Chapter 12

Linear Regression and Correlation

12.1 Linear Regression and Correlation[1]

12.1.1 Student Learning Objectives

By the end of this chapter, the student should be able to:

- Discuss basic ideas of linear regression and correlation.
- Create and interpret a line of best fit.
- Calculate and interpret the correlation coefficient.
- Calculate and interpret outliers.

12.1.2 Introduction

Professionals often want to know how two or more variables are related. For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is it and how strong is the relationship?

In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee. These are all examples in which regression can be used.

The type of data described in the examples is **bivariate** data - "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

12.2 Linear Equations[2]

Linear regression for two variables is based on a linear equation with one independent variable. It has the form:

$$y = a + bx \tag{12.1}$$

[1]This content is available online at <http://cnx.org/content/m17089/1.5/>.
[2]This content is available online at <http://cnx.org/content/m17086/1.4/>.

where a and b are constant numbers.

x is the independent variable, and y is the dependent variable. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

Example 12.1
The following examples are linear equations.

$$y = 3 + 2x \tag{12.2}$$

$$y = -0.01 + 1.2x \tag{12.3}$$

The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation.

Example 12.2

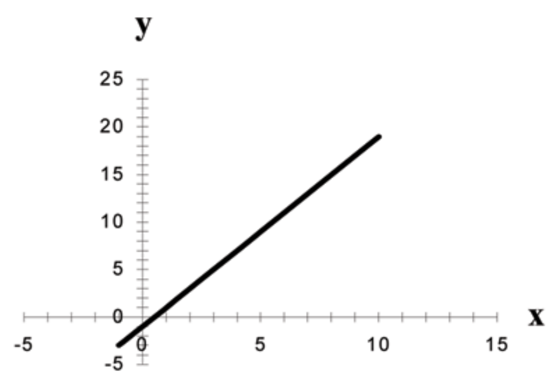


Figure 12.1: Graph of the equation $y = -1 + 2x$.

Linear equations of this form occur in applications of life sciences, social sciences, psychology, business, economics, physical sciences, mathematics, and other areas.

Example 12.3
Aaron's Word Processing Service (AWPS) does word processing. Its rate is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to do the word processing job.

Problem
Find the equation that expresses the **total cost** in terms of the **number of hours** required to finish the word processing job.

Solution
Let x = the number of hours it takes to get the job done.

Let y = the total cost to the customer.

The $31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is:

$y = 31.50 + 32x$

12.3 Slope and Y-Intercept of a Linear Equation[3]

For the linear equation $y = a + bx$, b = slope and a = y-intercept.

From algebra recall that the slope is a number that describes the steepness of a line and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis.

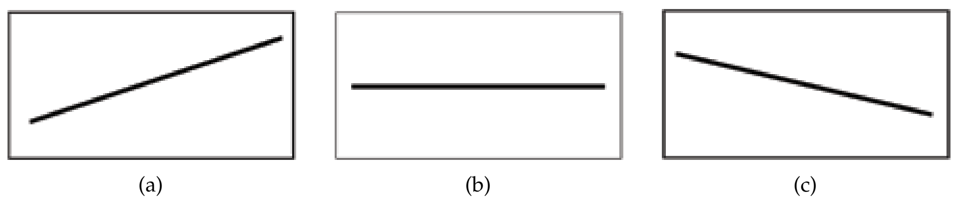


Figure 12.2: Three possible graphs of $y = a + bx$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Example 12.4
Svetlana tutors to make extra money for college. For each tutoring session, she charges a one time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

Problem
What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

Solution
The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 (a = 25). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when x = 0). The slope is 15 (b = 15). For each session, Svetlana earns $15 for each hour she tutors.

12.4 Scatter Plots[4]

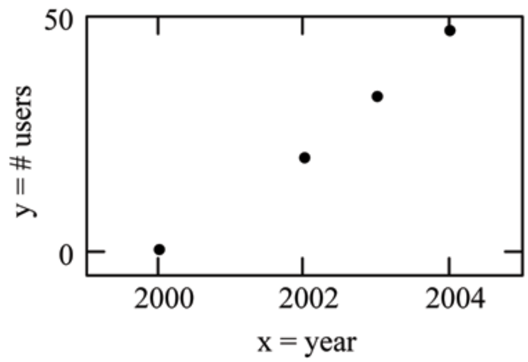
Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y. The most common and easiest way is a **scatter plot**. The following example illustrates a scatter plot.

[3] This content is available online at <http://cnx.org/content/m17083/1.5/>.
[4] This content is available online at <http://cnx.org/content/m17082/1.6/>.

Example 12.5
From an article in the *Wall Street Journal*: In Europe and Asia, m-commerce is becoming more popular. M-commerce users have special mobile phones that work like electronic wallets as well as provide phone and Internet services. Users can do everything from paying for parking to buying a TV set or soda from a machine to banking to checking sports scores on the Internet. In the next few years, will there be a relationship between the year and the number of m-commerce users? Construct a scatter plot. Let x = the year and let y = the number of m-commerce users, in millions.

x **(year)**	y **(# of users)**
2000	0.5
2002	20.0
2003	33.0
2004	47.0

(a)



(b)

Figure 12.3: (a) Table showing the number of m-commerce users (in millions) by year. (b) Scatter plot showing the number of m-commerce users (in millions) by year.

A scatter plot shows the **direction** and **strength** of a relationship between the variables. A clear direction happens when there is either:

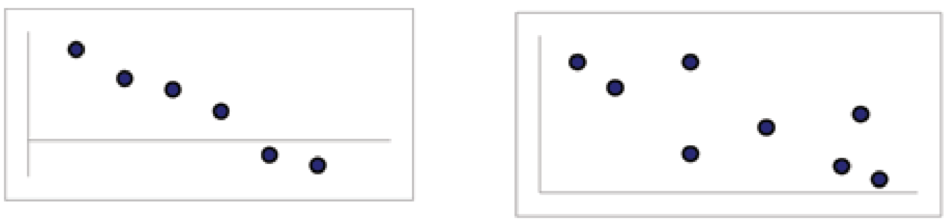
- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.
- High values of one variable occurring with low values of the other variable.

You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function.

When you look at a scatterplot, you want to notice the **overall pattern** and any **deviations** from the pattern. The following scatterplot examples illustrate these concepts.

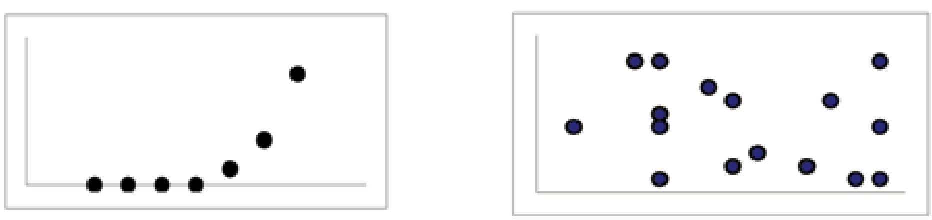
(a) Positive Linear Pattern (Strong) (b) Linear Pattern w/ One Deviation

Figure 12.4



(a) Negative Linear Pattern (Strong) (b) Negative Linear Pattern (Weak)

Figure 12.5



(a) Exponential Growth Pattern (b) No Pattern

Figure 12.6

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common. The linear relationship is strong if the points are close to a straight line. If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**. However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x.

12.5 The Regression Equation[5]

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to **"fit"** a straight line. This is called a **Line of Best Fit or Least Squares Line**.

12.5.1 Optional Collaborative Classroom Activity

If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, x, is pinky finger length and the dependent variable, y, is height.

For each set of data, plot the points on graph paper. Make your graph big enough and **use a ruler**. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the y-intercept of the line by extending your lines so they cross the y-axis. Using the slopes and the y-intercepts, write your equation of "best fit". Do you think everyone will have the same equation? Why or why not?

Using your equation, what is the predicted height for a pinky length of 2.5 inches?

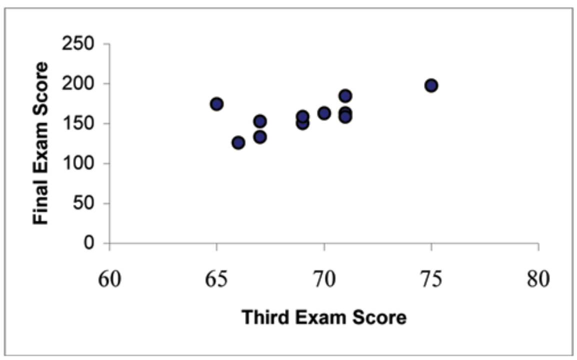
Example 12.6

A random sample of 11 statistics students produced the following data where x is the third exam score, out of 80, and y is the final exam score, out of 200. Can you predict the final exam score of a random student if you know the third exam score?

[5] This content is available online at <http://cnx.org/content/m17090/1.8/>.

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

(a)



(b)

Figure 12.7: (a) Table showing the scores on the final exam based on scores from the third exam. (b) Scatter plot showing the scores on the final exam based on scores from the third exam.

The third exam score, x, is the independent variable and the final exam score, y, is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye", you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

Consider the following diagram. Each point of data is of the the form $(x, y)$and each point of the line of best fit using least-squares linear regression has the form $\left(x, \hat{y}\right)$.

The $\hat{y}$ is read **"y hat"** and is the **estimated value of** y. It is the value of y obtained using the regression line. It is not generally equal to y from data.

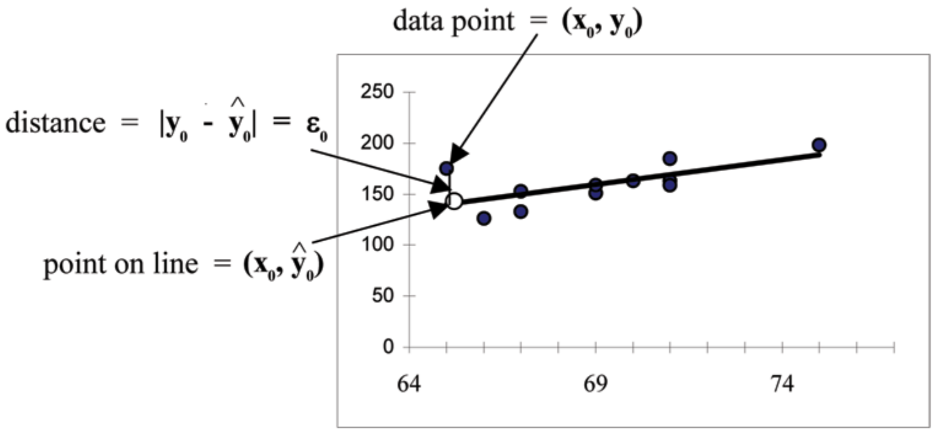


Figure 12.8

The term $|y_0 - \hat{y}_0| = \epsilon_0$ is called the **"error" or residual**. It is not an error in the sense of a mistake, but measures the vertical distance between the actual value of y and the estimated value of y.

ϵ = the Greek letter **epsilon**

For each data point, you can calculate, $|y_i - \hat{y}_i| = \epsilon_i$ for $i = 1, 2, 3, ..., 11$.

Each ϵ is a vertical distance.

For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 ϵ values. If you square each ϵ and add, you get

$$(\epsilon_1)^2 + (\epsilon_2)^2 + ... + (\epsilon_{11})^2 = \sum_{i=1}^{11} \epsilon^2$$

This is called the **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the values of a and b that make the **SSE** a minimum. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:

$$\hat{y} = a + bx \tag{12.4}$$

where $a = \overline{y} - b \cdot \overline{x}$ and $b = \frac{\Sigma(x-\overline{x})\cdot(y-\overline{y})^{2}}{\Sigma(x-\overline{x})}$.

$\overline{x}$ and $\overline{y}$ are the averages of the x values and the y values, respectively. The best fit line always passes through the point $(\overline{x}, \overline{y})$.

The slope b can be written as $b = r \cdot \left(\frac{s_y}{s_x}\right)$ where s_y = the standard deviation of the y values and s_x = the standard deviation of the x values. r is the correlation coefficient which is discussed in the next section.

NOTE: Many calculators or any linear regression and correlation computer program can calculate the best fit line. The calculations tend to be tedious if done by hand. **In the technology section, there are instructions for calculating the best fit line.**

The graph of the line of best fit for the third exam/final exam example is shown below:

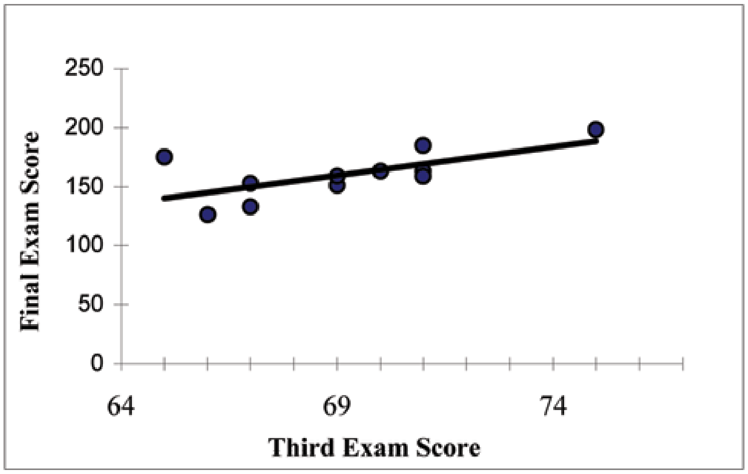


Figure 12.9

Remember, the best fit line is called the **least squares regression line** (it is sometimes referred to as the **LSL** which is an acronym for least squares line). The best fit line for the third exam/final exam example has the equation:

$$\hat{y} = -173.51 + 4.83x \tag{12.5}$$

The idea behind finding the best fit line is based on the assumption that the data are actually scattered about a straight line. Remember, it is always important to plot a scatter diagram first (which many calculators and computer programs can do) to see if it is worth calculating the line of best fit.

The slope of the line is 4.83 (b = 4.83). We can interpret the slope as follows: As the third exam score increases by one point, the final exam score increases by 4.83 points.

NOTE: If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for x-values outside that domain.**

12.6 The Correlation Coefficient[6]

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between x and y. The correlation coefficient, r, is defined as:

[6]This content is available online at <http://cnx.org/content/m17092/1.6/>.

$r = \frac{n \cdot \Sigma x \cdot y - (\Sigma x) \cdot (\Sigma y)}{\sqrt{\left[n \cdot \Sigma x^2 - (\Sigma x)^2\right] \cdot \left[n \cdot \Sigma y^2 - (\Sigma y)^2\right]}}$

where n = the number of data points.

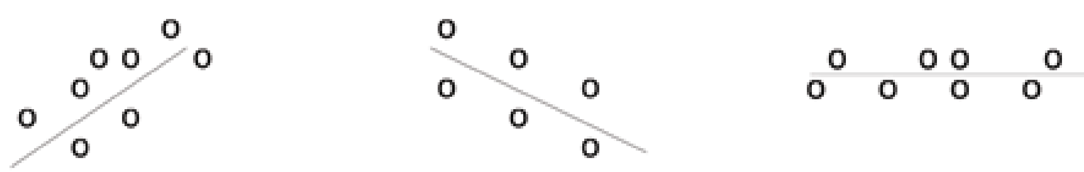
If you suspect a linear relationship between x and y, then r can measure how strong the linear relationship is.

One property of r is that $-1 \leq r \leq 1$. If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

The formula for r looks formidable. However, many calculators and any regression and correlation computer program can calculate r. The sign of r is the same as the slope, b, of the best fit line.

12.7 Facts About the Correlation Coefficient for Linear Regression[7]

- A positive r means that when x increases, y increases and when x decreases, y decreases **(positive correlation)**.
- A negative r means that when x increases, y decreases and when x decreases, y increases **(negative correlation)**.
- An r of zero means there is absolutely no linear relationship between x and y **(no correlation)**.
- High correlation does not suggest that x causes y or y causes x. We say **"correlation does not imply causation."** For example, every person who learned math in the 17th century is dead. However, learning math does not necessarily cause death!



(a) Positive Correlation (b) Negative Correlation (c) Zero Correlation

Figure 12.10: (a) A scatter plot showing data with a positive correlation. (b) A scatter plot showing data with a negative correlation. (c) A scatter plot showing data with zero correlation.

The 95% Critical Values of the Sample Correlation Coefficient Table (Section 12.10) at the end of this chapter (before the Summary (Section 12.11)) may be used to give you a good idea of whether the computed value of r **is significant or not**. Compare r to the appropriate critical value in the table. If r is significant, then you may want to use the line for prediction.

Example 12.7
Suppose you computed $r = 0.801$ using $n = 10$ data points. $df = n - 2 = 10 - 2 = 8$. The critical values associated with $df = 8$ are -0.632 and + 0.632. If $r <$ *negative critical value* or $r >$

[7] This content is available online at <http://cnx.org/content/m17077/1.7/>.

positive critical value, then r is significant. Since $r = 0.801$ and $0.801 > 0.632$, r is significant and the line may be used for prediction. If you view this example on a number line, it will help you.

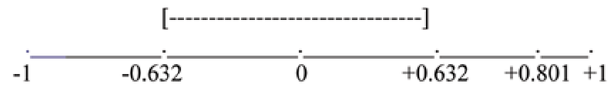


Figure 12.11: r is not significant between -0.632 and +0.632. $r = 0.801 > +0.632$. Therefore, r is significant.

Example 12.8

Suppose you computed $r = -0.624$ with 14 data points. $df = 14 - 2 = 12$. The critical values are -0.532 and 0.532. Since $-0.624 < -0.532$, r is significant and the line may be used for prediction

Figure 12.12: $r = -0.624 < -0.532$. Therefore, r is significant.

Example 12.9

Suppose you computed $r = 0.776$ and $n = 6$. $df = 6 - 2 = 4$. The critical values are -0.811 and 0.811. Since $-0.811 < 0.776 < 0.811$, r is not significant and the line should not be used for prediction.

Figure 12.13: $-0.811 < r = 0.776 < 0.811$. Therefore, r is not significant.

If $r = -1$ or $r = +1$, then all the data points lie exactly on a straight line.

If the line is significant, then **within the range of the x-values,** the line can be used to predict a y value.

As an illustration, consider the third exam/final exam example. The line of best fit is: $\hat{y} = -173.51 + 4.83x$ with $r = 0.6631$

Can the line be used for prediction? **Given a third exam score (x value), can we successfully predict the final exam score (predicted y value).** Test $r = 0.6631$ with its appropriate critical value.

Using the table with $df = 11 - 2 = 9$, the critical values are -0.602 and +0.602. Since $0.6631 > 0.602$, r is significant. **Because r is significant and the scatter plot shows a reasonable linear trend, the line can be used to predict final exam scores.**

Example 12.10
Suppose you computed the following correlation coefficients. Using the table at the end of the chapter, determine if r is significant and the line of best fit associated with each r can be used to predict a y value. If it helps, draw a number line.

- $r = -0.567$ and the sample size, n, is 19. The $df = n - 2 = 17$. The critical value is -0.456. $-0.567 < -0.456$ so r is significant.
- $r = 0.708$ and the sample size, n, is 9. The $df = n - 2 = 7$. The critical value is 0.666. $0.708 > 0.666$ so r is significant.
- $r = 0.134$ and the sample size, n, is 14. The $df = 14 - 2 = 12$. The critical value is 0.532. 0.134 is between -0.532 and 0.532 so r is not significant.
- $r = 0$ and the sample size, n, is 5. No matter what the dfs are, $r = 0$ is between the two critical values so r is not significant.

12.8 Prediction[8]

The exam scores **(x-values)** range from 65 to 75. Suppose you want to know the final exam score of statistics students who received 73 on the third exam. **Since 73 is between the x-values 65 and 75**, substitute $x = 73$ into the equation. Then:

$$\hat{y} = -173.51 + 4.83\,(73) = 179.08 \tag{12.7}$$

We predict that a statistics student who receives a 73 on the third exam will receive 179.08 on the final exam. **Remember, do not use the regression equation to predict values outside the domain of x.**

Example 12.11
Recall the third exam/final exam example.

Problem 1
What would you predict the final exam score to be for a student who scored a 66 on the third exam?

Solution
145.27

Problem 2 ***(Solution on p. 520.)***
What would you predict the final exam score to be for a student who scored a 78 on the third exam?

12.9 Outliers[9]

In some data sets, there are values **(points)** called **outliers**. **Outliers are points that are far from the least squares line.** They have large "errors." Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data. Other times, an outlier may hold valuable information about the population under study. The key is to carefully examine what causes a data point to be an outlier.

---

[8] This content is available online at <http://cnx.org/content/m17095/1.6/>.
[9] This content is available online at <http://cnx.org/content/m17094/1.7/>.

Example 12.12
In the third exam/final exam example, you can determine if there is an outlier or not. If there is one, as an exercise, delete it and fit the remaining data to a new line. For this example, the new line ought to fit the remaining data better. This means the **SSE** should be smaller and the correlation coefficient ought to be closer to 1 or -1.

Solution
Computers and many calculators can determine outliers from the data. However, as an exercise, we will go through the steps that are needed to calculate an outlier. In the table below, the first two columns are the third exam and the final exam data. The third column shows the y-hat values calculated from the line of best fit.

x	y	$\hat{y}$
65	175	140
67	133	150
71	185	169
71	163	169
66	126	145
75	198	189
67	153	150
70	163	164
71	159	169
69	151	160
69	159	160

Table 12.1

A **Residual** is the *Actual y value* $-$ *predicted y value* $= y - \hat{y}$

Calculate the absolute value of each residual.

Calculate each $|y - \hat{y}|$**:**

x	y	$\hat{y}$	$\lvert y-\hat{y}\rvert$
65	175	140	$\lvert 175-140\rvert = 35$
67	133	150	$\lvert 133-150\rvert = 17$
71	185	169	$\lvert 185-169\rvert = 16$
71	163	169	$\lvert 163-169\rvert = 6$
66	126	145	$\lvert 126-145\rvert = 19$
75	198	189	$\lvert 198-189\rvert = 9$
67	153	150	$\lvert 153-150\rvert = 3$
70	163	164	$\lvert 163-164\rvert = 1$
71	159	169	$\lvert 159-169\rvert = 10$
69	151	160	$\lvert 151-160\rvert = 9$
69	159	160	$\lvert 159-160\rvert = 1$

Table 12.2

Square each $\lvert y-\hat{y}\rvert$**:**

35^2; 17^2; 16^2; 6^2; 19^2; 9^2; 3^2; 1^2; 10^2; 9^2; 1^2

Then, add (sum) all the $\lvert y-\hat{y}\rvert$ **squared terms:**

$$\sum_{i=1}^{11}\left(\lvert y-\hat{y}\rvert\right)^2 = \sum_{i=1}^{11}\epsilon^2 \qquad \text{(Recall that } \lvert y_i-\hat{y}_i\rvert = \epsilon_i\text{.)}$$

$$= 35^2 + 17^2 + 16^2 + 6^2 + 19^2 + 9^2 + 3^2 + 1^2 + 10^2 + 9^2 + 1^2$$

$= 2440 =$ **SSE**

Next, calculate s**, the standard deviation of all the** $\lvert y-\hat{y}\rvert = \epsilon$ **values where** n **= the total number of data points.** (Calculate the standard deviation of 35; 17; 16; 6; 19; 9; 3; 1; 10; 9; 1.)

$$s = \sqrt{\frac{SSE}{n-2}}$$

For the third exam/final exam problem, $s = \sqrt{\frac{2440}{11-2}} = 16.47$

Next, multiply s by 1.9 and get $(1.9)\cdot(16.47) = 31.29$ (the value 31.29 is almost 2 standard deviations away from the mean of the $\lvert y-\hat{y}\rvert$ values.)

NOTE: The number $1.9s$ is equal to **1.9 standard deviations**. It is a measure that is almost 2 standard deviations. If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance was equal to $1.9s$ or greater, then we would consider the data point to be "too far" from the line of best fit. We would call that point a **potential outlier**.

For the example, if any of the $|y-\hat{y}|$ values are **at least** 31.29, the corresponding (x, y) point (data point) is a potential outlier.

Mathematically, we say that if $|y-\hat{y}| \geq (1.9) \cdot (s)$, then the corresponding point is an outlier.

For the third exam/final exam problem, all the $|y-\hat{y}|$'s are less than 31.29 except for the first one which is 35.

$35 > 31.29$ That is, $|y-\hat{y}| \geq (1.9) \cdot (s)$

The point which corresponds to $|y-\hat{y}| = 35$ is $(65, 175)$. **Therefore, the point** $(65, 175)$ **is an outlier.** For this example, we will delete it. (Remember, we do not always delete an outlier.) The next step is to compute a new best-fit line using the 10 remaining points. The new line of best fit and the correlation coefficient are:

$\hat{y} = -355.19 + 7.39x$ and $r = 0.9121$

If you compare $r = 0.9121$ to its critical value 0.632, $0.9121 > 0.632$. Therefore, r is significant. In fact, $r = 0.9121$ is a better r than the original (0.6631) because $r = 0.9121$ is closer to 1. This means that the 10 points fit the line better. The line can better predict the final exam score given the third exam score.

Example 12.13
Using the new line of best fit (calculated with 10 points), what would a student who receives a 73 on the third exam expect to receive on the final exam?

Example 12.14
(*From The Consumer Price Indexes Web site*) The Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for consumer goods and services. The CPI affects nearly all Americans because of the many ways it is used. One of its biggest uses is as a measure of inflation. By providing information about price changes in the Nation's economy to government, business, and labor, the CPI helps them to make economic decisions. The President, Congress, and the Federal Reserve Board use the CPI's trends to formulate monetary and fiscal policies. In the following table, x is the year and y is the CPI.

Data:

x	y
1915	10.1
1926	17.7
1935	13.7
1940	14.7
1947	24.1
1952	26.5
1964	31.0
1969	36.7
1975	49.3
1979	72.6
1980	82.4
1986	109.6
1991	130.7
1999	166.6

Table 12.3

Problem

- Make a scatterplot of the data.
- Calculate the least squares line. Write the equation in the form $\hat{y} = a + bx$.
- Draw the line on the scatterplot.
- Find the correlation coefficient. Is it significant?
- What is the average CPI for the year 1990?

Solution

- Scatter plot and line of best fit.
- $\hat{y} = -3204 + 1.662x$ is the equation of the line of best fit.
- $r = 0.8694$
- The number of data points is $n = 14$. Use the 95% Critical Values of the Sample Correlation Coefficient table at the end of Chapter 12. $n - 2 = 12$. The corresponding critical value is 0.532. Since $0.8694 > 0.532$, r is significant.
- $\hat{y} = -3204 + 1.662\,(1990) = 103.4$ CPI

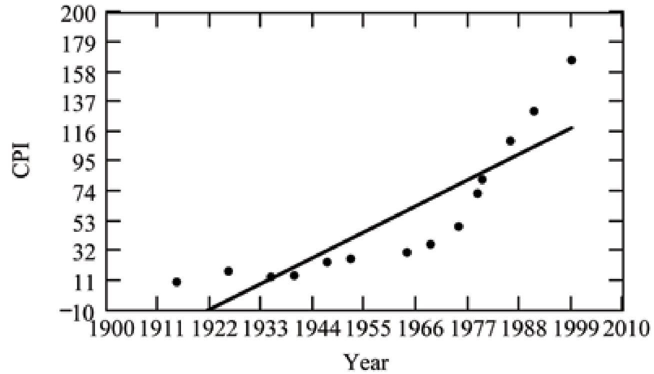


Figure 12.14

12.10 95% Critical Values of the Sample Correlation Coefficient Table[10]

Degrees of Freedom: $n-2$	Critical Values: $(+$ and $-)$
1	0.997
2	0.950
3	0.878
4	0.811
5	0.754
6	0.707
7	0.666
8	0.632
9	0.602
10	0.576
11	0.555
12	0.532
	continued on next page

[10]This content is available online at <http://cnx.org/content/m17098/1.5/>.

13	0.514
14	0.497
15	0.482
16	0.468
17	0.456
18	0.444
19	0.433
20	0.423
21	0.413
22	0.404
23	0.396
24	0.388
25	0.381
26	0.374
27	0.367
28	0.361
29	0.355
30	0.349
40	0.304
50	0.273
60	0.250
70	0.232
80	0.217
90	0.205
100 and over	0.195

Table 12.4

12.11 Summary[11]

Bivariate Data: Each data point has two values. The form is (x, y).

Line of Best Fit or Least Squares Line (LSL): $\hat{y} = a + bx$

x = independent variable; y = dependent variable

Residual: $Actual\ y\ value - predicted\ y\ value = y - \hat{y}$

Correlation Coefficient r:

1. Used to determine whether a line of best fit is good for prediction.
2. Between -1 and 1 inclusive. The closer r is to 1 or -1, the closer the original points are to a straight line.
3. If r is negative, the slope is negative. If r is positive, the slope is positive.
4. If $r = 0$, then the line is horizontal.

Sum of Squared Errors (SSE): The smaller the **SSE**, the better the original set of points fits the line of best fit.

Outlier: A point that does not seem to fit the rest of the data.

[11]This content is available online at <http://cnx.org/content/m17081/1.4/>.

12.12 Practice: Linear Regression[12]

12.12.1 Student Learning Outcomes

- The student will explore the properties of linear regression.

12.12.2 Given

The data below are real. Keep in mind that these are only reported figures. (*Source: Centers for Disease Control and Prevention, National Center for HIV, STD, and TB Prevention, October 24, 2003*)

Adults and Adolescents only, United States

Year	**# AIDS cases diagnosed**	**# AIDS deaths**
Pre-1981	91	29
1981	319	121
1982	1,170	453
1983	3,076	1,482
1984	6,240	3,466
1985	11,776	6,878
1986	19,032	11,987
1987	28,564	16,162
1988	35,447	20,868
1989	42,674	27,591
1990	48,634	31,335
1991	59,660	36,560
1992	78,530	41,055
1993	78,834	44,730
1994	71,874	49,095
1995	68,505	49,456
1996	59,347	38,510
1997	47,149	20,736
1998	38,393	19,005
1999	25,174	18,454
2000	25,522	17,347
2001	25,643	17,402
2002	26,464	16,371
Total	**802,118**	**489,093**

Table 12.5

[12]This content is available online at <http://cnx.org/content/m17088/1.8/>.

NOTE: We will use the columns "year" and "# AIDS cases diagnosed" for all questions unless otherwise stated.

12.12.3 Graphing

Graph "year" vs. "# AIDS cases diagnosed." **Plot the points on the graph located below in the section titled "Plot"** . Do not include pre-1981. Label both axes with words. Scale both axes.

12.12.4 Data

Exercise 12.12.1
Enter your data into your calculator or computer. The pre-1981 data should not be included. Why is that so?

12.12.5 Linear Equation

Write the linear equation below, rounding to 4 decimal places:

Exercise 12.12.2 *(Solution on p. 520.)*
Calculate the following:

a. $a =$
b. $b =$
c. $corr. =$
d. $n =$(# of pairs)

Exercise 12.12.3 *(Solution on p. 520.)*
equation: $\hat{y}=$

12.12.6 Solve

Exercise 12.12.4 *(Solution on p. 520.)*
Solve.

a. When $x = 1985$, $\hat{y}=$
b. When $x = 1990$, $\hat{y}=$

12.12.7 Plot

Plot the 2 above points on the graph below. Then, connect the 2 points to form the regression line.

Obtain the graph on your calculator or computer.

12.12.8 Discussion Questions

Look at the graph above.

Exercise 12.12.5
Does the line seem to fit the data? Why or why not?

Exercise 12.12.6
Do you think a linear fit is best? Why or why not?

Exercise 12.12.7
Hand draw a smooth curve on the graph above that shows the flow of the data.

Exercise 12.12.8
What does the correlation imply about the relationship between time (years) and the number of diagnosed AIDS cases reported in the U.S.?

Exercise 12.12.9
Why is "year" the independent variable and "# AIDS cases diagnosed." the dependent variable (instead of the reverse)?

Exercise 12.12.10 *(Solution on p. 520.)*
Solve.

- **a.** When $x = 1970$, $\hat{y}$=:
- **b.** Why doesn't this answer make sense?

12.13 Homework[13]

Exercise 12.13.1 *(Solution on p. 520.)*

For each situation below, state the independent variable and the dependent variable.

- **a.** A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than all other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
- **b.** A study is done to determine if the weekly grocery bill changes based on the number of family members.
- **c.** Insurance companies base life insurance premiums partially on the age of the applicant.
- **d.** Utility bills vary according to power consumption.
- **e.** A study is done to determine if a higher education reduces the crime rate in a population.

Exercise 12.13.2

In 1990 the number of driver deaths per 100,000 for the different age groups was as follows (Source: *The National Highway Traffic Safety Administration's National Center for Statistics and Analysis*):

Age	Number of Driver Deaths per 100,000
15-24	28
25-39	15
40-69	10
70-79	15
80+	25

Table 12.6

- **a.** For each age group, pick the midpoint of the interval for the x value. (For the 80+ group, use 85.)
- **b.** Using "ages" as the independent variable and "Number of driver deaths per 100,000" as the dependent variable, make a scatter plot of the data.
- **c.** Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + bx$
- **d.** Find the correlation coefficient. Is it significant?
- **e.** Pick two ages and find the estimated fatality rates.
- **f.** Use the two points in (e) to plot the least squares line on your graph from (b).
- **g.** Based on the above data, is there a linear relationship between age of a driver and driver fatality rate?
- **h.** What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.13.3 *(Solution on p. 520.)*

The average number of people in a family that received welfare for various years is given below. (Source: *House Ways and Means Committee, Health and Human Services Department*)

[13] This content is available online at <http://cnx.org/content/m17085/1.8/>.

Year	Welfare family size
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

Table 12.7

a. Using "year" as the independent variable and "welfare family size" as the dependent variable, make a scatter plot of the data.
b. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
c. Find the correlation coefficient. Is it significant?
d. Pick two years between 1969 and 1991 and find the estimated welfare family sizes.
e. Use the two points in (d) to plot the least squares line on your graph from (b).
f. Based on the above data, is there a linear relationship between the year and the average number of people in a welfare family?
g. Using the least squares line, estimate the welfare family sizes for 1960 and 1995. Does the least squares line give an accurate estimate for those years? Explain why or why not.
h. Are there any outliers in the above data?
i. What is the estimated average welfare family size for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.
j. What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.13.4

Use the AIDS data from the practice for this section (Section 12.12.2: Given), but this time use the columns "year #" and "# new AIDS deaths in U.S." Answer all of the questions from the practice again, using the new columns.

Exercise 12.13.5 ***(Solution on p. 520.)***

The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level). (Source: *Microsoft Bookshelf*)

Height (in feet)	Stories
1050	57
428	28
362	26
529	40
790	60
401	22
380	38
1454	110
1127	100
700	46

Table 12.8

a. Using "stories" as the independent variable and "height" as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables?
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated heights for 32 stories and for 94 stories.
f. Use the two points in (e) to plot the least squares line on your graph from (b).
g. Based on the above data, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
h. Are there any outliers in the above data? If so, which point(s)?
i. What is the estimated height of a building with 6 stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
j. Based on the least squares line, adding an extra story adds about how many feet to a building?
k. What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.13.6

Below is the life expectancy for an individual born in the United States in certain years. (Source: *National Center for Health Statistics*)

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7

Table 12.9

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
f. Why aren't the answers to part (e) the values on the above chart that correspond to those years?
g. Use the two points in (e) to plot the least squares line on your graph from (b).
h. Based on the above data, is there a linear relationship between the year of birth and life expectancy?
i. Are there any outliers in the above data?
j. Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.
k. What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.13.7 ***(Solution on p. 521.)***

The percent of female wage and salary workers who are paid hourly rates is given below for the years 1979 - 1992. (Source: *Bureau of Labor Statistics, U.S. Dept. of Labor*)

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

Table 12.10

a. Using "year" as the independent variable and "percent" as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated percents for 1991 and 1988.
f. Use the two points in (e) to plot the least squares line on your graph from (b).
g. Based on the above data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
h. Are there any outliers in the above data?
i. What is the estimated percent for the year 2050? Does the least squares line give an accurate estimate for that year? Explain why or why not?
j. What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.13.8
The maximum discount value of the Entertainment® card for the "Fine Dining" section, Edition 10, for various pages is given below.

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

Table 12.11

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated maximum values for the restaurants on page 10 and on page 70.
f. Use the two points in (e) to plot the least squares line on your graph from (b).
g. Does it appear that the restaurants giving the maximum value are placed in the beginning of the "Fine Dining" section? How did you arrive at your answer?
h. Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
i. Is the least squares line valid for page 200? Why or why not?
j. What is the slope of the least squares (best-fit) line? Interpret the slope.

The next two questions refer to the following data: The cost of a leading liquid laundry detergent in different sizes is given below.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

Table 12.12

Exercise 12.13.9 ***(Solution on p. 521.)***

a. Using "size" as the independent variable and "cost" as the dependent variable, make a scatter plot.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?

c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. If the laundry detergent were sold in a 40 ounce size, find the estimated cost.
f. If the laundry detergent were sold in a 90 ounce size, find the estimated cost.
g. Use the two points in (e) and (f) to plot the least squares line on your graph from (a).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?
j. Is the least squares line valid for predicting what a 300 ounce size of the laundry detergent would cost? Why or why not?
k. What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.13.10

a. Complete the above table for the cost per ounce of the different sizes.
b. Using "Size" as the independent variable and "Cost per ounce" as the dependent variable, make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is it significant?
f. If the laundry detergent were sold in a 40 ounce size, find the estimated cost per ounce.
g. If the laundry detergent were sold in a 90 ounce size, find the estimated cost per ounce.
h. Use the two points in (f) and (g) to plot the least squares line on your graph from (b).
i. Does it appear that a line is the best way to fit the data? Why or why not?
j. Are there any outliers in the above data?
k. Is the least squares line valid for predicting what a 300 ounce size of the laundry detergent would cost per ounce? Why or why not?
l. What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.13.11 ***(Solution on p. 521.)***

According to flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	Approximate Probate Fees and Taxes ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

Table 12.13

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make a scatter plot of the data.

c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is it significant?
f. Find the estimated total cost for a net taxable estate of $1,000,000. Find the cost for $2,500,000.
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?
j. Based on the above, what would be the probate fees and taxes for an estate that does not have any assets?
k. What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.13.12
The following are advertised sale prices of color televisions at Anderson's.

Size (inches)	Sale Price ($)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

Table 12.14

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is it significant?
f. Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?
j. What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.13.13 ***(Solution on p. 521.)***
Below are the average heights for American boys. (Source: *Physician's Handbook, 1990*)

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

Table 12.15

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is it significant?
f. Find the estimated average height for a one year–old. Find the estimated average height for an eleven year–old.
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?
j. Use the least squares line to estimate the average height for a sixty–two year–old man. Do you think that your answer is reasonable? Why or why not?
k. What is the slope of the least squares (best-fit) line? Interpret the slope.

Exercise 12.13.14
The following chart gives the gold medal times for every other Summer Olympics for the women's 100 meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64

Table 12.16

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. Is the decrease in times significant?
f. Find the estimated gold medal time for 1932. Find the estimated time for 1984.
g. Why are the answers from (f) different from the chart values?
h. Use the two points in (f) to plot the least squares line on your graph from (b).
i. Does it appear that a line is the best way to fit the data? Why or why not?
j. Use the least squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?

The next three questions use the following state information.

State	**# letters in name**	**Year entered the Union**	**Rank for entering the Union**	**Area (square miles)**
Alabama	7	1819	22	52,423
Colorado		1876	38	104,100
Hawaii		1959	50	10,932
Iowa		1846	29	56,276
Maryland		1788	7	12,407
Missouri		1821	24	69,709
New Jersey		1787	3	8,722
Ohio		1803	17	44,828
South Carolina	13	1788	8	32,008
Utah		1896	45	84,904
Wisconsin		1848	30	65,499

Table 12.17

Exercise 12.13.15 ***(Solution on p. 521.)***

We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union.

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.

g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Use the least squares line to estimate the number of letters a new state that enters the Union this year would have. Can the least squares line be used to predict it? Why or why not?

Exercise 12.13.16

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

a. Let rank be the independent variable and area be the dependent variable.
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers?
j. Use the least squares line to estimate the area of a new state that enters the Union. Can the least squares line be used to predict it? Why or why not?
k. Delete "Hawaii" and substitute "Alaska" for it. Alaska is the fortieth state with an area of 656,424 square miles.
l. Calculate the new least squares line.
m. Find the estimated area for Alabama. Is it closer to the actual area with this new least squares line or with the previous one that included Hawaii? Why do you think that's the case?
n. Do you think that, in general, newer states are larger than the original states?

Exercise 12.13.17 ***(Solution on p. 522.)***

We are interested in whether there is a relationship between the rank of a state and the year it entered the Union.

a. Let year be the independent variable and rank be the dependent variable.
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Why must the relationship be positive between the variables?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Let's say a fifty-first state entered the union. Based upon the least squares line, when should that have occurred?
g. Using the least squares line, how many states do we currently have?
h. Why isn't the least squares line a good estimator for this year?

Exercise 12.13.18

Below are the percents of the U.S. labor force (excluding self-employed and unemployed) that are members of a union. We are interested in whether the decrease is significant. (Source: *Bureau of Labor Statistics, U.S. Dept. of Labor*)

Year	Percent
1945	35.5
1950	31.5
1960	31.4
1970	27.3
1980	21.9
1986	17.5
1993	15.8

Table 12.18

a. Let year be the independent variable and percent be the dependent variable.
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Why will the relationship between the variables be negative?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Based on your answer to (e), do you think that the relationship can be said to be decreasing?
g. If the trend continues, when will there no longer be any union members? Do you think that will happen?

The next two questions refer to the following information: The data below reflects the 1991-92 Reunion Class Giving. (Source: *SUNY Albany alumni magazine*)

Class Year	Average Gift	Total Giving
1922	41.67	125
1927	60.75	1,215
1932	83.82	3,772
1937	87.84	5,710
1947	88.27	6,003
1952	76.14	5,254
1957	52.29	4,393
1962	57.80	4,451
1972	42.68	18,093
1976	49.39	22,473
1981	46.87	20,997
1986	37.03	12,590

Table 12.19

Exercise 12.13.19 *(Solution on p. 522.)*

We will use the columns "class year" and "total giving" for all questions, unless otherwise stated.

a. What do you think the scatter plot will look like? Make a scatter plot of the data.
b. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
c. Find the correlation coefficient. What does it imply about the significance of the relationship?
d. For the class of 1930, predict the total class gift.
e. For the class of 1964, predict the total class gift.
f. For the class of 1850, predict the total class gift. Why doesn't this value make any sense?

Exercise 12.13.20
We will use the columns "class year" and "average gift" for all questions, unless otherwise stated.

a. What do you think the scatter plot will look like? Make a scatter plot of the data.
b. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
c. Find the correlation coefficient. What does it imply about the significance of the relationship?
d. For the class of 1930, predict the average class gift.
e. For the class of 1964, predict the average class gift.
f. For the class of 2010, predict the average class gift. Why doesn't this value make any sense?

12.13.1 Try these multiple choice questions

Exercise 12.13.21 ***(Solution on p. 522.)***
A correlation coefficient of -0.95 means there is a ____________ between the two variables.

A. Strong positive correlation
B. Weak negative correlation
C. Strong negative correlation
D. No Correlation

Exercise 12.13.22 ***(Solution on p. 522.)***
According to the data reported by the New York State Department of Health regarding West Nile Virus for the years 2000-2004, the least squares line equation for the number of reported dead birds (x) versus the number of human West Nile virus cases (y) is $\hat{y} = -10.2638 + 0.0491x$. If the number of dead birds reported in a year is 732, how many human cases of West Nile virus can be expected?

A. 25.7
B. 46.2
C. -25.7
D. 7513

The next three questions refer to the following data: (showing the number of hurricanes by category to directly strike the mainland U.S. each decade) obtained from *www.nhc.noaa.gov/gifs/table6.gif*[14] A major hurricane is one with a strength rating of 3, 4 or 5.

[14]http://www.nhc.noaa.gov/gifs/table6.gif

Decade	Total Number of Hurricanes	Number of Major Hurricanes
1941-1950	24	10
1951-1960	17	8
1961-1970	14	6
1971-1980	12	4
1981-1990	15	5
1991-2000	14	5
2001 – 2004	9	3

Table 12.20

Exercise 12.13.23 ***(Solution on p. 522.)***

Using only completed decades (1941 – 2000), calculate the least squares line for the number of major hurricanes expected based upon the total number of hurricanes.

A. $\hat{y} = -1.67x + 0.5$
B. $\hat{y} = 0.5x - 1.67$
C. $\hat{y} = 0.94x - 1.67$
D. $\hat{y} = -2x + 1$

Exercise 12.13.24 ***(Solution on p. 522.)***

The correlation coefficient is 0.942. Is this considered significant? Why or why not?

A. No, because 0.942 is greater than the critical value of 0.707
B. Yes, because 0.942 is greater than the critical value of 0.707
C. No, because 0942 is greater than the critical value of 0.811
D. Yes, because 0.942 is greater than the critical value of 0.811

Exercise 12.13.25 ***(Solution on p. 522.)***

The data for 2001-2004 show 9 hurricanes have hit the mainland United States. The line of best fit predicts 2.83 major hurricanes to hit mainland U.S. Can the least squares line be used to make this prediction?

A. No, because 9 lies outside the independent variable values
B. Yes, because, in fact, there have been 3 major hurricanes this decade
C. No, because 2.83 lies outside the dependent variable values
D. Yes, because how else could we predict what is going to happen this decade.

12.14 Lab 1: Regression (Distance from School)[15]

Class Time:

Names:

12.14.1 Student Learning Outcomes:

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

12.14.2 Collect the Data

Use 8 members of your class for the sample. Collect bivariate data (distance an individual lives from school, the cost of supplies for the current term).

1. Complete the table.

Distance from school	**Cost of supplies this term**

Table 12.21

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph "distance" vs. "cost." Plot the points on the graph. Label both axes with words. Scale both axes.

[15]This content is available online at <http://cnx.org/content/m17080/1.10/>.

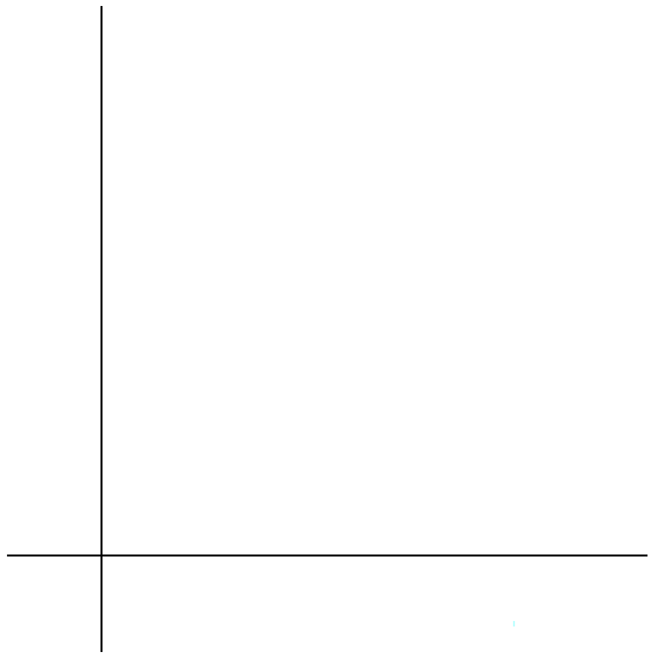


Figure 12.15

12.14.3 Analyze the Data

Enter your data into your calculator or computer. Write the linear equation below, rounding to 4 decimal places.

1. Calculate the following:
   a. a =
   b. b=
   c. correlation =
   d. n =
   e. equation: $\hat{y}$ =
   f. Is the correlation significant? Why or why not? (Answer in 1-3 complete sentences.)
2. Supply an answer for the following senarios:
   a. For a person who lives 8 miles from campus, predict the total cost of supplies this term:
   b. For a person who lives 80 miles from campus, predict the total cost of supplies this term:
3. Obtain the graph on your calculator or computer. Sketch the regression line below.

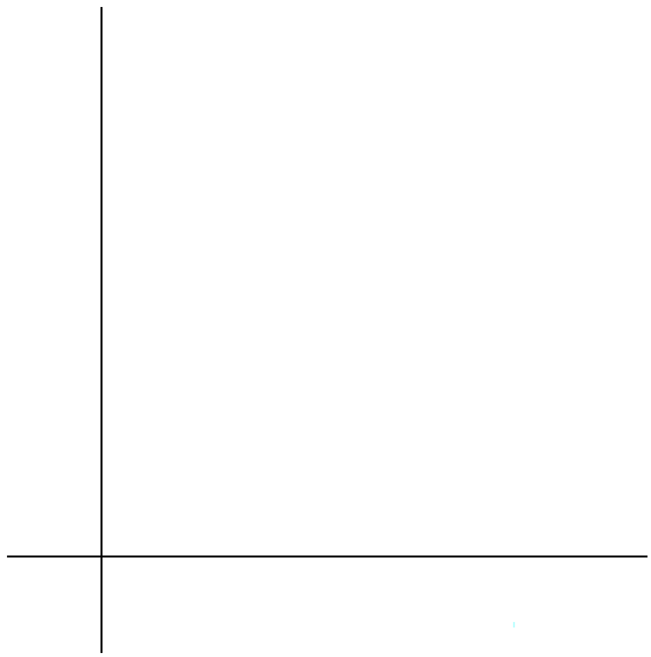


Figure 12.16

12.14.4 Discussion Questions

1. Answer each with 1-3 complete sentences.
   - **a.** Does the line seem to fit the data? Why?
   - **b.** What does the correlation imply about the relationship between the distance and the cost?
2. Are there any outliers? If so, which point is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

12.15 Lab 2: Regression (Textbook Cost)[16]

Class Time:

Names:

12.15.1 Student Learning Outcomes:

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

12.15.2 Collect the Data

Survey 10 textbooks. Collect bivariate data (number of pages in a textbook, the cost of the textbook).

1. Complete the table.

Number of pages	Cost of textbook

Table 12.22

2. Which variable should be the dependent variable and which should be the independent variable? Why?
3. Graph “distance” vs. “cost.” Plot the points on the graph in "Analyze the Data". Label both axes with words. Scale both axes.

12.15.3 Analyze the Data

Enter your data into your calculator or computer. Write the linear equation below, rounding to 4 decimal places.

1. Calculate the following:
   a. $a =$
   b. $b =$
   c. correlation =
   d. $n =$

[16]This content is available online at <http://cnx.org/content/m17087/1.9/>.

e. equation: y =
f. Is the correlation significant? Why or why not? (Answer in 1-3 complete sentences.)

2. Supply an answer for the following senarios:

   a. For a textbook with 400 pages, predict the cost:
   b. For a textbook with 600 pages, predict the cost:

3. Obtain the graph on your calculator or computer. Sketch the regression line below.

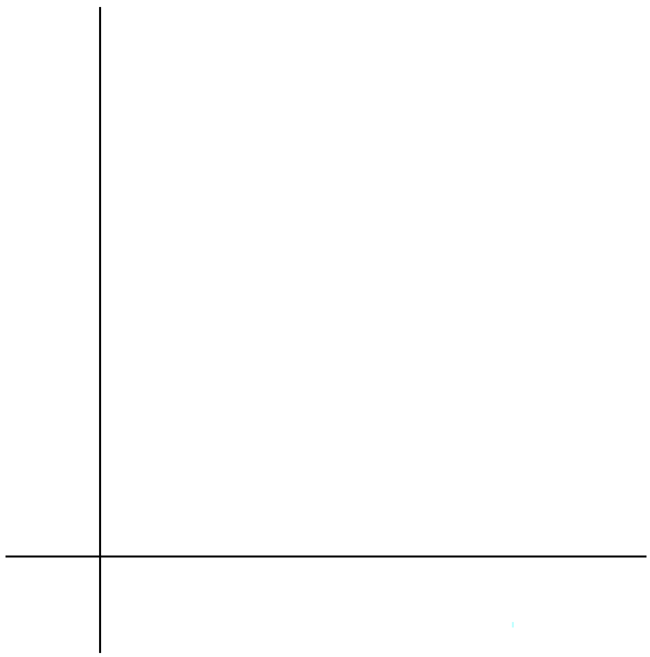


Figure 12.17

12.15.4 Discussion Questions

1. Answer each with 1-3 complete sentences.

   a. Does the line seem to fit the data? Why?
   b. What does the correlation imply about the relationship between the number of pages and the cost?

2. Are there any outliers? If so, which point(s) is an outlier?
3. Should the outlier, if it exists, be removed? Why or why not?

12.16 Lab 3: Regression (Fuel Efficiency)[17]

Class Time:

Names:

12.16.1 Student Learning Outcomes:

- The student will calculate and construct the line of best fit between two variables.
- The student will evaluate the relationship between two variables to determine if that relationship is significant.

12.16.2 Collect the Data

Use the most recent April issue of Consumer Reports. It will give the total fuel efficiency (in miles per gallon) and weight (in pounds) of new model cars with automatic transmissions. We will use this data to determine the relationship, if any, between the fuel efficiency of a car and its weight.

1. Which variable should be the independent variable and which should be the dependent variable? Explain your answer in one or two complete sentences.
2. Using your random number generator, randomly select 20 cars from the list and record their weights and fuel efficiency into the table below.

[17]This content is available online at <http://cnx.org/content/m17079/1.8/>.

Weight	Fuel Efficiency

Table 12.23

3. Which variable should be the dependent variable and which should be the independent variable? Why?
4. By hand, do a scatterplot of "weight" vs. "fuel efficiency". Plot the points on graph paper. Label both axes with words. Scale both axes accurately.

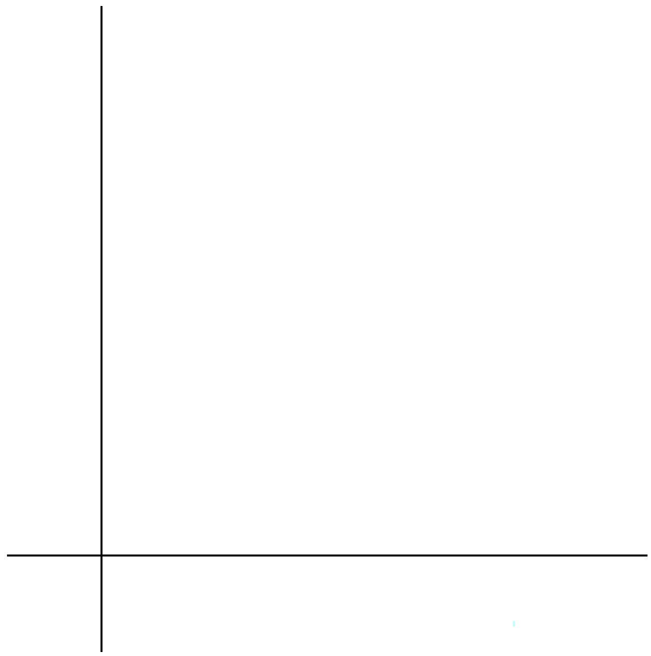


Figure 12.18

12.16.3 Analyze the Data

Enter your data into your calculator or computer. Write the linear equation below, rounding to 4 decimal places.

1. Calculate the following:
   - **a.** a =
   - **b.** b =
   - **c.** correlation =
   - **d.** n =
   - **e.** equation: $\hat{y}$ =
2. Obtain the graph of the regression line on your calculator. Sketch the regression line on the same axes as your scatterplot.

12.16.4 Discussion Questions

1. Is the correlation significant? Explain how you determined this in complete sentences.

2. Is the relationship a positive one or a negative one? Explain how you can tell and what this means in terms of weight and fuel efficiency.
3. In one or two complete sentences, what is the practical interpretation of the slope of the least squares line in terms of fuel efficiency and weight?
4. For a car that weighs 4000 pounds, predict its fuel efficiency. Include units.
5. Can we predict the fuel efficiency of a car that weighs 10000 pounds using the least squares line? Explain why or why not.
6. Questions. Answer each in 1 to 3 complete sentences.
   - **a.** Does the line seem to fit the data? Why or why not?
   - **b.** What does the correlation imply about the relationship between fuel efficiency and weight of a car? Is this what you expected?
7. Are there any outliers? If so, which point is an outlier?

** This lab was designed and contributed by Diane Mathios.

Solutions to Exercises in Chapter 12

Solution to Example 12.11, Problem 2 (p. 486)
78 is outside of the domain of x values (independent variables), so you cannot reliably predict the final exam score for this student.
Solution to Example 12.13 (p. 489)
184.28

Solutions to Practice: Linear Regression

Solution to Exercise 12.12.2 (p. 495)

- **a.** $a = -3{,}448{,}225$
- **b.** $b = 1750$
- **c.** $corr. = 0.4526$
- **d.** $n = 22$

Solution to Exercise 12.12.3 (p. 495)
$\hat{y} = -3{,}448{,}225 + 1750x$
Solution to Exercise 12.12.4 (p. 495)

- **a.** 25082
- **b.** 33,831

Solution to Exercise 12.12.10 (p. 496)

- **a.** -1164

Solutions to Homework

Solution to Exercise 12.13.1 (p. 497)

- **a.** Independent: Age; Dependent: Fatalities
- **d.** Independent: Power Consumption; Dependent: Utility

Solution to Exercise 12.13.3 (p. 497)

- **b.** $\hat{y} = 88.7206 - 0.0432x$
- **c.** -0.8533, Yes
- **g.** No
- **h.** No.
- **i.** 2.97, Yes
- **j.** slope = -0.0432. As the year increases by one, the welfare family size decreases by 0.0432 people.

Solution to Exercise 12.13.5 (p. 498)

- **b.** Yes
- **c.** $\hat{y} = 102.4287 + 11.7585x$
- **d.** 0.9436; yes
- **e.** 478.70 feet; 1207.73 feet
- **g.** Yes
- **h.** Yes; $(57, 1050)$
- **i.** 172.98; No
- **j.** 11.7585 feet

k. slope = 11.7585. As the number of stories increases by one, the height of the building increases by 11.7585 feet.

Solution to Exercise 12.13.7 (p. 500)

b. Yes
c. $\hat{y} = -266.8863 + 0.1656x$
d. 0.9448; Yes
e. 62.9206; 62.4237
h. No
i. 72.639; No
j. slope = 0.1656. As the year increases by one, the percent of workers paid hourly rates increases by 0.1565.

Solution to Exercise 12.13.9 (p. 502)

b. Yes
c. $\hat{y} = 3.5984 + 0.0371x$
d. 0.9986; Yes
e. $5.08
f. $6.93
i. No
j. Not valid
k. slope = 0.0371. As the number of ounces increases by one, the cost of the liquid detergent increases by $0.0371 (or about 4 cents).

Solution to Exercise 12.13.11 (p. 503)

c. Yes
d. $\hat{y} = -337,424.6478 + 0.5463x$
e. 0.9964; Yes
f. $208,872.49; $1,028,318.20
h. Yes
i. No
k. slope = 0.5463. As the net taxable estate increases by one dollar, the approximate probate fees and taxes increases by 0.5463 dollars (about 55 cents).

Solution to Exercise 12.13.13 (p. 504)

c. Yes
d. $\hat{y} = 65.0876 + 7.0948x$
e. 0.9761; yes
f. 72.2 cm; 143.13 cm
h. Yes
i. No
j. 505.0 cm; No
k. slope = 7.0948. As the age of an American boy increases by one year, the average height increases by 7.0948 cm.

Solution to Exercise 12.13.15 (p. 506)

c. No
d. $\hat{y} = 47.03 - 0.216x$

- **e.** -0.4280
- **f.** 6; 5

Solution to Exercise 12.13.17 (p. 507)

- **d.** $\hat{y} = -480.5845 + 0.2748x$
- **e.** 0.9553
- **f.** 1934

Solution to Exercise 12.13.19 (p. 508)

- **b.** $\hat{y} = -569{,}770.2796 + 296.0351$
- **c.** 0.8302
- **d.** $1577.48
- **e.** $11,642.68
- **f.** -$22,105.33

Solution to Exercise 12.13.21 (p. 509)
C
Solution to Exercise 12.13.22 (p. 509)
A
Solution to Exercise 12.13.23 (p. 510)
A
Solution to Exercise 12.13.24 (p. 510)
D
Solution to Exercise 12.13.25 (p. 510)
A