STATISTICS

INFORMED DECISIONS USING DATA

Fifth Edition

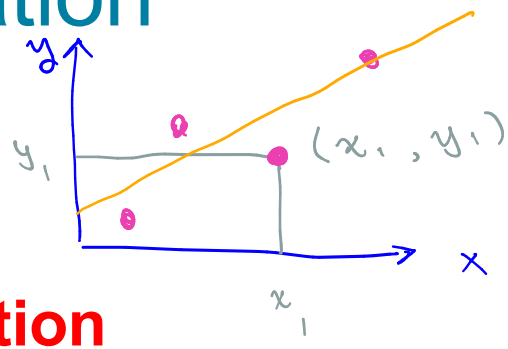
Chapter 4

Describing the Relation between Two Variables

Pearson

4.1 Scatter Diagrams and Correlation

Learning Objectives



1. Draw and interpret **scatter plots (diagrams)**
2. Describe the properties of the **linear correlation coefficient**
3. Compute and interpret the **linear correlation coefficient**
4. Determine whether a linear relation exists between two variables
5. Explain the difference between correlation and causation

4.1 Scatter Diagrams and Correlation

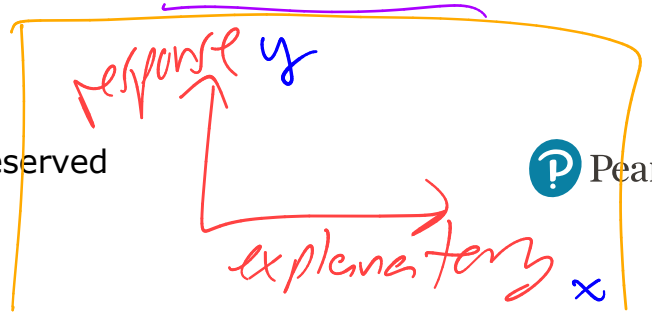
4.1.1 Draw and Interpret Scatter Diagrams (1 of 6)

The **response variable** is the variable whose value can be explained by the value of the **explanatory** or **predictor variable**.

A **scatter plot/diagram** is a graph that shows the relationship between two quantitative variables measured on the same individual.

Each individual in the data set is represented by a point in the scatter diagram.

The explanatory variable is plotted on the horizontal axis, and the response variable is plotted on the vertical axis.



4.1 Scatter Diagrams and Correlation

4.1.1 Draw and Interpret Scatter Diagrams (1 of 6)

The way to distinguish the difference between these two variables is by asking, "Which statement makes sense?"

Ex: *A researcher wants to examine whether babies fed on breast milk are more or less likely to be ill.*

(a) *Feeding a baby on breast milk causes resistance to disease.*

(b) *Resistance to disease causes a baby to feed on breast milk.*

Can you identify which variable is which? – presence or absence of breast milk – resistance to disease

(a) Exp. Var – feed baby w/ milk

Resp Var – resistance to disease

(b) Exp Var – resistance to disease

Resp. Var – feeding or milk

doesn't make sense

4.1 Scatter Diagrams and Correlation

4.1.1 Draw and Interpret Scatter Diagrams (2 of 6)

EXAMPLE Drawing and Interpreting a Scatter Diagram

The data shown to the right are based on a study for drilling rock. The researchers wanted to determine whether the time it takes to dry drill a distance of 5 feet in rock increases with the depth at which the drilling begins. So, depth at which drilling begins is the explanatory variable, x, and time (in minutes) to drill five feet is the response variable, y. Draw a scatter diagram of the data.

Source: Penner, R., and Watts, D.G. "Mining Information." *The American Statistician,* Vol. 45, No. 1, Feb. 1991, p. 6.

Depth at Which Drilling Begins, x (in feet)	Time to Drill 5 Feet, y (in minutes)
35	5.88
50	5.99
75	6.74
95	6.1
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.9

4.1 Scatter Diagrams and Correlation

4.1.1 Draw and Interpret Scatter Diagrams (3 of 6)

Calculator:

Enter into lists:

>> L1 = explanatory var;

>> L2 = response var

Plot:

>> Stat Plot (2^{nd}+Y=)

>> Turn Plot 1 ON

>> Select Scatter Plot icon

>> Select: Xlist = L1; Ylist = L2

>> ZOOM: 9. ZoomStat

Depth at Which Drilling Begins, *x* (in feet)	Time to Drill 5 Feet, *y* (in minutes)
35	5.88
50	5.99
75	6.74
95	6.1
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.9

4.1 Scatter Diagrams and Correlation

4.1.1 Draw and Interpret Scatter Diagrams (3 of 6)

is linear relationship but weak.

Calculator:

Enter into lists:

>> L1 = explanatory var;

>> L2 = response var

Plot:

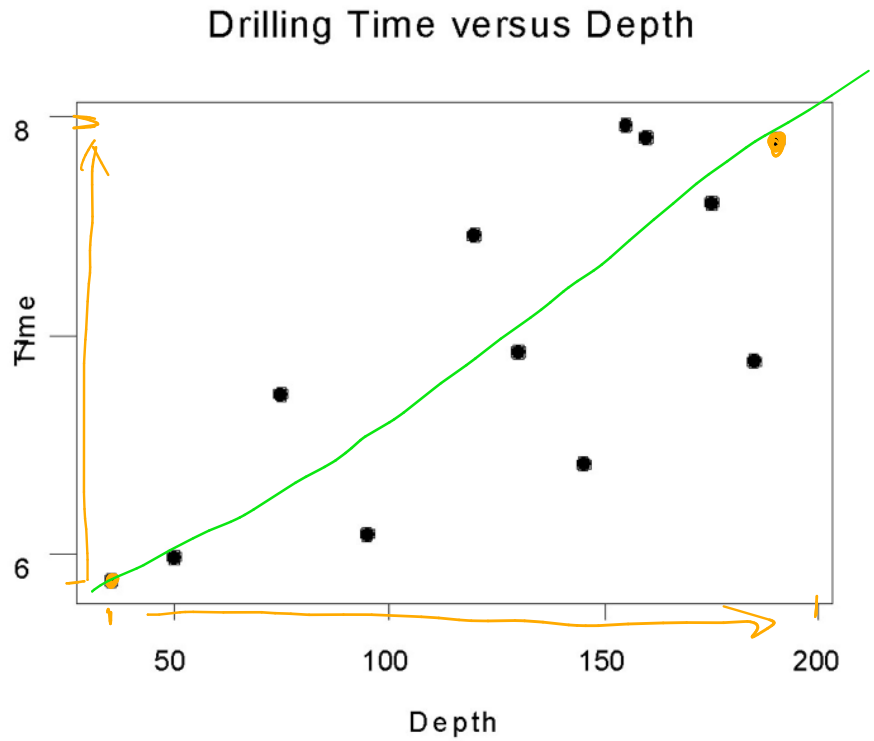
>> Stat Plot (2nd+Y=)

>> Turn Plot 1 ON

>> Select Scatter Plot icon

>> Select: Xlist = L1; Ylist = L2

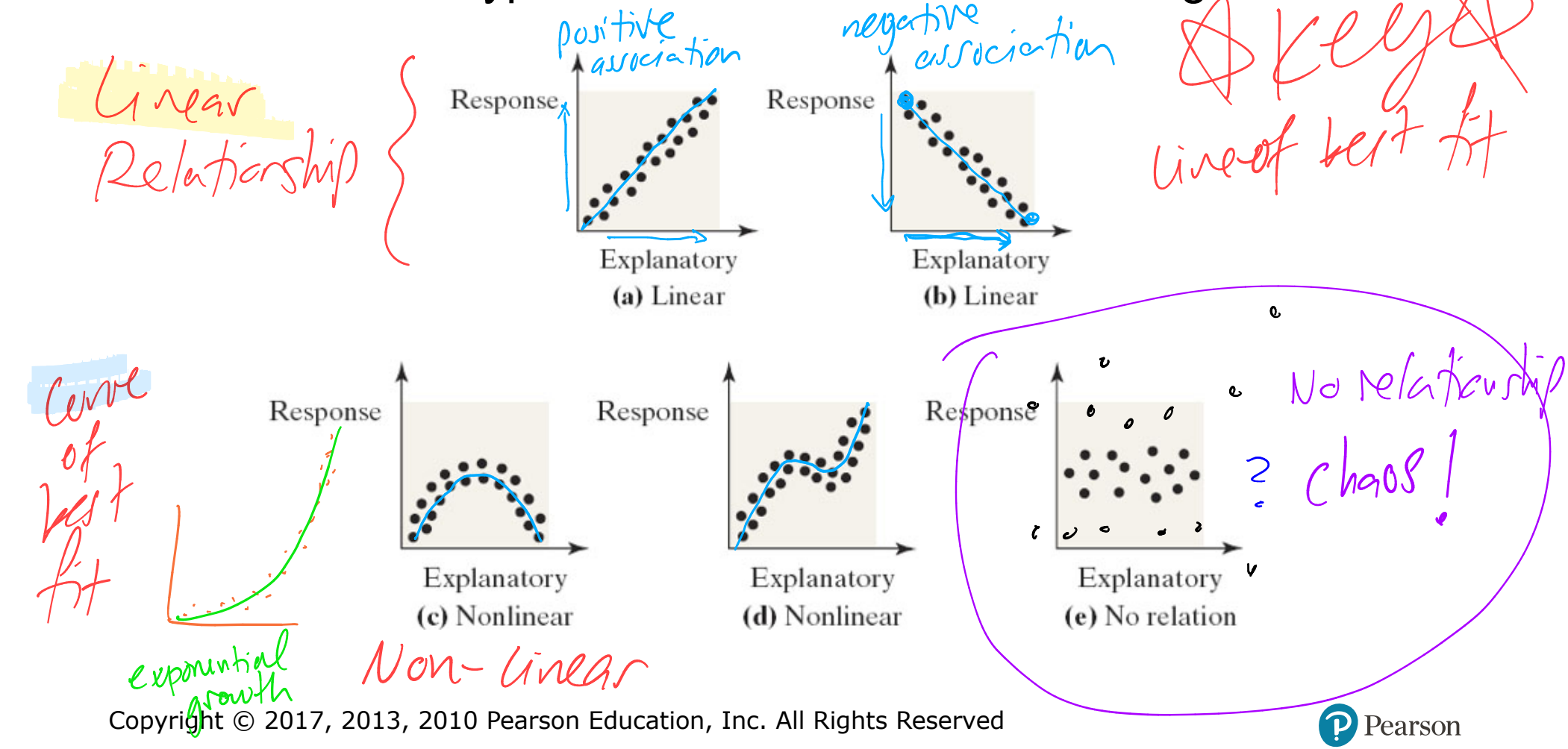
>> ZOOM: 9. ZoomStat



4.1 Scatter Diagrams and Correlation

4.1.1 Draw and Interpret Scatter Diagrams (4 of 6)

Various Types of Relations in a Scatter Diagram

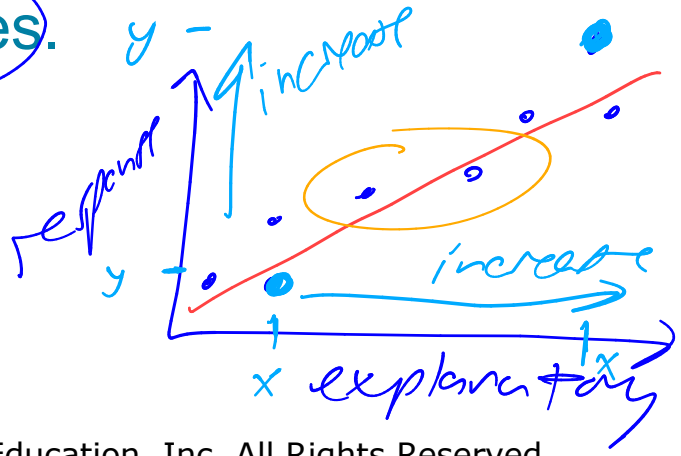


4.1 Scatter Diagrams and Correlation

4.1.1 Draw and Interpret Scatter Diagrams (5 of 6)

Two variables that are linearly related are **positively associated** when above-average values of one variable are associated with above-average values of the other variable and below-average values of one variable are associated with below-average values of the other variable.

That is, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable also increases.

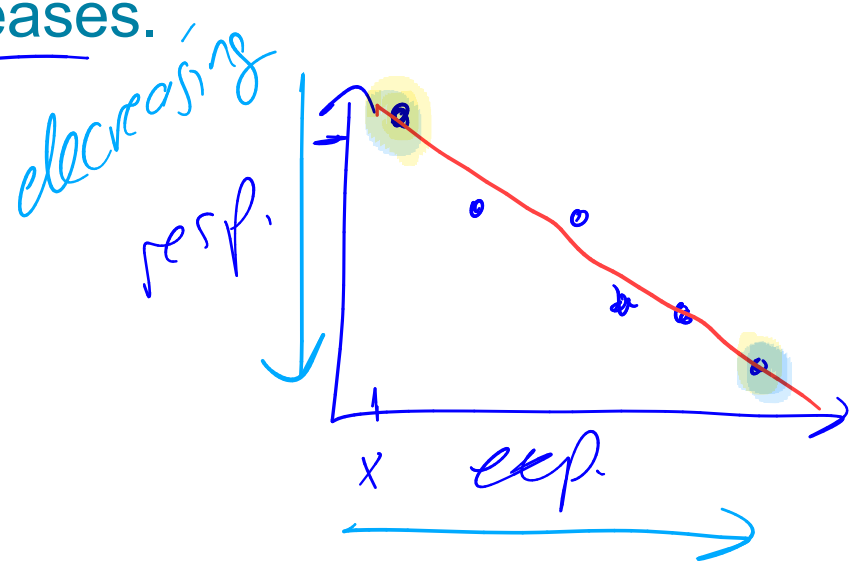


4.1 Scatter Diagrams and Correlation

4.1.1 Draw and Interpret Scatter Diagrams (6 of 6)

Two variables that are linearly related are **negatively associated** when above-average values of one variable are associated with below-average values of the other variable.

That is, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.



4.1 Scatter Diagrams and Correlation

4.1.2 Describe the Properties of the Linear Correlation Coefficient (1 of 6)

The **linear correlation coefficient** (or **Pearson product moment correlation coefficient)** is a measure of the strength and direction of the linear relation between two quantitative variables.

Notation:

Population correlation coefficient: The Greek letter ρ (rho)

Sample correlation coefficient: r

We present only the formula for the sample correlation coefficient.

4.1 Scatter Diagrams and Correlation

4.1.2 Describe the Properties of the Linear Correlation Coefficient (2 of 6)

Sample Linear Correlation Coefficient

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

where $\bar{x}$ is the sample mean of the explanatory variable
s_x is the sample standard deviation of the explanatory variable
$\bar{y}$ is the sample mean of the response variable
s_y is the sample standard deviation of the response variable
n is the number of individuals in the sample

Use Calculator Not Formula!

Enter into Lists:

L1 = explanatory var;

L2 = response var

>> Stats >> CALC

>> LinReg(a+bx)

Must set-up Diagnostics ON

$r = 0.773$

4.1 Scatter Diagrams and Correlation

4.1.2 Describe the Properties of the Linear Correlation Coefficient (3 of 6)

Properties of the Linear Correlation Coefficient

1. The linear correlation coefficient is always between −1 and 1, inclusive. That is, $-1 \le r \le 1$.
2. If $r = +1$, then a perfect positive linear relation exists between the two variables.
3. If $r = -1$, then a perfect negative linear relation exists between the two variables.
4. The closer r is to +1, the stronger the evidence is of a positive association between the two variables.
5. The closer r is to −1, the stronger the evidence is of a negative association between the two variables.

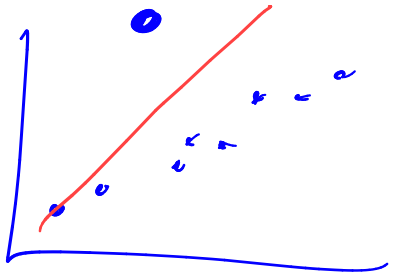
4.1 Scatter Diagrams and Correlation

4.1.2 Describe the Properties of the Linear Correlation Coefficient (4 of 6)

6. If r is close to 0, then little or no evidence exists of a linear relation between the two variables. So ***r* close to 0 does not imply no relation, just no linear relation**.

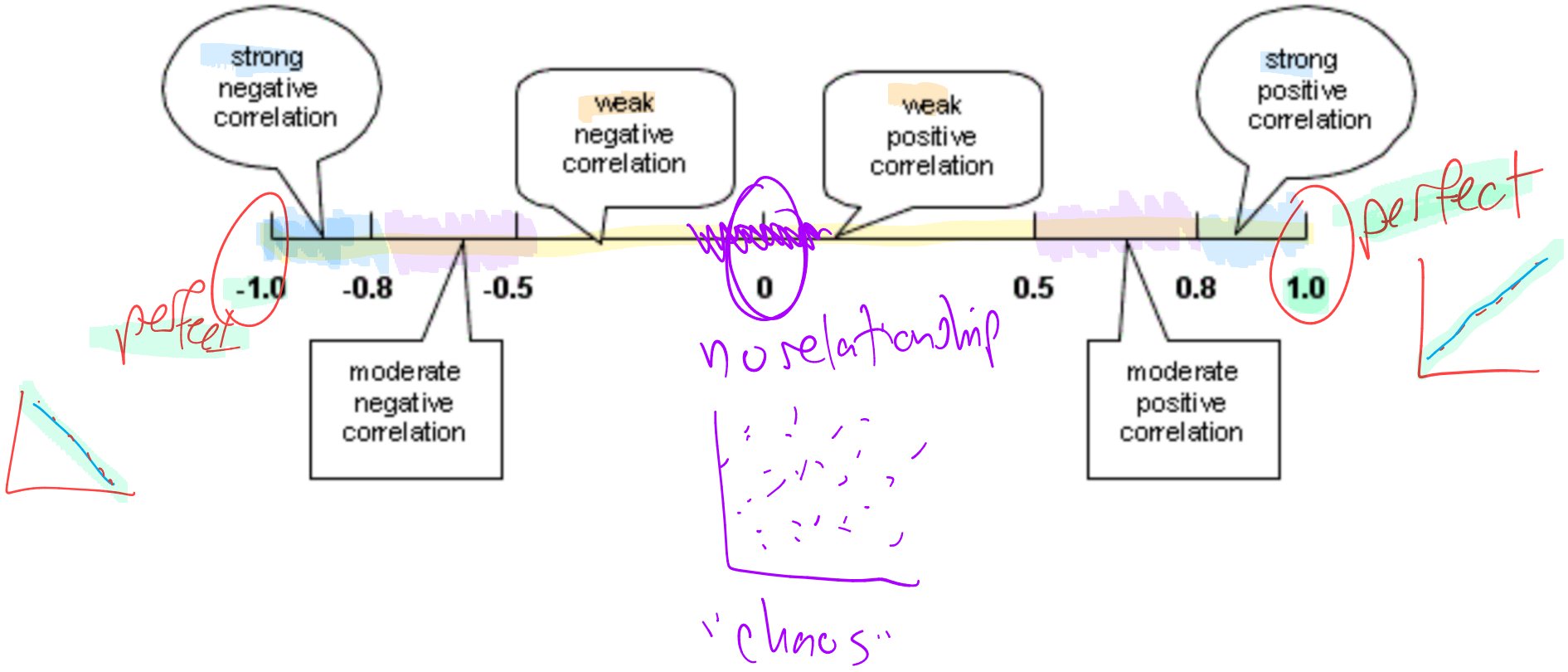
7. The linear correlation coefficient is a **unitless** measure of association. So the unit of measure for x and y plays no role in the interpretation of r.

8. The correlation coefficient is **not resistant**. Therefore, an observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.



4.1 Scatter Diagrams and Correlation

4.1.2 Describe the Properties of the Linear Correlation Coefficient (4 of 6)



Pearson

4.1 Scatter Diagrams and Correlation

4.1.2 Describe the Properties of the Linear Correlation Coefficient (4 of 6)

GTQ

Choose a word from each category to describe each scatter plot shown to the left and write near/next to each plot:

- Linear relationship
- Non-Linear Relationship
- No Relationship

- Perfect
- Strong
- Moderate
- Weak

- Positive Association
- Negative Association
- No Correlation

Match the r values with scatter plots on the next slide:

- r = 1, r = -1
- r = 0.9, r = -0.93
- r = 0.4, r = -0.4
- r = 0, r = 0.02

Choose a word from each category to describe each scatter plot shown to the left and write near/next to each plot:

- Linear relationship
- Non-Linear Relationship
- No Relationship

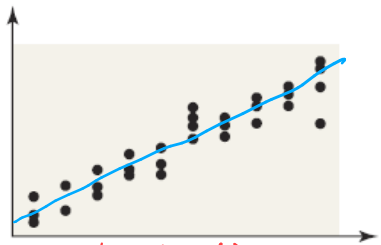
- Perfect
- Strong
- Moderate
- Weak

- Positive Association
- Negative Association
- No Correlation

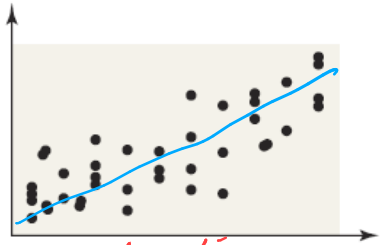
Match the r values with scatter plots on the next slide:

- r = 1, r = 0.4
- r = -0.4, r = 0
- r = 0.02, r = 0.9
- r = -0.93, r = -1

perfect linear
+ r = 1

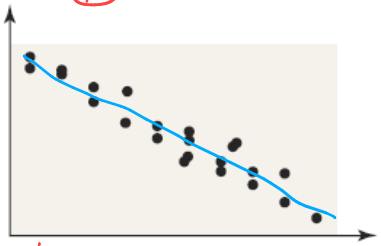


moderat linear
+ r = 0.9

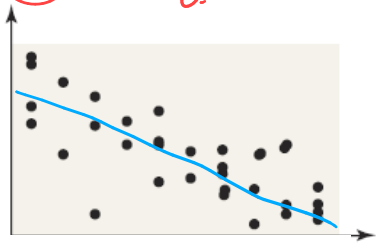


weak linear
+ r = 0.4

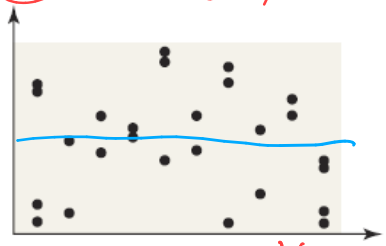
perfect linear
- r = -1



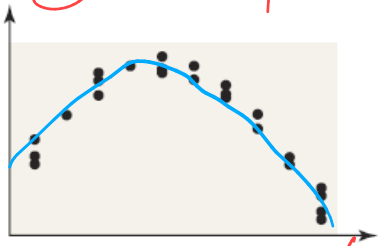
moderate linear
- r = -0.93



weak linear
- r = -0.4



No relationship
r = 0

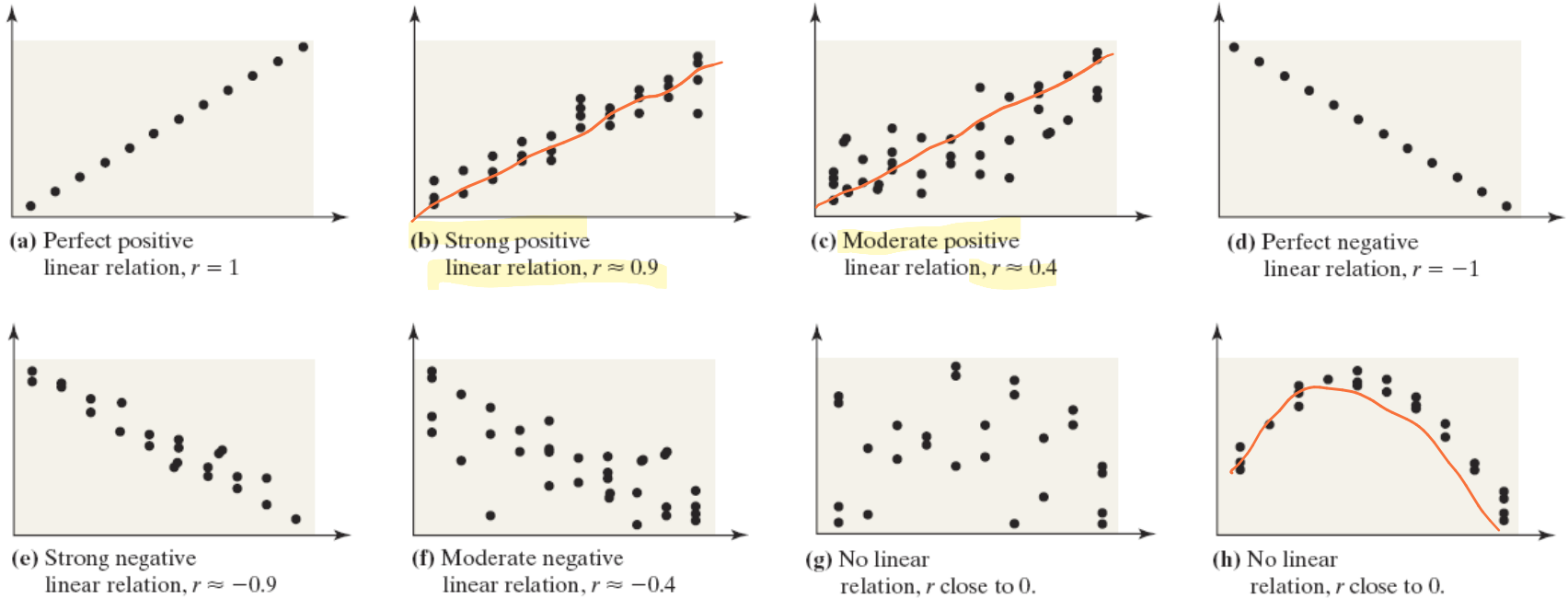


non linear rel.
r = 0.02

no correlation

4.1 Scatter Diagrams and Correlation

4.1.2 Describe the Properties of the Linear Correlation Coefficient (5 of 6)



(a) Perfect positive linear relation, $r = 1$

(b) Strong positive linear relation, $r \approx 0.9$

(c) Moderate positive linear relation, $r \approx 0.4$

(d) Perfect negative linear relation, $r = -1$

(e) Strong negative linear relation, $r \approx -0.9$

(f) Moderate negative linear relation, $r \approx -0.4$

(g) No linear relation, r close to 0.

(h) No linear relation, r close to 0.

non-linear

4.1 Scatter Diagrams and Correlation

4.1.3 Compute and Interpret the Linear Correlation Coefficient (1 of 5)

EXAMPLE Determining the Linear Correlation Coefficient

Determine the linear correlation coefficient of the drilling data.

Use Calculator Not Formula!

Enter into Lists: L1 = explanatory var; L2 = response var

>> Stats >> CALC

>> LinReg(a+bx)

$r = 0.773$

Depth at Which Drilling Begins, x (in feet)	Time to Drill 5 Feet, y (in minutes)
35	5.88
50	5.99
75	6.74
95	6.1
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.9

4.1 Scatter Diagrams and Correlation

4.1.3 Compute and Interpret the Linear Correlation Coefficient (3 of 5)

Rounding Rule for Correlation Coefficient: 3 sig figs

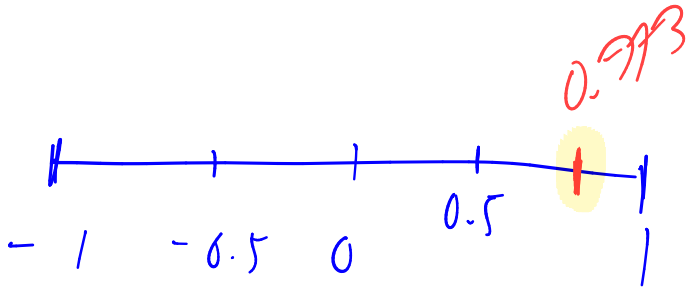
r = 0.773.

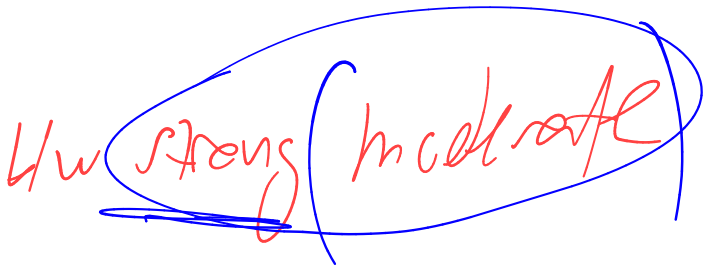
Use Calculator Not Formula!

Enter into Lists: L1 = explanatory var; L2 = response var

>> Stats >> CALC

>> LinReg(a+bx)





4.1 Scatter Diagrams and Correlation

4.1.4 Determine whether a Linear Relation Exists between Two Variables (1 of 2)

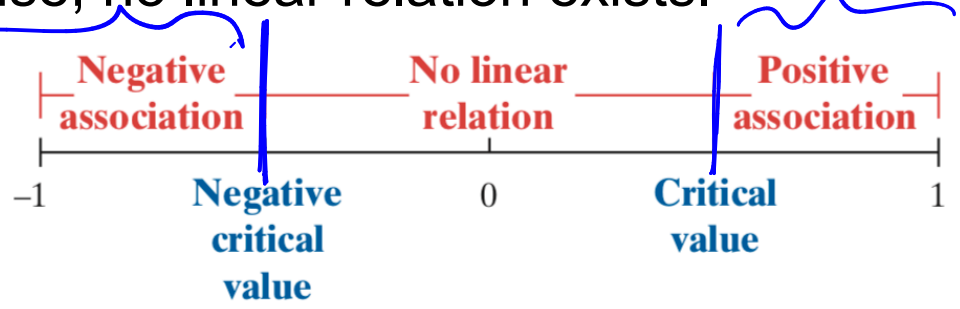
Testing for a Linear Relation

Step 1 Determine the absolute value of the correlation coefficient.

Step 2 Find the **critical value in Table II** for the given sample size.

Step 3 If the absolute value of the correlation coefficient is greater than the critical value, we say a linear relation exists between the two variables.

Otherwise, no linear relation exists.



4.1 Scatter Diagrams and Correlation

4.1.4 Determine whether a Linear Relation Exists between Two Variables (2 of 2)

EXAMPLE Does a Linear Relation Exist?

Determine whether a linear relation exists between time to drill five feet and depth at which drilling begins. Comment on the type of relation that appears to exist between time to drill five feet and depth at which drilling begins.

- $r = 0.773$
- $n = 12$
- $r = 0.773 > 0.576$ (critical value)

There is a linear relationship!

Table II

Critical Values for Correlation Coefficient

n	
3	0.997
4	0.950
5	0.878
6	0.811
7	0.754
8	0.707
9	0.666
10	0.632
11	0.602
12	0.576
13	0.553
14	0.532

4.1 Scatter Diagrams and Correlation

4.1.4 Determine whether a Linear Relation Exists between Two Variables (2 of 2)

EXAMPLE Does a Linear Relation Exist?

Determine whether a linear relation exists between time to drill five feet and depth at which drilling begins. Comment on the type of relation that appears to exist between time to drill five feet and depth at which drilling begins.

The correlation between drilling depth and time to drill is 0.773.

The critical value for $n = 12$ observations is 0.576.

Since 0.773 > 0.576, there is a positive linear relation between time to drill five feet and depth at which drilling begins.

Table II

Critical Values for Correlation Coefficient

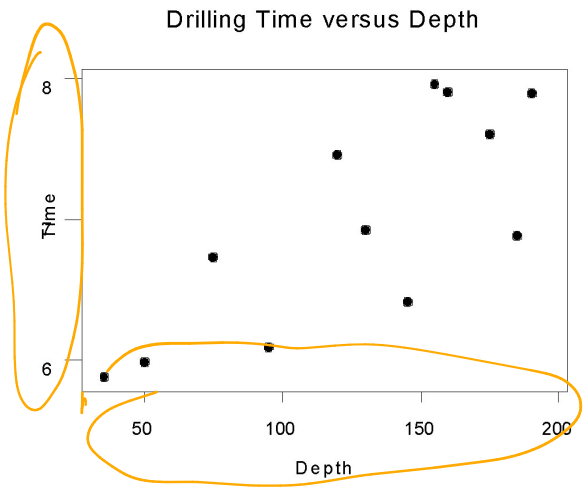
n	
3	0.997
4	0.950
5	0.878
6	0.811
7	0.754
8	0.707
9	0.666
10	0.632
11	0.602
12	0.576
13	0.553
14	0.532

4.1 Scatter Diagrams and Correlation

4.1.4 Determine whether a Linear Relation Exists between Two Variables (2 of 2)

EXAMPLE Does a Linear Relation Exist?

Determine whether a linear relation exists between time to drill five feet and depth at which drilling begins. Comment on the type of relation that appears to exist between time to drill five feet and depth at which drilling begins.



The correlation between drilling depth and time to drill is 0.773.

The critical value for $n = 12$ observations is 0.576.

Since 0.773 > 0.576, there is a positive linear relation between time to drill five feet and depth at which drilling begins.

Pearson

Thursday 3/12

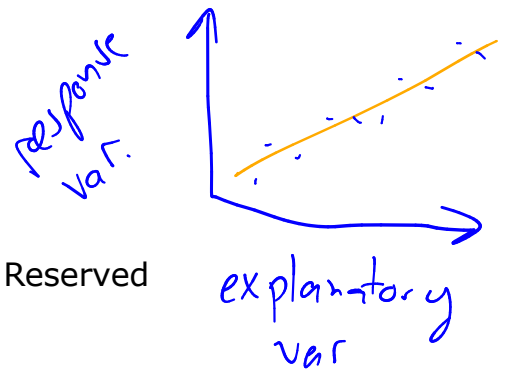
4.1 Scatter Diagrams and Correlation

4.1.5 Explain the Difference between Correlation and Causation (1 of 8)

According to data obtained from the Statistical Abstract of the United States, the correlation between the percentage of the female population with a bachelor's degree and the percentage of births to unmarried mothers since 1990 is 0.940.

Does this mean that a higher percentage of females with bachelor's degrees causes a higher percentage of births to unmarried mothers?

NO "correlation does not imply causation"



4.1 Scatter Diagrams and Correlation

4.1.5 Explain the Difference between Correlation and Causation (2 of 8)

Certainly not!

The correlation exists only because both percentages have been increasing since 1990. It is this relation that causes the high correlation. In general, *time series data (data collected over time)* may have high correlations because each variable is moving in a specific direction over time (both going up or down over time; one increasing, while the other is decreasing over time).

When data are observational, we cannot claim a causal relation exists between two variables. We can only claim causality when the data are collected through a designed experiment.

4.1 Scatter Diagrams and Correlation

4.1.5 Explain the Difference between Correlation and Causation (3 of 8)

Another way that two variables can be related even though there is not a causal relation is through a **lurking variable**.

A **lurking variable** is related to both the explanatory and response variable.

For example, ice cream sales and crime rates have a very high correlation. Does this mean that local governments should shut down all ice cream shops?

4.1 Scatter Diagrams and Correlation

4.1.5 Explain the Difference between Correlation and Causation (3 of 8)

Another way that two variables can be related even though there is not a causal relation is through a **lurking variable**.

A **lurking variable** is related to both the explanatory and response variable.

For example, ice cream sales and crime rates have a very high correlation. Does this mean that local governments should shut down all ice cream shops?

No! The lurking variable is temperature. As air temperatures rise, both ice cream sales and crime rates rise.