4. **(a)** H_1: The data does not fit the proposed model.
   (b) 30, 60, 90, 60, 60
   (c) p-value $= 0.339 > 0.05$. Fail to reject the null hypothesis. Data may fit the model.
5. **(a)** H_0: reports of side effects are not different among the two groups
   H_1: reports of side effects are different among the two groups
   (b) 12.5, 17.5, 10, 210
   (c) This is a GOF test. p-value $= 0.0475 < 0.05$. Reject the null hypothesis. There is evidence that the two groups differ.
6. **(a)** H_0: The collected data fits a fair die distribution.
   H_1: The collected data does not fit a fair die distribution.
   (b) 100 in each cell
   (c) Either $\chi^2_{calc} = 5.7 < 11.07$, or p-value $= 0.337 > 0.05$. We fail to reject the null hypothesis. The die appears to be fair.

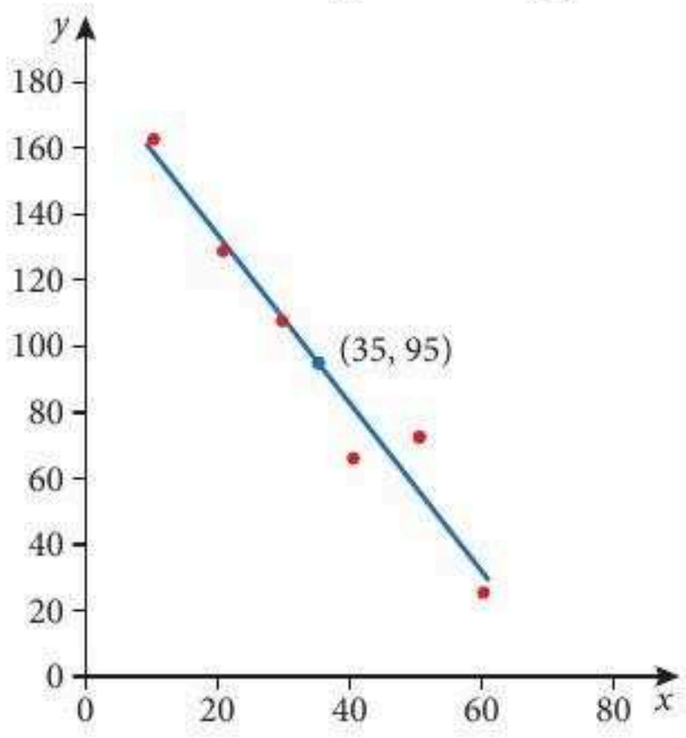
Chapter 13 Practice questions

1. **(a)** H_0: the type of Latin dance the viewer prefers is independent of their age.
   (b) 18
   (c) $p = 0.0876$
   (d) p-value > 0.05. The producer's claim may be justified.
2. **(a)** Conservatives 435, Liberals 615, Greens 225, rightists 225
   (b) H_0: voter support has not changed since election.
   H_1: voter support has changed since election.
   (c) GOF test with df $= 3$. P-value $= 0.009 \leqslant 0.10$. We reject the null hypothesis and conclude that voters support has changed.
3. **(a)** **(i)** H_0: age and opinion (about the reduction) are independent.
   (ii) H_1: age and opinion (about the reduction) are not independent.
   (b) 2
   (c) $\frac{80 \times 35}{130} = 21.5$
   (d) **(i)** 10.3
   (ii) 0.00573
   (e) Since p-value < 0.01, reject H_0, or χ^2statistic $> \chi^2$ critical, reject H_0.
4. This is a GOF test with 4 df. p-value $= 0.0230 < 0.10$. Reject H_0. Absences differ from one day to the other.
5. **(a)** This is a one tail t test of difference of means. The populations are approximately normal and with equal variances.
   (b) $H_0 : \mu_{device} = \mu_{usual}$; $H_1 : \mu_{device} > \mu$.
   (c) p-value $= 0.0213 < 0.10$. We reject the null hypothesis and conclude that users pay more with the device.
6. **(a)** Age and preferred destination are independent
   (b) $(4 - 1) \times (5 - 1) = 12$.
   (c) $\chi^2 > 21.026$
   (d) $\frac{285 \times 420}{1200} = 99.8$
7. **(a)** 27.9 **(b)** 0.0321
   (c) Reject the null hypothesis, TV show preference is not independent of gender.
8. This is a two-tail t test of difference of means. The populations are approximately normal and with equal variances.
   $H_0 : \mu_{violent} = \mu_{neutral}$; $H_1 : \mu_{violent} \neq \mu$.
   p-value $= 0.0000915 < 0.10$. We reject the null hypothesis and conclude that content of program affect viewers memory.

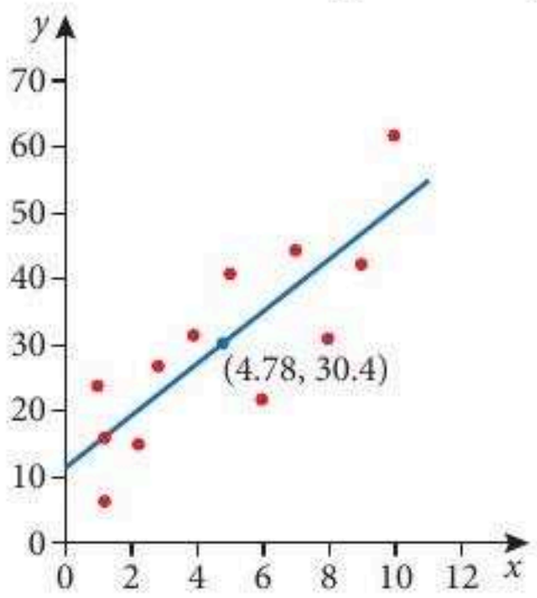
Chapter 14

Exercise 14.1

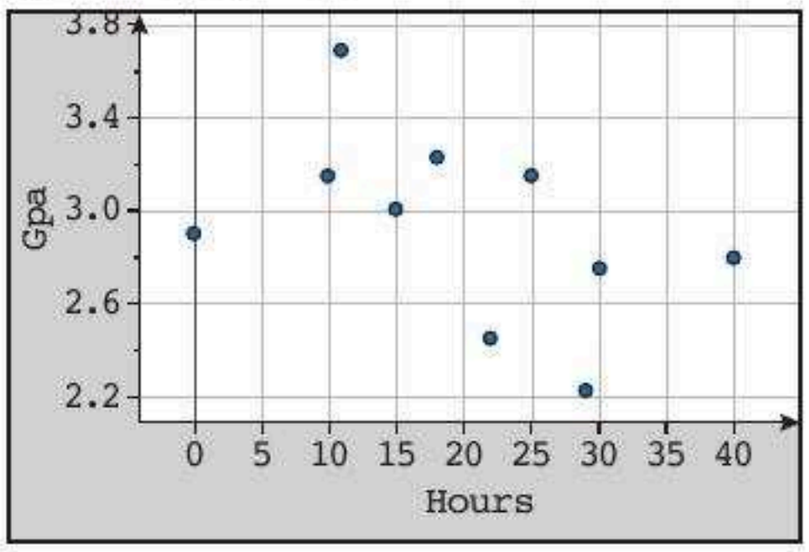
1. $(\bar{x}, \bar{y})$
2. form, direction, strength, unusual features
3. **(a)** Precipitation is the explanatory variable; autism prevalence rate is the response variable
   (b) No; no matter how strong the association is, it does not prove that precipitation causes autism, only that they are associated.
4. **(a)** No association.
   (b) Strong nonlinear association, possibly quadratic.
   (c) Nearly perfect negative linear association.
   (d) Strong positive linear association.
   (e) Moderate negative linear association.
   (f) No association.
   (g) Strong positive nonlinear association, possibly exponential.
   (h) Mostly strong positive linear association, but a cluster of outliers is a departure from the major pattern.
   (i) Very strong negative linear association with one outlier.
5. **(a)** The best fit line is approximately $y = -2.6x + 185$



(b) The best fit line is approximately $y = 4x + 12$



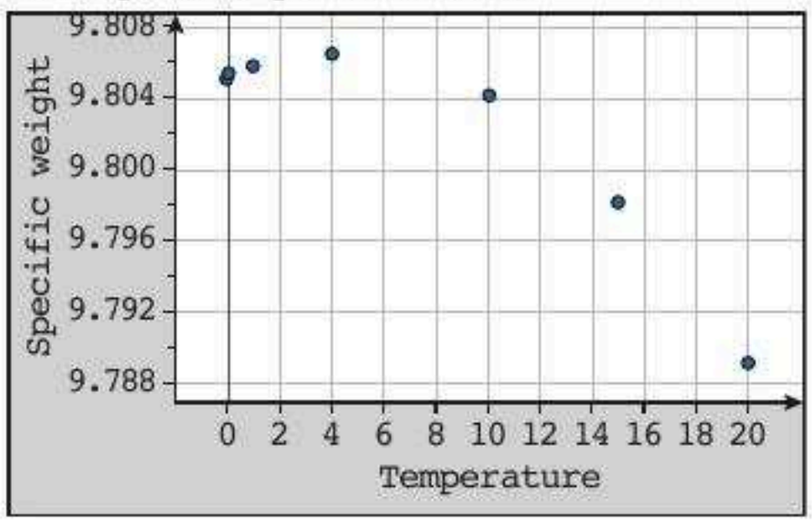
11. **(a)** The association appears moderate, negative, and approximately linear.



(b) $r_s = -0.533$; there is a moderate negative rank correlation between GPA and hours worked.

(c) $r = -0.461$; since the form of the association is approximately linear with no strong outliers, r is relatively close to r_s.

12. **(a)** The association is generally negative, but is nonlinear with a possible quadratic form.



(b) $r_s = -0.643$; there is a moderate negative rank correlation between temperature and the specific weight of water.

(c) The form does not appear monotonic.

Exercise 14.3

1. **(a)** $L = 0.00475h + 0.242$. On average, each additional hour increases mass loss by 0.00475%. The L intercept of 0.242 is not meaningful in this context (it is an extrapolation, and it suggests a mass loss of 0.242% for 0 hours).

(b) $k = 0.475\%$ **(c)** 2.14%

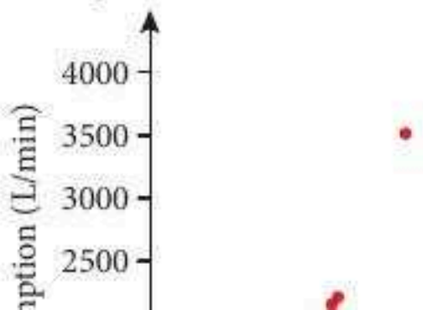
(d) Although the model appears quite good, the LSRL we generated should only be used to predict mass loss from hours in the acid bath. Using the model 'in reverse' lowers our confidence in the prediction significantly.

2. **(a)** $E = 1.85t + 16.4$. On average, for each additional second in 0–60 mph time, fuel efficiency increases my 1.85 miles gal^{-1}. The E-intercept of 16.4 is not meaningful in this context; it is an extrapolation and a car would have to have instantaneous acceleration to 60 mph!

(b) 26.9 miles gal^{-1}. Since the correlation appears moderate (not strong), there may be other factors.

(c) 9 seconds is beyond the range of the observed value of the explanatory variable; any prediction would be an extrapolation.

3. **(a)**



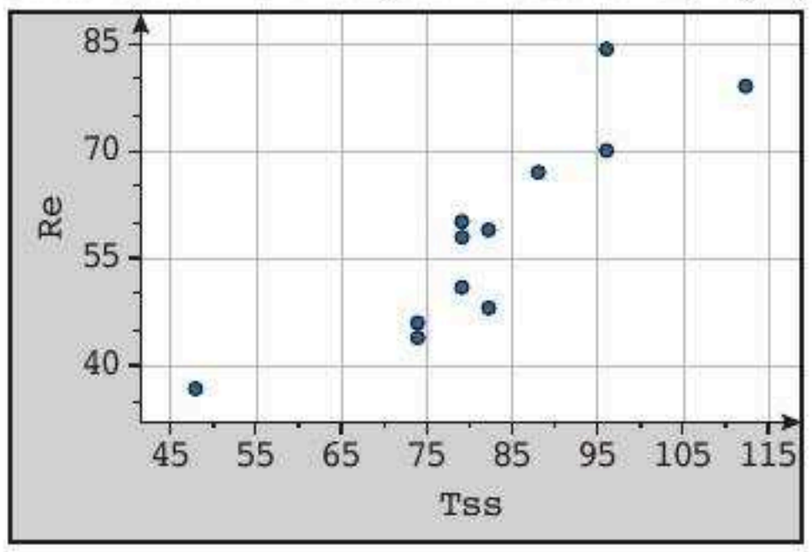
The association between the number of seats and fuel consumption is strong, positive, and approximately linear. There is one possible outlier but it appears to fit the general trend.

(b) $F = 0.462n + 3.65$. The fuel consumption increases by 0.462 L min^{-1} for each additional seat. The F-intercept of 3.95 could indcate fuel use beyond what is used to lift the weight of the seats.

(c) 166 L min^{-1}

(d) The data appears to have a slight nonlinear form in the scatter diagram. It would make sense that fuel consumption would increase in a faster-than-linear rate as the limits of current technology are reached, so a linear model may not be the most appropriate.

4. **(a)** The scatter diagram shows a strong, positive, linear association. There is a possible outlier at (48, 37).



(b) $E = 0.815(S) - 8.62$. For each additional unit increase in Training Stress Score, Relative Effort increases by 0.815 units. The E-intercept of -8.62 is not meaningful in this context.

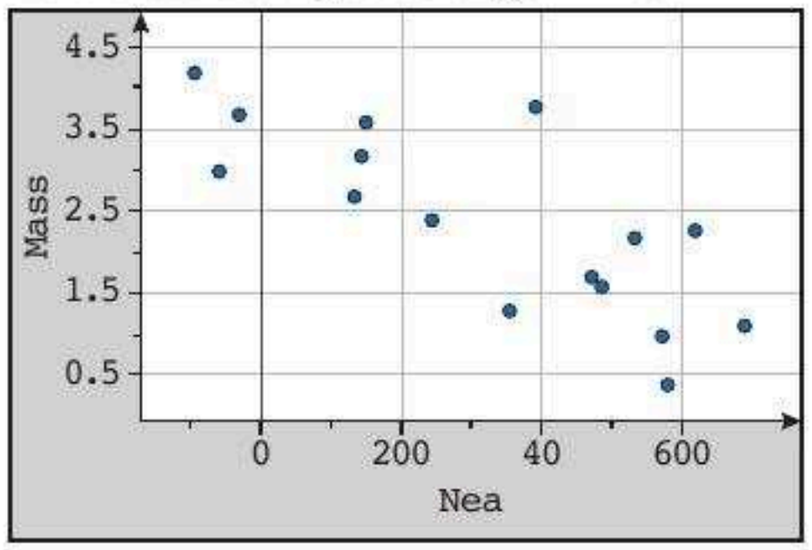
(c) 40.3

(d) $48 \leqslant S \leqslant 112$

(e) Since the low outlier is the minimum value in the domain, if we remove the outlier we would need to adjust the domain to $74 \leqslant S \leqslant 112$ and a prediction based on a Relative Effort of 60 would then become an extrapolation.

(f) Without (48,37), the LSRL is $E = 0.996(S) - 24.7$ with $R^2 = 0.756$. The LSRL has changed significantly.

5. **(a)** The scatter diagram shows a strong negative approximately linear association between change in nonexercise activity and change in mass.



(b) $M = -0.00344N + 3.51$. On average, for each additional 100 calories in change of NEA, change in mass decreases by 0.344 kg. The M-intercept suggests that when NEA does not change, mass will increase by 3.51 kg.

(c) 2.822 kg

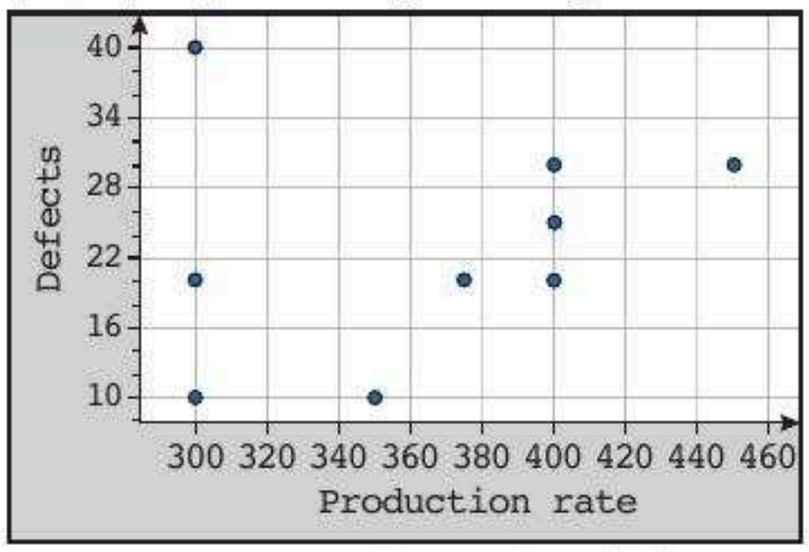
(d) A negative change in NEA suggests that the individual did less nonexercise activity after being overfed.

(e) $-94 \leqslant N \leqslant 690$

(f) The change in mass for $N = 1100$ would be negative, suggesting that the person was able to lose weight by eating more!

6. (a) Production Rate is the explanatory variable; Defects is the response variable.

(b) There appears to be a weak positive association between production rate and defects. However, an outlier at (300,40) may be affecting the strength of the association.

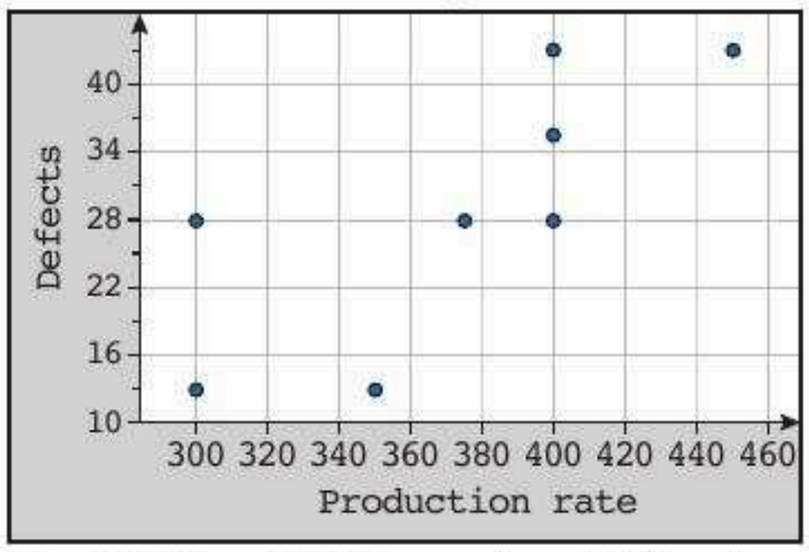


(c) $D = 0.0479R + 5.67$; the number of defects increases by about 5 for each 100 unit increase in production rate. The D-intercept of 5.67 is not meaningful in this context.

(d) $r = 0.295$; there is a weak positive linear correlation.

(e) The low value of r suggests that there are likely to be other factors determining the variation in the number of defects. It is better to try to find other factors instead.

(f) There appears to be a moderate positive approximately linear association between production rate and defects.



$D = 0.113R - 21.3$; The number of defects increases by about 11 for each 100 unit increase in production rate. The D-intercept of -21.3 is not meaningful in this context.

$r = 0.794$; there is a moderately strong positive linear correlation between defect and production rate.

The production rate appears to be correlated to the number of defects. Lower production rates should be attempted and further data can then be recorded.

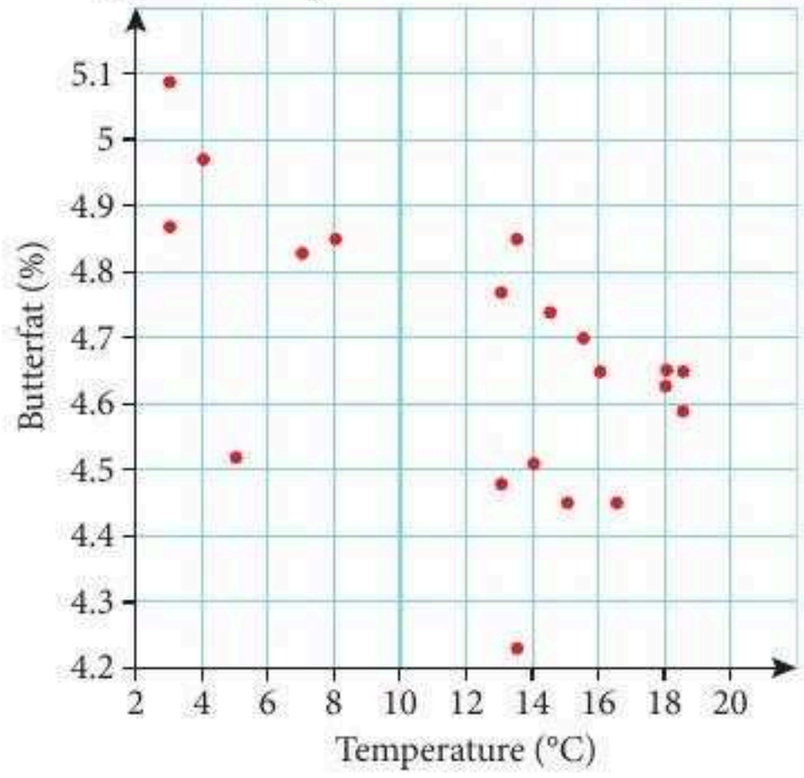
(g) It is OK to remove the outlier to investigate how influential it is. However, we must be careful to not put too much confidence in our predictions and recommendations using the revised data. Instead, we must investigate the outlier and see what caused it: was it a particularly unskilled worker? Mis-entered data? A power failure during production or other failure? Something else? If there are no unusual causes for the outlier, we should not remove it in our final analysis.

7. (a) There is a strong positive association between year and millions of active monthly users. The form appears nonlinear or piecewise linear.

(b) Suitable domains would be $0 \leqslant y \leqslant 5$ and $5 \leqslant y \leqslant 8$

8. (a) Temperature is the explanatory variable, butterfat is the response variable.

(b) The association appears to be negative and moderate and approximately linear.



There is a gap in the data; why are there no data values for temperatures in the interval $8 < T < 13$? We should proceed with caution.

(c) The LSRL is $F = -0.0216T + 4.94$. The butterfat content decreases by 0.02% for each increase of 1 °C. The F-intercept of 4.94 indicates that at a temperature of 0 °C, we predict that the butterfat content would be 4.94% – beware, this is an extrapolation.

(d) $r = -0.564$; there is a moderate negative linear correlation between butterfat and temperature.

(e) Correlation is not causation: while temperature and butterfat may have a moderate correlation, we cannot claim that lowering the temperature will cause increased butterfat content. Further research is needed; an investment in a climate-controlled barn may be an expensive experiment.

Chapter 14 Practice questions

1. (a) The value of r in the interval $-1 \leqslant r \leqslant 1$,

(b) If the association is negative, the value of r must be negative.

(c) If life expectancy increases as body mass increases, then r_s must be positive.

(d) The gradient in the LSRL is negative, but r is positive.

2. (a) 0 **(b)** -1 **(c)** $+1$ **(d)** $+1$
(e) 0 **(f)** -1

3. (a) The scatter diagram shows an strong, negative, approximately linear association, with $r = -0.984$.