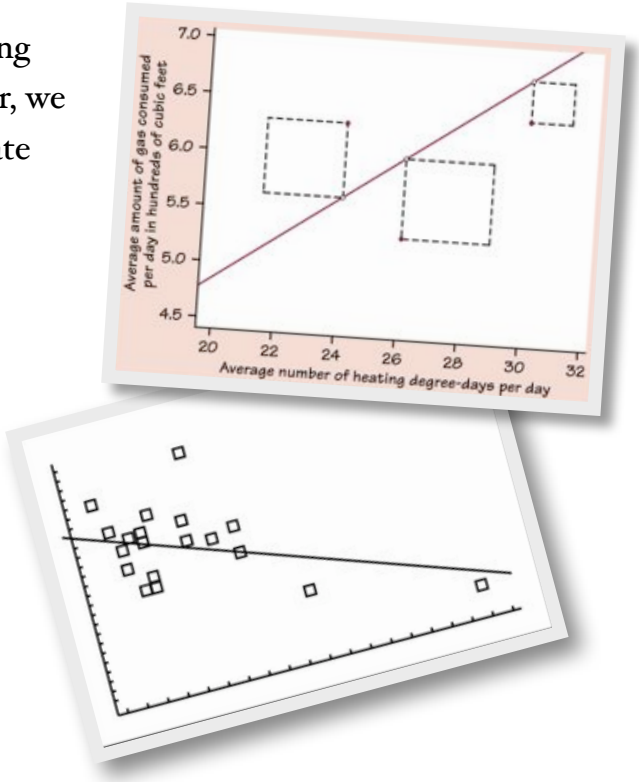
Chapter 3

Examining Relationships

Chapters 1 and 2 discussed methods for describing and summarizing univariate data. In this chapter, we are introduced to methods for describing bivariate relationships.

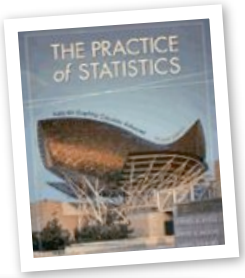


Examining Relationships:

- 3.1: Scatterplots
- 3.2: Correlation
- 3.3: Least-Squares Regression and Residuals

AP Stats Chapter 3: Examining Relationships

"Thou, nature, art my goddess; to thy laws my services are bound..." ~ Shakespeare's King Lear {Gauss' Motto}



Tentative Lesson Guide					
Date		**Stats**	**Lesson**	**Assignment**	**Done**
Tues	10/10	3.1	Scatterplots	Rd 121-134 Do 1-4,6-7,15-19, 22	
Wed	10/11	3.2	Correlation	Rd 140-145 Do 24-29, 33	
Thu	10/12	3.3	Least Squares Regression	Rd 149-156 Do 38-41	
Fri	10/13	3.3	Least Squares Regression	Rd 157-165 Do 42-43, 45	
Mon	10/16	3.3	Residuals	Rd 167-176, Do 46, 48	
Tues	10/17	**Rev**	**Review**	Rd 181-183 Do 62-65, 67-73	
Wed	10/18	**Exam**	**Chapter 3 Exam**	**Online Quiz Due**	
Thu	10/19	*EdMn*	*EdMn Break*		
Fri	10/20	*EdMn*	*EdMn Break*		

Note:
The purpose of this guide is to help you organize your studies for this chapter.
The schedule and assignments may change slightly.

Keep your homework organized and refer to this when you turn in your assignments at the end of the chapter.

Class Website:
Be sure to log on to the class website for notes, worksheets, links to our text companion site, etc.

http://web.mac.com/statsmonkey

Don't forget to take your online quiz!. Be sure to enter my email address correctly!

http://bcs.whfreeman.com/yates2e

My email address is:

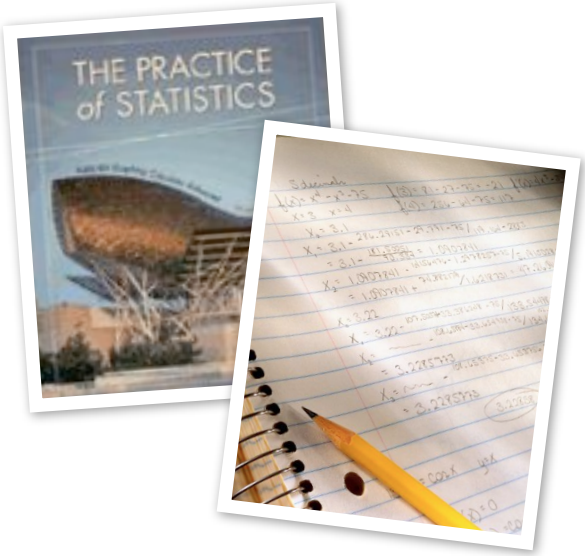
jmmolesky@isd194.k12.mn.us

Chapter 3 Objectives and Skills:

These are the expectations for this chapter. You should be able to answer these questions and perform these tasks accurately and thoroughly. Although this is not an exhaustive review sheet, it gives a good idea of the "big picture" skills that you should have after completing this chapter. The more thoroughly and accurately you can complete these tasks, the better your preparation.

SCATTERPLOTS

- ☐ Given a sample of data from a situation, and a few variables, be able to describe the association (direction, strength, and form), by using words, visual displays (both scatterplots and residual plots), and numerical measures of association. All of this must be in context of the data.
- ☐ Be able to spot and describe individual cases on a scatterplot.
- ☐ Recognize influential observations and outliers in a scatterplot. Remember that these aren't necessarily the same thing - outliers are values which "buck the trend." In most cases they have high residuals. Influential observations often "set the trend," but may throw off the association in the bulk of the data. They are often outliers in the x - direction.
- ☐ Given two variables in a scenario, be able to decide whether one should be the "explanatory variable" and the other the "response variable*," or if it doesn't matter.*



CORRELATION AND REGRESSION

- ☐ Understand the interpretation and properties of r.
- ☐ Use the TI-83/84 OR a computer printout to determine the least squares regression equation for predictions.
- ☐ Interpret the meaning of the numerical values of the slope and intercept of the regression equation, in proper context.
- ☐ Understand the technical meaning of r^2
- ☐ Use means and standard deviations of x and y to find the slope and the intercept of the regression line.
- ☐ Use the regression line to predict y values for a given x value.
- ☐ Recognize extrapolation, and be aware of its dangers.
- ☐ Calculate the residual for a given observation and Interpret residual plots.
- ☐ Recognize the fallacy: "correlation does not imply causation." Understand how to resolve the fallacy by explaining the role of lurking variables, common response, or confounding.
- ☐ TI 83 skills you must have: make a scatterplot, find LSRL, find r and r^2, make a residual plot.

3.1: Introduction to Bivariate Analysis

Like univariate data analyses, the exploration of bivariate data starts with an organized plot. While dotplots, stemplots, boxplots, and histograms were adequate tools for univariate datasets, a **scatterplot** is the most useful tool for assessing the relationship between two variables.

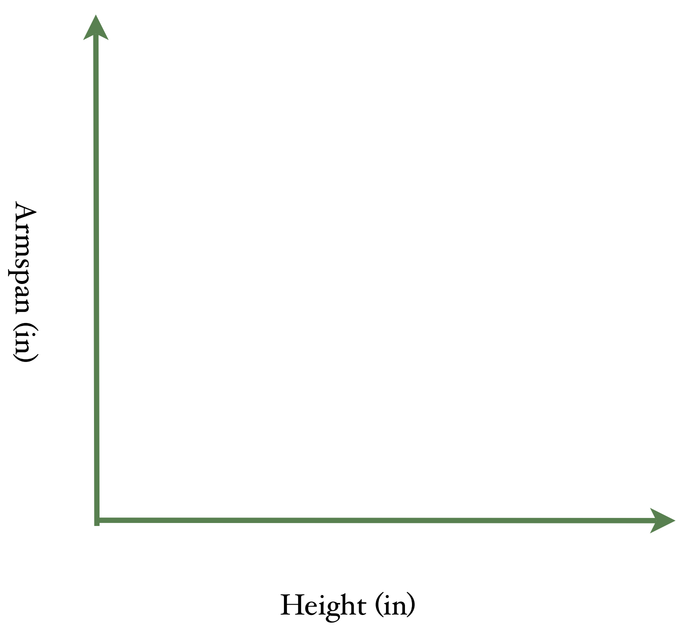
In this chapter, we will learn how to use a scatterplot to assess the strength, direction, and form of a bivariate set of data. We will quantify the strength of the relationship by discussing **correlation**, and we will learn how to model linear relationships by using the **method of least squares regression**.

Heights and Armspans:

Is there a relationship between your height and your armspan? Can one be predicted from the other? Let's gather data from the class and construct a plot to explore the relationship. Later in the chapter, we'll build a prediction model to calculate an armspan from a given height.

Height	Armspan

With the help of a partner, measure your height and armspan to the nearest 1/4 in. Record each pair of observations on the table on the board, and copy the class results below. Plot your (height, armspan) on the scatterplot on the board and sketch the class result below.

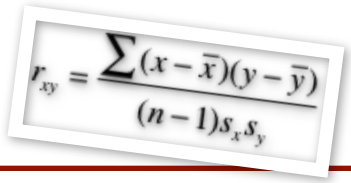


Describe the strength, direction, and form of the relationship *in context:*

Strength:
Direction:
Form:

Relationship *in context*:

3.2: Correlation



While a scatterplot can give us a general idea of the strength of a relationship between two variables, it is helpful to have a numeric measure that quantifies the relationship. The **correlation coefficient** is a quantitative assessment of the strength of the relationship between two variables.

Pearson's Product Moment Coefficient of Correlation: "r"

Like standard deviation, the calculation of the correlation coefficient by hand can be quite tedious. We will rely on technology to produce "r" for us. However, it is important to understand where the value comes from in order to interpret it properly.

Consider the following example adapted from "Statistics and Data Analysis" by Peck, Olsen, and Devore. The article "Effects of Humor on Test Anxiety and Performance" (*Psych. Reports* (1999): 1203-1212) examined the relationship between test anxiety (x) and exam score (y). Data consistent with summary quantities in the paper appear below. Higher levels of x indicate higher levels of anxiety.

x	23	14	14	0	7	20	20	15	21
y	43	59	48	77	50	52	46	51	51

Construct a scatterplot of anxiety vs. exam score. Describe the strength, direction and form of the relationship.

To calculate the correlation coefficient, we must calculate the mean and standard deviation for x and y, respectively. .

$\bar{x} =$	$s_x =$	$\bar{y} =$	$s_y =$

Next, we need to calculate a "z-score" for each observed x and y value. Use the list editor in your calculator to help automate the process. When z-scores are calculated for each observed x and y pair, we'll find the product $z_x z_y$ for each pair. Use your calculator to fill in the table on the next page.

As you fill in the table, think about what each z-score tells you about the observation. If the data are positively correlated, what should be true about the products of the z-scores? What if they are negatively correlated?

Anxiety vs. Exam Score

x	y	z_x	z_y	$z_x z_y$
23	43			
14	59			
14	48			
0	77			
7	50			
20	52			
20	46			
15	21			
21	51			
			$\sum z_x z_y =$	

The correlation coefficient is found by dividing the sum of the z-score products by (n-1).

That is, $r = \frac{\sum z_x z_y}{n-1} =$

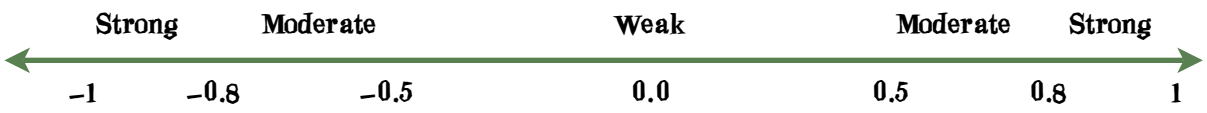
Calculate and interpret the value of r *in the context of the problem:*

Properties of r:

☑ The value of r does not depend on the unit of measurement for either variable.

☑ The value of r does not depend on which variable is the explanatory or which is the response.

☑ The value or r will always be between -1 and 1. The closer |r| is to 1, the stronger the linear relationship.



☑ The correlation coefficient r = 1 only when all the points on the scatterplot lie on a straight line with a positive slope. Likewise, r = -1 only when all points fall on a straight line with a negative slope.

☑ The value of r is a measure of the **linear** relationship between x and y.

3.3: The Method of Least Squares Regression

Humerus bones from the same species of animal tend to be good predictors of an animal's Femur length (and therefore of its Height). When fossils of humerus bones are discovered, archeologists can often determine the species of animal by examining these values and the ratios of one to the other. The species *Molekius Primatium* once inhabited the northern regions of Minnesota. Suppose 20 fossil pairs from this species are unearthed at a site on Minnesota's Iron Range. Use the following information to determine the relationship between humerus and femur length. Then, construct a model to predict Femur Length from Humerus Length based on the observed relationship.

Humerus	Femur
49	218.8
14.2	63.6
11.6	42.9
6.7	42.7
42.3	265.6
13.5	96.9
9.5	32.4
6.0	28.9
36.3	170.9
13.4	60.9
9.4	68.3
4.6	40.7
16.4	85.9
13.2	144.2
7.5	48.6
3.8	13.8
14.9	88.1
11.8	53.6
7.2	32.2
3.4	22.2

Construct and interpret a scatterplot of the relationship between the humerus and femur lengths of the fossil pairs.

Calculate and interpret the correlation coefficient r:

Calculate and interpret the coefficient of determination r^2:

Note there appears to be a strong positive linear relationship between humerus and femur length. Therefore, it appears a function of the form **y=a+bx** may exist that comes close to the pattern exhibited by the data. If we can find this equation, we will have a model that may be useful in predicting femur length from humerus length.

Since no line fits the data perfectly, each observation will exhibit a **deviation**. Each unique line placed on the scatterplot will produce a different set of deviations. The **line of best fit** is the one line that minimizes the sum of the squared deviations and is called the **least-squares regression line (LSRL)**.

Like many of the quantitative summaries we have observed so far, the calculations of the slope and intercept of the least squares regression line can be somewhat labor intensive. The formulas for the slope "**b**" and intercept "**a**" are below, but we will rely on our calculator to provide the equation for us.

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} \qquad a = \bar{y} - b\bar{x} \qquad \hat{y} = a + bx$$

The ^ above the "y" emphasizes that y-hat is a prediction of y for a particular x based on the model.

The Method of Least Squares Regression - Continued

Use the **STAT CALC 8:LinReg(a+bx)** feature on your calculator to find the equation of the line that best fits the (humerus, femur) data from the previous page. Interpret the equation in the context of the problem. What do the slope and intercept tell you about the relationship?

LSRL Equation:

Write in terms of the problem:

Interpret the slope "b":

Interpret the intercept "a":

Use the model to predict the femur length of a Moleskius Primatium whose humerus is 16.4 units long.

Note, we observed a fossil pair with a humerus=16.4. Did your prediction match? If not, how far off was it? We call the difference between the observed value and predicted value a **residual**. Calculate the residual for this observation. {residual = observed y - predicted y}

One way to assess whether or not we have an appropriate prediction model is to calculate and plot the residuals for all observed bivariate pairs. A **residual plot** of the (x,residual) pairs can be used to assess the appropriateness of a regression line and can be used to suggest better prediction models.

Use your calculator to construct and interpret a residual plot for the (humerus, femur) prediction line.

Examples and Practice

Consider the following data on x= height (in) and y= average weight (lb) for American females aged 30-39 from The World Almanac and Book of Facts.

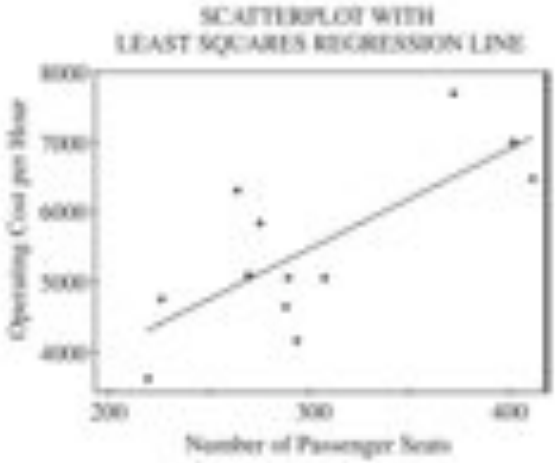
x	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
y	113	115	118	121	124	128	131	134	137	141	145	150	153	159	164

a. Construct a scatterplot and comment on the features of the plot (Strength, Direction, Form).
b. Compute the value of the correlation coefficient r and interpret it.
c. Compute the value of the coefficient of determination r^2 and interpret it. Does it appear that anxiety is the only factor that contributes to poor exam performance?
d. Calculate the formula for the least squares regression line and interpret the slope.
e. Use the LSRL to predict the average weight for a 63 in tall female in the 30-39 age range. Calculate the residual for x=63.
f. Assess whether or not the LSRL is suitable by constructing a residual plot.

Would you be fairly confident in your prediction from part e? Can you describe a model that may be a *better* predictor of average weight given height? What does this other model suggest about how weight and height are related?

2002 AP® FREE-RESPONSE PROBLEM #4

Commercial airlines need to know the operating cost per hour of flight for each plane in their fleet. In a study of the relationship between operating cost per hour and number of passenger seats, investigators computed the regression of operating cost per hour on the number of passenger seats. The 12 sample aircraft used in the study included planes with as few as 256 passenger seats and planes with as many as 400 passenger seats. Operating cost per hour ranged between $3,600 and $7,800. Some computer output from a regression analysis of these data is shown below.



Predictor	Coef	StDev	T	P
Constant	1136	1226	0.93	0.376
Seats	14.673	4.027	3.64	0.005
S = 845.3	R-Sq = 57.0%		R-Sq (adj) = 52.7%	

(a) What is the equation of the least squares regression line that describes the relationship between operating cost per hour and number of passenger seats in the plane? Define any variables used in this equation.

(b) What is the value of the correlation coefficient for operating cost per hour and number of passenger seats in the plane? Interpret this correlation.

(c) Suppose that you want to describe the relationship between operating cost per hour and number of passenger seats in the plane for planes only in the range of 250 to 350 seats. Does the line shown in the scatterplot still provide the best description of the relationship for data in this range? Why or why not?