Chapter 3

Can you plot a scatter plot? What goes on the x-axis? Y axis? Don't forget to label.

The explanatory variable always goes on the x-axis. You also want to make sure you keep your scale consistent (on graph paper if one box is 5 inches, that should be the same all the way along that axis). Always label your axes.

Know how to describe a scatterplot (direction, strength, linear/nonlinear).

Direction: Positive, negative, or no association

Form: Linear or non linear

Know how to graph a least squares regression line on a x and y plane.

The best suggestion would be to plug in two x values and find two different points. Those points form your line. It would probably be best two pick two points that are within the domain of the data (do not extrapolate)

Know what happens to the slope of a LSRL if new data points are added.

A LSRL will always be pulled towards a new point. If that point is an outlier, it could drastically change the slope. If that point is very near the line, the slope will probably not be greatly affected.

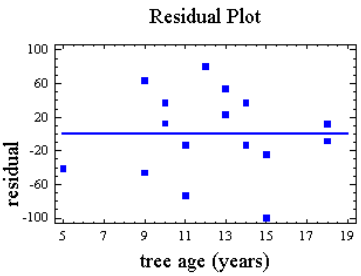
Know what happens to the correlation of a set of data if new data points are added.

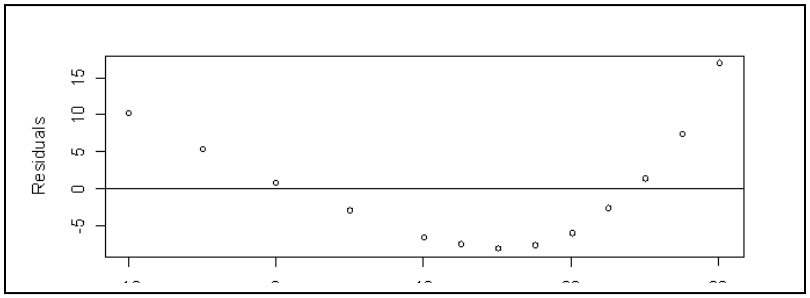
If the point is an outlier, that makes your data become less linear. This would pull your correlation closer to 0.

If the point falls on or very near your LSRL, that makes your data become more linear. This would pull your correlation towards 1 or -1 depending on the slope of your LSRL.

Know that you need a scattered residual plot to prove that something is linear (correlation does not prove linearity).

Even if r = 0.999999, that doesn't mean that the data is linear. In order for data to be considered linear, the residual plot needs to be scattered. The residual plot on the left implies that the data is linear. The residual plot on the right implies that the data is nonlinear





Know how correlation relates to the slope of a LSRL.

$b = r \cdot \frac{s_y}{s_x}$. If you have a correlation close to 0, your slope will be closer to 0.

Chapter 3 (Continued)

Windmills generate electricity by transferring energy from wind to a turbine. A study was conducted to examine the relationship between wind velocity in miles per hour (mph) and electricity production in amperes for one particular windmill. For the windmill, measurements were taken on twenty-five randomly selected days, and the computer output for the regression analysis for predicting electricity production based on wind velocity is given below. The regression model assumptions were checked and determined to be reasonable over the interval of wind speeds represented in the data, which were from 10 miles per hour to 40 miles per hour.

Predictor	Coef	SE Coef	T	P
Constant	0.137	0.126	1.09	0.289
Wind velocity	0.240	0.019	12.63	0.000

S = 0.237 R-Sq = 0.873 R-Sq (adj) = 0.868

Know how to get the equation of a LSRL from a Minitab printout and how to make it context specific.

Predicted amperes = 0.137 + 0.240(wind velocity)

Note: If you just write y and x, you will not get full credit. Fill in x and y with their context. You must also note that amperes are predicted by writing predicted or putting a ^ on top of it.

Know how to interpret the slope of a LSRL in the context of a problem.

For each 1 increase in___x___, the predicted ___y___increases/decreases by b (slope)

For each increase of 1 mph in wind velocity, the predicted production or electricity in amperes increases by 0.240

Know how to interpret the y-intercept of a LSRL in the context of a problem

When ___x___is 0, the predicted value of ___y___ is a

If there was no wind, the predicted amount of electricity produced would be 0.137 amperes

Know what a residual is, how to calculate it given data points and a LSRL, and know how to interpret it contextually. With that you should know how to find an expected number for a LSRL if you are given an x-value (plug it in).

A residual measures how far above or below the LSRL an observed point lies.
Let's pretend that on a day that the wind blew 10 miles per hour that 2.5 amperes of electricity was produced.

$\hat{y}(10) = 0.137 + 0.240(10) = 0.137 + 2.400 = 2.537$
We predict 2.537 amperes, but we only observed 2.5 amperes

Residual = 2.5 – 2.537 = -0.037

A residual of -0.037 means that our observation fell 0.037 amperes below what was predicted

Know that the s in the bottom left of a Minitab is the standard deviation of the residuals, or typical distance each observation is from the LSRL, and know how to interpret that in the context of a problem.

S = 0.237 is the standard deviation of the residuals. The typical observed ampere production is 0.237 away from the predicted amount of ampere production

Chapter 3 (Continued)

Windmills generate electricity by transferring energy from wind to a turbine. A study was conducted to examine the relationship between wind velocity in miles per hour (mph) and electricity production in amperes for one particular windmill. For the windmill, measurements were taken on twenty-five randomly selected days, and the computer output for the regression analysis for predicting electricity production based on wind velocity is given below. The regression model assumptions were checked and determined to be reasonable over the interval of wind speeds represented in the data, which were from 10 miles per hour to 40 miles per hour.

Predictor	Coef	SE Coef	T	P
Constant	0.137	0.126	1.09	0.289
Wind velocity	0.240	0.019	12.63	0.000

S = 0.237 R-Sq = 0.873 R-Sq (adj) = 0.868

Know what r-sq (coefficient of determination) is and how to interpret it in the context of the problem

____% of the variability in __y__ can be explained by x

$r^2 = 0.873$. DO NOT use R-Sq (Adj).

87.3% of the variation in electricity production can be explained by its linear relationship with wind velocity

Know what extrapolation is and when it is and is not appropriate.

It would not be appropriate to predict the electrical output if the wind is 0 mph or 50 mph, because we only collected data between 10 and 40 mph. We can only make predictions between 10 and 40.

Know what happens to correlation if you change the units of measurement (Change weight from lbs to kgs).

In this case r = $+\sqrt{0.873}$ = 0.934. We know that it is positive because the slope of the LSRL is positive. If you collected the same data, but instead measured data in kilometers per hour, the correlation would still be 0.934. Since correlation is based on z-scores, even if you change the units (mph to kph), the relative variability remains the same

Chapter 3 Multiple Choice

1) A teacher raised each student's grade by 10 points on an algebra exam. Which of the following describes the correlation between students' original grades and their adjusted grades.
   a. Slightly negative
   b. Slightly positive
   c. Close to 0
   d. -1
   e. 1

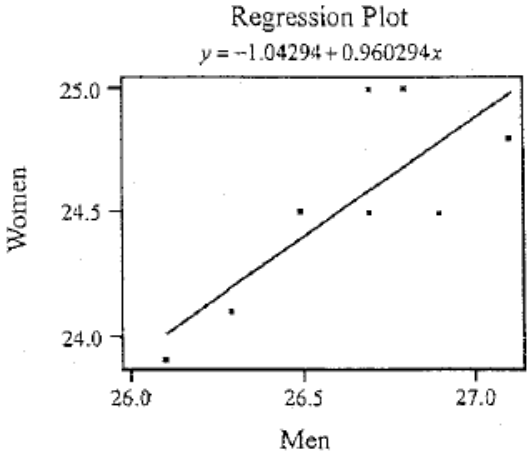
2) The equation of the LSRL for a set of points in a scatterplot is given by $\hat{y} = 2.2 + 0.81x$. The point (5, 7) is one point on this scatterplot. Which of the following is the residual for the point (5, 7)?
   a. 0.71
   b. 0.75
   c. 4.05
   d. 6.25
   e. 7.87

3) The correlation between height and weight among men between the ages of 18 and 70 in US is approximately 0.42. Which of the following conclusions does NOT follow from the data.
   a. Taller men tend to be heavier
   b. Changing the units of weight and height would still yield the same correlation
   c. Heavier men tend to be taller
   d. If a man in this group changes his diet and gains ten pounds, he is likely to get taller
   e. There is a mediocre association between a man's height and weight

4) The point (x, y) on a scatterplot form an ellipse. As the ellipse becomes thinner (points pulled toward the major axis), what can be concluded about the correlation r between the variables x and y?
   a. R decreases in absolute value
   b. R increases in absolute value
   c. R remains constant
   d. The value of r changes sign
   e. No conclusion can be drawn

5) The correlation r between the magnitude of an earthquake and the depth below the surface of the earth at which the quake occurs has been determined experimentally to be about 0.51. Suppose that we use the magnitude of the earthquake (x) to predict the depth below the surface (y) at which the quake occurs. It has been shown that the data has a linear relationship. We can infer that
   a. The LSRL of y on x has a slope equal to 0.51
   b. The fraction of the variation in depths explained by the LSRL of y on x is 0.26
   c. About 51% of the time, the magnitude of an earthquake will accurately predict the depth at which the earthquake occurs
   d. The numerical value of the depth is usually 51% of the numerical value of the earthquake
   e. 26% of the data values lie on the LSRL

6) Using computer software and data for the years 1990-1998, a least-squares regression line was computed to predict the median age of women (y) given the median age of men (x) at first marriage.



Variable	N	Mean	Median	TrMean	StDev	SE Mean
Men	9	26.622	26.700	26.622	0.307	0.102
Women	9	24.522	24.500	24.522	0.373	0.124

The regression equation is Women = –1.04 + 0.960Men

The percent variation of y that is explained by the least-squares regression of y on x is:

A.) 0.04 B.) 0.08 C.) 0.62 D.) 0.79 E.) 0.96

7) Which of the following statements related to residuals is true?

I. The mean of the LSRL residuals is always 0

II. If one tries to fit a linear model to bivariate data, a curved pattern in a residual plot shows that the relationship between two variables is not linear

III. A residual plot can be a scatterplot of the regression residuals against the explanatory variable

a. I only
b. II only
c. III only
d. II and III only
e. I, II, and III

8) A regression line includes the point (2, 14) and has the equation $\hat{y} = mx + 4$. If $\bar{x}$ *and* $\bar{y}$ are the sample means of the x and y values, then $\bar{y} =$

a. $\bar{x} + 4$
b. $5\bar{x} + 4$
c. $\bar{x}$
d. $\bar{x} - 1/7$
e. $\bar{x} - 2$

9) Which of the following statements is true
   a. Values of r near 0 indicate a strong linear relationship
   b. Changing the measurements units of x and y may affect the correlation between x and y.
   c. Strong correlation means that there is a definite cause and effect relationship between x and y
   d. Correlations changes when the x and y variables are reversed
   e. The correlation can be strongly affected by a few outying observation

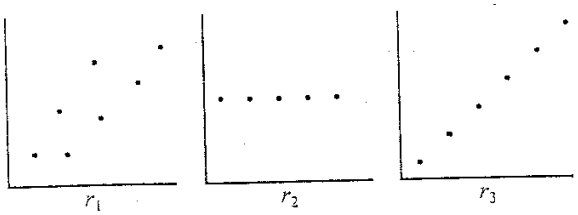
10) Data are obtained from a group of high school seniors comparing age and the number of hours spent on the telephone. The resulting regression equation is

Predicted hours = 0.123(age) + 2.57 and r = 0.866

What percent of the variation in the number of hours spent on the telephone can be explained by the LSRL?

a. 0.75%
b. 0.866%
c. 75%
d. 86.6%
e. This value cannot be determined

11) Which of the following statements are true about the correlations?



A.) $r_1 < r_2 < r_3$
B.) $r_1 < r_3 < r_2$
C.) $r_2 < r_3 < r_1$
D.) $r_2 < r_1 < r_3$
E.) $r_3 < r_2 < r_1$

12) A researcher wishes to examine the relationship between years of schooling completed and the number of pregnancies in young women. Her research discovers a linear relationship, and a least squares fit of her data results in

$$\hat{y} = 6.4 - 0.12x$$

Where x is the number of years completed in school and y is the number of pregnancies. What is the estimated change in the number of pregnancies that corresponds to the completion of an additional 10 years of school?

   a. A decrease of 5.2
   b. An increase of 7.6
   c. A decrease of 7.6
   d. An increase of 1.2
   e. A decrease of 1.2

Chapter 3 Multiple Choice Solutions

1) E. Obviosly there is a positive relationship. If you got a high schore on the test, if you increase it by 10, you will still have a high score. Since the data follows exactly the equation y = x + 10, the data is perfectly linear, making the correlation exactly 1.

2) B. Observed – Expected = 7 - $\hat{y}(5)$ = 7 - 6.25 = 0.75

3) D. Both A and C are true because the slope is positive (we know this because correlation is positive). B is true because changing units does not change the relative relationship between the data. E is true because the correlation is 0.42. D is false.

4) B. Imagine an ellipse that is formed around the equation y = -3x + 4. As the ellipse is getting thinner, all of the data points are being drawn towards the line. As the data is being pulled towards the line, the data is appearing more linear, making our correlation get closer to -1, thus increasing the absolute value of r.

5) B. Both C and D are false, because correlation is not a percent. A is false because $b = r \cdot \frac{s_y}{s_x}$. There is a relationship between r and slope, but they are not necessarily equal. R squared explains the amount of variability. It does not explain how many points are correct, as is implied in e.

6) C. First you must realize that this is not a usual minitab printout. This give the summative data for each variable, not the summative data of the line. Using our packet, we know that $b = r \cdot \frac{s_y}{s_x}$. This means that $0.960 = r \cdot \frac{.373}{.307}$. This makes r = 0.79013. So, $r^2 = 0.624$

7) E. Since the residuals have to add to 0, the average will also be 0. If data were linear, the residual plot would be scattered. Statement III is essentially the definition of a residual plot

8) B. On every least squares regression line $(\bar{x}, \bar{y})$ must be on the line. So, plugging in the points, we get 14 = m(2) + 4. This gives us a slope of 5.

9) E. Obviously a is false. B is false because if you change units, that will not change the relationship. Whether you measure Mr. Merlo's age in days or years, he is still older than you. Strong correlation is good evidence of cause and effect, but only a hypothesis test or other inference procedure can produce cause and affect. D is false, because no matter which variable is x, the relationship will still be the same.

10) C. If r = 0.866, then $r^2 = 0.749956$. So, 75% of the variation in y can be explained by x.

11) D. The third scatterplot has a correlation of r = 1 because it is positive and perfectly linear. The second scatterplot has a correlation of r = 0, because its slope is 0. The first would have a slope somewhere between 0 and 1.

12) E. If the slope of the equation is 0.12, that means that as school goes up by a year, the number of births decreases by 0.12 because the slope is negative 0.12. That means that if school goes up by 10 years, pregnancies will decrease by 1.2

Chapter 15

Windmills generate electricity by transferring energy from wind to a turbine. A study was conducted to examine the relationship between wind velocity in miles per hour (mph) and electricity production in amperes for one particular windmill. For the windmill, measurements were taken on twenty-five randomly selected days, and the computer output for the regression analysis for predicting electricity production based on wind velocity is given below. The regression model assumptions were checked and determined to be reasonable over the interval of wind speeds represented in the data, which were from 10 miles per hour to 40 miles per hour.

Predictor	Coef	SE Coef	T	P
Constant	0.137	0.126	1.09	0.289
Wind velocity	0.240	0.019	12.63	0.000

S = 0.237 R-Sq = 0.873 R-Sq (adj) = 0.868

Know how to write a null and alternative hypothesis for testing slope and when this is an appropriate test. Know what the t score and P-value on the Minitab printout are the results of. Know how to find a t-statistic, degrees of freedom (n-2), and P-value for a test on slope.

The row containing the wind velocity statistics displays the results of the hypothesis test

$$H_0: \beta = 0, \quad H_a: \beta \neq 0$$

$t = \frac{0.240-0}{0.019} = 12.63$, means that our estimated slope (0.240) is 12.63 standard deviations above a slope β = 0

With a t statistic of 12.63 and df = n – 2 = 25 – 2 = 23, we would get a P-value close to 0.000

Know how to draw a conclusion from a p-value for a test on slope.

We reject the null hypothesis in favor of the alternative at the 1% significance level, because our p-value of 0.000 < 0.01. There is significant evidence that β ≠ 0 and there is a relationship between the wind velocity and the amperes produced

Know how to calculate a confidence interval for a slope when given a Minitab printout.

CI = Statistic ± (Critical Value)(Standard Deviation)

95% confidence interval CI = 0.240 ± (2.069)(0.019) →Note: 2.069 comes from df = 23 because n = 25

CI = (0.200689, 0.279311)

We are 95% confident that the slope of the true least squares regression line comparing wind velocity and electricity output is between 0.200689 and 0.279311

Know what it means if 0 is a possibility in a confidence interval or if you fail to reject the test againt B = 0.

Since β = 0 in not in our interval, we are confident that our relationship does have a slope that is not 0. This implies that there is a relationship, between the data. In other words, electricity output is dependent on wind speed. If β = 0, then there is no relationship between the variable. They are independent.

Know that the SE in a Minitab printout is the standard error of the slope, and know how to interpret that in the context of a problem.

The standard error of the estimated slope parameter is 0.019 amperes per mph. Thus, the standard deviation of the estimated slope for predicting the amount of electricity based on wind speed is estimated to be 0.019 amperes per mph. This value estimates the variability in the sampling distribution of the estimated slope (i.e., how much we would expect sample slopes to vary from experiment to experiment).