It appears that a simple linear regression model is not appropriate because there is curvature in the plot.

b. The MINITAB output is shown below:

```
The regression equation is
2005% = 17.1 + 3.15 1999% - 0.0445 1999%Sq

Predictor    Coef      SE Coef       T       p

Constant    17.099     4.639       3.69   0.003
1999%        3.1462    0.4971      6.33   0.000
1999%Sq     -0.04454   0.01018    -4.37   0.001

S = 5.667   R-sq = 89.7%   R-sq(adj) = 88.1%

Analysis of Variance

SOURCE            DF     SS       MS      F      p

Regression         2  3646.3  1823.2  56.78  0.000
Residual Error    13   417.4    32.1
Total             15  4063.8
```

c. The MINITAB output is shown below:

```
The regression equation is
Log2000% = 1.17 + 0.449 Log1999%

Predictor      Coef     SE Coef     T        p

Constant     1.17420   0.07468   15.72   0.000
Log1999%     0.44895   0.05978    7.51   0.000

S = 0.08011   R-sq = 80.1%   R-sq(adj) = 78.7%

Analysis of Variance

SOURCE            DF     SS        MS        F      p

Regression         1  0.36199  0.36199  56.40  0.000
Residual Error    14  0.08985  0.00642
Total             15  0.45184
```

d. The estimated regression in part (b) is preferred because it explains a higher percentage of the variability in the dependent variable.

**10** a. SSE = SST − SSR = 1805 − 1760 = 45

MSR = 1760/4 = 440   MSE = 45/25 = 1.8

$F = 440/1.8 = 244.44$

$F_{0.05} = 2.76$ (4 degrees of freedom numerator and 25 denominator)

Since 244.44 > 2.76, variables $x_1$ and $x_4$ contribute significantly to the model

b. $SSE(x_1, x_2, x_3, x_4) = 45$

c. $SSE(x_2, x_3) = 1805 - 1705 = 100$

d. $F = \dfrac{(100 - 45)/2}{1.8} = 15.28$

$F_{0.05} = 3.39$ (2 numerator and 25 denominator DF) Since 15.28 > 3.39 we conclude that $x_1$ and $x_3$ contribute significantly to the model.

**12** For the first model featuring the five predictors $x_1$, $x_2$, $x_3$, $x_4$ and $x_5$, the significant F ratio from the ANOVA table ($p$-value = 0.005 < $\alpha$ = 0.05) suggests that the overall model is a significant fit to the data. Yet none of the individual t tests associated with each of the regression slopes beforehand are significant except that for $x_4$ ($p$-value = 0.005 < $\alpha$ = 0.05). From the VIF's which are all close to 1, multicollinearity would not appear to be a problem for the data. The $R$ Square of 66.3 per cent indicates that the multiple regression model explains 66.3 per cent of the variation in the response variable and this might be regarded as quite favourable. On the down side the model suffers from a single outlier according to MINITAB but for a sample of size 20 this does not seem unreasonable. The Durbin-Watson statistic is 1.72 but for a two-sided Durbin-Watson test the relevant $d_L$ and $d_U$ values (based on $n = 20$ and $k = 5$ predictors) are 0.70 and 1.87. As $d_L < 1.72 < d_U$ we deduce the test is inconclusive.

The second model is a simple regression with just $x_4$ as the predictor. The model is significant according to both the overall $F$ test and the specific t tests associated with the regression slope for $x_4$. As would be expected the $R$ square value has dropped – in this case to 51.7 per cent. Again there is an outlier (albeit for observation 12 now instead of observation 1 previously but with a corresponding standardized residual of −2.07 this does not look too serious.)

To check if the earlier five predictor model is a significant improvement on this one predictor model, a partial $F$ test can be undertaken. The relevant calculation using equation (16,11) is as follows (note that $p = 5, q = 1$):

$$F = \frac{\dfrac{S\overline{S}E(x_1, x_2, \ldots, x_q) - SSE(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{p - q}}{\dfrac{SSE(x_1, x_2, \ldots, x_q, x_{q+1}, \ldots, x_p)}{n - p - 1}}$$

$$= \frac{\dfrac{23.002 - 16.032}{5 - 1}}{\dfrac{16.032}{14}}$$

$$= 1.52 < 3.11 = F_{0.05}(4,14)$$

Hence we are not able to reject $H_0$: $\beta_1 = \beta_2 = \beta_3 = \beta_5 = 0$ at a 5 per cent significance level and deduce that the five predictor model is not a significant improvement on the corresponding 1 predictor equivalent.

Note that because of the 'ln' transformation on y the relationship between $y$ and $x_4$ has an essentially exponential character despite the fact that we have effectively used a linear modelling formulation for the analysis.

**14** Let Health = 1 if health-drugs

Health = 0 if energy-international or other

```
The regression equation is
P/E = 10.8 + 0.430 Sales% + 10.6 Health

Predictor    Coef     SE Coef      T       p

Constant    10.817    3.143      3.44   0.004
Sales%       0.4297   0.1813     2.37   0.034
Health      10.600    2.750      3.85   0.002

S = 5.012    R-sq = 55.4%    R-sq(adj) = 48.5%

Analysis of Variance

SOURCE            DF     SS       MS      F      p

Regression         2  405.69  202.85  8.07  0.005
Residual Error    13  326.59   25.12
Total             15  732.28

Source    DF    Seq SS

Sales%     1     32.36

Health     1    373.33
```



**16** a. From the best subsets regression summary, the five predictor model with an $R$ Square of 86.2 per cent is almost as good on all measures as the full six predictor model represented by the bottom line of the table. The same five predictor model is described in detail after the correlation matrix and can be seen from the ANOVA $F$ statistic to be significant overall. Corresponding $t$ statistics are also significant (though technically the $p$-value (of 0.054) associated with the regression slope for the *pop* variable is just slightly above the test size of 5 per cent).

b. Clearly multicollinearity is a problem here. This is confirmed by significant correlations between predictor variables e.g. *pr* and *con*. Also the sign of the coefficient of the *con* predictor in the detailed regression output is opposite to that of the corresponding correlation between *con* and *ao*.

c. Yes. In these circumstances the two predictor model from Stepwise now looks technically more appealing.

**18** a. Because of the significant correlations between predictors in the summary table at the beginning of the output the possibility of multicollinearity looms large for any subsequent regression modelling. From the Best Subsets table the three predictor model with an $R$ Square of 86.2 per cent compares well with the four predictor baseline model. This is essentially the same model picked out by the Stepwise (backward elimination) analysis afterwards. According to this, all predictors accept Support look as if they could be usefully retained in for regression modelling analysis. Following on, the detailed output for a three predictor regression model shows that Retired, Unemployment and Total Staff are all significant predictors of the response variable, Crimes. Because of relatively low VIF values the model does not seem to suffer from multicollinearity problems mentioned earlier. With an $R$ Square of 84.6 per cent it compares with the three predictor model described earlier fairly well. The Durbin-Watson statistic of 2.22 is not problematic since if $n = 26$ and $k = 3$ then $d_L = 1.04$, $d_U = 1.54$. As $1.54 = d_U < 2.22 < 4\text{-}d_U = 2.46$ we deduce there is no evidence of first order serial correlation of residuals present.

b. For the relevant two predictor model the root mean square error $s = 89.041$. Correspondingly the error sums of squares would be $23 \times 89.041^2 = 182\,350.9$. By comparing this model with the last three predictor model in a. A partial $F$ test statistic from equation (16.11) can be calculated as follows:

$$F = \frac{\dfrac{SSE(x_1, x_2) - SSE(x_1, x_2, x_3)}{p - q}}{\dfrac{SSE(x_1, x_2, x_3)}{n - p - 1}}$$

$$= \frac{\dfrac{182\,350.9 - 150\,468}{3 - 2}}{\dfrac{150\,468}{22}}$$

$$= 26.66 > 4.30 = F_{.95}(1,22)$$

Hence we reject $H_0$: $\beta_3 = 0$ at the 5 per cent significance level and deduce that the three predictor model is a significant improvement on the corresponding two predictor one.

c. Following on from *b*. the three predictor model based on Retired, Unemployment and Total Staff would be preferred.

## Chapter 17

### Solutions

**2** a. From the price relative we see the percentage increase was (132 − 100) = 32 per cent.

b. Divide the current cost by the price relative and multiply by 100.

$$2000 \text{ cost} = \frac{€10.75}{132}(100) = €8.14$$

**4** a. A, 110: B, 106: C,113

b. 110

c. 109. This implies a 9 per cent increase over the two-year period.

**6** $I = \dfrac{0.19(500) + 1.80(50) + 4.20(100) + 13.20(40)}{0.15(500) + 1.60(50) + 4.50(100) + 12.00(40)}(100)$

$= 104$

Paasche index

**8** 164

**10** a. Price Relatives A = (3.95 / 2.50) 100 = 158

B = (9.90 / 8.75) 100 = 113

C = (0.95 /0.99) 100 = 96

b.

| Item | Price relatives | Base-period price | Quantity | Weight $P_{i0}Q_i$ | Weighted price relatives |
|---|---|---|---|---|---|
| A | 158 | 2.50 | 25 | 62.5 | 9 875 |
| B | 113 | 8.75 | 15 | 131.3 | 14 837 |
| C | 96 | 0.99 | 60 | 59.4 | 5 702 |
| | | | | **253.2** | **30 414** |