ANSWERS CHAPTER 10

THINK IT OVER

think it over

TIO 10.1: (a) The result is extreme but it indicates that initially money motivates but once the extra money becomes the 'norm' people tend to revert back to previous behaviour.

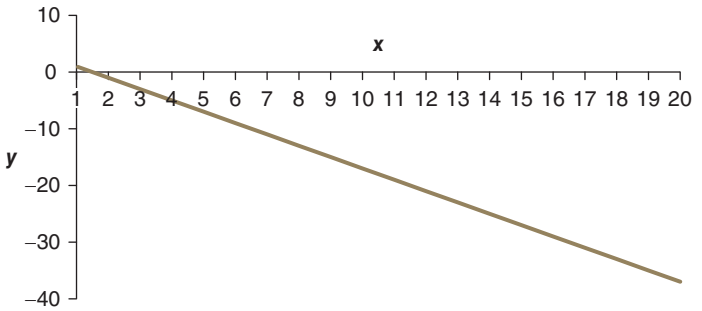
(b) If people feel they are getting a decent wage for the work they do and live comfortably, then they tend to perform to the best of their abilities (in most cases!). Beyond this level, the extra money is nice but soon becomes normal.

TIO 10.2: It is cold in the UK during the winter months. Days are relatively short and people like something to do. A common phrase is 'retail therapy', in other words, people think that by spending money they will feel better – short term perhaps!

TIO 10.3: A model is a replica of a real thing, e.g. a model aeroplane is a replica of a real aeroplane but with less functionality. It still gives an 'idea' of what a real plane looks like. Similarly in mathematics and statistics, a model attempts to replicate a physical system or behaviour. In order to be useful, assumptions are normally made since it is virtually impossible to model a complex system or behaviour accurately.

TIO 10.4: (a)

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
y	1	–1	–3	–5	–7	–9	–11	–13	–15	–17	–19	–21	–23	–25	–27	–29	–31	–33	–35	–37



(b) As the value of *x* increases, the value of *y* decreases. The *y* intercept (where the curve crosses the vertical axis) has a value of 3, i.e. *x* = 0.

TIO 10.5: At *x* = 50, *y* = –97. Mathematically it is accurate, but if the prediction was based on a statistical analysis, then it is assuming that the initial conditions remain constant, the relationship between the variables remains linear and nothing 'funny' happens at the higher values for *x*.

TIO 10.6:

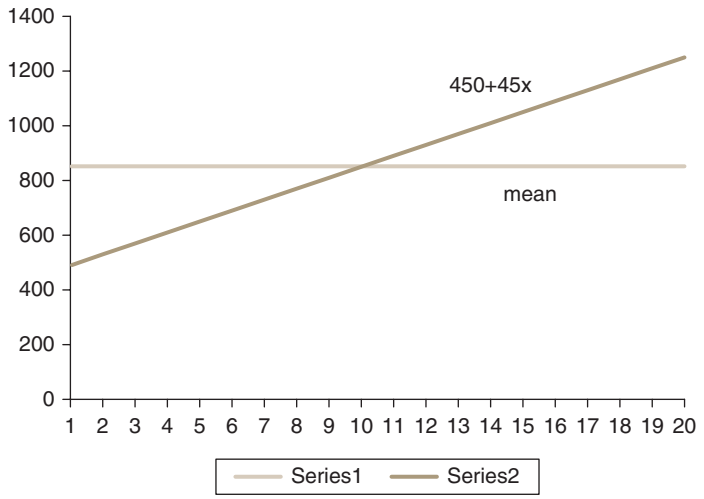
$$\text{deviation} = \sum(\text{observed} - \text{model})^2$$

$$\text{sum of squares of residuals} = \sum(y_i - \hat{y}_i)^2$$

If the SSE value = 0 it means every point fits exactly onto the regression line, i.e. no errors. Highly unlikely, and if you ever come across such a case be wary!

TIO 10.7: The mean is used as a model and the distance squared of some points away from the mean which cannot be accounted for divided by the number of degrees of freedom is a measure of how much the error varies.

TIO 10.8: (a)



(b) The mean weekly sales curve is not really very useful. Provided the regression equation was a reasonably good fit to the data, it can be seen there would be large differences between it and the mean.

TIO 10.9: The null hypothesis states that the predictor value is equal to zero whereas the alternative hypothesis states the predictor value is not equal to zero, i.e. different from the mean. Since we didn't know before the regression equation was formulated if the difference between the mean and predicted value would be greater or less than the zero, then a two-tailed test is appropriate.

TIO 10.10: (a) Weekly bar sales can vary by 24% each week with average sales of £18,130.
(b) The 95% confidence interval states that we are 95% confident that weekly bar sales will be somewhere between £13,790 and £22,470.

TIO 10.11: It greatly depends on what you are modelling. The best advice is 'KISS – Keep It Simple Stupid'. In other words; the more predictors you have the more complicated the model which increases the probability of misinterpretation and errors.

TIO 10.12:

B_1	B_2	B_3	B_4
1	0	0	0
0	1	0	0
1	1	0	0
0	0	1	0
1	0	1	0
0	1	1	0
1	1	1	0
0	0	0	1
1	0	0	1
0	1	0	1
1	1	0	1
0	0	1	1
1	0	1	1
0	1	1	1
1	1	1	1

TIO 10.13: No it does not. What the *t*-test indicates is which of the predictors have the most 'influence' on the model. It could be, to simplify the model, that the least significant predictors can be left off and the analysis rerun to see how the model changes.

TIO 10.14: (1) The data has been recorded correctly. (2) Scan through the data and investigate any points that do not 'look right'.

TIO 10.15: Investigate the data point further. Do the checks mentioned in TIO 10.14. Run the analysis excluding this data point and see the difference in the regression model.

TIO 10.16: Not necessarily. The more complicated the model the more difficult accurate interpretation is.

TIO 10.17: covariance.s returns the sample covariance whereas covariance.p returns the population covariance.

TIO 10.18: Plot the data and see if a model could be constructed. It could be the data is completely random and therefore a valid model is not possible.

EXERCISES

1. $\left(x_i - \bar{x}\right)^2$ is shorthand for $\left(x_i - \bar{x}\right) \times \left(x_i - \bar{x}\right)$.

2. (a) The covariance is positive.
   (b) The covariance is approximately 12.3. This means that as the customer variable deviates from its mean so does the sales variable in the same direction.

3. (a) There is a positive relationship, i.e. as the temperature increases so do the sales of ice creams.
   (b) Positive correlation.
   (c) The sum of the product of the variances will always be greater than the products of the standard deviations

$$r = \frac{\text{cov}_{xy}}{S_x S_y} = \frac{\Sigma\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{(N-1)S_x a}$$

4. The correlation coefficient, the *F*-test and the *t*-test on individual predictors.

5. (a) The regression model.
   (b) Using Excel, the *R* squared value is 0.91 which means the model is a good fit. Excel gives the SST value as 1576.7.
   (c) The total sum of squares is equal to the model sum of squares plus the error sum of squares.

6. Because with *n* number of parameters once you know *n* – 2 you have enough information to know the remaining values of the parameters.

7. The mean square due to the model is equal to the model sum of squares divided by the number of predictors. This gives an estimate of the variance of the error.

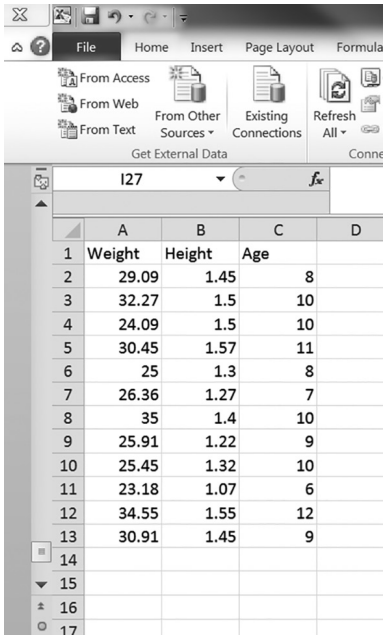
8. A value of 10 since $F = \frac{MSM}{MSE}$

9. A large value for MSM and a low value for MSE would give a large value for *F* indicating a good model.

10. The horizontal line represents the mean, the filled circles the observed data and the solid angled line the regression equation.

11. No. The *F*-test will tell you that at least one parameter is not equal to zero.

12. $w = 1.66 + 15.294b_1 + 0.685b_2$.



	A	B	C	D
1	Weight	Height	Age	
2	29.09	1.45	8	
3	32.27	1.5	10	
4	24.09	1.5	10	
5	30.45	1.57	11	
6	25	1.3	8	
7	26.36	1.27	7	
8	35	1.4	10	
9	25.91	1.22	9	
10	25.45	1.32	10	
11	23.18	1.07	6	
12	34.55	1.55	12	
13	30.91	1.45	9	
14				
15				
16				
17				

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.678400744
R Square	0.46
Adjusted R Square	0.34
Standard Error	3.32
Observations	12

ANOVA

	df	SS	MS	F	Significance F	
Regression	2.00	84.50	42.25	3.84	0.06	
Residual	9.00	99.11	11.01			
Total	11.00	183.61				

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	6.16	9.85	0.63	0.55	–16.12	28.43
Height	11.55	11.14	1.04	0.33	–13.65	36.75
Age	0.70	0.99	0.71	0.50	–1.53	2.92

RESIDUAL OUTPUT

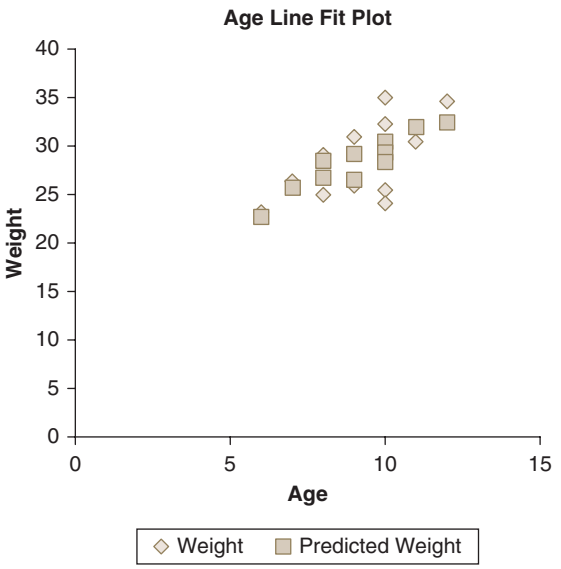
Observation	Predicted Weight	Residuals
1	28.48	0.61
2	30.45	1.82
3	30.45	–6.36
4	31.95	–1.50
5	26.75	–1.75
6	25.70	0.66
7	29.29	5.71
8	26.52	–0.61
9	28.37	–2.92
10	22.70	0.48
11	32.42	2.13
12	29.18	1.73

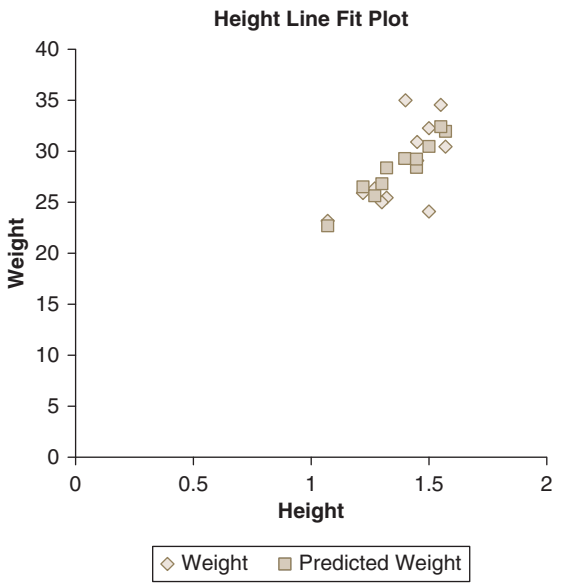
(b) This equation tell us that if the height is held constant a yearly increase in age results in an average increase of 0.685 kg in weight. Similarly, if the age is held constant, a 1 metre increase in height accounts for, on average, an increase of 15.29 kg in weight. This equation can be used to predict the weight of a boy given his height and age.

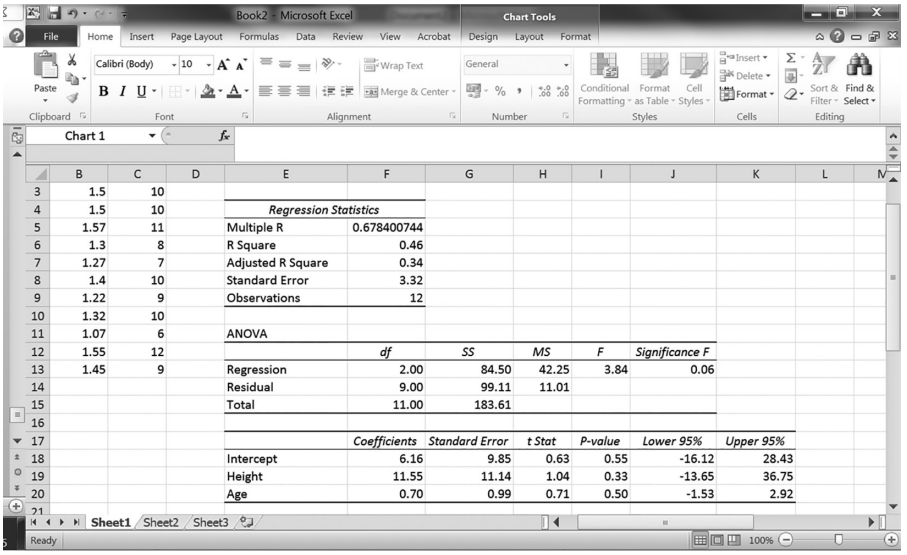
(c) Estimated weight is 63 lb.

(d) 28 kg.

13. (a)







The line fit plots produced by Excel, indicate the 'line of best fit' produced by the regression equation is not very good. This is confirmed by the summary statistics: *R* squared 0.46 and *F*-value of 3.84 with a significance of 0.06. Notice Excel gives a different regression equation. You can check the validity of all three equations by estimating the weight of, say, a boy who weighs 54 lb and is 9 years old. You should get very similar answers of approximately 63 lb or 28 kg.

(b) $s = 2.87$.

(c)

RESIDUAL OUTPUT

Observation	Predicted Weight	Residuals
1	28.48	0.61
2	30.45	1.82
3	30.45	–6.36
4	31.95	–1.50
5	26.75	–1.75
6	25.70	0.66
7	29.29	5.71
8	26.52	–0.61
9	28.37	–2.92
10	22.70	0.48
11	32.42	2.13
12	29.18	1.73

From the residual data given by Excel, i.e. actual weight – predicted weight (see Table), and knowing that ± 1 *SE*. on a normalised standard normal distribution is given by ± 2.87, we can see eight of the points lie between these two values, which gives a value of 67%.

(d) Same as the standard deviation.

(e) It tells us that 67% of the predicted weights lie between the mean ±the standard error, i.e. between 28.55 – 2.87. and 28.55 + 2.87 which equals a lower limit of 25.68 kg and an upper limit of 31.42 kg. Hopefully, we can predict the weight within a range of 5.74 kg, which is not particularly good given the weight range of our sample was 11.73 kg.

(f) Output from SPSS. SPSS generates similar coefficients to Excel.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	6.157	9.848		.625	.547	-16.120	28.435
	Height	11.553	11.139	.424	1.037	.327	-13.646	36.752
	Age	.696	.985	.289	.707	.498	-1.532	2.925

a. Dependent Variable: Weight

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.678[a]	.460	.340	3.31842

a. Predictors: (Constant), Age, Height

b. Dependent Variable: Weight

(g) It should reduce.