Chapter 3: Examining Relationships

Intro:

- This section is going to focus on relationships among several variables for the same group of individuals. In these relationships, does one variable cause the other variable to change?
- **Explanatory Variable**:
  - Attempts to explain the observed outcomes
  - Independent variable
- **Response Variable**:
  - Measures an outcome of a study
  - Dependent variable

2

Principles That Guide Examination of Data

- Same as one-variable methods from Ch. 1 and 2

1. Plot the data, then add numerical summaries.
2. Look for overall patterns and deviation from those patterns.
3. When the overall pattern is quite regular, use a compact mathematical model to describe it.

3

3.1 EXPLANATORY AND RESPONSE VARIABLES In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?

(a) The amount of time a student spends studying for a statistics exam and the grade on the exam

(b) The weight and height of a person

(c) The amount of yearly rainfall and the yield of a crop

(d) A student's grades in statistics and in French

(e) The occupational class of a father and of a son

A. Explanatory – Time studying
Response – Exam grade

B. Explore the relationship

C. Explanatory – Rainfall
Response – Crop yield

D. Explore the relationship

E. Explanatory – Father's class
Response – Son's class

4

3.1: Scatterplots

- Most effective way to display relationship between two quantitative variables
- Shows the relationship between two quantitative variables measured on the same individuals
  - Each individual in the data appears as the point in the plot
- Plot the explanatory variable on the horizontal axis
- Plot the response variable on the vertical axis.

5

Examining scatterplots:

- Describe the overall pattern of a scatterplot by:
  - **Form** – linear, quadratic, logarithmic, etc.
  - **Direction** – positive or negative.
  - **Strength of the relationship** – weak, moderate, or strong.
- An important kind of deviation is an outlier.
- Variable Association:
  - Positively associated
    - Direct Variation
  - Negatively Associated
    - Inverse Variation

6

Tips for drawing scatterplots:

1) Scale the vertical and horizontal axes.
   a. Intervals must be uniform
   b. Use a symbol to indicate a break in the scale

2) Label both axes, and title the graph.

3) If you are given a grid, try to adopt a scale that makes your plot use the whole grid.

7

3.6 THE ENDANGERED MANATEE Manatees are large, gentle sea creatures that live along the Florida coast. Many manatees are killed or injured by powerboats. Here are data on powerboat registrations (in thousands) and the number of manatees killed by boats in Florida in the years 1977 to 1990:

Year	Powerboat registrations (1000)	Manatees killed	Year	Powerboat registrations (1000)	Manatees killed
1977	447	13	1984	559	34
1978	460	21	1985	585	33
1979	481	24	1986	614	33
1980	498	16	1987	645	39
1981	513	24	1988	675	43
1982	512	20	1989	711	50
1983	526	15	1990	719	47

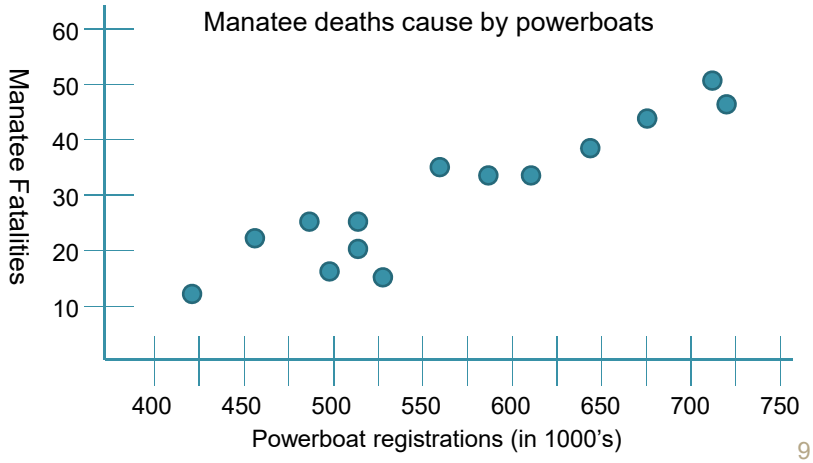
(a) We want to examine the relationship between number of powerboats and number of manatees killed by boats. Which is the explanatory variable?

(b) Make a scatterplot of these data. (Be sure to label the axes with the variable names, not just x and y.) What does the scatterplot show about the relationship between these variables?

8

Manatees

- A) The explanatory variable is the number of powerboat registrations.
- B) Make a scatterplot.



9

3.9 MORE ON THE ENDANGERED MANATEE In Exercise 3.6 (page 125) you made a scatterplot of powerboats registered in Florida and manatees killed by boats.

(a) Describe the direction of the relationship. Are the variables positively or negatively associated?

(b) Describe the form of the relationship. Is it linear?

(c) Describe the strength of the relationship. Can the number of manatees killed be predicted accurately from powerboat registrations? If powerboat registrations remained constant at 719,000, about how many manatees would be killed by boats each year?

A. The direction of the relationship is positive, because the trend shows that as powerboat registrations increased there were more deaths to manatees.

B. The form of the relationship looks to be linear.

C. The strength of the relationship looks strong, because there aren't to many points that deviate away from the occurring trend.

10

Section 3.1 Complete

- Homework: #'s 1-4,10 (scatterplot by hand), 12
- Any questions on pg. 1-4 in additional notes packet

11

3.2: Correlation

- Measures the direction and strength of a linear relationship between two variables.
- Usually written as r.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- The formula is a little complex and most of the time we will use our calculators.

12

Facts about correlation:

- Makes no distinction between *explanatory* and *response* variables
- Requires that both variables be quantitative
  - The correlation between the incomes of a group of people and what city they live in cannot be calculated because city is a categorical variable.
- ***r*** does not change when we change the units of measurement of x, y, or both
  - ***r*** has no unit of measurement; it is just a number.
- Positive ***r*** indicates *positive* association
- Negative ***r*** indicates *negative* association.

13

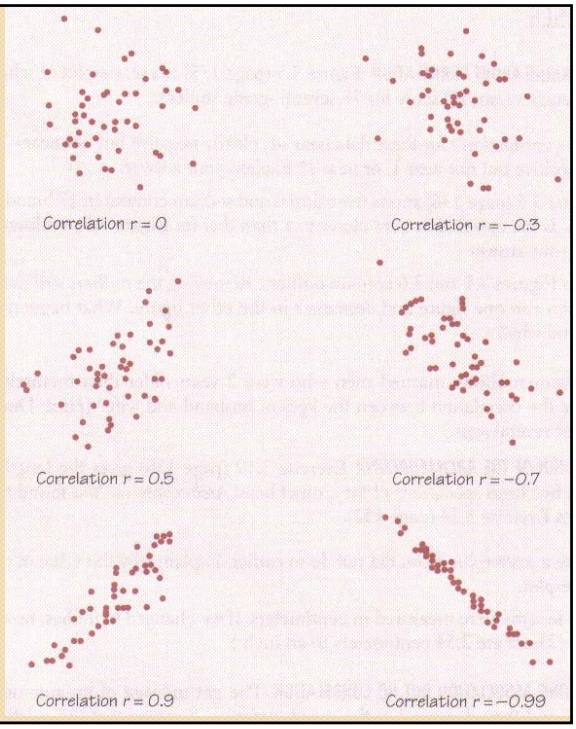
Facts about correlation:

- ***r*** is always a number between -1 and 1
  - Values near 0 indicate a very weak linear relationship
    - The strength increases as ***r*** moves toward -1 or 1.
- Measures the strength of only a linear relationship
  - **Does not** describe curved relationships
- ***r*** is not a resistant measurement
  - Use ***r*** with caution when there are outliers
- ***r*** is not a complete description of two-variable data
  - Need to use the **means** and **standard deviations** of BOTH x and y along with the correlation when describing the data.

14

Correlation Charts

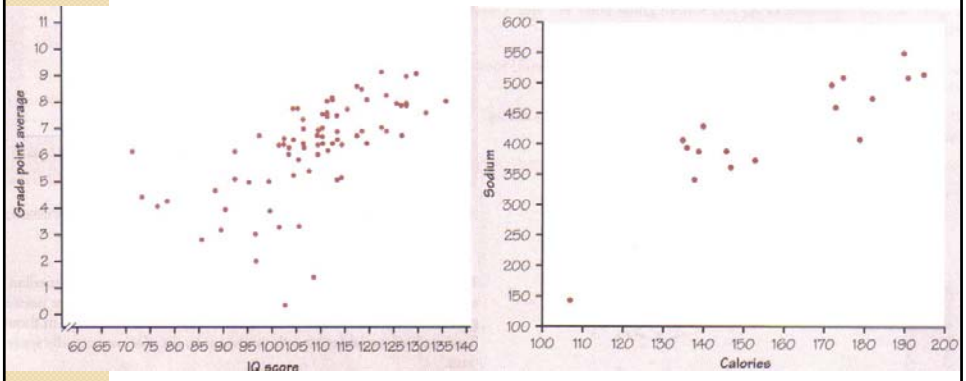
Correlation measures how closely related the data is to a linear approximation. The slope of the correlation gives the sign of the value.



3.25 THINKING ABOUT CORRELATION Figure 3.5 (page 135) is a scatterplot of school grade point average versus IQ score for 78 seventh-grade students.

(a) Is the correlation r for these data near –1, clearly negative but not near –1, near 0, clearly positive but not near 1, or near 1? Explain your answer.

(b) Figure 3.6 (page 136) shows the calories and sodium content in 17 brands of meat hot dogs. Is the correlation here closer to 1 than that for Figure 3.5, or closer to zero? Explain your answer.



16

Calculator Problem

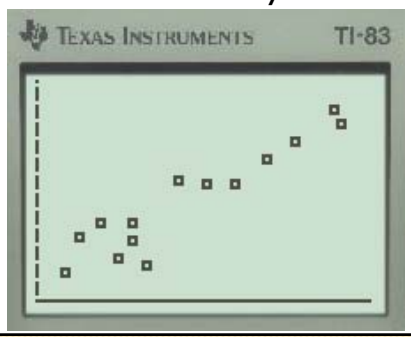
- Take yesterday's example of Manatee deaths and put the data into your calculator's lists
  - $List_1$ – Powerboat registration (explanatory)
  - $List_2$ – Manatees killed (response)

Year	Powerboat registrations (1000)	Manatees killed	Year	Powerboat registrations (1000)	Manatees killed
1977	447	13	1984	559	34
1978	460	21	1985	585	33
1979	481	24	1986	614	33
1980	498	16	1987	645	39
1981	513	24	1988	675	43
1982	512	20	1989	711	50
1983	526	15	1990	719	47

17

Make a Scatterplot

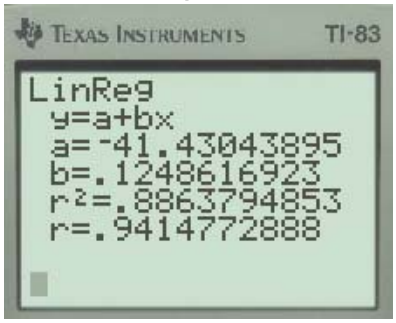
- Use 2nd Y=
  - Turn plot 1 on
  - The first type of graph is a scatterplot
  - Xlist = L_1
  - Ylist = L_2
  - Press the zoom key then number 9



18

Find the Correlation

- Press 2nd 0
  - Brings up catalog
    - Find DiagnosticOn and press enter twice
- Press the STAT key
  - Scroll over to CALC
  - Use either option 4 or 8



19

Calculator Problem

3.28 STRONG ASSOCIATION BUT NO CORRELATION The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon):

Speed:	20	30	40	50	60
MPG:	24	28	30	28	24

- Make a scatterplot on your calculator.
- Does there appear to be a strong relationship between speed and MPG?
- Calculate ***r***.
- Why is ***r*** = 0, when there is a strong relationship?
  - ***r*** only measures the strength of a **linear** relationship

20

Section 3.2 Complete

- Homework: #'s 15, 17, 19(calculator for b), 21, 23
- Any questions on pg. 5-8 in additional notes packet

21

3.3: Least-Squares Regression

- Correlation measures the strength and direction of the linear relationship
- **Least-squares regression**
  - Method for finding a line that summarizes that relationship in a specific setting.
  - Describes how a response variable **y** changes as an explanatory variable ***x*** changes
  - Used to predict the value of **y** for a given value of ***x***
  - Unlike correlation, requires an *explanatory* and *response* variable.

22

Least-squares regression line (LSRL).

- The equation is $\hat{y} = a + bx$
- $\hat{y}$ is used because the equation is a prediction
- The slope is ***b*** and the *y*-intercept is ***a***

$$b = r\frac{s_y}{s_x} \qquad a = \bar{y} - b\bar{x}$$

- Every least-squares regression line passes through the point $(\bar{x}, \bar{y})$

23

Facts about least-squares regression.

1. Distinction between explanatory and response variables is essential
   a. If we reversed the roles of the two variables, we get a different LSRL
2. LSRL is calculated by minimizing the sum of the squares of $y - \hat{y}$
3. There is a close connection between correlation and the slope of the regression line
   a. As ***r*** gets closer to 0, $\hat{y}$ moves less in response to changes in *x*.

$$b = r\frac{s_y}{s_x}$$

24

Interpretation of LSRL

- Slope
  - For every unit increase in *x*, there is ***on average*** a change of ***b*** units in $\hat{y}$
- *y*-intercept
  - Value of $\hat{y}$ when $x = 0$
  - Only meaningful when *x* can actually take values close to zero.

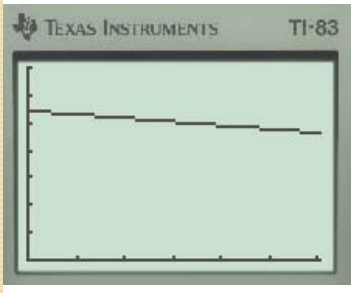
25

3.40 ACID RAIN Researchers studying acid rain measured the acidity of precipitation in a Colorado wilderness area for 150 consecutive weeks. Acidity is measured by pH. Lower pH values show higher acidity. The acid rain researchers observed a linear pattern over time. They reported that the least-squares regression line

$$\text{pH} = 5.43 - (0.0053 \times \text{weeks})$$

(a) Draw a graph of this line. Is the association positive or negative? Explain in plain language what this association means.

A. y=5.43-.0053x
Window
0 < X < 151
0 < Y < 7



By looking at the equation or the graph, the association between time and pH is negative.

This means that there is an inverse relationship between weeks and pH levels. As more weeks go by the pH level of the rain gets lower, meaning the rain is becoming more acidic.

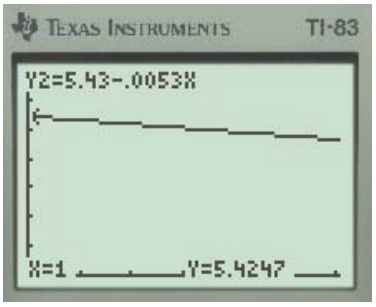
26

3.40 ACID RAIN Researchers studying acid rain measured the acidity of precipitation in a Colorado wilderness area for 150 consecutive weeks. Acidity is measured by pH. Lower pH values show higher acidity. The acid rain researchers observed a linear pattern over time. They reported that the least-squares regression line

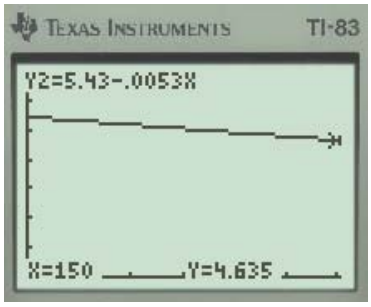
$$\text{pH} = 5.43 - (0.0053 \times \text{weeks})$$

(b) According to the regression line, what was the pH at the beginning of the study (weeks = 1)? At the end (weeks = 150)?

B. x = 1, y = 5.4247



x = 150, y = 4.635



27

3.40 ACID RAIN Researchers studying acid rain measured the acidity of precipitation in a Colorado wilderness area for 150 consecutive weeks. Acidity is measured by pH. Lower pH values show higher acidity. The acid rain researchers observed a linear pattern over time. They reported that the least-squares regression line

$$\text{pH} = 5.43 - (0.0053 \times \text{weeks})$$

(c) What is the slope of the regression line? Explain clearly what this slope says about the change in the pH of the precipitation in this wilderness area.

C. The slope of the regression line is -.0053. This means that with the passing of each week, **on average**, the pH level of rain in the Colorado wilderness decreases by .0053.

28

3.44 PREDICTING THE STOCK MARKET Some people think that the behavior of the stock market in January predicts its behavior for the rest of the year. Take the explanatory variable x to be the percent change in a stock market index in January and the response variable y to be the change in the index for the entire year. We expect a positive correlation between x and y because the change during January contributes to the full year's change. Calculation from data for the years 1960 to 1997 gives

$$\bar{x} = 1.75\% \qquad s_x = 5.36\% \qquad r = 0.596$$
$$\bar{y} = 9.07\% \qquad s_y = 15.35\%$$

(b) What is the equation of the least-squares line for predicting full-year change from January change?

$$\hat{y} = a + bx \qquad b = .596\frac{15.35}{5.36} \qquad a = 9.07 - 1.7068(1.75)$$

$$b = r\frac{s_y}{s_x} \qquad b = 1.7068 \qquad a = 6.0831$$

$$a = \bar{y} - b\bar{x} \qquad \hat{y} = 6.08\% + 1.71\%x$$

29

Calculator Problem Continued

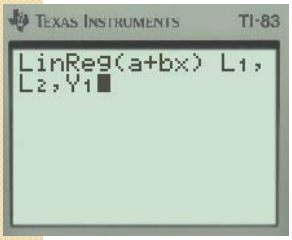
- Take yesterday's example of Manatee deaths and put the data into your calculator's lists
  - $List_1$ – Powerboat registration (explanatory)
  - $List_2$ – Manatees killed (response)

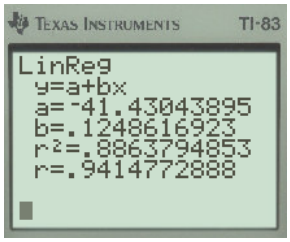
Year	Powerboat registrations (1000)	Manatees killed	Year	Powerboat registrations (1000)	Manatees killed
1977	447	13	1984	559	34
1978	460	21	1985	585	33
1979	481	24	1986	614	33
1980	498	16	1987	645	39
1981	513	24	1988	675	43
1982	512	20	1989	711	50
1983	526	15	1990	719	47

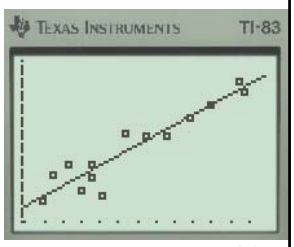
30

Find the LSRL and Overlay it on your Scatterplot

- Press the STAT key
  - Scroll over to CALC
  - Use either option 4 or 8
  - After the command is on your home screen:
    - Put the following L_1, L_2, Y_1
    - To get Y_1, press VARS, Y-VARS, Function



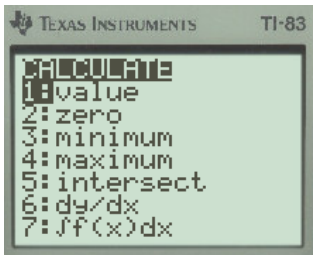


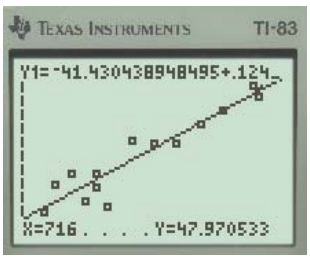


31

Use the LSRL to Predict

- With an equation stored on the calculator it makes it easy to calculate a value of y for any known x.
  - Use the LSRL to predict the number of manatee deaths for a year that had 716,000 powerboat registrations.
    - 2nd Trace, Value
      - x = 716 (remember scale)





32

The role of r^2 in regression.

- **Coefficient of determination**
  - The fraction of the variation in the values of y that is explained by least-squares regression of y on x.
  - Measurement of the contribution of x in predicting y.
  - Equation is tedious

$$\frac{\sum\left(y-\bar{y}\right)^2-\sum\left(y-\hat{y}\right)^2}{\sum\left(y-\bar{y}\right)^2}$$

  - But it can be shown algebraically to be equal to the correlation coefficient squared (r^2)

33

Example - Calculation

3.42 **CLASS ATTENDANCE AND GRADES** A study of class attendance and grades among first-year students at a state university showed that in general students who attended a higher percent of their classes earned higher grades. Class attendance explained 16% of the variation in grade index among the students. What is the numerical value of the correlation between percent of classes attended and grade index?

- The wording of this question gives the value of r^2.
  - Take the square root

$$\sqrt{.16}$$

$$= .4$$

34

Example – Sentence Structure

- Data from problem # 44 shows:
  - x = January stock change
  - y = Entire year stock change

$\bar{x} = 1.75\%$ $s_x = 5.36\%$ $r = 0.596$
$\bar{y} = 9.07\%$ $s_y = 15.35\%$

- What is the coefficient of determination and what does it mean in the context of this problem?
  - $r^2 = .335$
    - This means that 33.5% of the variation in the change in stock index for the entire year can be explained by the least-squares regression of entire year stock change on January's stock change.

35

Section 3.3 Day 1

- Homework: #'s 35, 37, 38, 47(a-c), 48(a-c)
- Any questions on pg. 9-12 in additional notes packet

36

3.3: Least-Squares Regression – Day 2

- **Residuals**
  - Deviations from the overall pattern
    - Measured as vertical distances
  - Difference between an observed value of the response variable and the value predicted by the regression line
    - Observed y – predicted $\hat{y}$

$$residual = y - \hat{y}$$

  - The sum of the least-squares residuals are always zero

37

Continue with Manatees

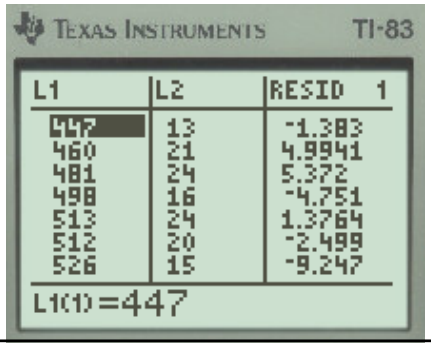
- Find the residual for the observed value of 447,000 powerboats
  - From previous $\hat{y} = -41.4304 + .1249x$
  - Substitution of 447 gives $\hat{y} = 14.39$
  - Observed data at 447 is 13
  - $y - \hat{y} = 13 - 14.39 = -1.39$

Year	Powerboat registrations (1000)	Manatees killed	Year	Powerboat registrations (1000)	Manatees killed
1977	447	13	1984	559	34
1978	460	21	1985	585	33
1979	481	24	1986	614	33
1980	498	16	1987	645	39
1981	513	24	1988	675	43
1982	512	20	1989	711	50
1983	526	15	1990	719	47

38

See all of the residuals at once

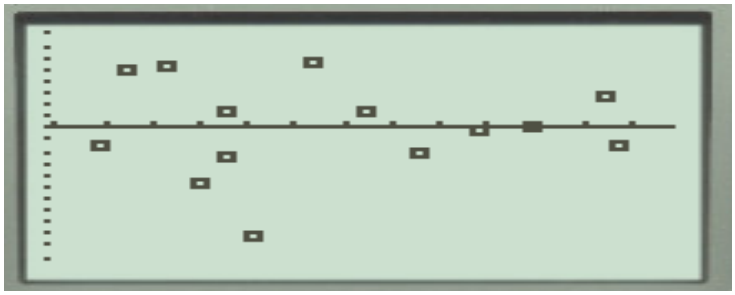
- The calculator calculates the residuals for all points every time it runs a linear regression command
  - To see this, press 2nd STAT and under NAMES scroll down to RESID
    - The residuals will be in the order of the data
- Can also set up your "screen of lists" to always have residuals showing



39

What to do with all the residuals

- **Residual Plot**
  - Scatterplot of the regression residuals against the explanatory variable
    - Help to assess the fit of a regression line
      - If the regression line captures the overall relationship between x and y, the residuals should have no systematic pattern
- Residual Plot for the Manatees



40

Good Fit

- Below is a residual plot that shows a linear model is a good fit to the original data
  - **Reason**
    - There is a uniform scatter of points

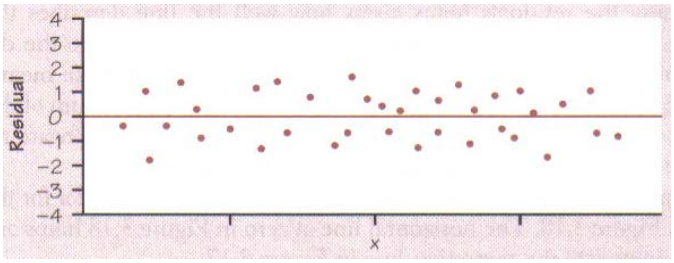


FIGURE 3.19(a) The uniform scatter of points indicates that the regression line fits the data well, so the line is a good model.

41

Poor Fit

- Below are two residual plots that show a linear model is not a good fit to the original data
  - **Reasons**
    - Curved pattern
    - Residuals get larger with larger values of x

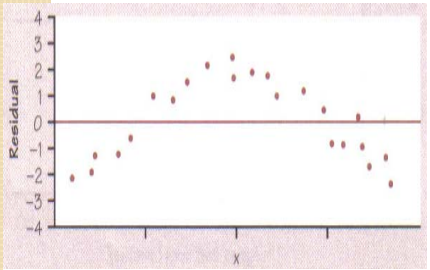


FIGURE 3.19(b) The residuals have a curved pattern, so a straight line is an inappropriate model.

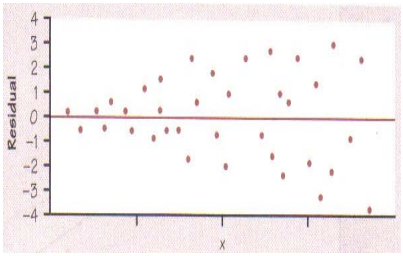


FIGURE 3.19(c) The response variable y has more spread for larger values of the explanatory variable x, so prediction will be less accurate when x is large.

42

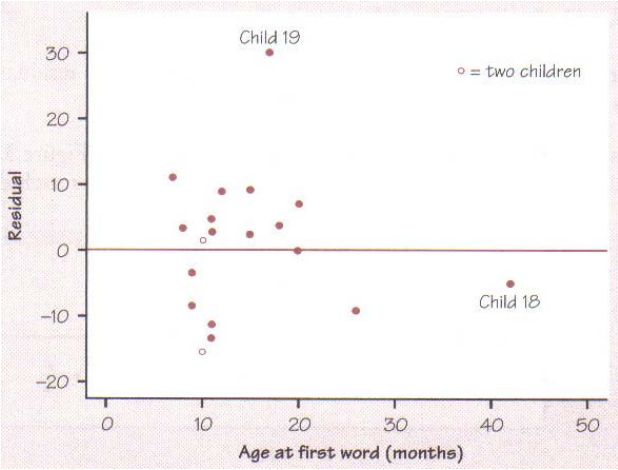
Influential observations:

- **Outlier**
  - An observation that lies outside the overall pattern in the y direction of the other observations.
- **Influential Point**
  - An observation is influential if removing it would markedly change the result of the LSRL
  - Are outliers in the x direction of a scatterplot
  - Have small residuals, because they pull the regression line toward themselves.
    - If you just look at residuals, you will miss influential points.
  - Can greatly change the interpretation of data.

43

Location of Influential observations

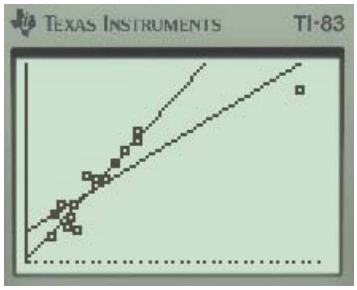
- Child 19
  - Outlier
- Child 18
  - Influential Point

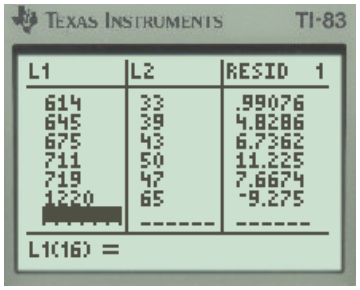


44

Manatee

- Add the point, (year 2013, boats 1220, manatees 65) to the data we have.
- Run a LSRL command and store the new line in L_2
- Look how the point drastically changed the LSRL
  - But it does not have the largest absolute residual





45

Section 3.3 Day 2 Complete

- Homework: #'s 49, 51, 52, 55, 56, 61(use line for part b), 64, 69
- Any questions on pg. 13-16 in additional notes packet

46

MINITAB printout

- A healthy cereal should be low in both calories and sodium. Data for 77 cereals were examined and judged acceptable for inference. The 77 cereals had between 50 and 160 calories per serving and between 0 and 320 mg of sodium per serving. The regression analysis is shown.

R-squared = 9.0%
s = 80.49 with 77 – 2 = 75 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	Prob
Constant	21.4143	51.47	0.416	0.6706
Calories	1.29357	0.4738	?	?

MINITAB printout

- Find the following:
  a) Line of best fit.
  b) Interpret the slope in context.
  c) Interpret the y-int in context.
  d) Correlation coefficient, what does it tell you in context?
  e) Interpret r^2 in context.

R-squared = 9.0%
s = 80.49 with 77 – 2 = 75 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	Prob
Constant	21.4143	51.47	0.416	0.6706
Calories	1.29357	0.4738	?	?

Line of best fit.

	Variable	Coefficient	SE(Coeff)	t-ratio	Prob
y-axis →	Constant	21.4143	51.47	0.416	0.6706
x-axis →	Calories	1.29357	0.4738	?	?

In the variable column there are always two entrees, one is your explanatory variable and one is your response variable. The explanatory variable is named with the label for your x-axis and the response variable is called constant.

$$\hat{y} = 21.4143 + 1.29357x$$

$\hat{y}$ = the amount of sodium, (mg)

x = the amount of calories, (calories)

Interpret the slope in context.

- The slope equals 1.29357.
  - This means that for every additional calorie increase in a serving of cereal, **on average**, the amount of sodium will also increase by 1.3 mg per serving.

Interpret the y-int in context.

- The y-int equals 21.4143.
  - This means that for a serving of cereal with zero calories, there are still 21.4 mg of sodium per serving.

Correlation coefficient, what does it tell you in context?

- The correlation coefficient is the square-root of r^2.
  $r = \sqrt{.09} = .3$
- This means there is a **weak**, **positive**, **linear** association between calorie count and sodium amount in a serving of cereal.

The meaning of r^2.

- In regression, R^2 (coefficient of determination), is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1.0 indicates that the regression line perfectly fits the data. R^2 has a range of values from 0 to 1. It is the proportion of variability in a data set that is accounted for by the statistical model.

Interpret r^2 in context.

- R^2 = 9%.
  - 9% of the variability in the amount of sodium per serving of cereal is (explained or accounted) by the regression model with calories per serving of cereal. This proportion provides a measure of how well future predictions of sodium content can be made from calorie count by the model. In this case it shows a weak relationship.

Reading MINITAB Complete

- Homework: Finish worksheet on Minitab
- Any questions on pg. 17-20 in additional notes packet

54

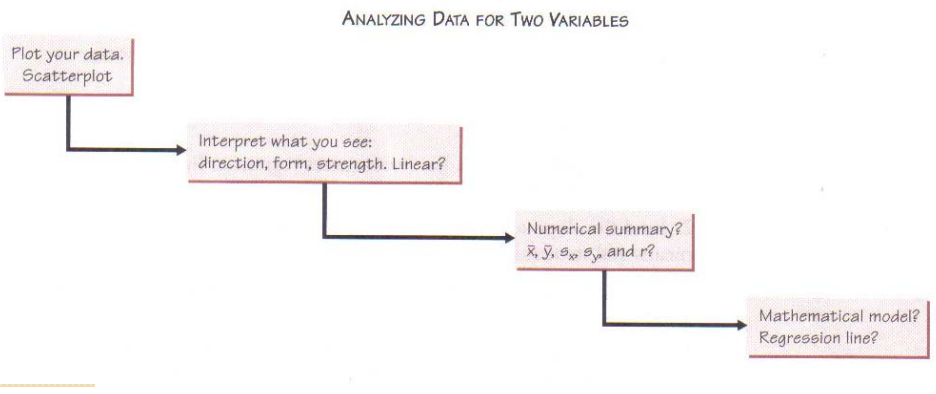
Chapter Review

Chapters 1 and 2 dealt with data analysis for a single variable. In this chapter, we have studied analysis of data for two or more variables. The proper analysis depends on whether the variables are categorical or quantitative and on whether one is an explanatory variable and the other a response variable.

Data analysis begins with graphs and then adds numerical summaries of specific aspects of the data.

This chapter concentrates on relations between two quantitative variables. Scatterplots show the relationship, whether or not there is an explanatory-response distinction. Correlation describes the strength of a linear relationship, and least-squares regression fits a line to data that have an explanatory-response relation.

55



A. DATA

1. Recognize whether each variable is quantitative or categorical.
2. Identify the explanatory and response variables in situations where one variable explains or influences another.

56

B. SCATTERPLOTS

1. Make a scatterplot to display the relationship between two quantitative variables. Place the explanatory variable (if any) on the horizontal scale of the plot.

2. Add a categorical variable to a scatterplot by using a different plotting symbol or color.

3. Describe the form, direction, and strength of the overall pattern of a scatterplot. In particular, recognize positive or negative association and linear (straight-line) patterns. Recognize outliers in a scatterplot.

57

C. CORRELATION

1. Using a calculator, find the correlation r between two quantitative variables.

2. Know the basic properties of correlation: r measures the strength and direction of only linear relationships; $-1 \leq r \leq 1$ always; $r = \pm 1$ only for perfect straight-line relations; r moves away from 0 toward ± 1 as the linear relation gets stronger.

D. STRAIGHT LINES

1. Explain what the slope b and the intercept a mean in the equation $y = a + bx$ of a straight line.

2. Draw a graph of the straight line when you are given its equation.

58

E. REGRESSION

1. Using a calculator, find the least-squares regression line of a response variable y on an explanatory variable x from data.
2. Find the slope and intercept of the least-squares regression line from the means and standard deviations of x and y and their correlation.
3. Use the regression line to predict y for a given x. Recognize extrapolation and be aware of its dangers.
4. Use r^2 to describe how much of the variation in one variable can be accounted for by a straight-line relationship with another variable.
5. Recognize outliers and potentially influential observations from a scatterplot with the regression line drawn on it.
6. Calculate the residuals and plot them against the explanatory variable x or against other variables. Recognize unusual patterns.

59

Chapter 3 Complete

- Homework: #'s 62, 63, 68a&b, 71
- Any questions on pg. 21-24 in additional notes packet

60