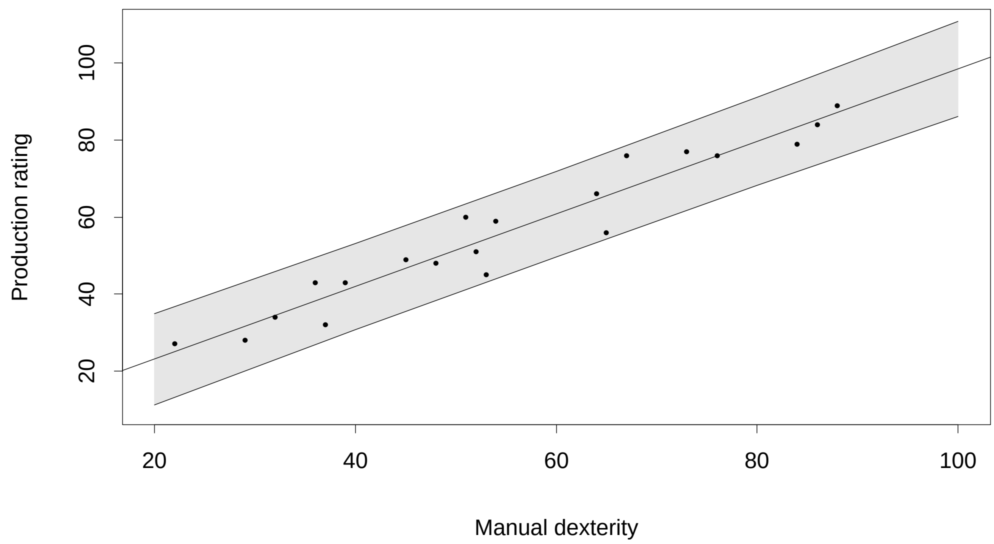
5.4 **(a)** See plot on the next page and computer output on page 103.

(b) Coefficients: $a = 4.184$, $b = 0.9431$. Only b is significant, showing the relationship is significant. (We could refit the model without the intercept term.)

(c) If $X = 80$, $\hat{Y} = 4.184 + 0.9431(80) = 79.63$. Standard error of forecast is 1.88 (from computer output).



Exercise 5.4(a): *Scatterplot of production rating against manual dexterity test scores.*



Exercise 5.4(e): *95% prediction intervals for production rating.*

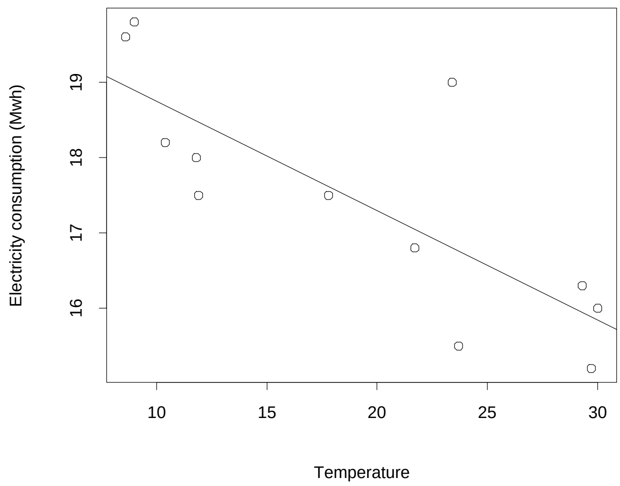
Output using Minitab for Exercise **5.4**

```
MTB > Regress 'Y' 1 'X';
SUBC> Predict 'newX'.

The regression equation is
Y = 4.18 + 0.943 X

Predictor       Coef       StDev          T        P
Constant       4.184       3.476       1.20    0.244
X            0.94306     0.05961      15.82    0.000

S = 5.126       R-Sq = 93.3%     R-Sq(adj) = 92.9%

Analysis of Variance

Source        DF          SS          MS         F         P
Regression     1      6576.8      6576.8    250.29     0.000
Error         18       473.0        26.3
Total         19      7049.8

  Fit  StDev Fit       95.0% CI              95.0% PI
23.05       2.38   (   18.04,   28.05)  (   11.17,   34.92)
41.91       1.46   (   38.85,   44.97)  (   30.71,   53.10)
60.77       1.18   (   58.28,   63.26)  (   49.71,   71.82)
79.63       1.88   (   75.68,   83.58)  (   68.16,   91.10)
```

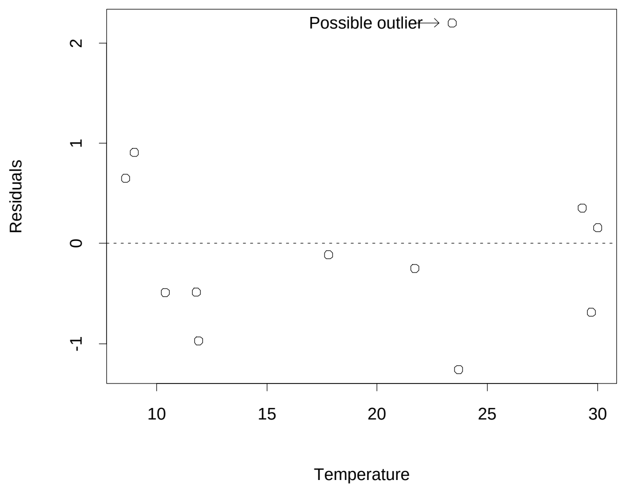
(d) For confidence and prediction intervals, use Table B with 18 df. 95% CI for β is $0.94306 \pm 2.10(0.05961) = [0.82, 1.07]$.

(e) See output. Again it is the prediction interval (PI) we want here, not the confidence interval (CI). The prediction intervals are shown in the plot on the previous page.

5.5 **(a)** See the plot on the following page. The straight line regression model is $\hat{Y} = 20.2 - 0.145X$ where Y = electricity consumption and X = temperature. There is a negative relationship because heating is used for lower temperatures, but there is no need to use heating for the higher temperatures. The temperatures are not sufficiently high to warrant the use of air conditioning. Hence, the electricity consumption is higher when the temperature is lower.



Exercise 5.5(a): *Electricity consumption (Mwh) plotted against temperature (degrees Celsius).*



Exercise 5.5(c): *Residual plot for the straight line regression of electricity consumption against temperature.*

(b) $r = -0.791$

(c) See the plot on the previous page. Apart from the possible outlier, the model appears to be adequate. There are no highly influential observations.

(d) If $X = 10$, $\hat{Y} = 20.2 - 0.145(10) = 18.75$. If $X = 35$, $\hat{Y} = 20.2 - 0.145(35) = 15.12$. The first of these predictions seems reasonable. The second is unlikely. Note that $X = 35$ is outside the range of the data making prediction dangerous. For temperatures above about 20°C, it is unlikely electricity consumption would continue to fall because no heating would be used. Instead, at high temperatures (such as $X = 35$°C), electricity consumption is likely to increase again due to the use of air-conditioning.

5.6 (a) When $H = 130$ and $W = 45$, $r = 0.553$.

(b) When $H = 40$ and $W = 150$, $r = -0.001$.

(c) The following table shows the influence of outliers at various positions.

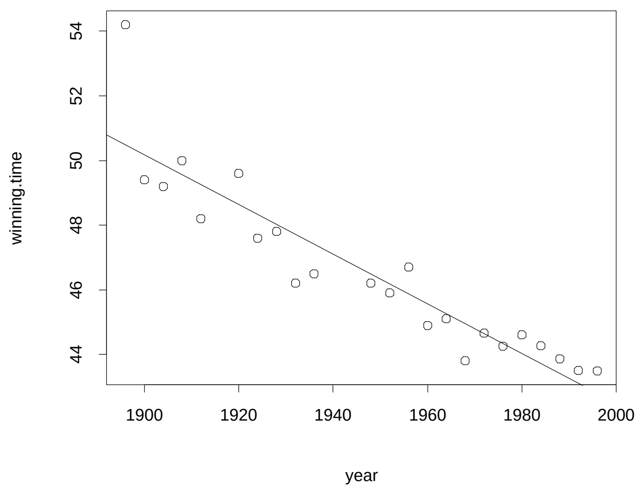
H	W	r
129	0	-0.393
128	22	0.032
122	44	0.527
112	64	0.773
99	83	0.846
83	99	0.810
65	112	0.627
44	122	0.151
22	128	-0.365
0	129	-0.624

The point about all this is that an outlier (and skewness in general) can seriously affect the correlation coefficient. It is a good idea to look at the scatterplot before computing any correlation.

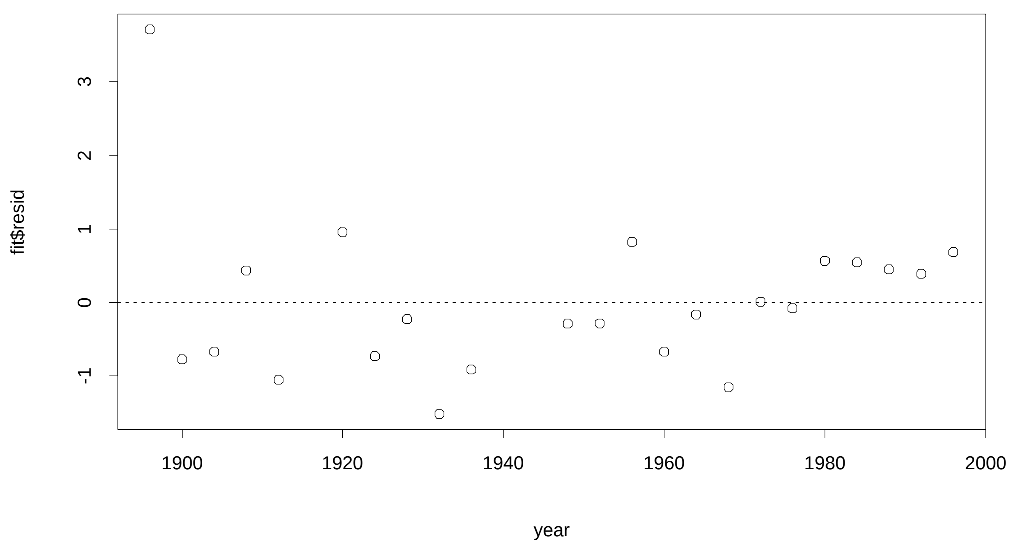
5.7 (a) See the plot on the next page. The winning time has been decreasing with year. There is an outlier in 1896.

(b) The fitted line is $\hat{Y} = 196 - 0.0768X$ where X denotes the year of the Olympics. Therefore the winning time has been decreasing an average 0.0768 seconds per year.

(c) The residuals are plotted on the next page. The residuals show random scatter about 0 with only one usual point (the outlier in 1896). But note that the last five residuals are positive. This suggests that the straight line is "levelling out"—the winning time is decreasing at a slower rate now than it was earlier.



Exercise 5.7(a): *Scatterplot of winning times against year.*



Exercise 5.7(c): *Residual plot for linear regression model of winning times.*

(d) The predicted winning time in the 2000 Olympics is

$$\hat{Y} = 196 - 0.0768(2000) = 42.50 \text{ seconds.}$$

This would smash the world record. But given the previous five results (with positive residuals), it would seem more likely that the actual winning time would be higher. A prediction interval is

$$42.50 \pm 2.0796(1.1762) = 42.50 \pm 2.45 = [40.05, 44.95].$$

5.8 **(a)** There is strong seasonality with peaks in November and December and a trough in January. The surfing festival shows as a smaller peak in March from 1988. The variation in the series is increasing with the level and there is a strong positive trend due to sales growth.

(b) Logarithms are necessary to stabilize the variance so it does not increase with the level of the series.

(c) See the plot on the next page and the computer output on page 109. The fitted line is $\hat{Y} = -526.57 + 0.2706X$ where X is the year and Y is the logged annual sales.

(d)

$$\begin{aligned} X = 1994: &\quad \hat{Y} = -526.57 + 0.2706(1994) = 12.98 \\ X = 1995: &\quad \hat{Y} = -526.57 + 0.2706(1995) = 13.25 \\ X = 1996: &\quad \hat{Y} = -526.57 + 0.2706(1996) = 13.52 \end{aligned}$$

Prediction intervals (from computer output):

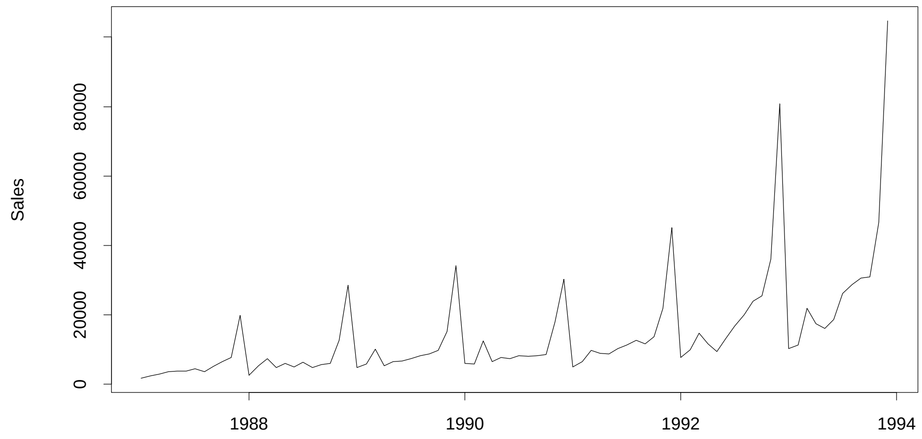
$$\begin{aligned} X = 1994: &\quad [12.57, 13.40] \\ X = 1995: &\quad [12.80, 13.71] \\ X = 1996: &\quad [13.03, 14.02] \end{aligned}$$

(e) We transform the forecasts and intervals with the exponential function:

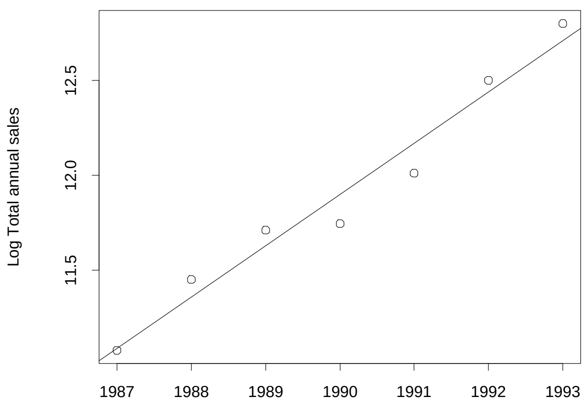
Total annual sales for 1994	$\exp(12.98) = \$434,443$
Total annual sales for 1995	$\exp(13.25) = \$569,439$
Total annual sales for 1996	$\exp(13.52) = \$746,383$

Prediction intervals:

$$\begin{aligned} X = 1994: &\quad [e^{12.57}, e^{13.40}] = [286673, 658385] \\ X = 1995: &\quad [e^{12.80}, e^{13.71}] = [361994, 895764] \\ X = 1996: &\quad [e^{13.03}, e^{14.02}] = [455060, 1224208] \end{aligned}$$



Exercise 5.8(a): *Time plot of sales figures.*



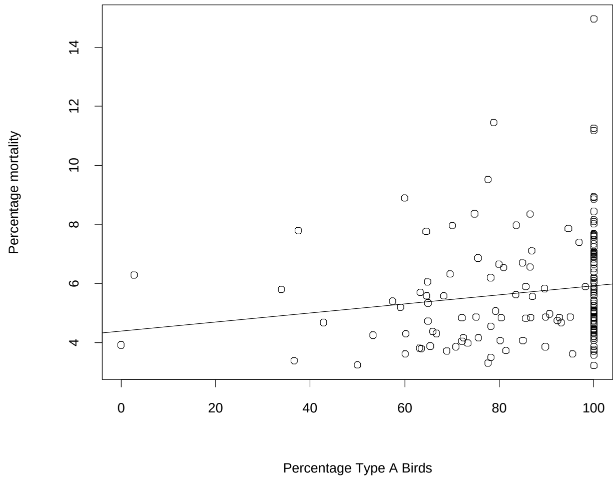
Exercise 5.8(c): *Regression line fitted to the logged sales data.*

Output using Minitab for Exercise **5.8**:

```
MTB > regress 'Log Sales' 1 'Year';
SUBC> predict 'new years';

The regression equation is
Log Sales = - 527 + 0.271 Year

Predictor       Coef       StDev          T        P
Constant     -526.57       46.44     -11.34    0.000
Year         0.27059     0.02334      11.60    0.000

S = 0.1235      R-Sq = 96.4%     R-Sq(adj) = 95.7%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    1      2.0501      2.0501    134.45    0.000
Error         5      0.0762      0.0152
Total         6      2.1263

     Fit  StDev Fit       95.0% CI            95.0% PI
 12.9818     0.1044   ( 12.7135, 13.2502)  ( 12.5661, 13.3975)
 13.2524     0.1257   ( 12.9293, 13.5755)  ( 12.7994, 13.7054) X
 13.5230     0.1476   ( 13.1435, 13.9025)  ( 13.0282, 14.0178) X
X  denotes a row with X values away from the center
```

These prediction intervals are very wide because we are only using annual totals in making these predictions. A more accurate method would be to fit a model to the monthly data allowing for the seasonal patterns. This is discussed in Chapter 7.

(f) One way would be to calculate the proportion of sales for each month compared to the total sales for that year. Averaging these proportions will give a rough guide as to how to split the annual totals into 12 monthly totals.



Exercise 5.9(a): *Scatterplot of percentage mortality against percentage of Type A birds.*

5.9 **(a)** The plot is shown above. The fitted line is

$$\hat{Y} = 4.38 + 0.0154X$$

where X = percentage of type A birds and Y = percentage mortality.

(b) From the computer output:

Predictor	Coef	StDev	T	P
Constant	4.3817	0.6848	6.40	0.000
% Type A	0.015432	0.007672	2.01	0.046

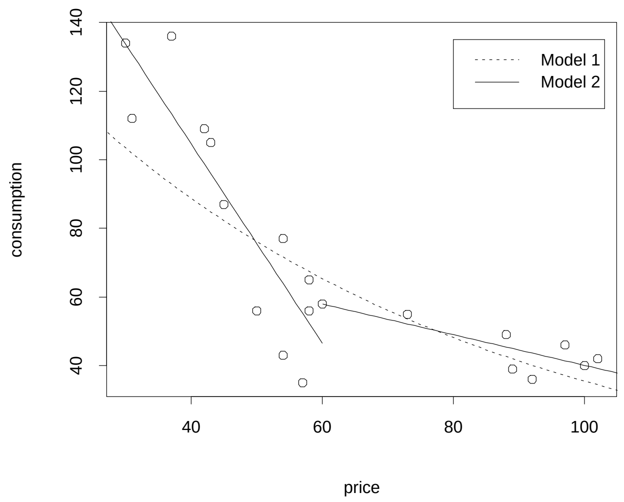
So the t-test is significant (since $P < 0.05$). A 95% confidence interval for the slope is

$$0.01543 \pm 1.976(0.007672) = 0.01543 \pm 0.01516 = [0.003, 0.031].$$

This suggests that the Type A birds have a higher mortality than the Type B birds, the opposite to what the farmers claim.

(c) For a farmer using all Type A birds, $X = 100$. So $\hat{Y} = 4.38 + 0.0154(100) = 5.92\%$. For a farmer using all Type B birds, $X = 0$. So $\hat{Y} = 4.38\%$. Prediction intervals for these are [2.363, 9.487] and [0.587, 8.177] respectively.

(d) $R^2 = 2.6$. So only 2.6% of the variation in mortality is due to bird type.



Exercise 5.10(b): *Scatterplot of gas consumption against price.*

(e) This information suggests that heat may be a lurking variable. If Type A birds are being used more in summer and the mortality is higher in summer, than the increased mortality of Type A birds may be due to the summer rather than the bird type. A proper randomized experiment would need to be done to properly assess whether bird type is having an effect here.

5.10 **(a)** Cross sectional data. There is no time component.

(b) See the plot above.

(c) When the price is higher, the consumption may be lower due to the pressure of increased cost. Therefore, we would expect $b_1 < b_2 < 0$.

(d) Model 1: First take logarithms of Y_i, then use simple linear regression to obtain

$$a = 5.10, \quad b = -0.0153, \quad \sigma_e^2 = 0.0735.$$

Model 2: Split data into two groups. Fit each group separately using simple linear regression to obtain

$$a_1 = 221, \quad b_1 = -2.91 \qquad \text{and} \qquad a_2 = 84.8, \quad b_2 = -0.447.$$

Using the equation given in the question, we obtain

$$\sigma_e^2 = 2913.7/16 = 182.06.$$

The fitted lines are shown on the graph above.

(e) Model 1: $R^2 = r^2_{Y\hat{Y}_1} = 0.721$.
Model 2: $R^2 = r^2_{Y\hat{Y}_2} = 0.859$. The second model is better with higher R^2 value. The residual plots are given on the following page. Again, the second model is much better showing random scatter about zero. The first model show pattern in the residuals.

(f) The graph on page 114 shows a local linear regression through the data. The fitted curve resembles the fitted lines for model 2. This suggests that model 2 is a reasonable model for the data. However, our approach has also meant the two lines do not join at $X = 60$. A better model would force them to join. This means the parameters must be restricted which makes the estimation much harder.

(g) and (h) Using model 2, forecasts are obtained by

$$\hat{Y} = \begin{cases} 220.9 - 2.906X & \text{when } X \leq 60 \\ 84.8 - 0.447X & \text{when } X > 60. \end{cases}$$

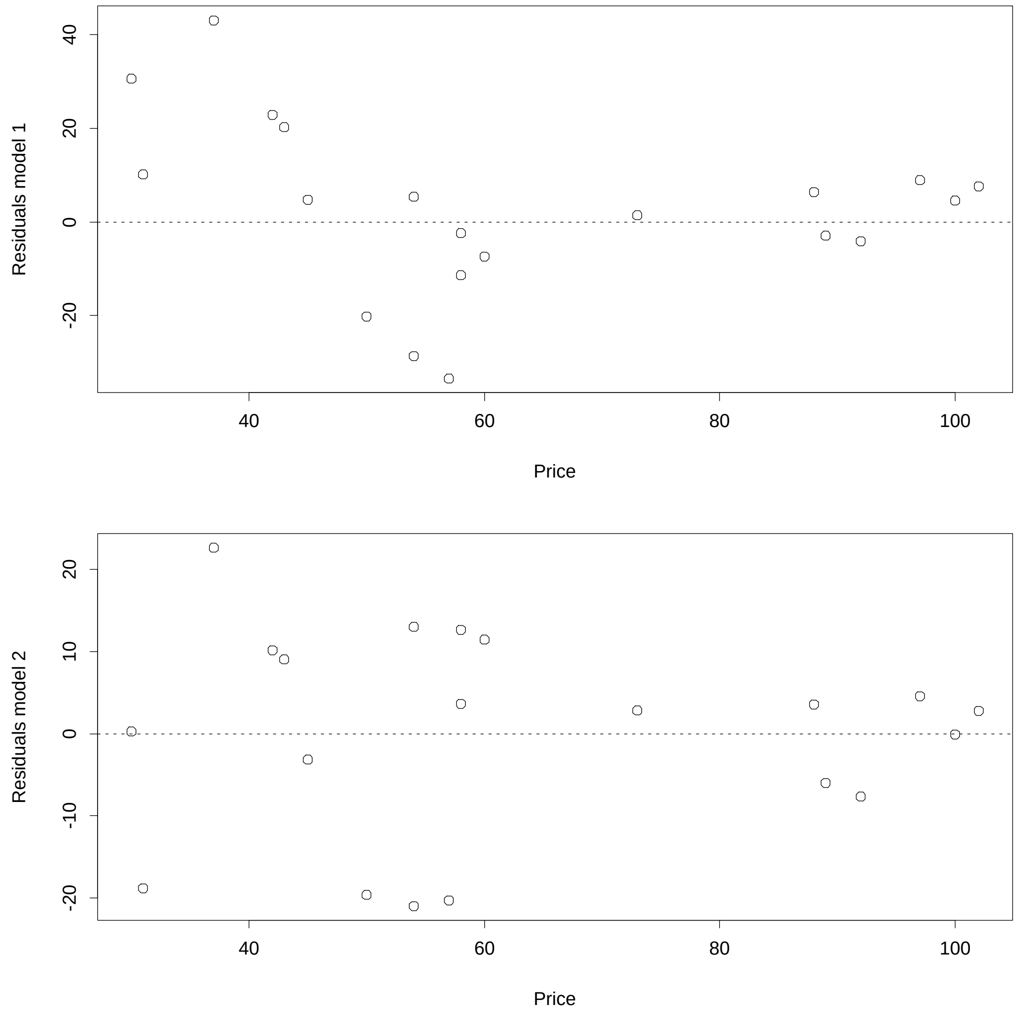
and standard errors are obtained from (5.19):

$$\text{s.e.}(\hat{Y}) = \sqrt{182.06}\sqrt{1 + \frac{1}{20} + \frac{(X-63)^2}{10672.11}}.$$

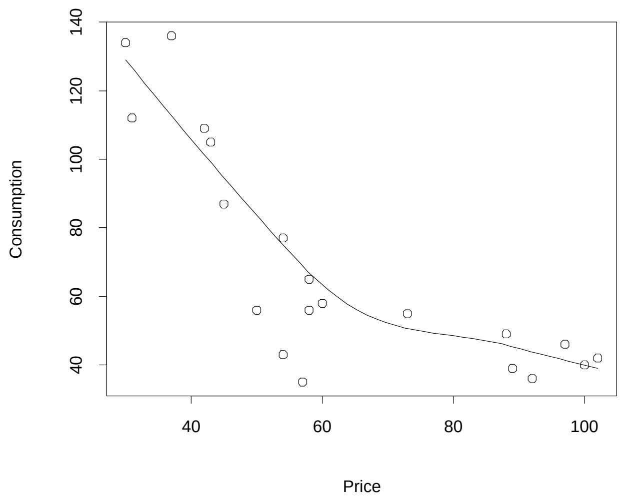
The 95% PI are obtained using $\hat{Y} \pm t^*(\text{s.e.})$ where $t^* = 2.12$ (from Table B with 16 df). Hence, we obtain the following values.

X	$\hat{Y}$	s.e.	[95% PI]
40	104.67	14.15	[74.7 , 134.7]
60	46.55	13.83	[17.2 , 75.9]
80	49.03	14.00	[19.3 , 78.7]
100	40.09	14.65	[9.0 , 71.1]
120	31.15	15.70	[-2.1 , 64.4]

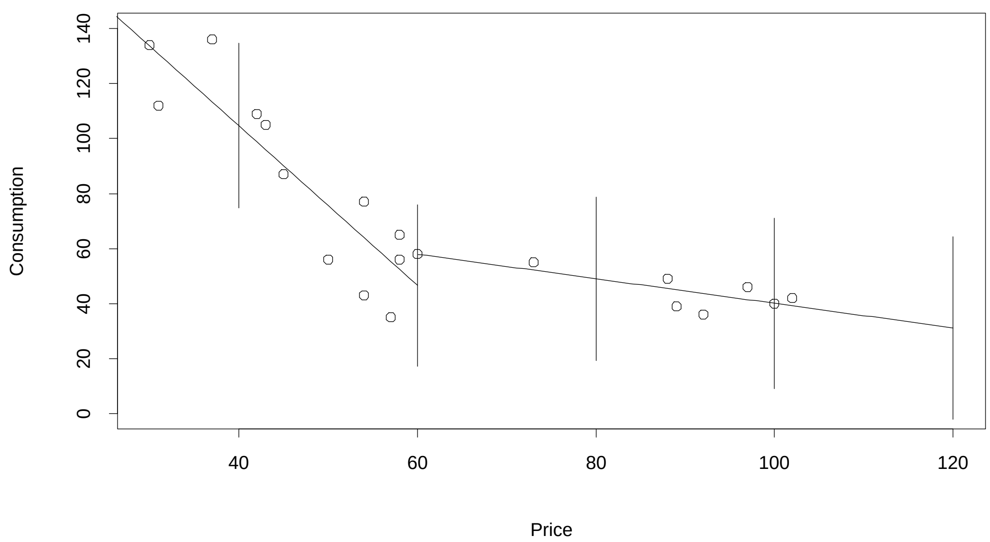
For example, at a price of 80c, the gas consumption will lie between 19.3 and 78.7 for 95% of towns.



Exercise 5.10(e): *Residual plots for the two models.*



Exercise 5.10(f): *Local linear regression through the gas consumption data. The fitted line suggests that model 2 is more appropriate.*



Exercise 5.10(h): *95% prediction intervals for gas consumption.*

Chapter 6: Multiple regression

6.1 **(a)** df for numerator = k and for denominator = $n-k-1$ where n = number of observations and k = number of explanatory variables. Here, $k = 16$ so that $n - 16 - 1 = 30$. Hence, $n = 30 + 16 + 1 = 47$.

(b)

$$\bar{R}^2 = 1 - (1 - R^2)\tfrac{n-1}{n-k-1} = 1 - (1 - 0.943)\frac{47-1}{48-16-1} = 0.913.$$

(c) $F = 31.04$ on (17,30) df. From Table C in Appendix III, the P-value is much smaller than 0.01. So the regression is highly significant.

(d) The coefficients should be compared with a t_{30} distribution. From Table B in Appendix III, any value greater than 2.04 in absolute value will be significant at the 5% level. So the constant and variables 4, 8, 12, 13, 14, 15 and 17 are significant in the presence of other explanatory variables. Note that the significance level of 5% is arbitrary. There is no reason why some other significance level (e.g. 2%) could not be used.

(e) The next stage would be to reduce the number of variables in the model by removing some of the least significant variables and re-fitting the model.

6.2 **(a)** The fitted model is $\hat{C} = 273.93 - 5.68P + 0.034P^2$. For this model, $R^2 = 0.8315$. [Recall: in exercise 5.6, model 1 had $R^2 = 0.721$ and model 2 had $R^2 = 0.859$.] So the $\bar{R}^2$ values for each model are:

Model 1 $$\bar{R}^2 = 1 - (1 - 0.721)\tfrac{n-1}{n-k-1} = 1 - (1 - 0.721)\frac{46}{45} = 0.715.$$

Model 2 $$\bar{R}^2 = 1 - (1 - 0.859)\tfrac{n-1}{n-k-1} = 1 - (1 - 0.859)\frac{46}{43} = 0.849.$$

Model 3 $$\bar{R}^2 = 1 - (1 - 0.832)\tfrac{n-1}{n-k-1} = 1 - (1 - 0.832)\frac{46}{44} = 0.824.$$

These values show that model 2 is the best model, followed by model 3. The t values for the coefficients are:

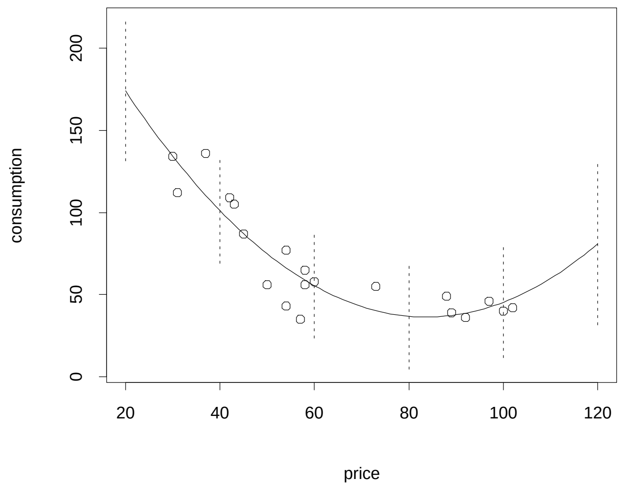
Model 1	$\alpha : t = 10.22$	$\beta : t = -5.47$		
Model 2	$\alpha_1 : t = 10.33$	$\beta_1 : t = -6.61$	$\alpha_2 : t = 4.11$	$\beta_2 : t = -1.99$
Model 3	$\beta_0 : t = 8.83$	$\beta_1 : t = -5.62$	$\beta_2 : t = 4.57$	

Of these, only β_2 from model 2 is not significantly different from zero. This suggests that a better model would be to allow the second part of model 2 to be a constant rather than a linear function.

(b) From the computer output the following 95% prediction intervals are obtained.

Output using Minitab for Exercise **6.2**:

```
MTB > regress 'C' 2 'P' 'Psq';
SUBC> predict 'newP' 'newPsq'.

The regression equation is
C = 274 - 5.68 P + 0.0339 Psq

Predictor       Coef       StDev          T        P
Constant      273.93       31.03       8.83    0.000
P             -5.676       1.009      -5.62    0.000
Psq         0.033904    0.007412       4.57    0.000

S = 14.37       R-Sq = 83.2%     R-Sq(adj) = 81.2%

Analysis of Variance
Source       DF          SS          MS         F        P
Regression    2     17327.0      8663.5     41.95    0.000
Error        17      3511.0       206.5
Total        19     20838.0

Source       DF      Seq SS
P             1     13005.7
Psq           1      4321.3

     Fit  StDev Fit       95.0% CI              95.0% PI
  173.97      14.29   (  143.82,  204.13)  (  131.21,  216.74) XX
  101.14       4.77   (   91.08,  111.21)  (   69.19,  133.10)
   55.43       4.91   (   45.07,   65.80)  (   23.38,   87.48)
   36.85       4.95   (   26.40,   47.29)  (    4.77,   68.92)
   45.38       7.14   (   30.31,   60.46)  (   11.52,   79.25)
   81.04      18.49   (   42.02,  120.06)  (   31.62,  130.46) XX
X  denotes a row with X values away from the center
XX denotes a row with very extreme X values
```



Exercise 6.2: *Quadratic regression of gas consumption against price. 95% prediction intervals shown.*

P	$\hat{C}$	[95% PI]
20	173.97	[131.21 , 216.74]
40	101.14	[69.19 , 133.10]
60	55.43	[23.38 , 87.48]
80	36.85	[4.77 , 68.92]
100	45.38	[11.52 , 79.25]
120	81.04	[31.62 , 130.46]

It is clear from the plot that it is dangerous predicting outside the observed price range. In this case, the predictions at $P = 20$ and $P = 120$ are almost certainly wrong. Predicting outside the range of the explanatory variable is always dangerous, but much more so when a quadratic (or higher-order polynomial) is used.

(c) $r_{PP^2} = 0.990$. If we were to use P, P^2 and P^3, the correlations among these explanatory variables would be very high and we would have a serious multicollinearity problem on our hands. The coefficients estimates would be unstable (i.e. have large standard errors). Multicollinearity will often be a problem with polynomial regression.

6.3 **(a)** From Table 6-15, we obtain the following values

Period	Actual	Forecast
54	4.646	1.863
55	1.060	1.221
56	-0.758	0.114
57	4.702	2.779
58	1.878	1.959
59	6.620	5.789

Analysis of errors: periods 54 through 59.

ME	MAE	MSE	MPE	MAPE	ACF1	Theil's U
0.74	1.11	2.15	34.82	41.32	-0.35	0.34

Strictly speaking, we should not compute relative measures when the data cross the zero line (i.e., when there are positive and negative values) because relative measures will "blow up" if divided by zero.

(b) and (c) Optimizing the coefficients for Holt's method will give better forecasts. Another approach is to use a simple MA forecast. An MA(2) forecast actually works better than Holt's method for both series. Other approaches are also possible.

Calculate accuracy statistics for your forecasts and compare them with the forecasts in Table 6-14.

6.4 **(a)** The fitted equation is

$$\hat{Y} = 73.40 + 1.52X_1 + 0.38X_2 - 0.27X_3.$$

95% confidence intervals for the parameters are calculated using a t_6 distribution. So the multiplier is 2.45:

$$\begin{aligned} 73.40 \pm 2.45(14.687) &= [37.46, 109.3] \\ 1.52 \pm 2.45(0.1295) &= [1.20, 1.84] \\ 0.38 \pm 2.45(0.1941) &= [-0.09, 0.85] \\ -0.27 \pm 2.45(0.1841) &= [-0.72, 0.18] \end{aligned}$$

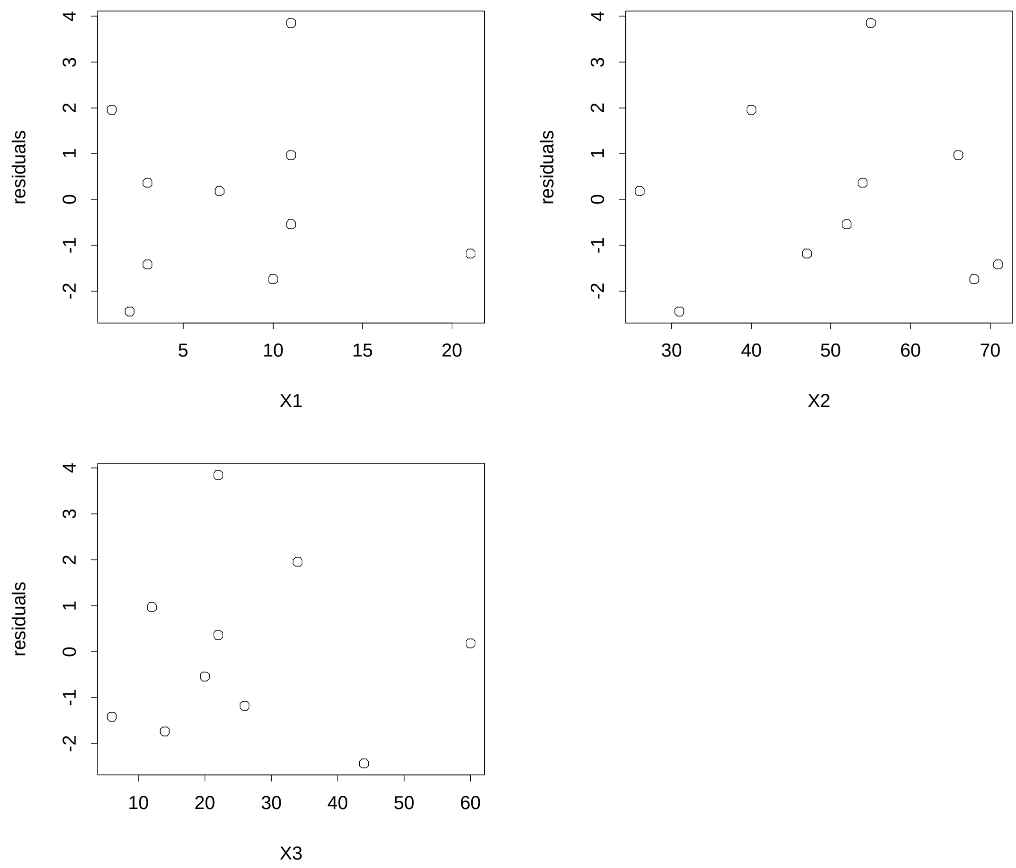
(b) $F = 123.3$ on (3,6) df. $P = 0.000$. This means that the probability of results like this, if the three explanatory variables were not relevant, is very small.

(c) The residual plots on page 120 show the model is satisfactory. There is no pattern in any of the residual plots.

(d) $R^2 = 0.984$. Therefore 98.4% of the variation in Y is explained by the regression relationship.

Output using Minitab for Exercise **6.4**:

```
MTB > Regress 'Y' 3 'X1' 'X2' 'X3';
SUBC>   Predict 10 40 30;
SUBC>   Confidence 90.

The regression equation is
Y = 73.4 + 1.52 X1 + 0.381 X2 - 0.268 X3

Predictor       Coef       StDev          T        P
Constant       73.40       14.69       5.00    0.002
X1            1.5162      0.1295      11.71    0.000
X2            0.3815      0.1941       1.97    0.097
X3           -0.2685      0.1841      -1.46    0.195

S = 2.326       R-Sq = 98.4%     R-Sq(adj) = 97.6%

Analysis of Variance
Source       DF          SS          MS         F        P
Regression    3     2001.54      667.18    123.32    0.000
Error         6       32.46        5.41
Total         9     2034.00

Source       DF      Seq SS
X1            1     1118.36
X2            1      871.67
X3            1       11.51

     Fit  StDev Fit       90.0% CI             90.0% PI
  95.762      1.632   (  92.590,  98.934)  (  90.239, 101.285)
```



Exercise 6.4(c): *Residual plots for the cement data.*

(e) The signs of the coefficients indicate the direction of the effect of each variable. X_1 increases heat and has the greatest effect (the largest coefficient). The other variables are not significant, so they may not have any effect. If they do, the coefficients suggest that X_2 might increase heat and X_3 might decrease heat.

(f) For $X_1 = 10$, $X_2 = 40$ and $X_3 = 30$, $\hat{Y} = 73.40+1.52(10)+0.38(40)-0.27(30) = 95.76$. 90% Prediction interval: [90.24,101.29]

6.5 The data for this exercise were taken from McGee and Carleton (1970) "Piecewise regression", *Journal of the American Statistical Association*, **65**, 1109–1124. It might be worthwhile to get this paper to compare what conventional regression can accomplish when there are special features in the data. In this case, the relationship between the Boston dollar volume and the NYSE-AME dollar volume underwent a series of changes over the time period of interest. In this paper, the solution was as follows:

from Jan '67 through Oct '67	$\hat{Y} =$	$8.748 + 0.0061X$
from Nov '67 through Jul '68	$\hat{Y} =$	$-20.905 + 0.0114X$
from Aug '68 through Nov '68	$\hat{Y} =$	$-79.043 + 0.0205X$
from Dec '68 through Nov '69	$\hat{Y} =$	$11.075 + 0.0067X$

Notice the slope coefficients in these four equations. They are small (because Boston's dollar volume is small relative to the big board volumes) but they get increasingly stronger (from6 1 to 114 to 205) in successive periods of commission splitting. Then in Dec '68, the SEC said "no more commission splitting" and it hurt the Boston dollar volume. The slope went back to 67, which is almost where it started.

(a) The fitted equation is $\hat{Y} = -66.2 + 0.014X$. The following output was obtained from a computer package.

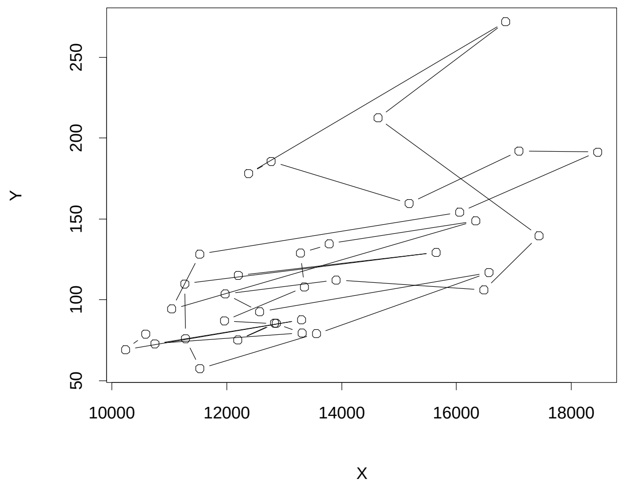
```
                Value Std. Error  t value Pr(>|t|)
 (Intercept) -66.2193  39.6809    -1.6688   0.1046
           X   0.0138   0.0029     4.7856   0.0000
F statistic: 22.9 on 1 and 33 degrees of freedom
             the p-value is 0.00003465
R-sq = 0.4097  Rbar-sq = 0.3918  D-W = 0.694
```

Clearly, the regression is significant, although the intercept is not significant.

(b) Output from computer package:

```
               Value Std. Error  t value Pr(>|t|)
(Intercept) -67.2116  40.2550    -1.6696   0.1047
          X   0.0135   0.0030     4.5025   0.0001
       time   0.2737   0.6518     0.4199   0.6773
F statistic: 11.25 on 2 and 32 degrees of freedom
             the p-value is 0.0001992
R-sq = 0.4129  Rbar-sq = 0.3762  D-W = 0.6814
```

Here, the regression is significant, but time is not significant. In fact, comparing these two models shows that adding time to the regression equation is actually worse than not adding it. See the $\bar{R}^2$ values. And for both analyses, the D-W



Exercise 6.5(c): *Connected scatterplot for the Boston and American stock exchanges.*

statistic shows that there is a lot of pattern left in the residuals. A piecewise regression approach does far better with this data set.

(c) See the plot above.

6.6 **(a) and (b)** Here are the seasonality indices based on the regression equations (6.10) and (6.12). They represent the intercept term in the regression for each of the 12 first differences.

	Using (6.10)	Using (6.12)
Mar-Feb	-2.6	-6.2
Apr-Mar	-6.7	-10.6
May-Apr	-3.5	-7.4
Jun-May	-5.3	-9.2
Jul-Jun	-3.6	-7.4
Aug-Jul	-5.2	-9.2
Sep-Aug	-5.9	-9.7
Oct-Sep	-6.9	-10.7
Nov-Oct	-4.1	-7.9
Dec-Nov	-4.7	-8.5
Jan-Dec	-0.8	-4.6
Feb-Jan	-2.2	-6.2

These two sets of seasonal indices are not quite the same. In the first equation (6.10), all eleven dummy variables for seasonality were allowed to be in the regression. In the second equation (6.12), the best subsets regression procedure did not allow the first seasonal dummy into the final equation. The absolute values are not so important because, in the presence of different sets of explanatory variables, we expect the intercept terms to be different.

(c) The seasonal indices should be the same regardless of which month is used as a base.

6.7 **(a)** $Y_t = 78.7 + 0.534x_t + e_t$

(b) DW = 0.57. $d_L = 1.04$ at 1% level. Therefore there is significant positive autocorrelation.

Chapter 7: The Box-Jenkins methodology for ARIMA models

7.1 **(a)** In general, the approximate standard error of the sample autocorrelations is $1/\sqrt{n}$. So the larger the value of n, the smaller the standard error. Therefore, the ACF has more variation for small values of n than for large values of n. All three series show the autocorrelations mostly falling with the 95% bands. The few that lie just outside the bands are not of concern since we would expect about 5% of spikes to cross the bands. There is no reason to think these series are anything but white noise.

(b) The lines shown are 95% critical values. These are calculated as $\pm 1.96/\sqrt{n}$. So they are closer to zero when n is larger. The autocorrelations vary randomly, but they mostly stay within the bounds.

7.2 The time plot shows the series as a non-stationary level. It wanders up and down over time in a similar way to a random walk. The ACF decays very slowly which also indicates non-stationarity in the level. Finally, the PACF has a very large value at lag 1, indicating the data should be differenced.

7.3 The five models are

AR(1)	$Y_t = 0.6Y_{t-1} + e_t.$
MA(1)	$Y_t = e_t + 0.6e_{t-1}.$
ARMA(1,1)	$Y_t = 0.6Y_{t-1} + e_t + 0.6e_{t-1}.$
AR(2)	$Y_t = -0.8Y_{t-1} + 0.3Y_{t-2} + e_t.$
MA(2)	$Y_t = e_t + 0.8e_{t-1} - 0.3e_{t-2}.$

In each case, we assume $Y_t = 0$ and $e_t = 0$ for $t \leq 0$. The generated data are shown on the following two pages. There is a lot of similarity in the shapes of the series because they are based on exactly the same errors.

7.4 **(a)** The ACF is slow to die out and the time plot shows the series wandering in a non-stationary way. So we take first differences. The ACF of the first differences show one significant spike at lag 1 indicating an MA(1) is appropriate. So the model for the raw data is ARIMA(0,1,1).

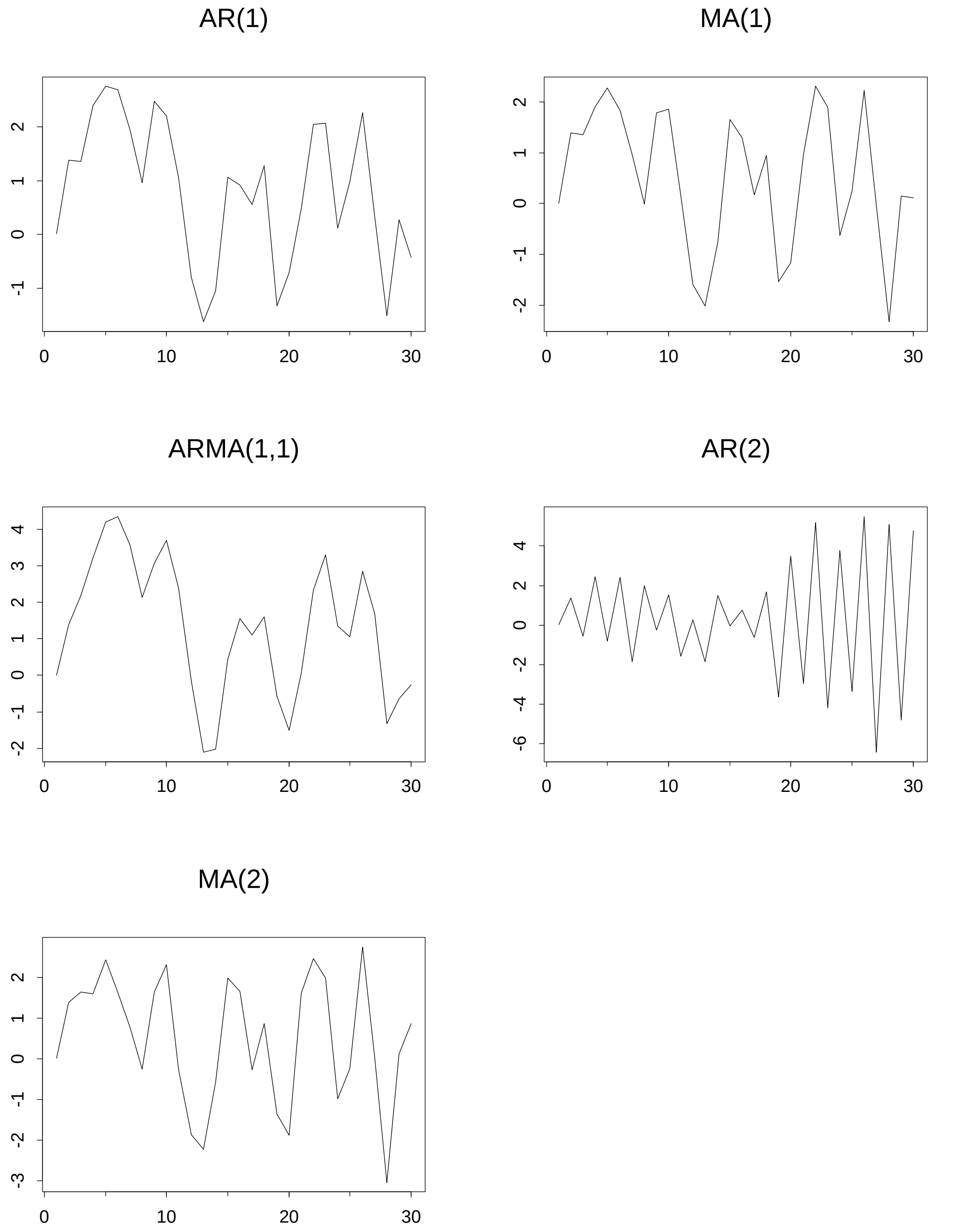
(b) There is not consistent trend in the raw data and the differenced data have mean close to zero. Therefore, there is no need to include a constant term.

(c) $(1 - B)Y_t = (1 - \theta_1 B)e_t.$

(d) See the output on page 127. There may be slight differences with different software packages and even different versions of the same package. The Ljung-Box statistics are not significant and the ACF and PACF of residuals show no significant differences from white noise.

t	AR(1)	MA(1)	ARMA(1,1)	AR(2)	MA(2)
1	0.010	0.010	0.010	0.010	0.010
2	1.386	1.386	1.392	1.372	1.388
3	1.362	1.358	2.193	-0.565	1.631
4	2.397	1.898	3.214	2.443	1.590
5	2.758	2.268	4.196	-0.804	2.425
6	2.695	1.832	4.350	2.416	1.622
7	1.947	0.954	3.564	-1.844	0.766
8	0.968	-0.002	2.136	2.000	-0.248
9	2.481	1.780	3.062	-0.253	1.641
10	2.209	1.860	3.697	1.523	2.300
11	1.055	0.162	2.380	-1.564	-0.264
12	-0.797	-1.592	-0.164	0.278	-1.862
13	-1.628	-2.008	-2.106	-1.842	-2.213
14	-1.047	-0.760	-2.024	1.487	-0.561
15	1.062	1.648	0.434	-0.052	1.979
16	0.917	1.294	1.554	0.768	1.653
17	0.560	0.178	1.111	-0.620	-0.273
18	1.276	0.946	1.612	1.666	0.864
19	-1.334	-1.536	-0.569	-3.619	-1.351
20	-0.711	-1.170	-1.511	3.485	-1.872
21	0.484	0.964	0.057	-2.964	1.612
22	2.050	2.306	2.340	5.176	2.461
23	2.070	1.896	3.300	-4.190	1.975
24	0.112	-0.626	1.354	3.775	-0.986
25	0.987	0.242	1.054	-3.357	-0.236
26	2.262	2.222	2.855	5.488	2.745
27	0.327	-0.028	1.685	-6.428	0.030
28	-1.514	-2.328	-1.317	5.079	-3.035
29	0.272	0.154	-0.636	-4.811	0.121
30	-0.427	0.118	-0.264	4.783	0.867

Generated data for Exercise 7.3



Exercise 7.3: *Simulated ARMA series.*

Output using Minitab for Exercise **7.4**:

```
MTB > ARIMA 0 1 1 'Strikes';
SUBC> NoConstant;
SUBC> Forecast 3.

Final Estimates of Parameters
Type          Coef       StDev         T
MA   1      0.3174      0.1886      1.68

Differencing: 1 regular difference
Number of observations:  Original series 30, after differencing 29
Residuals:    SS =  9256634  (backforecasts excluded)
              MS =   330594  DF = 28

Modified Box-Pierce (Ljung-Box) Chi-Square statistic
Lag                 12            24
Chi-Square    8.1(DF=11)   34.1(DF=23)

Forecasts from period 30
                             95 Percent Limits
Period      Forecast        Lower        Upper
  31         4164.87      3037.70      5292.04
  32         4164.87      2800.11      5529.63
  33         4164.87      2598.14      5731.60
```

(e) The last observation is $y_t = 3885$; the last residual in series is $e_t = -881.87$ (obtained from the computer package). Now

$$\begin{aligned} Y_t &= Y_{t-1} + e_t - 0.3174e_{t-1}. \\ \text{So} \quad \hat{Y}_{31} &= Y_{30} + \hat{e}_{31} - 0.3174\hat{e}_{30} \\ &= 3885 + 0 - 0.3174(-881.87) = 4164.9 \\ \hat{Y}_{32} &= \hat{Y}_{31} + 0 - 0.3174(0) = 4164.9 \\ \hat{Y}_{33} &= \hat{Y}_{32} + 0 - 0.3174(0) = 4164.9 \end{aligned}$$

(f) See the graph on the following page.

7.5 **(a)** The monthly data show strong seasonality and the seasonal pattern is reasonably stable. There is no trend in the data (this is a mature product).

(b) The pattern in the ACF plot shows the dominance of the seasonality. The autocorrelations at lags 6, 18 and 30 are negative (because we are correlating

Exercise 7.4(f): *Predicted number of strikes in USA. 95% prediction intervals shown.*

the high periods with the low periods) and at lags 12, 14 and 36 they are positive (because we are correlating high periods with high periods).

(c) The pattern in the PACF plot is not particularly revealing. However, there is little need to try to interpret this plot when the analysis clearly shows the dominance of the seasonality. The best approach would be to difference the series to reduce the effect of the seasonality and then see what is left over.

(d) These graphs suggest a seasonal MA(1) because of the spike at lag 12 in the ACF and the decreasing spikes at lags 12 and 24 in the PACF. Overall, the suggested model is ARIMA(0,1,0)(0,1,1)$_{12}$.

(e) Using the backshift operator: $(1 - B)(1 - B^{12})Y_t = (1 - \Theta B^{12})e_t$. Rewriting gives

$$Y_t - Y_{t-12} - Y_{t-1} + Y_{t-13} = e_t - \Theta e_{t-12}.$$

7.6 **(a)** ARIMA(3,1,0).

(b) For the differenced data, the PACF has a significant spikes at lags 1, 2 and 3 and a spike at lag 17 which is marginally significant. The spike at lag 17 is probably due to chance. Therefore an AR(3) is an appropriate model for the differenced data. Consequently, an ARIMA(3,1,0) model is suitable for the original data.

(c) Now

$$(Y_t - Y_{t-1}) = 0.42(Y_{t-1} - Y_{t-2}) - 0.20(Y_{t-2} - Y_{t-3}) - 0.30(Y_{t-3} - Y_{t-4}) + e_t.$$

Therefore $\quad Y_t = 1.42Y_{t-1} - 0.62Y_{t-2} - 0.10Y_{t-3} + 0.30Y_{t-4} + e_t \quad$ and

$$\begin{aligned}
\hat{Y}_{1940} &= 1.42(1797) - 0.62(1791) - 0.10(1627) + 0.30(1665) = 1778.1 \\
\hat{Y}_{1941} &= 1.42(1778.1) - 0.62(1797) - 0.10(1791) + 0.30(1627) = 1719.8 \\
\hat{Y}_{1942} &= 1.42(1719.8) - 0.62(1778.1) - 0.10(1797) + 0.30(1791) = 1697.3
\end{aligned}$$

7.7 **(a)** ARIMA(4,0,0).

(b) The model was chosen because the last significant spike in the PACF was at lag 4. Note that the spikes at lags 2 and 3 were not significant. This makes no difference. It is the *last* significant spike which determines the order of the model.

(c) The model is

$$Y_t = 146.1 + 0.891Y_{t-1} - 0.257Y_{t-2} + 0.392Y_{t-3} - 0.333Y_{t-4} + e_t.$$

So

$$\begin{aligned}
\hat{Y}_{1969} &= 146.1 + 0.891(545) - 0.257(552) + 0.392(534) - 0.333(512) = 528.7 \\
\hat{Y}_{1970} &= 146.1 + 0.891(528.7) - 0.257(545) + 0.392(552) - 0.333(534) = 515.7 \\
\hat{Y}_{1971} &= 146.1 + 0.891(515.7) - 0.257(528.7) + 0.392(545) - 0.333(552) = 499.5
\end{aligned}$$

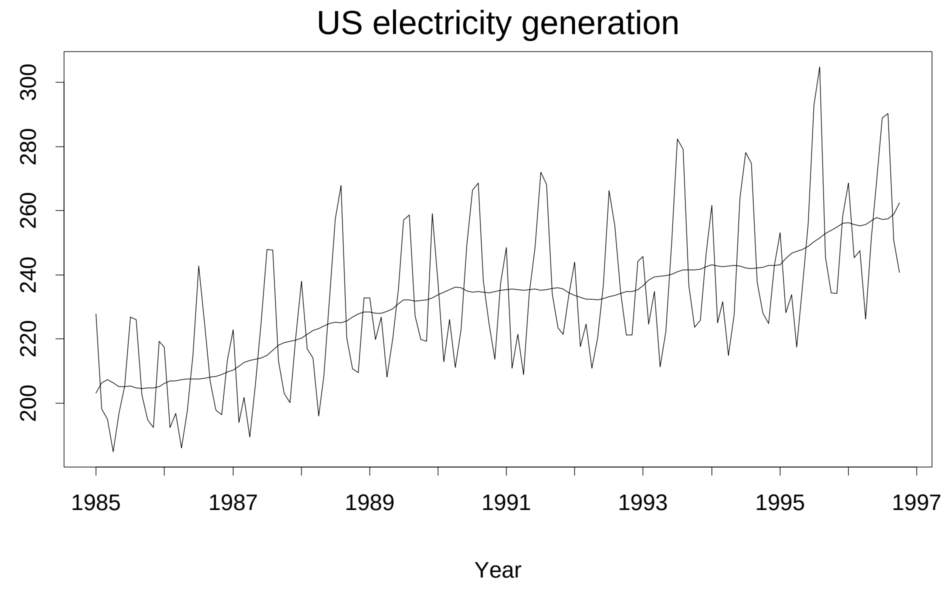
7.8 **(a)** The centered 12-MA smooth is shown in the plot on the next page. The trend is generally linear and increasing with a flat period between 1990 and 1993.

(b) The variation does not change much with the level, so transforming will not make much difference.

(c) The data are not stationary. There is a trend and seasonality in the data. Differencing at lag 12 gives the data shown in the plot on page 131. These appear stationary although it is possible another difference at lag 1 is needed.

(d) From the plots on page 131 it is clear there is a seasonal MA component of order 1. In addition there is a significant spike at lag 1 in both the ACF and PACF. Hence plausible models are ARIMA(1,0,0)$(0,1,1)_{12}$ and ARIMA(0,0,1)$(0,1,1)_{12}$. Comparing the two models we have the following results

ARIMA(1,0,0)$(0,1,1)_{12}$	AIC=900.2
ARIMA(0,0,1)$(0,1,1)_{12}$	AIC=926.9



Exercise 7.8(a): *Total net generation of electricity in USA.*

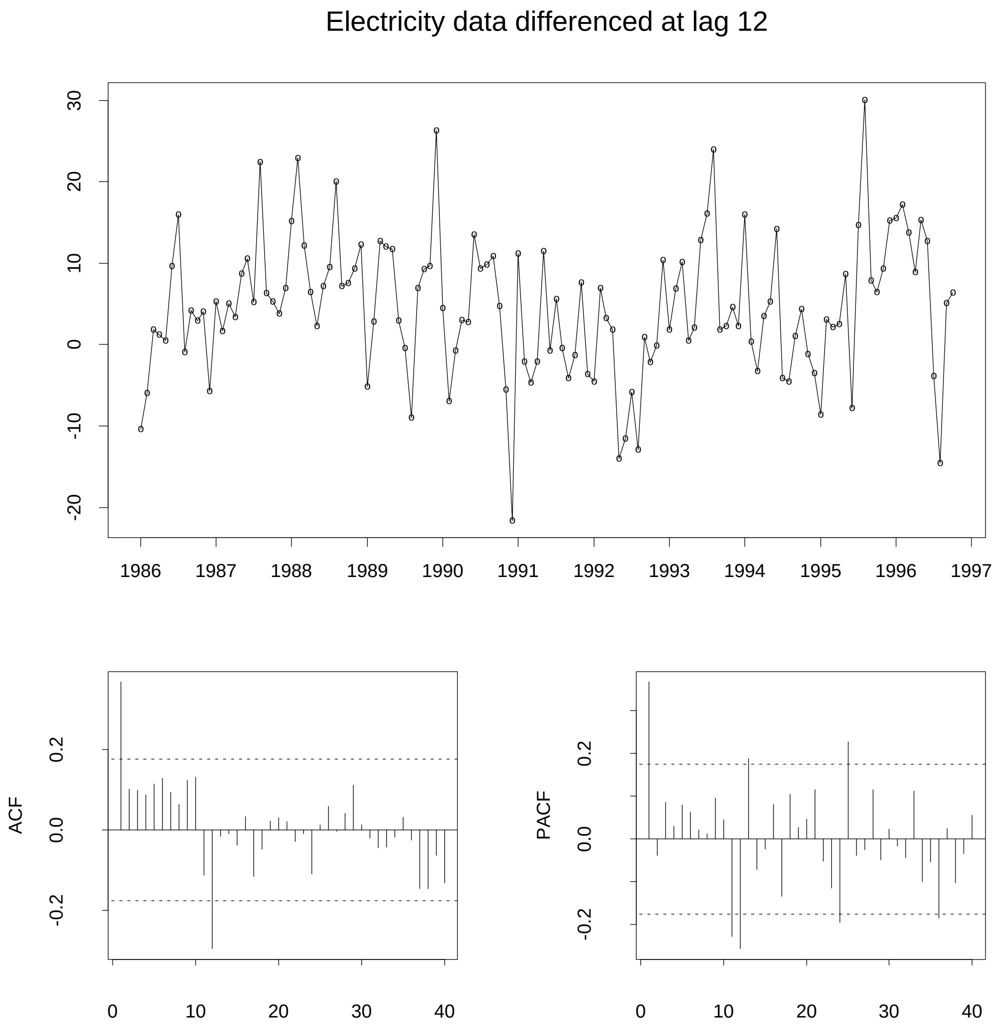
Hence the better model is the first one. Note that different packages will give different values for the AIC depending on how it is calculated. Therefore the same package should be used for all calculations.

(e) The residuals from the ARIMA(1,0,0)(0,1,1)$_{12}$ are shown in the plots on page 132. Because there are significant spikes in the ACF and PACF, the model is not adequately describing the series. These plots suggest we need to add an MA(1) term to the model. So we fit the revised model ARIMA(1,0,1)(0,1,1)$_{12}$. This time, the residual plots (not shown here) look like white noise. The AIC is 876.7. Part of the computer output for fitting the revised model is shown below.

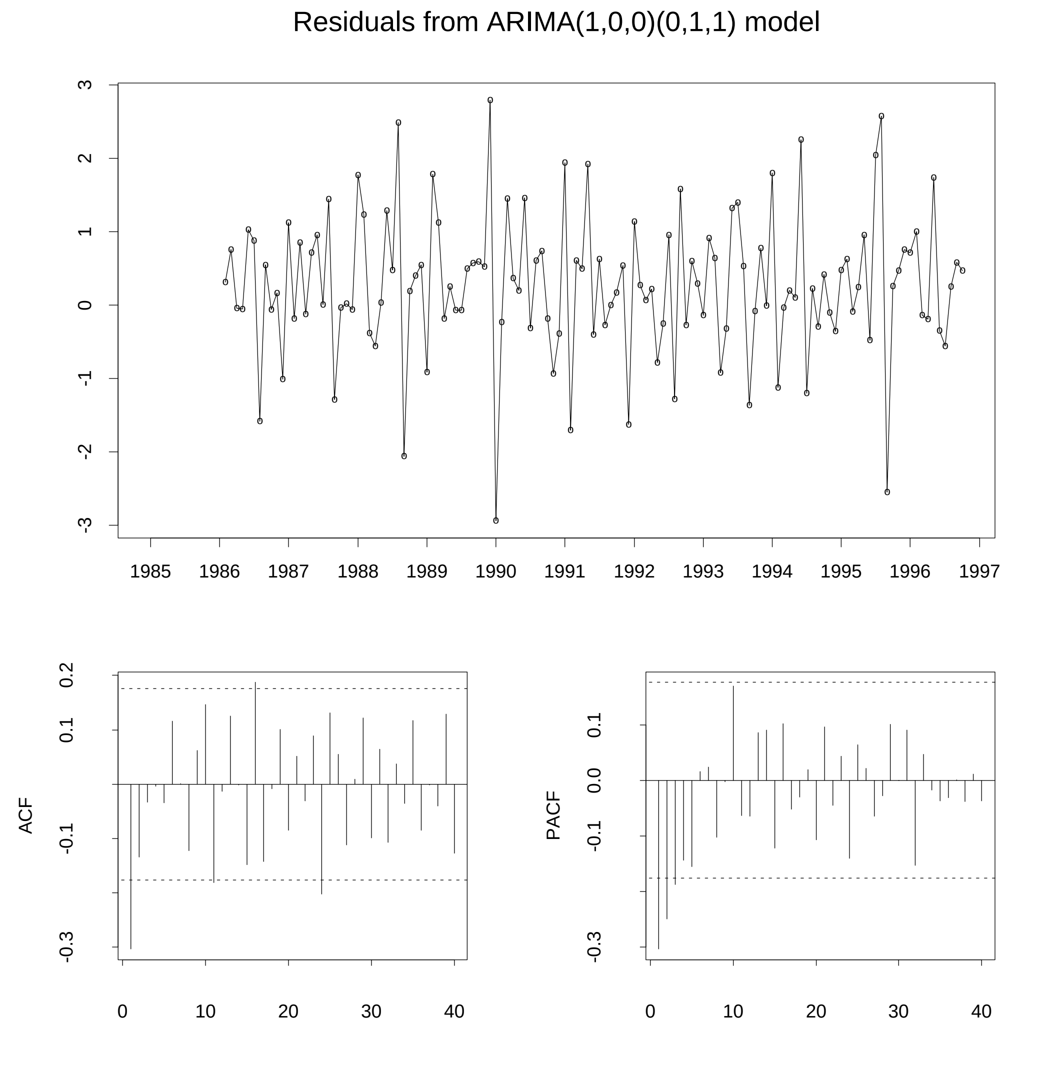
```
                                Approx.
Parameter    Estimate     Std Error    T Ratio  Lag
MA1,1         0.74427       0.05887      12.64    1
MA2,1         0.77650       0.09047       8.58   12
AR1,1         0.99566     0.0070613     141.00    1
```

So the fitted model is

$$(1 - 0.996B)(1 - B^{12})Y_t = (1 - 0.744B)(1 - 0.777B^{12})e_t.$$



Exercise 7.8(c): *Seasonally differenced electricity generation.*



Exercise 7.8(e): *Residuals from ARIMA(1,0,0)(0,1,1)*$_{12}$ *model fitted to the electricity data.*

Output using SAS for Exercise **7.8**:

```
                         Approx.
Parameter   Estimate    Std Error   T Ratio  Lag
MA1,1        0.86486      0.06044     14.31   1
MA2,1        0.80875      0.09544      8.47  12
AR1,1        0.27744      0.10751      2.58   1

Variance  Estimate = 41.8498466
Std Error Estimate = 6.46914574
AIC                = 864.616345
SBC                = 873.195782
Number of Residuals=        129

 To   Chi                       Autocorrelations of Residuals
Lag  Square DF   Prob
  6     1.60  3  0.659  0.028 -0.036 -0.020 -0.010 -0.014  0.095
 12     7.67  9  0.568  0.004 -0.082  0.073  0.128 -0.095  0.072
 18    15.32 15  0.429  0.126  0.016 -0.091  0.125 -0.105 -0.020
 24    18.67 21  0.607  0.065 -0.048  0.051  0.005  0.069 -0.085
```

Note that the first term on the left is almost the same as differencing $(1 - B)$. This suggests that we probably should have taking a first difference as well as a seasonal difference. We repeated the above analysis and arrived at the following model: ARIMA(1,1,1)(0,1,1)$_{12}$ which has AIC=864.6.

The computer output for the final model is shown above. The figures under the heading `Chi Square` concern the Ljung-Box test. Clearly the model passes the test (see Table E in Appendix III).

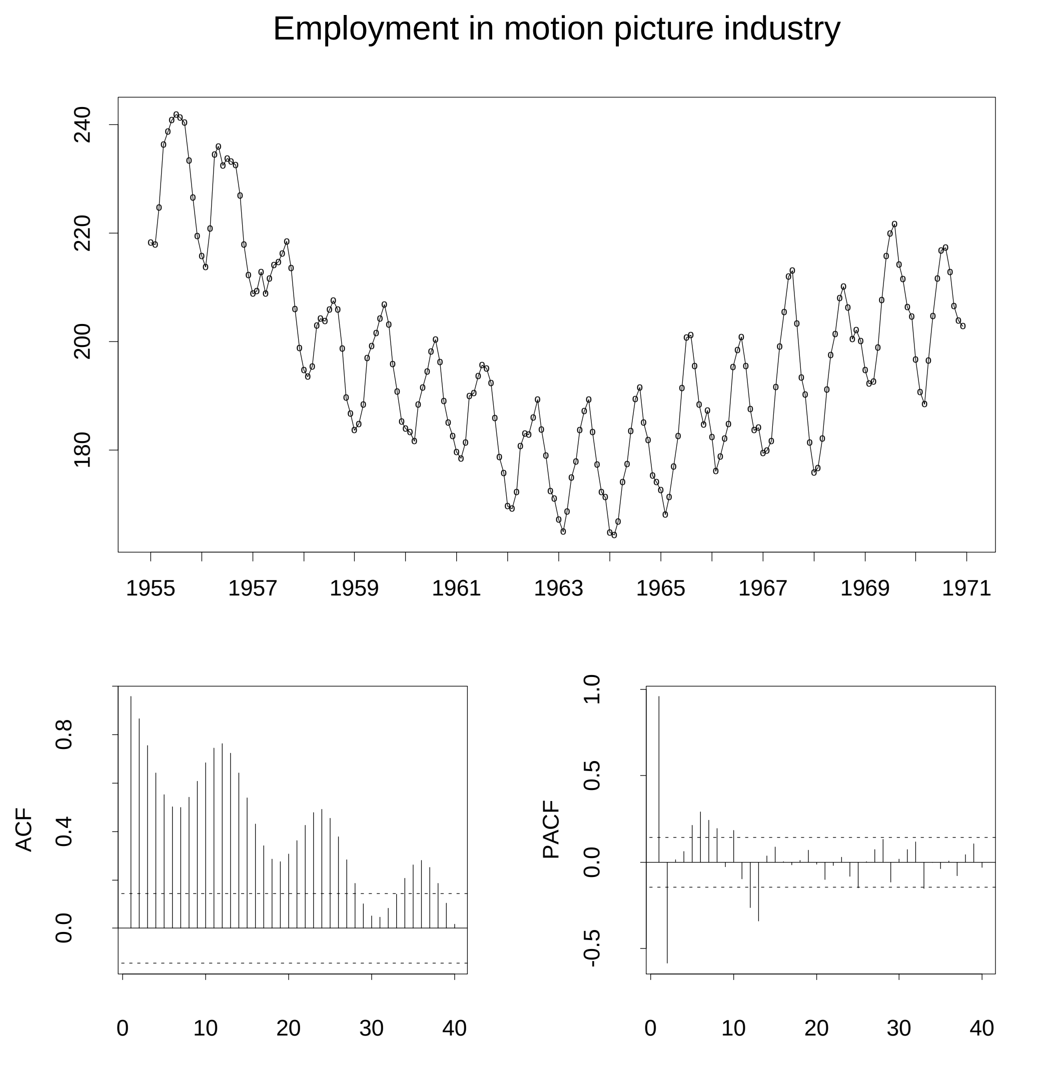
(f) Forecasts for the next 24 months are given on the following page.

Month	Obs	Forecast	Std Error	Lower 95%	Upper 95%
Nov 96	143	240.1614	6.4691	227.4821	252.8407
Dec 96	144	262.5516	6.9981	248.8356	276.2677
Jan 97	145	270.2423	7.1820	256.1659	284.3187
Feb 97	146	244.0064	7.3027	229.6934	258.3194
Mar 97	147	249.8899	7.4074	235.3718	264.4081
Apr 97	148	232.7683	7.5069	218.0550	247.4816
May 97	149	249.3720	7.6042	234.4680	264.2759
Jun 97	150	270.7257	7.6999	255.6341	285.8173
Jul 97	151	295.5439	7.7944	280.2671	310.8207
Aug 97	152	295.6598	7.8878	280.2000	311.1196
Sep 97	153	257.1358	7.9800	241.4952	272.7764
Oct 97	154	246.4526	8.0712	230.6332	262.2719
Nov 97	155	245.0224	8.4340	228.4920	261.5528
Dec 97	156	267.1930	8.6077	250.3222	284.0638
Jan 98	157	274.8228	8.7406	257.6914	291.9541
Feb 98	158	248.5699	8.8622	231.2003	265.9395
Mar 98	159	254.4488	8.9796	236.8491	272.0484
Apr 98	160	237.3258	9.0948	219.5004	255.1513
May 98	161	253.9291	9.2083	235.8811	271.9772
Jun 98	162	275.2828	9.3205	257.0150	293.5506
Jul 98	163	300.1009	9.4313	281.6160	318.5858
Aug 98	164	300.2168	9.5407	281.5173	318.9163
Sep 98	165	261.6929	9.6490	242.7812	280.6045
Oct 98	166	251.0096	9.7560	231.8881	270.1311

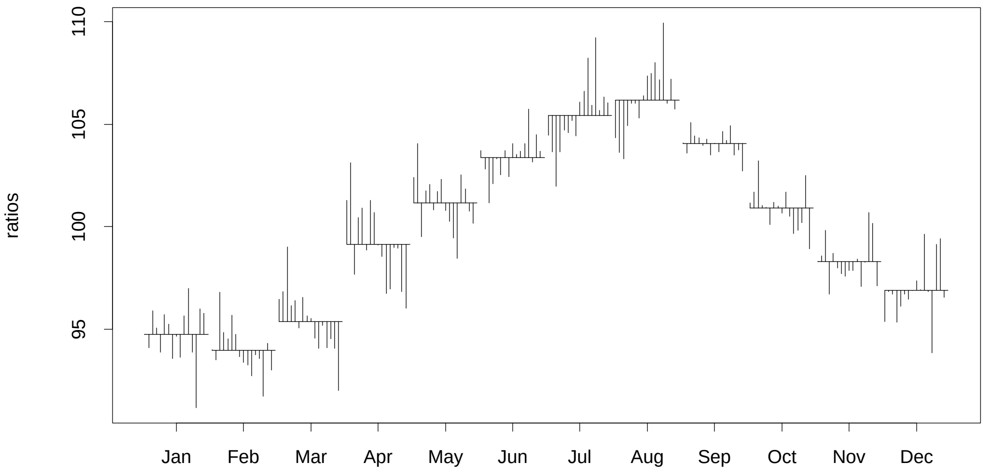
7.9 **(a)** See the plot on the following page. Note that there is strong seasonality and a pronounced trend-cycle. One way to study the consistency of the seasonal pattern is to compute the seasonal sub-series and see how stable each month is. The results are given below.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1955:	94.7	94.0	96.5	101.3	102.4	103.7	104.5	104.3	104.1	101.2	98.3	95.4
1956:	94.1	93.5	96.8	103.1	104.1	102.8	103.7	103.6	103.6	101.7	98.6	96.8
1957:	95.9	96.8	99.0	97.7	99.5	101.1	102.0	103.3	105.1	103.2	99.8	96.7
1958:	95.0	94.8	96.1	100.4	101.7	102.1	103.6	104.9	104.4	101.1	96.7	95.3
1959:	93.9	94.5	96.4	100.9	102.1	103.3	104.7	106.0	104.4	100.9	98.7	96.1
1960:	95.7	95.7	95.1	98.9	100.8	102.5	104.6	106.0	104.0	100.1	98.0	96.7
1961:	95.2	94.8	96.5	101.3	101.7	103.7	105.2	105.3	104.3	101.2	97.7	96.5
1962:	93.5	93.7	95.6	100.7	102.3	102.5	104.4	106.4	103.5	101.0	97.6	96.9
1963:	94.6	93.4	95.5	99.1	100.8	104.1	106.1	107.4	104.1	100.7	97.9	97.4
1964:	93.6	93.2	94.6	98.6	100.2	103.5	106.6	107.5	103.6	101.7	97.9	96.9
1965:	95.6	92.7	94.0	96.7	99.4	103.7	108.2	108.0	104.7	100.5	98.4	99.6
1966:	97.0	93.7	95.2	97.0	98.4	104.1	105.9	107.2	104.2	99.7	97.1	96.8
1967:	93.9	93.6	94.1	99.0	102.5	105.7	109.2	109.9	104.9	99.8	98.3	93.8
1968:	91.2	91.7	94.5	99.0	101.9	103.1	105.7	106.0	103.5	100.2	100.7	99.1
1969:	96.0	94.3	94.1	96.8	100.7	104.5	106.3	107.2	103.7	102.5	100.2	99.4
1970:	95.8	93.0	92.0	96.0	100.2	103.7	106.0	105.8	102.7	98.9	97.1	96.5

These detrended data are relatively consistent from year to year with only minor



Exercise 7.9(a): *Employment in the motion picture industry*



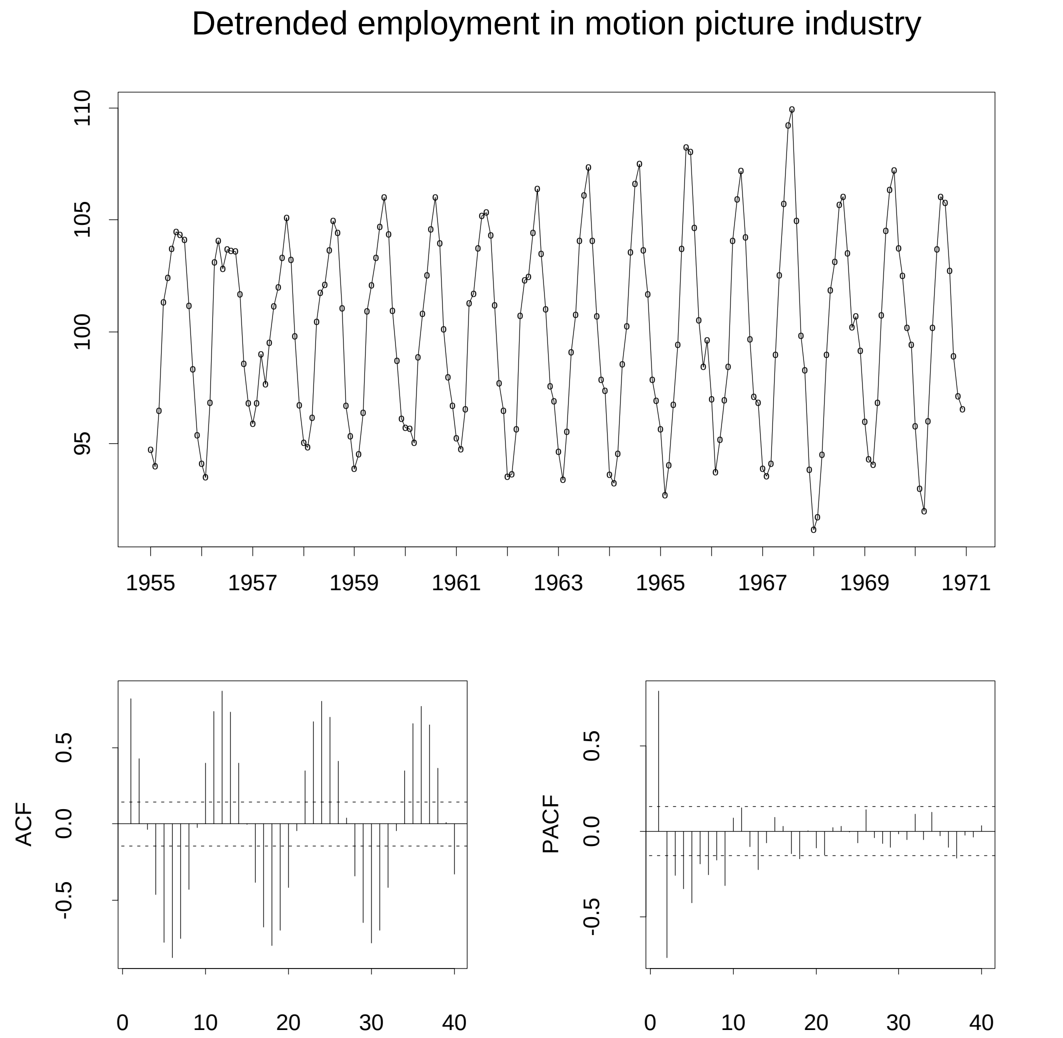
Exercise 7.9(a): *Sub-series of detrended data*

variations occurring here and there. For example, December 1967 and January and February 1968 were noticeably lower than surrounding years.

Another way to look at seasonal patterns is via autocorrelation functions. Note that for the raw data, the ACF shows strong seasonality over several seasonal lags. This is further evidence of the consistency of the seasonal pattern. The plot on the previous page shows the detrended data. Again, the seasonal pattern is very consistent although the amplitude of the pattern each year varies. Unusual results in early 1968 and early 1970 are seen.

(b) For the first 96 months, we identified an ARIMA(0,1,0)(0,1,1)$_{12}$. For the second 96 months, we identified an ARIMA(0,1,0)(1,1,0)$_{12}$: In practice, there is little difference between these models. This means that once the trend has been eliminated (by differencing), the seasonal patterns are very similar.

(c) Using the above ARIMA(0,1,0)(0,1,1)$_{12}$ model, we obtained the following forecasts.



Exercise 7.9(a): *Seasonal differences of employment in the motion picture industry*

Month		Actual	Forecast	Upper (95%)	Lower (95%)	Error
Jan	1963	167.2	167.3286	172.4228	162.2345	-0.1286
Feb	1963	165.0	166.7030	171.7902	161.6159	-1.7030
Mar	1963	168.8	167.7141	172.8012	162.6269	1.0859
Apr	1963	175.0	176.5972	181.6844	171.5100	-1.5972
May	1963	177.9	176.9498	182.0369	171.8626	0.9502
Jun	1963	183.7	179.2212	184.3084	174.1340	4.4788
Jul	1963	187.2	186.1708	191.2579	181.0836	1.0292
Aug	1963	189.3	188.7598	193.8470	183.6727	0.5402
Sep	1963	183.4	186.1678	191.2550	181.0806	-2.7678
Oct	1963	177.3	177.2392	182.3264	172.1520	0.0608
Nov	1963	172.3	170.7711	175.8583	165.6839	1.5289
Dec	1963	171.4	168.8708	173.9580	163.7837	2.5292
Jan	1964	164.9	167.5935	172.6807	162.5063	-2.6935
Feb	1964	164.4	163.9412	169.0247	158.8577	0.4588
Mar	1964	166.9	167.4086	172.4920	162.3251	-0.5086
Apr	1964	174.2	174.2641	179.3475	169.1806	-0.0641
May	1964	177.5	176.4075	181.4909	171.3240	1.0925
Jun	1964	183.6	180.0358	185.1193	174.9523	3.5642
Jul	1964	189.5	186.3499	191.4333	181.2664	3.1501
Aug	1964	191.6	191.2063	196.2898	186.1228	0.3937
Sep	1964	185.1	187.7172	192.8007	182.6338	-2.6172
Oct	1964	181.9	178.9557	184.0392	173.8722	2.9443
Nov	1964	175.4	175.7857	180.8692	170.7022	-0.3857
Dec	1964	174.2	172.6567	177.7402	167.5732	1.5433

(d) For the second half of the data we used the ARIMA$(0,1,0)(1,1,0)_{12}$ to obtain the forecasts at the top of the following page. The actual 1971–1972 figures are also shown. The source is "Employment and Earnings, US 1909–1978", published by the Department of Labor, 1979.

A good exercise would be to take these forecasts and check the MAPE for 1971 and 1972 separately. The MAPE for the first forecast year should be smaller than the MAPE for the second year.

(e) If the objective is to forecast the next 12 months then the latest data is obviously the most relevant but to get seasonal indices we have to go back several years and to anticipate what the next move the large cycle is going to be, we really need to look at as much data as possible. So a good strategy would be

i. study the trend-cycle by looking at the 12-month moving average;
ii. remove the trend-cycle and study the consistency of the seasonality;
iii. decide how much of the data series to retain for the ARIMA modeling;
iv. forecast the next 12 months and use some judgment as to how to modify the ARIMA forecasts on the basis of anticipated trend-cycle movements.

Month		Actual	Forecast	Upper (95%)	Lower (95%)	Error
Jan	1971	194.5	196.0141	201.4418	190.5864	-1.5141
Feb	1971	187.9	191.2939	198.9698	183.6180	-3.3939
Mar	1971	187.7	189.9446	199.3456	180.5436	-2.2446
Apr	1971	198.3	197.3595	208.2149	186.5042	0.9405
May	1971	202.7	205.7424	217.8790	193.6057	-3.0424
Jun	1971	204.2	213.1446	226.4396	199.8495	-8.9446
Jul	1971	211.7	217.8789	232.2392	203.5186	-6.1789
Aug	1971	213.4	218.8543	234.2061	203.5025	-5.4543
Sep	1971	212.0	213.2939	229.5769	197.0109	-1.2939
Oct	1971	203.4	208.3371	225.5009	191.1733	-4.9371
Nov	1971	199.5	204.7595	222.7611	186.7580	-5.2595
Dec	1971	199.3	203.4670	222.2690	184.6650	-4.1670
Jan	1972	191.3	196.2469	217.0365	175.4573	-4.9469
Feb	1972	192.1	191.0587	213.6617	168.4557	1.0413
Mar	1972	193.3	189.3618	213.6432	165.0804	3.9382
Apr	1972	203.4	196.9907	222.8417	171.1396	6.4093
May	1972	205.5	205.3066	232.6373	177.9760	0.1934
Jun	1972	218.2	212.5618	241.2960	183.8276	5.6382
Jul	1972	220.3	217.4298	247.5022	187.3575	2.8702
Aug	1972	219.9	218.2314	249.5848	186.8780	1.6686
Sep	1972	211.9	213.0587	245.6428	180.4746	-1.1587
Oct	1972	204.5	207.6473	241.4174	173.8773	-3.1473
Nov	1972	198.5	204.3907	239.3064	169.4750	-5.8907
Dec	1972	200.5	203.2051	239.2300	167.1802	-2.7051

Forecasts for Exercise 7.9(d)

7.10 **(a)** There is strong seasonality as can be seen from the time plot and the seasonal peaks in the ACF.

(b) The trend in the series is small compared to the seasonal variation. However, there is a period of downward trend in the first four years, followed by an upward trend for four years. At the end the trend seems to have levelled off.

(c) The one large spike in the PACF of Figure 7-34 suggests the series needs differencing at lag 1. This is also apparent from the slow decay in the ACF and the non-stationary mean in the time plot.

(d) You would need to difference again at lag 1 and plot the ACF and PACF of the new series (differenced at lags 12 and 1). It is not possible to identify a model from Figures 7-33 and 7-34.

Chapter 8: Advanced forecasting models

8.1 **(a)** The fitted model in Exercise 6.7 (using OLS) was

$$Y_t = 78.7 + 0.534x_t + N_t.$$

The computer output below shows the results for fitting the straight line regression with AR(1) errors. Hence the new model is

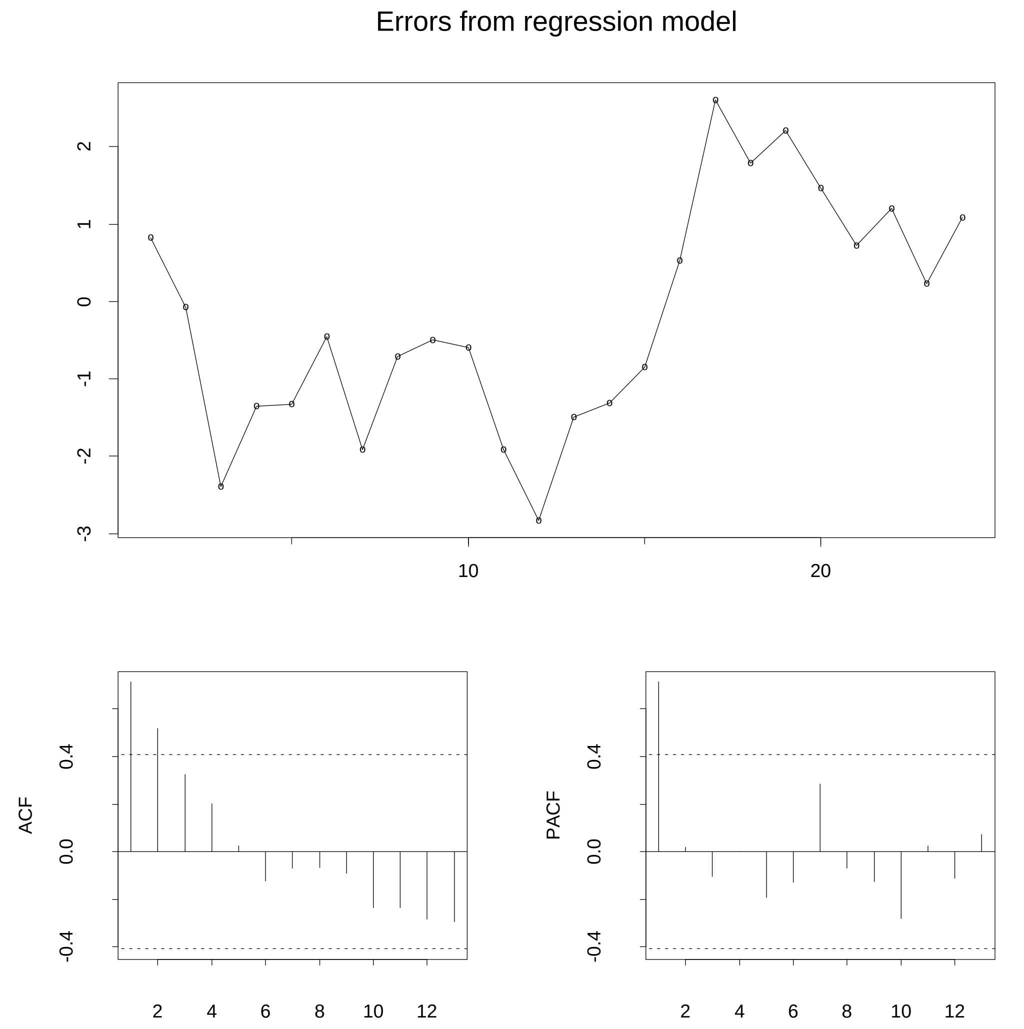
$$Y_t = 79.3 + 0.508x_t + N_t \qquad \text{where} \quad N_t = 0.72N_{t-1} + e_t.$$

In this case, the error model makes very little difference to the parameters.

Output from SAS for Exercise **8.1**:

```
                          Approx.
Parameter   Estimate    Std Error   T Ratio  Lag  Variable Shift
MU          79.27236      0.76093    104.18   0   SALES        0
AR1,1        0.72469      0.14647      4.95   1   SALES        0
NUM1         0.50801      0.02318     21.91   0   ADVERT       0

Constant Estimate  = 21.8242442
Variance  Estimate = 1.11639088
Std Error Estimate =   1.056594
AIC                = 74.2915405
SBC                =  77.825702
Number of Residuals=         24

                      Autocorrelation Check of Residuals
 To   Chi                             Autocorrelations
Lag  Square DF    Prob
  6    3.46  5  0.630  0.027  0.099 -0.037  0.111 -0.060 -0.274
 12    9.31 11  0.593  0.055  0.126  0.229 -0.227  0.060 -0.095
 18   16.39 17  0.497 -0.117 -0.238 -0.080  0.054 -0.108  0.101
```

(b) The ACF and PACF of the errors is plotted on the following page. An AR(1) model for the errors is appropriate since there is a single significant spike at lag 1 in the PACF and geometric decay in the ACF. This is confirmed by the Ljung-Box test in the computer output above. The Q^* values are given under the column `Chi Square`. None are significant showing the residuals from the full model are white noise.



Exercise 8.1: *Errors from regression model with AR(1) error term.*

Output from SAS for Exercise **8.2(a)**:

```
                         Approx.
Parameter   Estimate    Std Error   T Ratio  Lag  Variable Shift
MU           9.56328      0.40537     23.59   0   HURON        0
AR1,1        0.78346      0.06559     11.94   1   HURON        0
NUM1        -0.02038      0.01066     -1.91   0   YEAR         0

Constant Estimate  = 2.07087134

Variance  Estimate = 0.51219788
Std Error Estimate = 0.71568001
AIC                = 216.450147
SBC                = 224.205049
Number of Residuals=         98

                    Autocorrelation Check of Residuals
 To   Chi                        Autocorrelations
Lag  Square DF   Prob
  6    8.35  5  0.138  0.222 -0.100 -0.133 -0.056 -0.007 -0.042
 12   15.01 11  0.182 -0.051  0.009  0.175  0.017 -0.121 -0.107
 18   16.36 17  0.499 -0.053  0.014  0.019  0.058  0.006 -0.067
 24   25.47 23  0.326 -0.071 -0.166 -0.043  0.051  0.160  0.092
```

8.2 **(a)** To reduce numerical error, we subtracted 1900 from the year to create an explanatory variable. Hence the year ranged from -25 (1875) to 72 (1972). The computer output above shows the fitted model to be

$$Y_t = 9.56 - 0.02x_t + N_t \qquad \text{where} \quad N_t = 0.78N_{t-1} + e_t$$

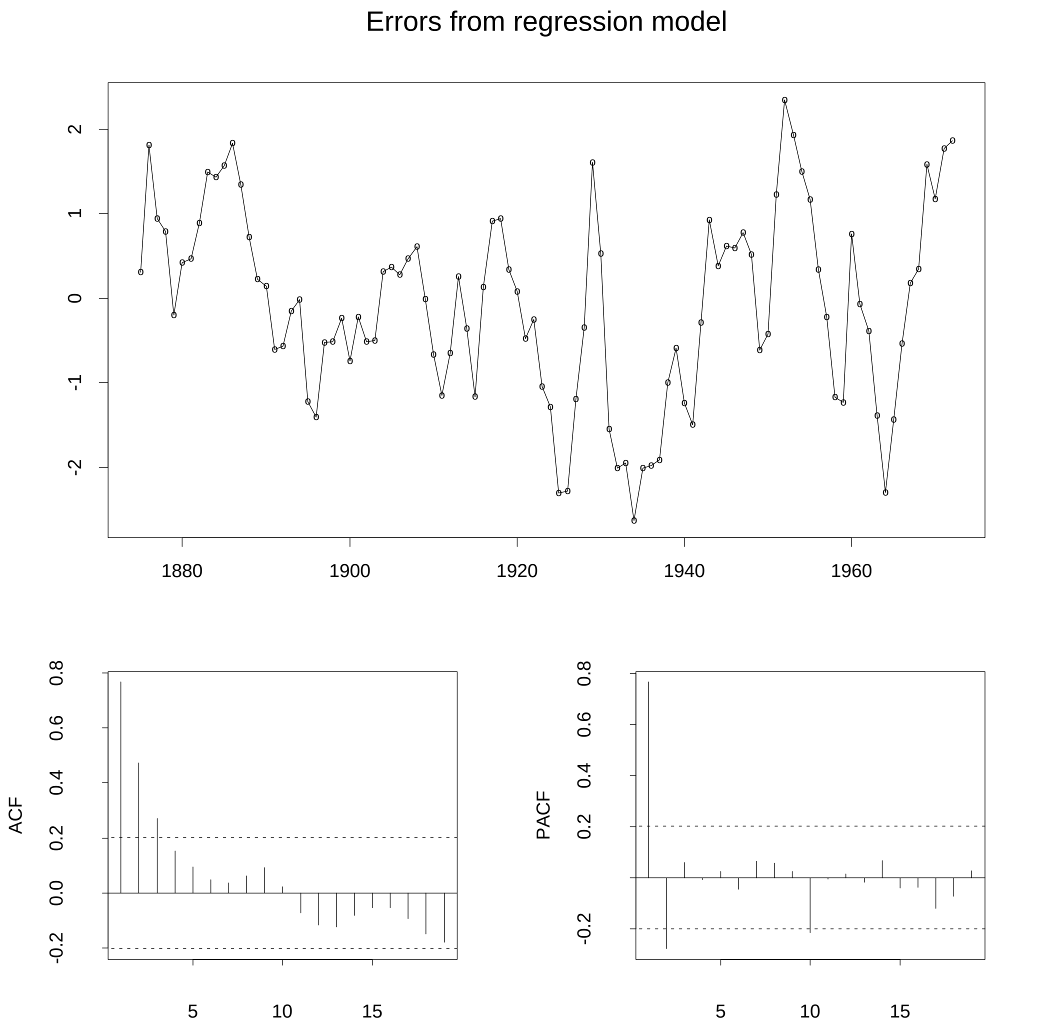
where x_t is the year -1900.

(b) The errors are shown in the plot on the following page. This demonstrates that a better model would have an AR(2) error term since the PACF has two significant spikes at lags 1 and 2. The spike at lag 10 is probably due to chance. The ACF shows geometric decay which is possible with an AR(2) model. So the full regression model is

$$Y_t = \beta_0 + \beta_1 x_t + N_t \qquad \text{where} \quad N_t = \phi_1 N_{t-1} + \phi_2 N_{t-2} + e_t.$$

Fitting this model gives the output shown on page 144. So the fitted model is

$$Y_t = 9.53 - 0.02x_t + N_t \qquad \text{where} \quad N_t = N_{t-1} - 0.29N_{t-2} + e_t.$$



Exercise 8.2(b): *Errors from regression model with AR(1) error term.*

Output from SAS for Exercise **8.2(b)**:

```
                                        Approx.
Parameter   Estimate    Std Error   T Ratio  Lag  Variable Shift
MU           9.53078      0.30653     31.09   0   HURON         0
AR1,1        1.00479      0.09839     10.21   1   HURON         0
AR1,2       -0.29128      0.10030     -2.90   2   HURON         0
NUM1        -0.02157    0.0082537     -2.61   0   YEAR          0

Constant Estimate  = 2.73048107

Variance  Estimate =  0.4760492
Std Error Estimate = 0.68996319
AIC                = 210.396534
SBC                = 220.736404
Number of Residuals=         98

                   Autocorrelation Check of Residuals
 To   Chi                       Autocorrelations
Lag  Square DF   Prob
  6    0.60  4  0.964  0.018 -0.028 -0.003  0.040  0.054 -0.007
 12    5.35 10  0.867 -0.032 -0.037  0.167 -0.007 -0.098 -0.055
 18    6.21 16  0.986 -0.036  0.005 -0.025  0.035 -0.006 -0.063
 24   10.49 22  0.981 -0.003 -0.141 -0.007  0.006  0.116  0.008
```

8.3 **(a)** ARIMA(0,1,1)$(2,1,0)_{12}$. This model would have been chosen by first identifying that differences at lags 12 and 1 are necessary to make the data stationary. Then looking at the ACF and PACF of the differenced data would have shown two significant spikes in the PACF at lags 12 and 24. There would have also been a significant spike in the ACF at lag 1 and geometric decay in the early lags of the PACF.

(b) Since both parameter estimates are positive (and significantly different from zero), we can conclude that electricity consumption increases with both heating degrees and cooling degrees. Because b_2 is larger, we know that there is a greater increase in electricity usage for each heating degree than for each cooling degree.

(c) To use this model for forecasting, we would first need forecasts of both $X_{1,t}$ and $X_{2,t}$ into the future. These could be obtained by taking averages of these variables over the equivalent months of the previous few decades. Then the model can be used to forecast electricity demand over the next 12 months by forecasting the N_t series using the method discussed in chapter 7 and plugging the forecasts of $X_{1,t}$, $X_{2,t}$ and N_t into the formula for Y_t.

(d) If the model was fitted using a standard regression package (thus modeling N_t as white noise), then the seasonality and autocorrelation in the data would have been ignored. This would result in less efficient parameter estimates and invalid estimates of their standard errors. In particular, tests for significance would be incorrect, as would prediction intervals. Also, when producing forecasts of Y_t, the forecasts of N_t would be all be zero. Hence, the model would not adequately allow for the seasonality or autocorrelation in the data.

8.4 **(a)** $b = 3$, $r = 1$, $s = 2$.

(b) ARIMA(2,0,0)

(c) $\omega_0 = -0.53$, $\omega_1 = -0.37$, $\omega_2 = -0.51$, $\delta_1 = 0.57$, $\delta_2 = 0$, $\theta_1 = \theta_2 = 0$, $\phi_1 = 1.53$, $\phi_2 = -0.63$.

(d) 27 seconds.

8.5 See the graphs on the following page.

8.6 **(a)** The three series are shown on page 147. For Set 1, four X_t values are needed (since v_1, v_2, v_3 and v_4 are all non-zero). Therefore 27 Y_t values can be produced. Similarly 26 Y_t values for Set 2 and 24 Y_t values for Set 3 can be calculated.

(b) The first model is

$$Y_t = \frac{2.0B}{1 - 0.7B} X_t + N_t.$$

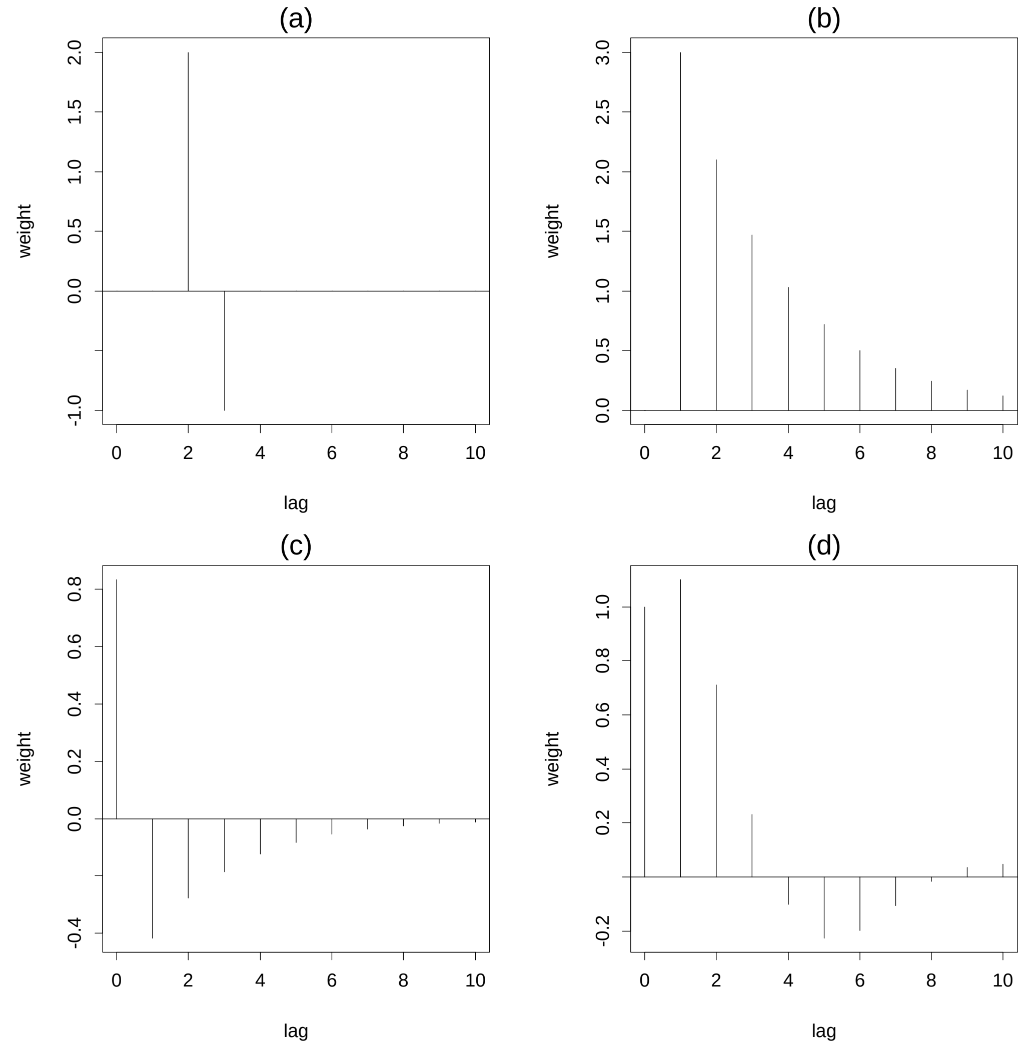
The simplest way to generate data for this transfer function is to rewrite it as follows

$$(1 - 0.7B)Y_t = (1 - 0.7B)B(2 - 1.4B)X_t + (1 - 0.7B)N_t$$

$$\text{so that} \qquad Y_t = 0.7Y_{t-1} + 2.0X_{t-1} - 1.4X_{t-2} + N_t - 0.7N_{t-1}.$$

Thus Y_t values can only be generated for times $t = 3, 4, \ldots$ since we need at least two previous X_t values. However, for $t = 3$, we also need Y_2. To start the process going, we have assumed here that $Y_2 = 0$. Other values could also have been used. The effect of this initialization is negligible after a few time periods. The second model is easier to generate as we can write it

$$Y_t = 1.2X_t + 2.0X_{t-1} - 0.8X_{t-2} + N_t.$$



Exercise 8.5: *Impulse response weights for the four different transfer functions.*

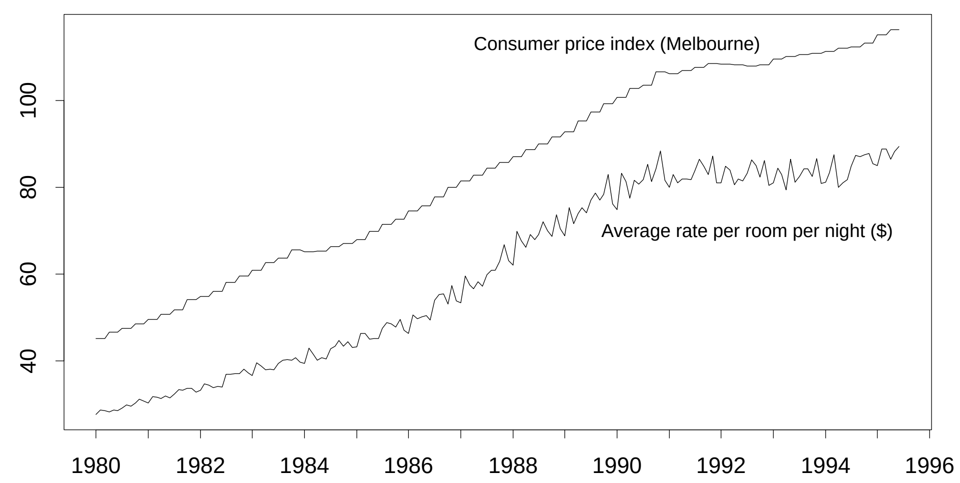
			Set 1	Set 2	Set 3	Set 4	Set 5
t	N_t	X_t	Y_t	Y_t	Y_t		
1	-0.8003	50					
2	0.8357	90				0.0	
3	1.4631	50				110.9	201.5
4	0.7332	30	58.7			51.3	64.7
5	0.3260	80	52.3	58.3		25.7	116.3
6	-0.7442	80	61.3	51.3		135.0	231.3
7	0.7362	30	65.7	62.7	74.7	143.8	132.7
8	1.1931	70	59.2	66.2	88.2	49.3	81.2
9	-1.4681	60	55.5	56.5	87.5	130.2	186.5
10	-0.5285	10	49.5	56.5	79.5	113.7	75.5
11	0.4314	40	37.4	50.4	83.4	16.4	20.4
12	-1.6341	20	27.4	35.4	72.4	75.5	94.4
13	0.8198	40	29.8	29.8	54.8	38.8	56.8
14	0.4183	20	30.4	29.4	47.4	79.0	88.4
15	-0.4065	10	23.6	29.6	41.6	38.6	19.6
16	-0.0615	30	19.9	23.9	40.9	19.3	39.9
17	0.1432	60	29.1	20.1	36.1	59.7	124.1
18	-1.0747	70	46.9	27.9	27.9	118.6	178.9
19	-0.5355	40	56.5	47.5	38.5	139.2	139.5
20	-0.1454	70	56.9	56.9	59.9	79.7	107.9
21	0.2088	10	49.2	57.2	74.2	140.1	120.2
22	-0.6854	30	34.3	48.3	76.3	19.2	-0.7
23	0.1182	30	28.1	35.1	73.1	60.1	88.1
24	0.6971	40	30.7	28.7	54.7	60.7	84.7
25	0.3698	30	34.4	30.4	43.4	80.3	92.4
26	-0.0802	100	46.9	33.9	43.9	59.9	147.9
27	-0.9202	60	64.1	46.1	44.1	199.1	247.1
28	1.1483	90	76.1	66.1	61.1	121.1	149.1
29	-0.1663	60	75.8	74.8	84.8	179.8	203.8
30	-0.5461	100	76.5	75.5	97.5	119.4	167.5

Generated data for Exercise 8.6

8.7 **(a)** The average cost of a night's accommodation is C/R.

(b) There are a number of ways this could be done. The simplest is to define the monthly CPI to be the same as that of the quarter. For example, January, February and March of 1980 would each have a CPI of 45.2; April, May and June 1980 would each have a CPI of 46.6; and so on. Other methods might involve fitting a smooth curve through the quarterly figures and using the curve to predict the CPI at other points along the time axis.

(c) See figure below.

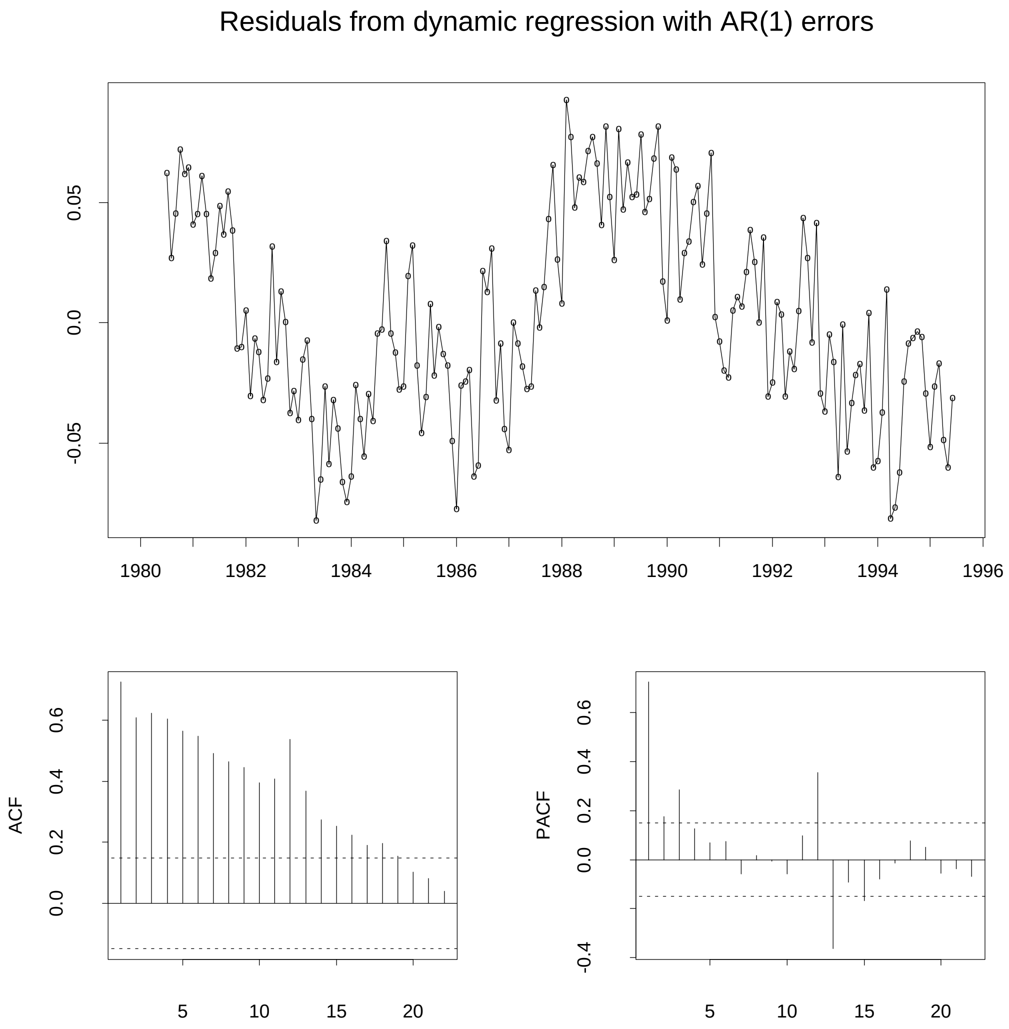


Exercise 8.7(c): *Time plots of average room rate and CPI.*

(d) Our preliminary model is

$$Y_t = a + (\nu_0 + \nu_1 B + \cdots + \nu_6 B^6)X_t + N_t$$

where Y_t denotes the average room rate, X_t denotes the CPI and N_t is an AR(1) process. The estimated errors from this model are shown in the figure on the previous page. They are clearly non-stationary and have some seasonality. So we difference both Y_t and X_t and refit the model with N_t specified as an ARIMA(1,0,0)$(1,0,0)_{12}$. The parameter estimates are shown below (as given by SAS).



Exercise 8.7(d): *Errors from regression model with AR(1) error term.*

Parameter	Estimate	s.e.	P-value
a	0.20200	0.2848	0.4791
ν_0	0.20730	0.2602	0.4267
ν_1	-0.41687	0.2634	0.1154
ν_2	0.23165	0.2655	0.3842
ν_3	0.32048	0.2716	0.2396
ν_4	-0.72093	0.2665	0.0075
ν_5	0.74707	0.2633	0.0051
ν_6	-0.36272	0.2656	0.1739

Thus the intercept and first four coefficients are not significant and can be omitted. Hence we select $b = 4$. We shall retain the last three coefficients for the moment. Since they show no clear pattern, we select $r = 0$ and $s = 3$ giving the model

$$Y_t = (\omega_0 + \omega_1 B + \omega_2 B^2)B^4 X_t + N_t.$$

Looking at the ACF and PACF of the error series (not shown) and trying a number of alternative models led us to the model ARIMA(2,1,0)(2,0,0)$_{12}$ for N_t. That is

$$(1 - \phi_1 B - \phi_2 B^2)(1 - \Phi_1 B^{12} - \Phi_2 B^{24})(1 - B)N_t = e_t.$$

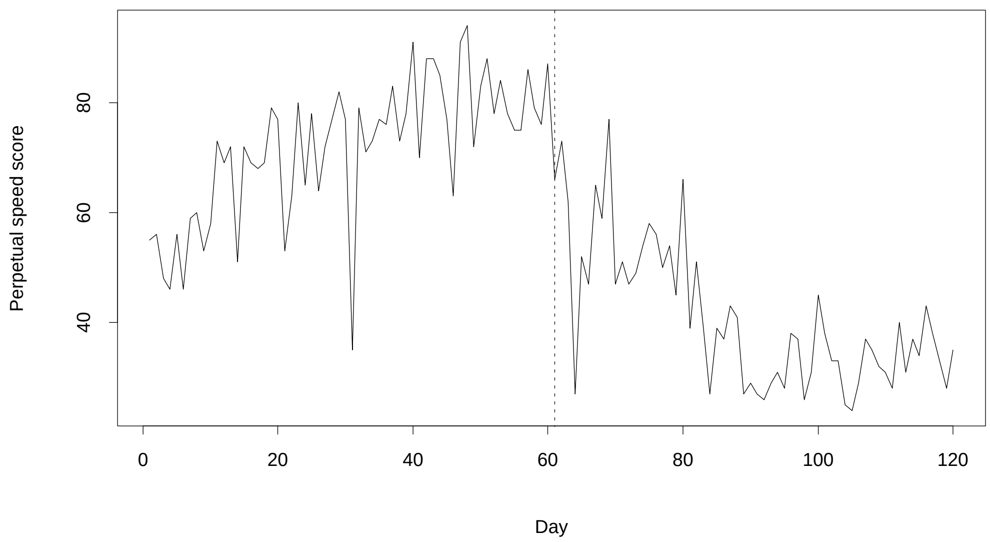
The parameter values (all significant) were

Parameter	ω_0	ω_1	ω_2	ϕ_1	ϕ_2	Φ_1	Φ_2
Estimate	0.52	0.61	-0.47	-0.49	-0.33	0.37	0.41

The model suggests that there is a lag of four months between changes in the CPI and changes in the price of travel accommodation. The seasonality inherent in the model may be due to seasonal price variation or due to the way CPI was estimated from quarterly data.

(e) Forecasts of CPI were obtained using Holt's method. These are only needed from November 1995 because of the time lag of 4 months. Actual data beyond June 1995 are given in the second column for comparison.

Month	Predicted Y_t	Actual Y_t	X_{t-4}	X_{t-5}	X_{t-6}
Jul 1995	90.4	94.0	115.0	115.0	115.0
Aug 1995	91.8	96.7	116.2	115.0	115.0
Sep 1995	92.0	94.8	116.2	116.2	115.0
Oct 1995	91.6	89.6	116.2	116.2	116.2
Nov 1995	93.4	95.8	116.9	116.2	116.2
Dec 1995	90.3	91.5	117.3	116.9	116.2
Jan 1996	90.4	92.0	117.7	117.3	116.9
Feb 1996	92.6	95.5	118.1	117.7	117.3
Mar 1996	94.7	100.6	118.5	118.1	117.7
Apr 1996	90.7	94.1	118.9	118.5	118.1
May 1996	91.5	97.2	119.3	118.9	118.5
Jun 1996	93.0	102.9	119.8	119.3	118.9



Exercise 8.8: *Time plot of daily perceptual speed scores for a schizophrenic patient. The drug intervention is shown at day 61.*

8.8 **(a)** See the figure above.

(b) The step intervention model with an ARIMA(0,1,1) error was used:

$$Y_t = \omega X_t + N_t \qquad \text{where} \quad (1 - B)N_t = \theta e_{t-1} + e_t$$

where Y_t denotes the perceptual speed score and X_t denotes the step dummy variable. The estimated coefficients were

Parameter	ω	θ
Estimate	-22.1	0.76

(c) The drug has lowered the perceptual speed score by about 22.

(d) The new model is

$$Y_t = \frac{\omega}{1 - \delta B} X_t + N_t \qquad \text{where} \quad (1 - B)N_t = \theta e_{t-1} + e_t$$

(An ARIMA(0,1,1) error was found to be the best model again.) Here the estimated coefficients were

Parameter	ω	δ	θ
Estimate	-13.21	0.54	0.76

The following accuracy measures show that the delayed effect model fits the data better.

Model	Step	Delayed step
MAPE	15.1	15.0
MSE	92.5	91.1
AIC	542.8	538.4

The forecasts for the two models are very similar. This is because the effect of the step in the delayed step model is almost complete at the end of the series, 60 days after the drug intervention.

(e) The best ARIMA model we found was an ARIMA(0,1,1) with $\theta = 0.69$. This gave MAPE=15.4, MSE=100.8 and AIC=550.9.

This model gives a flat forecast function (since we did not include a constant term). The forecast values are 33.9. Because the step effect is almost complete in the delayed step model, it also gives a virtually flat forecast function with forecast values of 34.1. Hence there is virtually no difference. If forecasts had been made earlier (for example, at day 80), there would have been a difference because the step effect would still be in progress and so the delayed step model would have showed a continuing decline in perceptual speed. The real advantage of the intervention model over the ARIMA model is that the intervention model provides a way of measuring and evaluating the effect of an intervention.

(f) If the drug varied from day to day and the reaction times depended on dose, then a better model would be a dynamic regression model with the the quantity of drug as an explanatory variable.

8.9 **(a)**

$$\begin{bmatrix} Y_t - Y_{t-1} \\ X_t - X_{t-1} \end{bmatrix} = \Phi_1 \begin{bmatrix} Y_{t-1} - Y_{t-2} \\ X_{t-1} - X_{t-2} \end{bmatrix} + \Phi_2 \begin{bmatrix} Y_{t-2} - Y_{t-3} \\ X_{t-2} - X_{t-3} \end{bmatrix} + \cdots + \Phi_{12} \begin{bmatrix} Y_{t-12} - Y_{t-13} \\ X_{t-12} - X_{t-13} \end{bmatrix} + \boldsymbol{Z}_t.$$

(b)

$$\begin{aligned} Y_t &= Y_{t-1} - 0.38(Y_{t-1} - Y_{t-2}) + 0.15(X_{t-1} - X_{t-2}) \\ &\quad - 0.37(Y_{t-2} - Y_{t-3}) + 0.13(X_{t-2} - X_{t-3}) + \cdots \\ &= 0.62Y_{t-1} + 0.01Y_{t-2} + 0.15X_{t-1} - 0.02X_{t-2} + \cdots \end{aligned}$$

(c)

- Multivariate model assumes feedback. That is, X_t depends on past values of Y_t. But regression does not allow this.
- Regression model does not assume X_t is random.
- Regression model allows Y_t to depend on X_t as well as past values $X_{t-1}, X_{t-2}, \ldots$. Multivariate AR only allows dependence on *past* values of $\{X_t\}$.
- For these data, it is unlikely room rates will substantially affect Y_t although it is possible. Small values in lower left of coefficient matrices suggest that X_t is not affecting Y_t. Y_t *should* depend on X_t. So regression is probably better.

8.10 **(a)** An AR(3) model can be written using the same procedure as the AR(2) model described in Section 8/5/1. Thus we define $X_{1,t} = Y_t$, $X_{2,t} = Y_{t-1}$ and $X_{3,t} = Y_{t-2}$. Then write

$$\boldsymbol{X}_t = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \boldsymbol{X}_{t-1} + \begin{bmatrix} a_t \\ 0 \\ 0 \end{bmatrix}$$

$$\text{and} \qquad Y_t = [1 \;\; 0 \;\; 0]\boldsymbol{X}_t.$$

This is now in state space form with

$$F = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, G = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, H = [1 \;\; 0 \;\; 0], \boldsymbol{e}_t = \begin{bmatrix} a_t \\ 0 \\ 0 \end{bmatrix} \text{ and } z_t = 0.$$

(b) An MA(1) can be written as $Y_t = \theta a_{t-1} + a_t$ where a_t is white noise. We can write this in state space form by letting $F = 0$, $G = \theta$, $e_t = a_{t-1}$, $z_t = a_t$ and $H = 1$. Thus

$$Y_t = X_t + a_t \qquad \text{and} \qquad X_t = \theta a_{t-1}.$$

(c) Holt's method is defined in Chapter 4 as

$$\begin{aligned} L_t &= \alpha Y_t + (1-\alpha)(L_{t-1} + b_{t-1}), \\ b_t &= \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}, \end{aligned}$$

with the one-step forecast as $F_{t+1} = L_t + b_t$. Hence the one-step error is $e_t = Y_t - L_{t-1} - b_{t-1}$. The first row can be written

$$\begin{aligned} L_t &= \alpha(Y_t - L_{t-1} - b_{t-1}) + L_{t-1} + b_{t-1} \\ &= \alpha e_t + L_{t-1} + b_{t-1} \end{aligned}$$

and the second row can be written

$$\begin{aligned} b_t &= \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1} \\ &= b_{t-1} + \beta(L_t - L_{t-1} - b_{t-1}) \\ &= b_{t-1} + \beta\alpha e_t \end{aligned}$$

using the first equation.

Now let $X_{t,1} = L_t$ and $X_{t,2} = b_t$. Then the state space form of the model is

$$\begin{aligned} \boldsymbol{X}_t &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \boldsymbol{X}_{t-1} + \begin{bmatrix} \alpha \\ \beta\alpha \end{bmatrix} e_t \\ Y_t &= [1 \quad 1]\, \boldsymbol{X}_{t-1} + e_t. \end{aligned}$$

(d) The state space form might be preferable because

- it allows missing values to be handled easily;
- it is easy to generalize to allow the parameters to change over time;
- the Kalman recursion equations can be used to calculate the forecasts and likelihood.

Chapter 9: Forecasting the long term

9.1 There is little doubt that the trends in computer power and memory show a very clear exponential growth while that of price is declining exponentially. It is therefore a question of time until computers that cost only a few hundred dollars will exist that can perform an incredible array of tasks which until now have been the sole prerogative of humans, for example playing chess (a high-power judgmental and creative process). It is therefore up to our imaginations to come up with future scenarios when such computers will be used as extensively as electrical appliances are used today. The trick is to free our thinking process so that we can come up with scenarios that are not constrained by our perception of the present when computers are being used mostly to make calculations.

9.2 As the cost of computers (including all of the peripherals such as printers and scanners) is being reduced drastically, and at the same time we will be getting soon to devices that will perform a great number of functions now done by separate machines, it will become more practical and economical to work at home. Furthermore, the size of these all-purpose machines is being continuously reduced. In the next five to ten years we will be able to have everything that is provided to us now in an office at home with two machines: one a powerful all-inclusive computer and the other a printer-scanner-photocopier-fax machine. Moreover these two machines will be connected to any network we wish via modems so that we can communicate and get information from anywhere.

9.3 As it was also mentioned in Exercise 9.1, there is no doubt that the trend in computer and equipment prices are declining exponentially at a fast rate. This would make it possible for everyone to be able to afford them and be able to have an office not only at home but at any other place he or she wishes, including one's car, a hotel room, a summer vacation residence, or a sail boat.

9.4 Statements like those referred to in Exercise 9.4 abound and demonstrate the short-sightedness of peoples' ability to predict the future. As a matter of fact as late as the beginning of our century people did not predict all four major inventions of the Industrial Revolution (cars, telephones, electrical appliances and television) that have dramatically changed our lives. Moreover, they did not predict the huge impact of computers even as late as the beginning of the 1950s. This is why we must break from our present mode of thinking and see things in a different, new light. This is where scenarios and analogies can be extremely useful.

Chapter 10: Judgmental forecasting and adjustments

10.1 Phillips' problems have to do with the management bias of overoptimism, that is believing that all changes will be successful and that they can overcome peoples' resistance to change. This is not true, but we tend to believe that most organisational changes are successful because we hear and we read about the successful ones while there is very little mention of those that fail. Introducing changes must be considered, therefore, in an objective manner and our ability to succeed estimated correctly.

10.2 The quote by Glassman illustrates the extent to which professional, expert investment managers underperform the average of the market. Business Week, Fortune and other business journals regularly publish summary statistics of the performance of mutual funds and other professionally-managed funds, benchmarking them with the Standard & Poor or other appropriate indexes. The instructor can therefore get some more recent comparisons than those shown in Chapter 10 and show them in class.

10.3 Assuming that the length of cycles varies considerably we have little way to say how long it will take until the expansion started in May 1991 will be interrupted. Unfortunately the length (and depth) of cycles varies a great deal making it extremely difficult to say how long an expansion will last. It will all depend on the specific situation involved that will require judgmental inputs, structured in such a way as to avoid biases and other problems.

10.4 There are twenty 8s that one will encounter when counting from 0–100. When given this exercise most people say nine or ten because they are not counting the eights coming from 81 to 87, and 89 (they usually count the 8s in 88 often one time).

Chapter 11: The use of forecasting methods in practice

11.1 The results of Table 1 are very similar to those of the previous M-Competitions. As a matter of fact the resemblance is phenomenal given the fact that the series used and the time horizon they refer to are completely different.

11.2 In our view the combined method will do extremely well. More specifically its accuracy will be higher than the individual methods being combined while its variance of forecasting error will be smaller than that of the methods involved.

11.3 It seems that proponents of new forecasting methods usually exaggerate their benefits. This has been the case with methods under the banner of neural networks, machine learning and expert systems. These methods did not do well in the M3-IJF Competition. In addition only few experts participated in the competition using such methods, even though more than a hundred were contacted (and invited to participate) and more than fifty expressed an interest in the M3-IJF Competition, indicating that they would “possibly” participate . In the final analysis it seems that it is not so simple to run a great number of series by such methods resulting in not too many participants from such methods.

Chapter 12: Implementing forecasting: its uses, advantages, and limitations

The exercises for Chapter 12 are general and can be answered by referring to the text of Chapter 12 which covers each one of the topics. Each instructor can therefore form his/her way of answering these exercises.