Section 4.1

Scatter Diagrams and Correlation

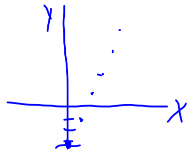
Objectives

1. Draw and interpret scatter 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

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

$y = 2x \cdot 3$

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



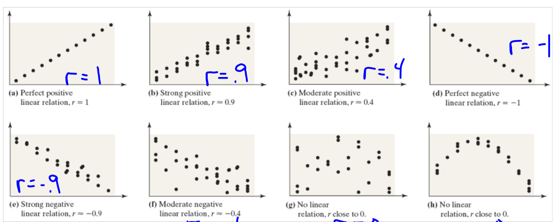
The data shown to the right are based on a study for drilling rock. The researchers wanted to determine whether the time it takes to dry drill a distance of 5 feet in rock increases with the depth at which the drilling begins.

Depth at Which Drilling Begins, x (in feet) L_1	Time to Drill 5 Feet, y (in minutes) L_2
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

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.

r represents the sample correlation coefficient.

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



(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.

Strong neg Linear   r = -.4   r = 0   r = 0

Determine the linear correlation coefficient of the drilling data.

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

$n = 12$

Fix Calc:
2nd
Catalog
x⁻¹ D
↓ Diagnostic ON

Find r:
Stat
Calc
#4 LinReg
$r = .773$

Direction   Strength
Positive strong Linear

Testing for a Linear Relation

Step 1 Determine the absolute value of the correlation coefficient. r=.773

Step 2 Find the critical value in Table II from Appendix A for the given sample size. CV=.576

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.

.773>.576
yes
It is Linear

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

Difference between Correlation and Causation

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?

↑ Population

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?

Warm temp.

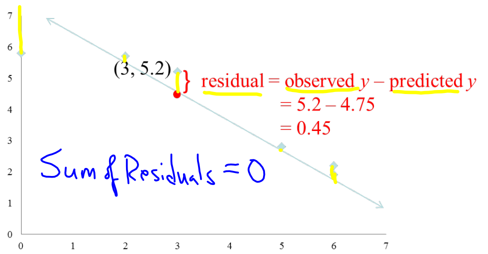
Section 4.2

Least-squares Regression

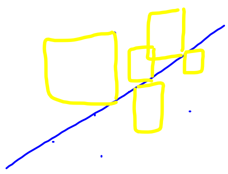
Objectives

- Find the least-squares regression line and use the line to make predictions
- Interpret the slope and the *y*-intercept of the least-squares regression line
- Compute the sum of squared residuals

The difference between the observed value of *y* and the predicted value of *y* is the error, or **residual**.



Least-Squares Regression Criterion
The **least-squares regression line** is the line that minimizes the sum of the squared errors (or residuals)



Using the drilling data

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

(a)Find the least-squares regression line.

(b)Predict the drilling time if drilling starts at 130 feet.

(c)Is the observed drilling time at 130 feet above, or below, average.

(d)Draw the least-squares regression line on the scatter diagram of the data.

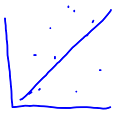
a) $\hat{y} = .012x + 5.53$

b) $y = .012(130) + 5.53$

$y = 7.09$

c) Below

d)



Residual for 130 : $6.93 - 7.09$

$-.16$

Interpret the of Slope

For every addition foot of depth the drilling time increases by .012.

Interpret the *y*-intercept

The drilling at 0 feet is 5.53 minutes.

Extrapolation

If the least-squares regression line is used to make predictions based on values of the explanatory variable that are much larger or much smaller than the observed values, we say the researcher is working **outside the scope of the model**. Never use a least-squares regression line to make predictions outside the scope of the model because we can't be sure the linear relation continues to exist.

Predicting: 130 feet

Good predition because it is within the data set