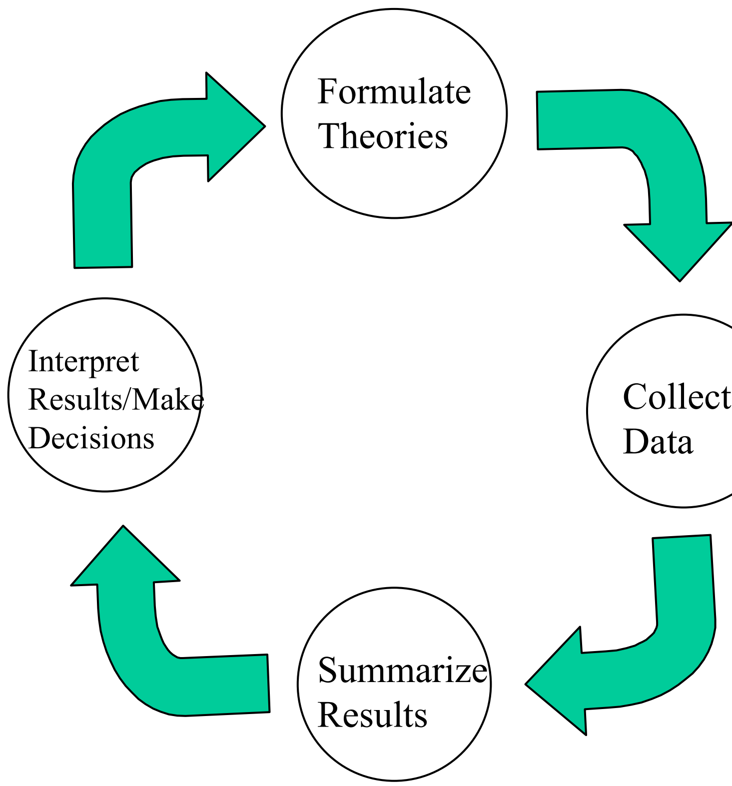
Chapter Goals

To understand the methods for displaying and describing relationship among variables.



Definition *Data for a single variable is called* ***univariate data****, for two variables,* ***bivariate data****, and for more than two, it is called* ***multivariate data****.*

Methods for studying relationships:

- Graphical
  - Scatterplots
  - Line plots
  - 3-D plots
- Models
  - Linear regression
  - Correlations
  - Frequency tables

Two Quantitative Variables

Definition *The **response variable**, also called the **dependent variable**, is the variable we want to predict, and is usually denoted by y.*

Definition *The **explanatory variable**, also called the **independent variable**, is the variable that attempts to explain the response, and is denoted by x.*

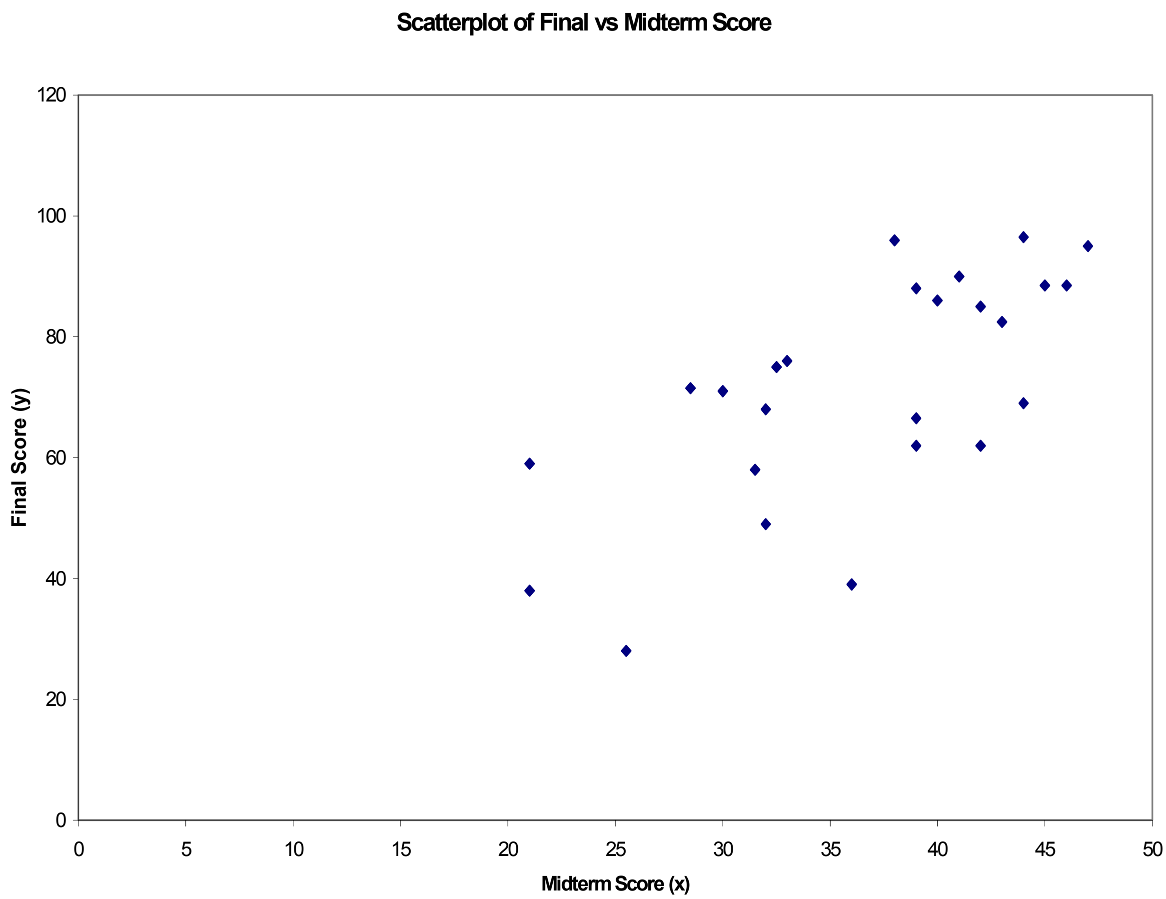
Let's do it! 7.1

Response	Explanatory
Height of son	Height of the father, height of the mother, age
Weight	

Scatterplots

Example *The following table gives the partial list of midterm score* (x) *and the final score* (y) *for 25 students in a particular class.*

Student	x	y
1	39	62
2	44	69
3	32	68
4	40	86
5	45	88.5
6	46	88.5
7	33	76
8	39	66.5
9	32.5	75
10	21	38
:	:	:


Scatterplot of Final vs Midterm Score
120
100
80
60
40
20
0
Final Score (y)
0
5
10
15
20
25
30
35
40
45
50
Midterm Score (x)

Trends

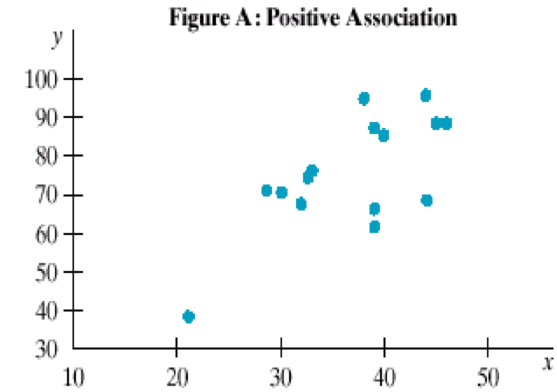


Figure A: Positive Association
y
100
90
80
70
60
50
40
30
10
20
30
40
50
x

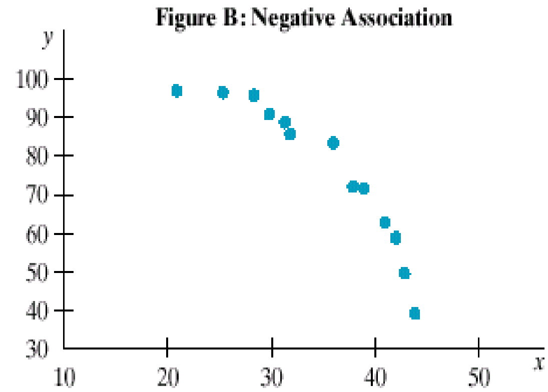


Figure B: Negative Association
y
100
90
80
70
60
50
40
30
10
20
30
40
50
x

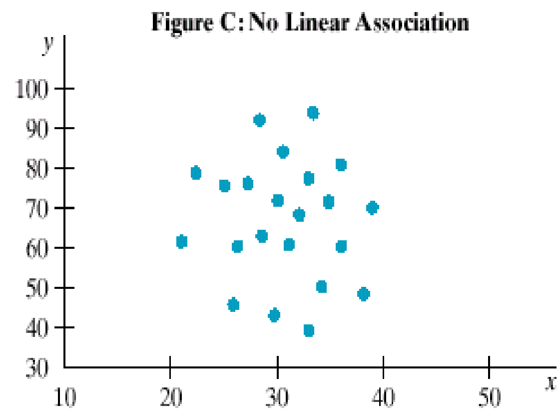


Figure C: No Linear Association
y
100
90
80
70
60
50
40
30
10
20
30
40
50
x

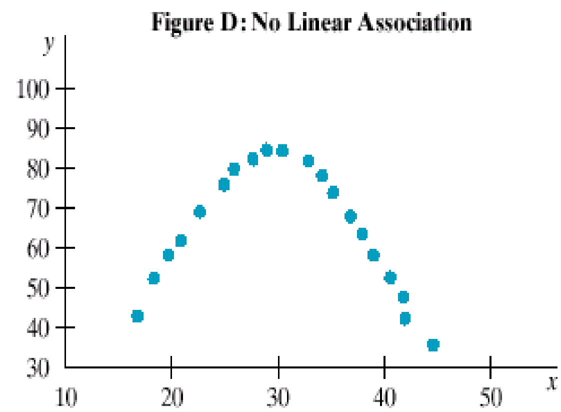


Figure D: No Linear Association
y
100
90
80
70
60
50
40
30
10
20
30
40
50
x

Let's do it! 7.3

What kind of relationship would you expect in the following situations:

1. age (in years) of a car, and its price.

2. number of calories consumed per day and weight.

3. height and IQ of a person.

Let's do it! 7.4

For each of the two descriptions given,

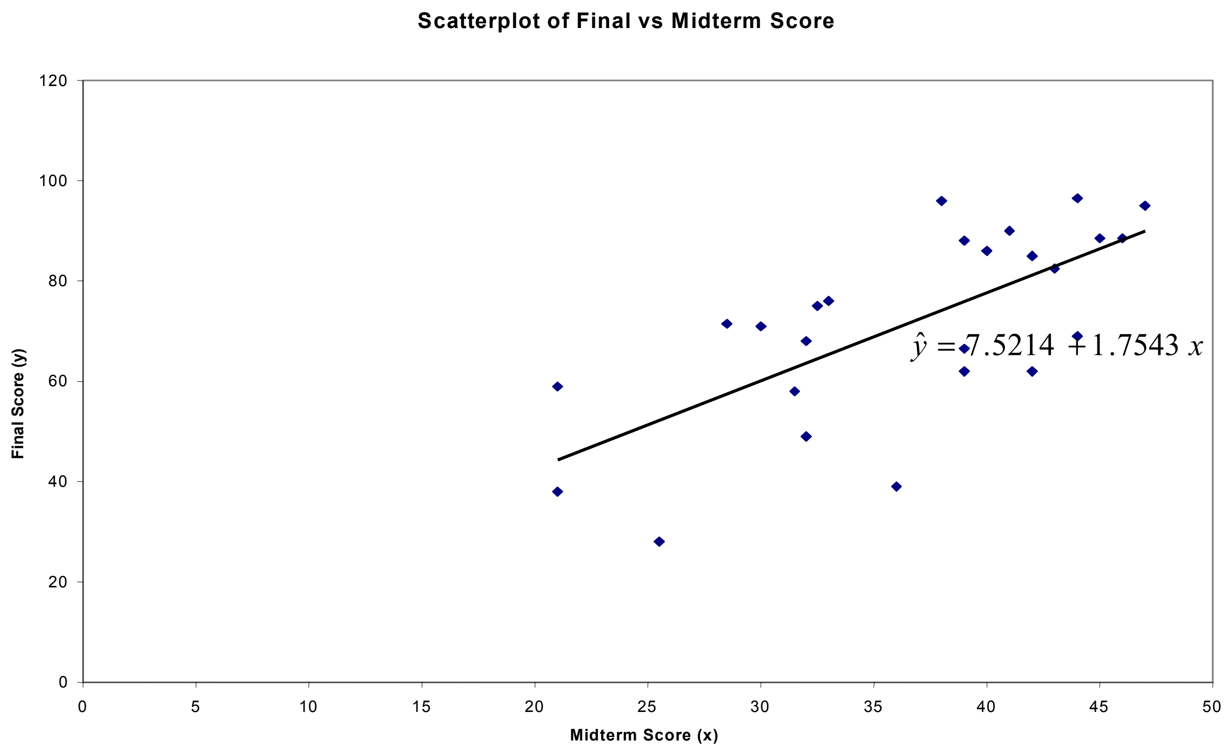
- Identify the two variables that vary and decide which should be the independent variable and which should be the dependent variable.
- Sketch a graph that you think best represents the relationship between the two variables
- Write a sentence that explains the shape and behavior shown in your graph.

1. The size of a persons vocabulary over his or her lifetime.

2. The distance from the ceiling to the tip of the minute hand of a clock hung on the wall.

Simple Linear Regression

Objective: To find the best fit to the data.

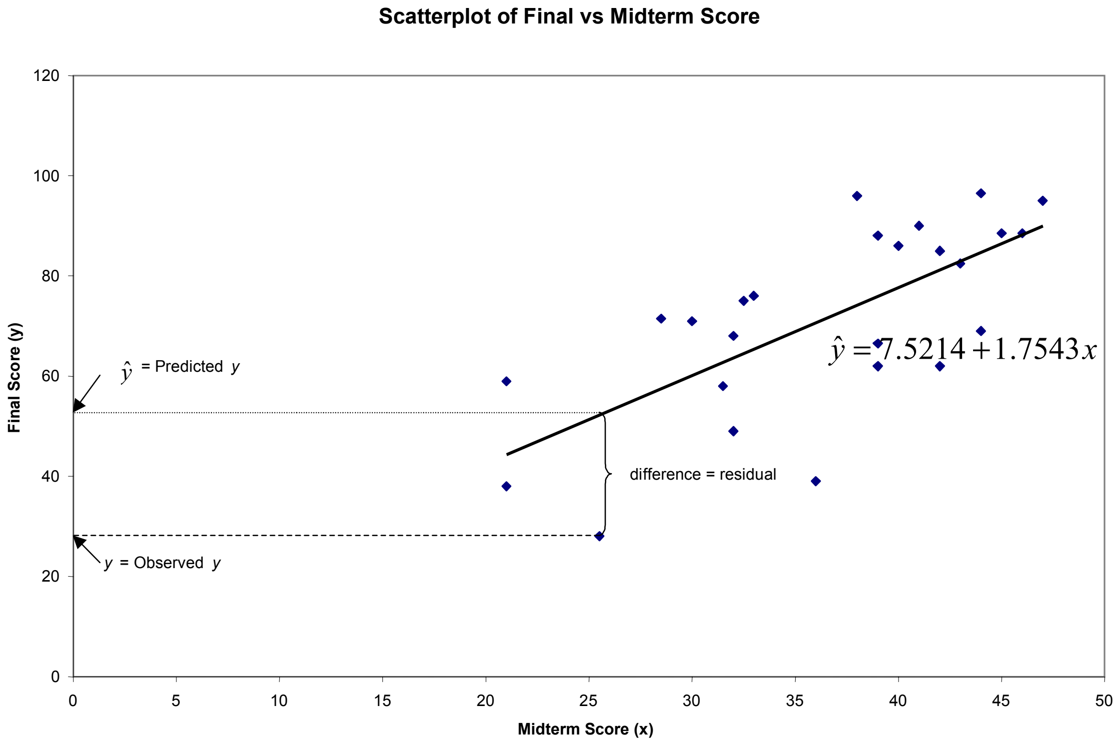


Equation of a line: $\hat{y} = a + bx$

where,

b = slope; the change in y for a unit change in x

a = y intercept; the value of y when $x = 0$



Scatterplot of Final vs Midterm Score
120
100
80
60
40
20
0
Final Score (y)
0
5
10
15
20
25
30
35
40
45
50
Midterm Score (x)
$\hat{y}$ = Predicted y
y = Observed y
difference = residual
$\hat{y} = 7.5214 + 1.7543x$

Definition *A **residual** is the difference between the observed response y and the predicted response $\widehat{y}$ (determined using the regression line). Thus, for each pair of observations (x_i, y_i), the i^{th} residual is*

$$e_i = y_i - \widehat{y}_i = y_i - (a + bx)$$

Definition *The **least squares regression line**, given by $\widehat{y} = a + bx$, is the line that minimizes (makes as small as possible) the **sum of squared vertical deviations (i.e., the square of the residuals)** of the observed points from the line. This is often stated as regress y on x.*

Let's do it! 7.6

The growth of children from early childhood through adolescence generally follows a linear pattern. Data on the heights of female Americans during childhood, from four to nine years old, were compiled and the least squares regression line was obtained as $\widehat{y} = 32 + 2.4x$ where $\widehat{y}$ is the predicted height in inches, and x is age in years.

- Interpret the value of the estimated slope $b = 2.4$.

- Would interpretation of the value of the estimated y-intercept, $a = 32$, make sense here?

- What would you predict the height to be for a female American at 8 years old?

- What would you predict the height to be for a female American at 25 years old? How does the quality of this answer compare to the previous question?

Calculating the Least Squares Regression Line

Find a solution to

$$\min \sum_i (y_i - \widehat{y_i})^2$$

i.e.,

$$\min \sum_i e_i^2$$

where,

$$\widehat{y_i} = a + bx,$$
$$e_i = y_i - \widehat{y_i}$$

The solution is

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2},$$

$$a = \overline{y} - b\overline{x}$$

Student	x_i	y_i	x_i^2	$x_i y_i$
1	39	62	1521	2418
2	44	69		
3	32	68		
4	40	86		
5	45	88.5		
Total	$\sum_i x_i =$	$\sum_i y_i =$	$\sum_i x_i^2 =$	$\sum_i x_i y_i =$

$\sum_i x_i =$

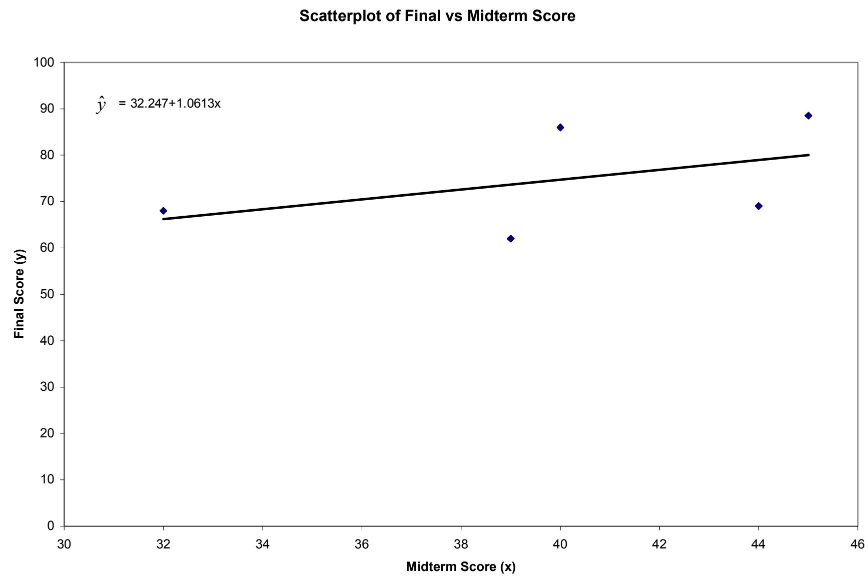
$\sum_i y_i =$

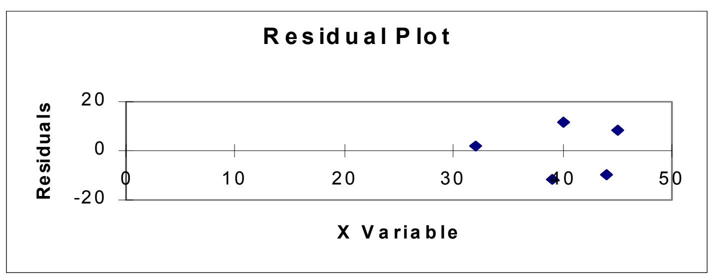
$\sum_i x_i^2 =$

$\sum_i x_i y_i =$

$$b = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2} =$$

$$a = \overline{y} - b\overline{x} =$$


Scatterplot of Final vs Midterm Score
$\hat{y}$ = 32.247+1.0613x
Final Score (y)
Midterm Score (x)
100
90
80
70
60
50
40
30
20
10
0
30
32
34
36
38
40
42
44
46


Residual Plot
Residuals
20
0
-20
0
10
20
30
40
50
X Variable

Statistically Significant?

Consider the results of regressing the midterm scores (x) against the final score (y) for the earlier example with 25 students.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.691[a]	.478	.455	14.0211

a. Predictors: (Constant), X

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	7.521	14.235		.528	.602	-21.926	36.969
	X	1.754	.383	.691	4.586	.000	.963	2.546

a. Dependent Variable: Y

Let's do it! 7.9

There are many factors that affect the selling price of a home. The total dwelling size and the assessed value are just two factors. Data were gathered on homes in a Milwaukee, Wisconsin, neighborhood. A scatterplot revealed a linear relationship between the total dwelling size of a home in 100 square feet and its selling price in dollars. The following is the regression output for the least squares regression of selling price on total dwelling size.

Dep var: PRICE N: 20 Multiple R: .913 Squared multiple R: .834
Adjusted squared multiple R: .825 Standard error of estimate: 3377.192

Variable	Coefficient	Std error	Std coef	Tolerance	T	P(2 tail)
CONSTANT	11947.010	4748.133	0.000	.	2.516	0.022
SIZE	2749.622	288.980	0.913	.100E+01	9.515	0.000

Analysis of Variance

Source	Sum-of-squares	DF	Mean-square	F-ratio	P
Regression	.103257E+10	1	.103257E+10	90.533	0.000
Residual	.205298E+09	18	.114054E+08		

1. How many homes were included in this study?

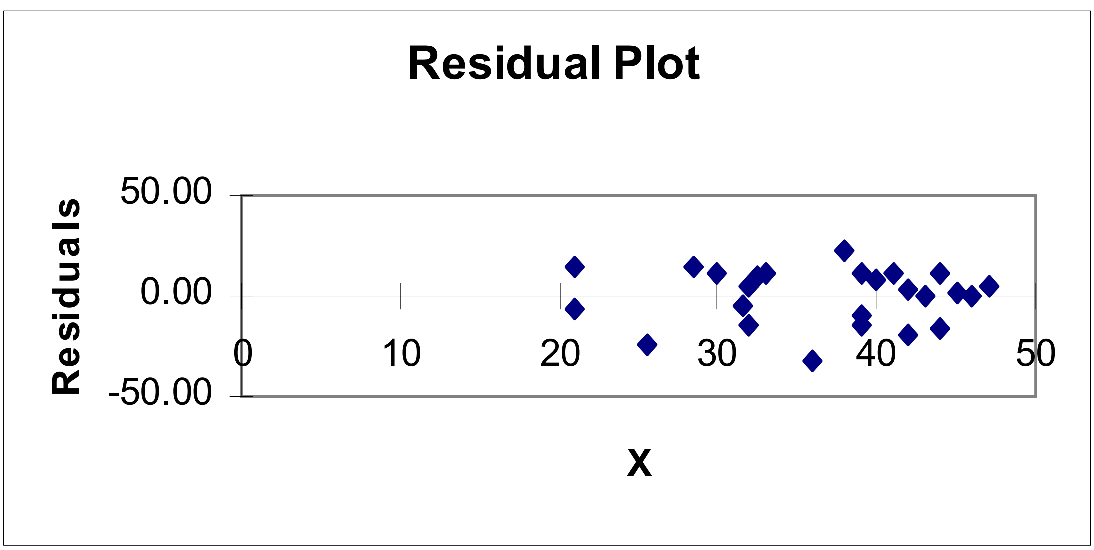
2. Obtain the least squares regression line for predicting selling price from the size of the home.

3. Is there evidence of a significant linear relationship between price and size?

4. The total dwelling size for another home in this neighborhood is 1620 square feet. Use the least squares line to estimate the selling price of this home.

Residual Analysis

Student Number	Midterm	Final	Predicted Y	Residuals
1	39	62	75.94	-13.94
2	44	69	84.71	-15.71
3	32	68	63.66	4.34
4	40	86	77.70	8.30
5	45	88.5	86.47	2.03
6	46	88.5	88.22	0.28
7	33	76	65.41	10.59
8	39	66.5	75.94	-9.44
9	32.5	75	64.54	10.46
10	21	38	44.36	-6.36
11	30	71	60.15	10.85
12	39	88	75.94	12.06
13	44	96.5	84.71	11.79
14	28.5	71.5	57.52	13.98
15	38	96	74.19	21.81
16	43	82.5	82.96	-0.46
17	42	85	81.20	3.80
18	25.5	28	52.26	-24.26
19	47	95	89.98	5.02
20	36	39	70.68	-31.68
21	31.5	58	62.78	-4.78
22	32	49	63.66	-14.66
23	42	62	81.20	-19.20
24	21	59	44.36	14.64
25	41	90	79.45	10.55



Residual Analysis (continued)

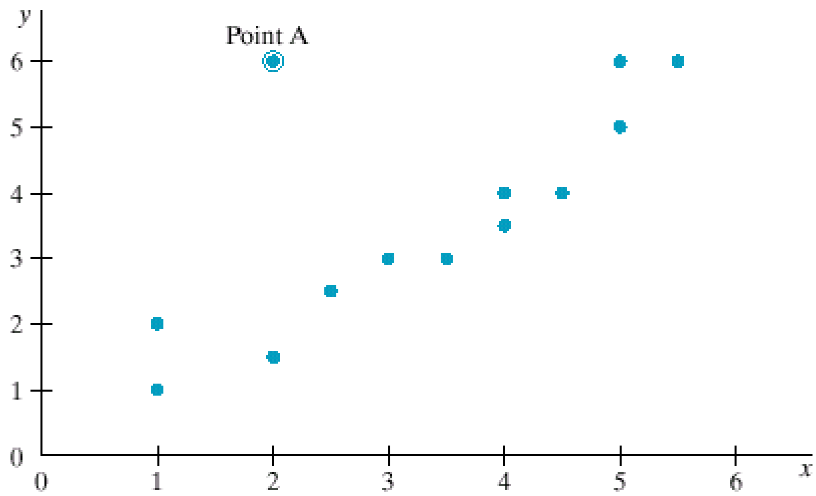
Steps to take

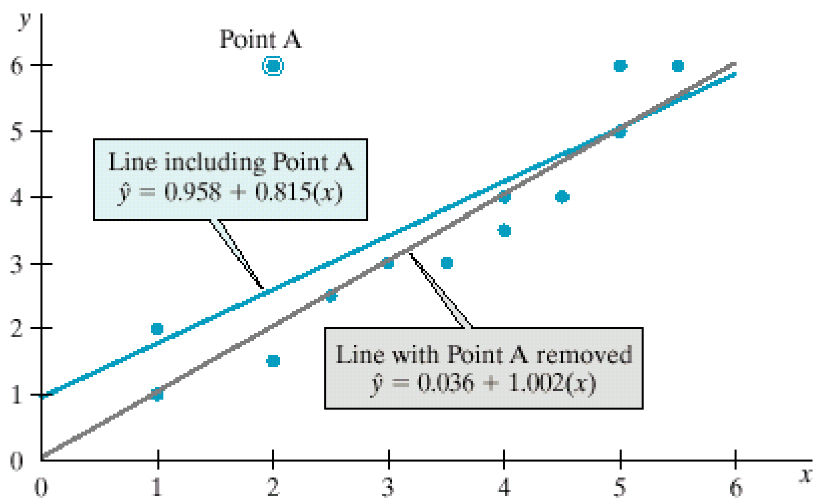
1. Plot the residuals e_i versus the predicted values $\hat{y}_i$.
    - If you see **NO** systematic patterns then the regression model appears appropriate for the data. Systematic patterns indicate that the model may **not** be appropriate for the data.
2. Plot the residuals e_i versus the x_i.
    - Again, if you see systematic patterns, the linear regression model may **not** be appropriate for the data.

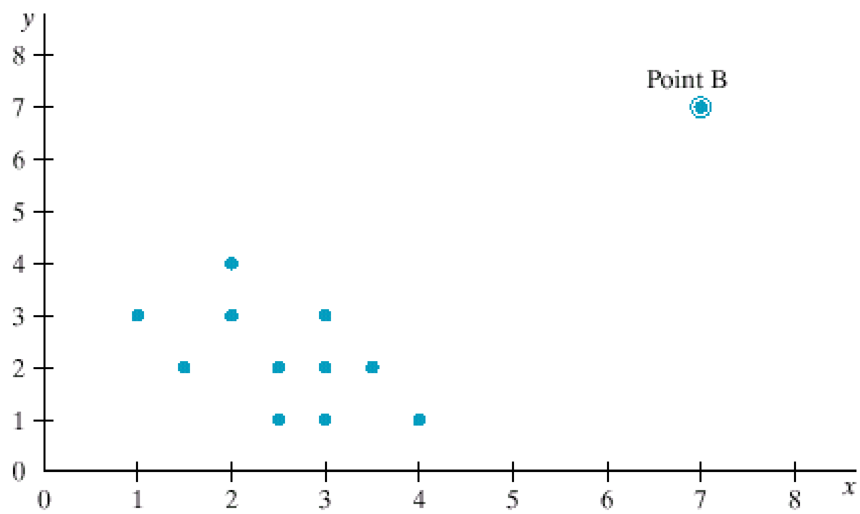
Influential Points and Outliers

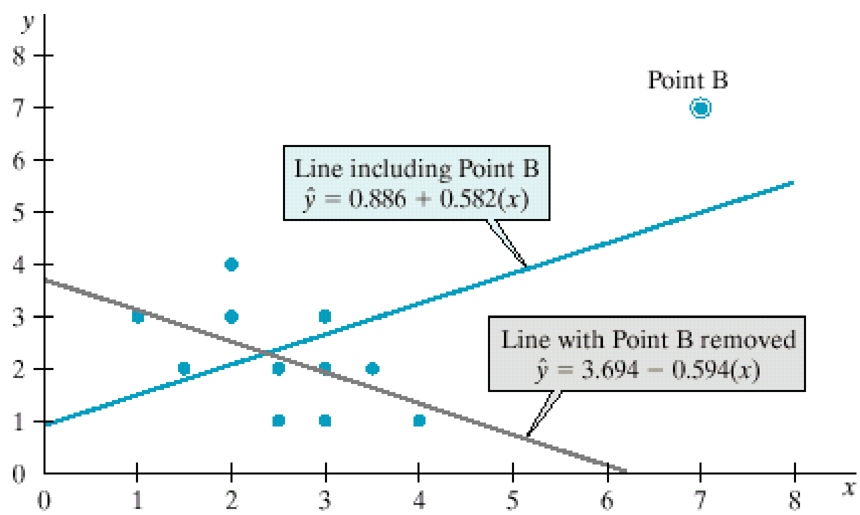
Definition *An **outlier** in regression is an observation with a residual that is unusually large (positive or negative) as compared to the other residuals.*

Definition *An **influential point** in regression is an observation that has a great deal of influence in determining the regression equation. Removing such a point would markedly change the position of the regression line. Observations that are somewhat extreme for the value of x are often influential.*


Point A


Point A
Line including Point A
ŷ = 0.958 + 0.815(x)
Line with Point A removed
ŷ = 0.036 + 1.002(x)


y
8
7
6
5
4
3
2
1
0
Point B
0
1
2
3
4
5
6
7
8
x


y
8
7
6
5
4
3
2
1
0
Point B
Line including Point B
$\hat{y} = 0.886 + 0.582(x)$
Line with Point B removed
$\hat{y} = 3.694 - 0.594(x)$
0
1
2
3
4
5
6
7
8
x

Correlation

Definition ***Correlation*** *measures the strength of* ***linear*** *relationship between two variables. It is usually denoted by r.*

Properties

Range $-1 \le r \le 1$.

Sign The sign indicates direction of association — Negative $[-1, 0)$ or positive $(0, 1]$.

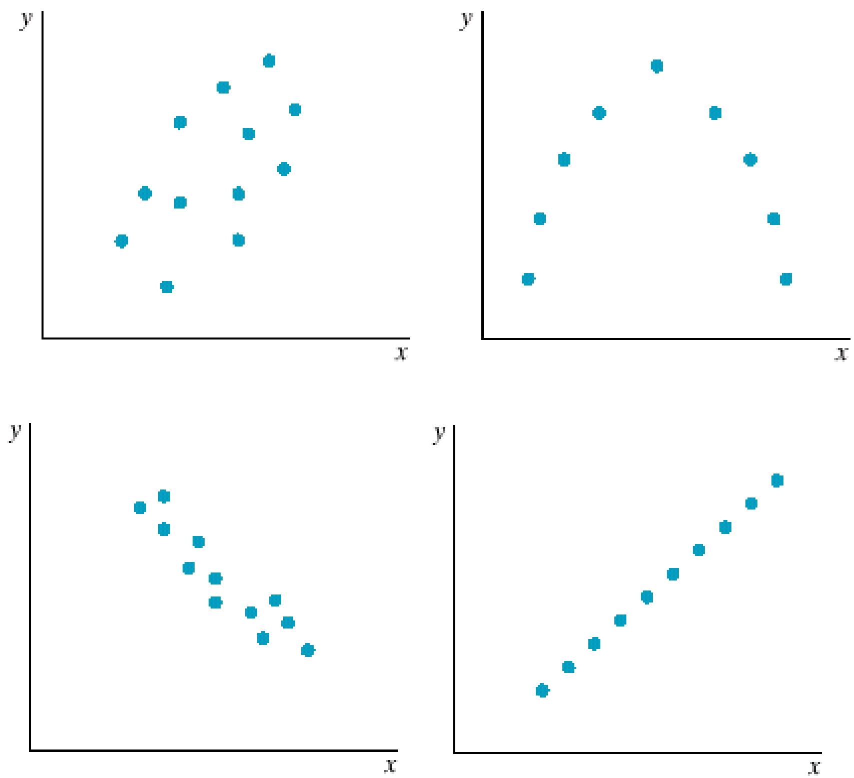
Magnitude The magnitude indicates the strength of the relationship. A $r = \pm 1$ indicates a straight line, while $r = 0$ indicates no linear relationship.

Units None.

Let's do it! 7.13

Match the following correlation values to the graphs.

r	0	1.0	–1.0	0.6	–0.2	–0.8	0.1



Two Qualitative Variables

Relationships between two qualitative variables are usually described using frequency or contingency tables.

Example *Consider the following table:*

	Nutritional Status			
Academic Performance	Poor	Adequate	Excellent	Total
Below Average	70	95	35	200
Average	130	450	30	610
Above Average	90	30	70	190
Total	290	575	135	1000

Questions:

Marginal and Conditional Distributions

Definition *The **marginal** distribution is found by computing the percentage of each row or column total based on the grand total.*

Definition *The **conditional** distribution of the row variable given the column variable is found by expressing the entries in the original table as percentages of the column total. Similarly, the conditional distribution of the column variable given the row variable is found by expressing the entries in the original table as percentages of the row total.*

Is There a Significant Relationship?

Consider the above example.

Crosstabs

Nutritional Status

Academic Performance

Count	1	2	3	
1	70	95	35	200
2	130	450	30	610
3	90	30	70	190
	290	575	135	1000

Tests

Source	DF	-LogLikelihood	RSquare (U)
Model	4	121.54736	0.1295
Error	994	817.39991	
C Total	998	938.94727	
Total Count	1000		

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	243.095	<.0001
Pearson	238.405	<.0001

Kappa	Std Err
0.275106	0.024136

Kappa measures the degree of agreement.