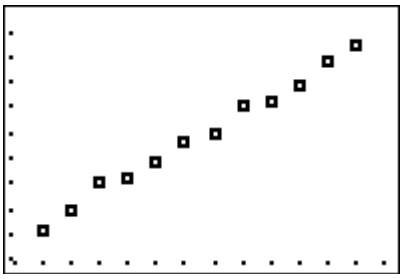
Unit 2 (Chapters 7 – 9) Review Packet – Answer Key

Use the data below for questions 1 through 14
Below is data concerning the mean height of Kalama children. A scientist wanted to look at the effect that age had on the mean height of the children.

Age (months)	18	19	20	21	22	23	24	25	26	27	28	29
Height (cm)	76.1	77	78.1	78.2	78.8	79.7	79.9	81.1	81.2	81.8	82.8	83.5

1. Determine the explanatory and response variables
   The explanatory variable is age and the response is Height.
2. Create a scatterplot of the data. Be sure to label the axes. DESCRIBE the plot.

.



The scatterplot shows a strong(r = 0.994), positive linear association with no apparent outliers.

3. Find the LSR line and the correlation coefficient. Add the line to your plot.
   $y = 64.928 + 0.635x$ $r = 0.994$

4. What proportion of the variability in the height of Kalama children is explained by the variability in their age?
   R^2 = 98.9%; 98.9% of the variability in the height of Kalama children is explained by the variability in their age.

5. Interpret the slope of the LSR line in a complete sentence
   On average for every increase of 1 month in height there tends to be an increase of 0.635 cm in height.

6. Predict the mean height of a child that is 42 months old (show work!). Are you confident in your prediction? Why or why not?
   $y = 64.928 + 0.635(42)$
   $y = 91.597cm$
   I would not be confident in this prediction since 42 is far above the available data set. This is extrapolation.

7. Predict the mean height for a child who is 24 months old (show work!)
   $y = 64.928 + 0.635(24)$
   $y = 80.167cm$
   I would be very confident in this prediction since it is within the data set and R^2 shows such a strong predictive model.

8. Find the error of your predicted value for a child who is 24 months old

$e = 79.7 - 80.167 = -0.267$ cm

9. Was your prediction an overestimate or an underestimate?

This was an overestimation since the actual value was less than the predicted value.

10. How old is a child expected to be if they are 100cm long? (show work!)

$y = -100.841 + 1.557x$

$y = -100.841 + 1.557(100)$

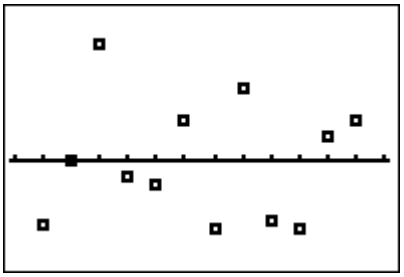
$y = 54.859$ *months*

11. What is the sum of the residuals for this data?

The sum of the residuals from a LSRL is always Zero.

12. Create a residual plot below. Describe the FORM of the plot.

The residuals look scattered with no obvious curve or change in the spread.



13. What does the plot tell you about your linear model? Explain BRIEFLY.

Since there is no apparent form to the residual plot the linear model is an appropriate one.

14. What conclusions can be made from the previous questions? Does the age of a child CAUSE the height of the child? Why or why not?

We cannot assume there is a cause and effect relationship. Correlation does not equal causation.

15. Given the following data about variables x and y calculate by hand (using AP formulas) the LSR line. Show all work! Write the line in the form y = b0 + b1x .

	X		**Y**		
Mean	45.6		37.2		**r =** 0.765
St. Dev		3.2		2.1	

$b_1 = 0.765\left(\frac{2.1}{3.2}\right) = 0.502$

$b_0 = 37.2 - 0.502(45.6) = 14.309$

$\hat{y} = 14.209 + 0.502x$

16. Below is a Minitab statistical analysis. The data is looking at clothes salespersons and examining the effect that the number of minutes spent with a customer has on the total dollar amount that the customer buys. In other words, if a salesperson spends more time with a customer, does the customer buy more clothing (increasing the commission of the salesperson)?

Predictor	**Coeff**	**s.e.**	**T**	**P**
Constant	-1.731	2.4065	-0.876	0.4561
Minutes	0.5679	0.00456	6.6898	1.2358

S = 1.3425 R-Sq = 0.7896 R-Sq (adj) = 0.7748

(a) What is the equation of the LSR line?

$y = -1.731 + 0.5679x$

(b) What is the value of the correlation coefficient?

$r = 0.8886$

(c) What does the correlation tell you about the relationship of your two variables?

The scatterplot would be strong, positive linear relationship. We can't say for certain though without being able to look at the actual scatterplot and residual plot to see if it truly is a linear relationship.

(d) Interpret the slope in the context of the problem

On average for every increase of 1 minute of time spent with a customer, there tends to be an increase of $0.5679 in the total amount spent by the customer.

(e) What is the coefficient of determination? Interpret this value in context of the problem.

$R^2 = 78.96\%$. 78.96% of the variability in the total amount spent by the customer is explained by variations in the amount of time spent with the customer.

(f) How much is a customer expected to buy if a salesperson spends 45 minutes with them?

$y = -1.731 + 0.5679(45)$

$y = \$23.82$

(g) A salesperson spent 35 minutes with a customer and the total sale was $78.50. What is the residual?

$y = -1.731 + 0.5679(35)$

$y = \$18.15$

$e = 78.50 - 18.15 = \$60.35$

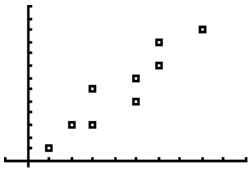
17. What does a residual plot tell us? What do we look for in a residual plot?

A residual plot will tell us whether the linear model is an appropriate one for the data. We look for a completely scattered plot for a good linear model. If we see a curve or a change in the spread we then say the linear model is not appropriate.

18. What type of relationship does r measure?

Linear only!

19. For the graph below, what would be the closest approximation to the correlation coefficient?



(a) 0.2 (b) 0.88 (c) –0.9 (d) –0.2 (e) 0
(f) 0.5

20. What is the difference between outliers, influential points, and high leverage points?

Outliers are any points that deviate from the overall pattern of the data or fall outside the data set.

Influential points are points that if removed would dramatically change the slope of the LSRL.

High leverage points are any points that are outliers in the x variable.

24. Describe the following plots:

(a)

Negative, moderately strong, curved

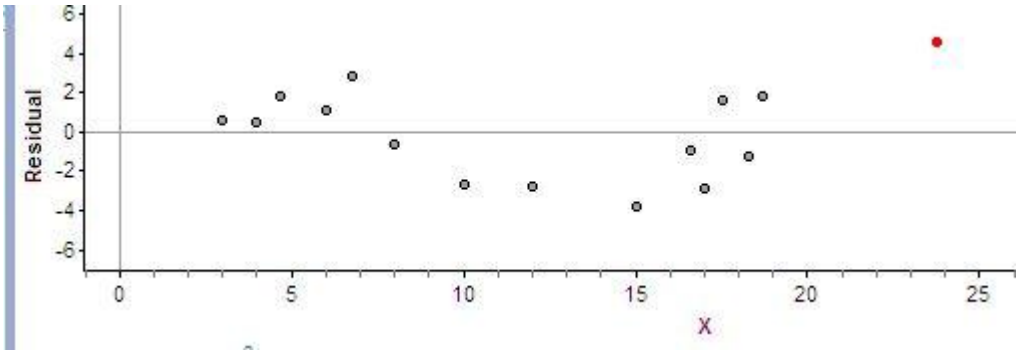
(b)

2 clusters of data, positive direction.

(c)

Strong curved association that is negative but then turns positive.

25. Look at the following residual plot.



(a) Would you expect the model to overestimate or underestimate for a prediction from an x-value of 13? Explain.

Since most of the residuals around 13 are negative I would expect the model to overestimate the prediction for x = 13.

(b) Are there any outliers, high leverage points, or influential points? Identify any, and tell what type of point it is.

There is an outlier at x = 25. This has high leverage and is likely influential.