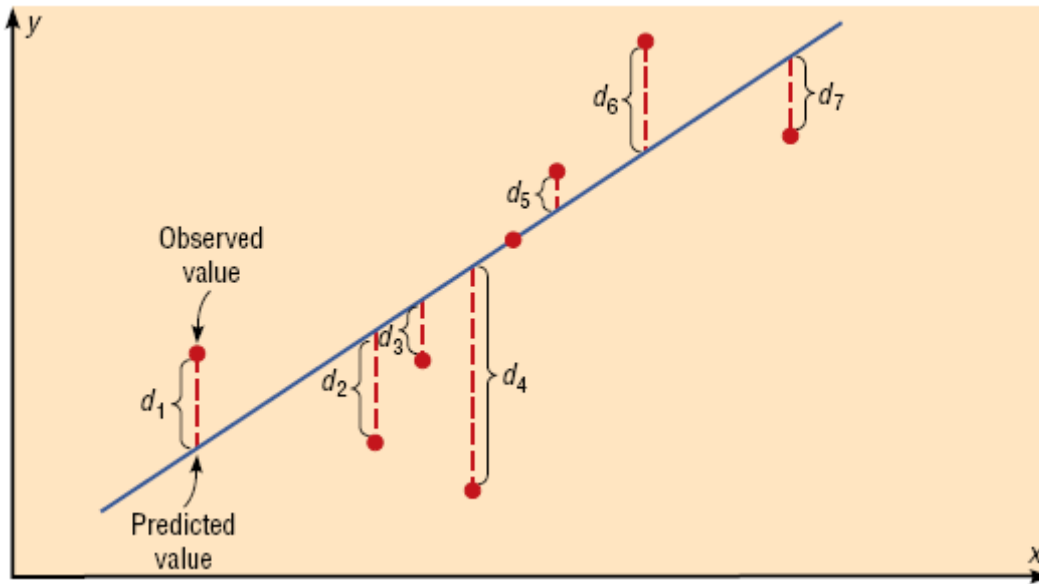


10A Linear Regression

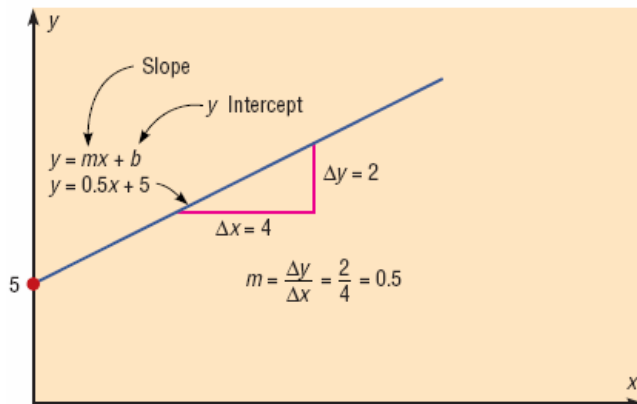
Sometimes there is a **direct** or **linear** relationship between data values/ variables.

We wish to find the "best" line to describe the data, using a **least square criterion**. We want the straight line for which the *sum of the squared errors* is smallest.

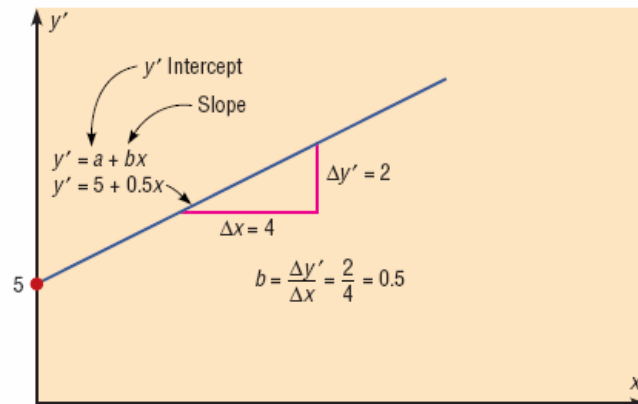


A **regression line** is the straight line that best fits a data set according to the least square criterion. A **regression equation** is the equation of the regression line.

Since not all sets of data will be approximately linear, plot the data first. When there doesn't seem to be a linear relationship, we say that there is little or no correlation.



(a) Algebra of a line



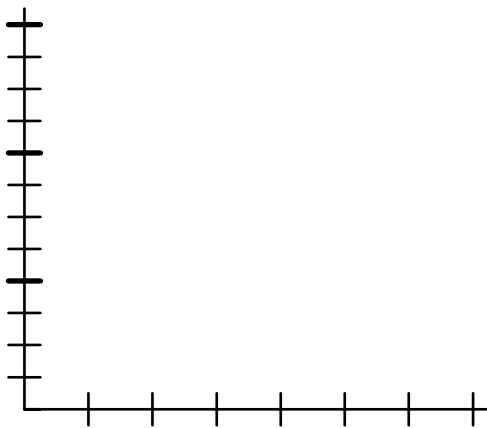
(b) Statistical notation for a regression line

example: Ten Corvettes between 1 and 6 years old were randomly selected from the classified ads of *The Arizona Republic*. The following data were obtained, where x denotes age, in years, and y denotes price, in hundreds of dollars.

X	6	6	6	2	2	5	4	5	1	4
Y	270	260	275	405	364	295	335	308	405	305

a) Determine the regression equation for the data.

b) Graph the regression equation along with the data.



c) Describe the apparent relationship between age and price for Corvettes.

d) What does the slope of the regression line represent in terms of Corvette prices?

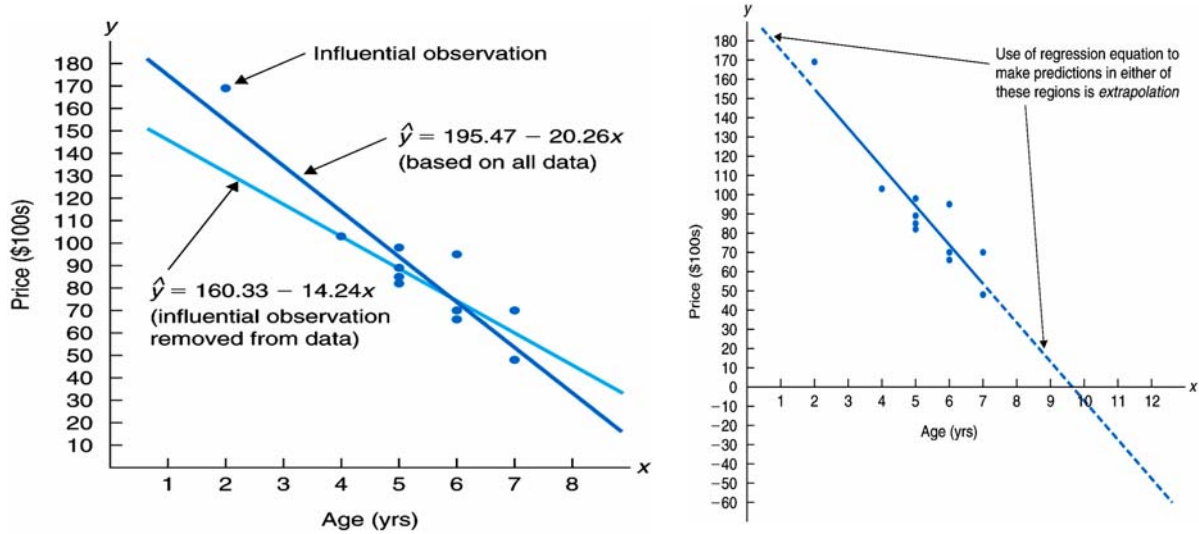
e) Use the regression equation to predict the price of a 2-year-old Corvette; a 3-year-old Corvette.

f) Identify the predictor (independent) and response (dependent) variables.

g) Identify outliers and potential influential observations.

An **influential observation** is a data value, that if removed, causes the regression line to change considerably. Usually the data value is separated in the x-direction from the other data values, thus "pulling" the regression line toward it without being counteracted by other data points.

Figure: Regression lines with and without the influential observation removed

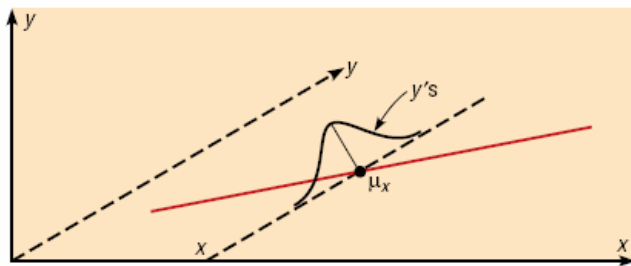


Assumptions for Valid Predictions in Regression

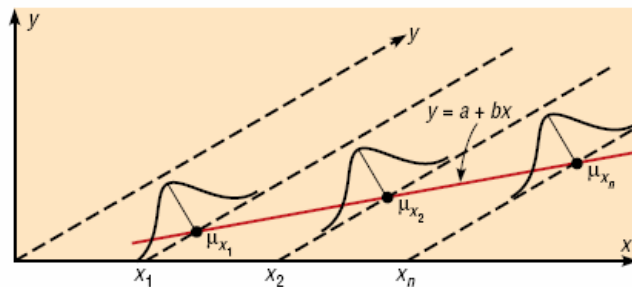
1. For any specific value of the independent variable x , the value of the dependent variable y must be normally distributed about the regression line. See Figure 10–16(a).
2. The standard deviation of each of the dependent variables must be the same for each value of the independent variable. See Figure 10–16(b).

Figure 10–16

Assumptions for Predictions



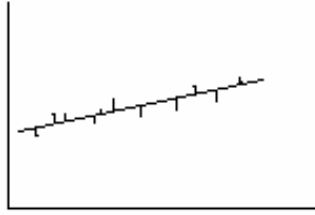
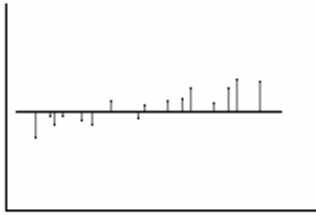
(a) Dependent variable y normally distributed



(b) $\sigma_1 = \sigma_2 = \dots = \sigma_n$

10B The Coefficient of Determination r^2

There are 2 ways to measure error involved with regression lines:



SST (Total Sum of Squares) **SSE** (Error Sum of Squares)

$$SST = \sum (y - \bar{y})^2$$

$$SSE = \sum (y - \hat{y})^2$$

r^2 is a descriptive measure of the utility of the regression equation for making predictions. It is the percentage reduction obtained in the total squared error by using the regression equation, \hat{y} , instead of the sample mean, \bar{y} , to predict observed y-values.

r^2 is always between 0 and 1. Values near 0 indicate that the regression is not very useful for making predictions. Values near 1 indicate that the regression equation is extremely useful for making predictions.

example: Refer back to the Corvette problem...

a) Compute the coefficient of determination r^2 .

b) Determine the percentage of the total variation in the observed y-values that is explained by the regression equation and interpret your results.

c) State how useful the regression equation appears to be for making predictions.

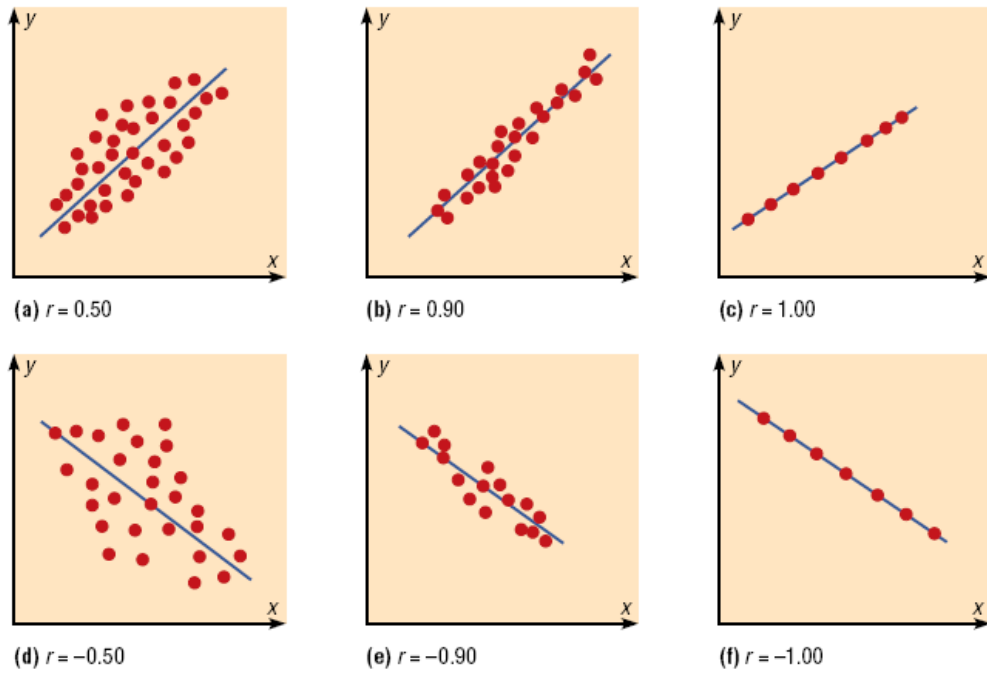
The **linear coefficient r** ranges from -1 to +1. It is the square root of r^2 , the coefficient of determination (to be learned later).

-1 means that there is a strong negative linear correlation.

+1 means that there is a strong positive linear correlation.

0 means that there is a weak linear relationship.

Figure 10-12
Relationship Between the Correlation Coefficient and the Line of Best Fit



example: Referring back to the Corvette problem...

X	6	6	6	2	2	5	4	5	1	4
Y	270	260	275	405	364	295	335	308	405	305

a) compute the linear coefficient r .

b) interpret the meaning of r in terms of the linear relationship between age and value of Corvettes.

Hypothesis Test for the Usefulness of the Slope for making Predictions:

Formally defined, the **population correlation coefficient** ρ is the correlation computed by using all possible pairs of data values (x, y) taken from a population.

In hypothesis testing, one of these is true:

$H_0: \rho = 0$ This null hypothesis means that there is no correlation between the x and y variables in the population.

$H_1: \rho \neq 0$ This alternative hypothesis means that there is a significant correlation between the variables in the population.

- STEP 1** State the hypotheses.
- STEP 2** Find the test value. (In this case, use the t test.)
- STEP 3** Find the P -value. (In this case, use Table F.)
- STEP 4** Make the decision.
- STEP 5** Summarize the results.

example: Referring to the Corvette problem, does the data provide sufficient evidence to conclude that the slope of the regression line is not zero, and hence that the age of Corvettes is useful for making predictions about its price?

Standard Error of the Estimate

Objective 6. Compute the standard error of the estimate.

When a y' value is predicted for a specific x value, the prediction is a point prediction. However, a prediction interval about the y' value can be constructed, just as a confidence interval was constructed for an estimate of the population mean. The prediction interval uses a statistic called the *standard error of the estimate*.

The **standard error of the estimate**, denoted by s_{est} , is the standard deviation of the observed y values about the predicted y' values. The formula for the standard error of estimate is

$$s_{\text{est}} = \sqrt{\frac{\sum (y - y')^2}{n - 2}}$$

Example: Referring to the Corvette problem, find and interpret the standard error of the estimate.

Example: Find and interpret a 95% confidence interval for the slope of the regression line:

Example: Find and interpret a 95% confidence interval for the mean price of 4-year-old Corvettes:

Example: Find and interpret a 95% prediction interval for the price of a randomly selected 4-year-old Corvette:

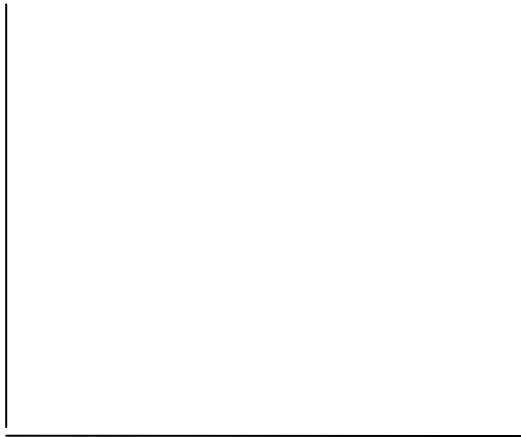
o Multiple Regression

In the last section, we only used age as an independent variable. If we also looked at the condition of the car, its mileage, etc., we would be doing multiple regression.

In recent years, physicians have used the “diving reflex to reduce abnormally rapid heartbeats in humans by briefly submerging the patient’s face in cold water. The reflex, triggered by cold water temperatures, is an involuntary neural response that shuts off circulation to the skin, muscles, and internal organs to divert extra oxygen-carrying blood to the heart, lungs, and brain. A research physician conducted an experiment to investigate the effects of various cold temperatures on the pulse rate of small children. The data for seven 6-year-old children are shown below in the table:

Child #	Temperature of Water x ($^{\circ}\text{F}$)	Decrease in Pulse rate y (beats/minute)
1	68	2
2	65	5
3	70	1
4	62	10
5	60	9
6	55	13
7	58	10

- a) The regression equation is: _____
- b) Graph the data and the regression equation on the graph provided. Label the graph.



- c) Describe the apparent relationship between temperature and pulse rate.

- d) What does the **slope** of the regression equation represent in terms of a decrease in pulse rate?
- e) Identify any outliers or potential influential observations.
- f) Use the regression equation to predict the decrease in pulse rate when a child’s face is briefly submerged in 60-degree water.
- g) Identify the predictor and response variables.
- h) $r^2 =$ _____ I) $r =$ _____
- j) Interpret the meaning of r^2 and how useful the regression equation is for making predictions.
- k) Interpret the meaning of r in terms of the linear relationship between temperature and decreases in pulse rate.

l) Find the standard error of the estimate Se and interpret its meaning in terms of a decrease in pulse rate.

m) At the 10% significance level, do the data provide sufficient evidence to conclude that the slope of the population regression line is not 0 and hence is useful as a predictor of a decrease in pulse rate?

n) Obtain a 90% confidence interval for the slope of the regression equation. Interpret its meaning in context.

o) Find a 90% confidence interval for the **mean** decrease in pulse rate of children that are submerged in 60-degree water. Interpret this interval.

p) Determine a 90% confidence prediction interval for the decrease in pulse rate when a child’s face is briefly submerged in 60-degree water. Interpret this interval.

11.1 Chi-Square Goodness-of-Fit Test

example: According to the Census Bureau, the marital-status distribution of the U.S. adult population is as follows:

Marital Status:	Single	Married	Widowed	Divorced
Percentage:	21.5	63.9	7.7	6.9

A random sample of 750 U.S. males, 25-29 years old, yielded the following frequency distributions:

Marital Status:	Single	Married	Widowed	Divorced
Frequency:	289	408	0	53

At the 1% significance level, does it appear that the marital-status distribution of all 25-29 year-old U.S. males is different from that of the U.S. adult population as a whole?

To do this problem, we need to discover the following:

1. How do the actual numbers differ from what is expected?
2. Once we know the difference, how do we judge goodness-of-fit?

Goodness-of-Fit Process:

Assumptions: 1. all expected frequencies are at least 1.

2. at most 20% of the expected frequencies are less than 5.

step 1: Ho: The distribution of marital status is the same for 25-29 year-old males as it is for the adult population as a whole.

H₁: The distribution is different.

step 2: Enter observed values into L1 and expected values into L2, using the formula $E=np$ to calculate the expected values.

step 3: Check assumptions...

step 4: Significance level: $\alpha =$ _____

step 5: Run the program GFIT.

step 6: Test statistic and P-value:

step 7: Decision:

step 8: Conclusion:

example: A roulette wheel contains 18 red numbers, 18 black numbers, and 2 green numbers. The table below shows the frequency with which the ball landed on each color in 200 trials.

Color	Red	Black	Green
Frequency	88	102	10

At the 5% significance level, do the data suggest that the wheel is out of balance?

Example: Toner Toys, a world-wide chain with stores in all fifty states, wanted to know whether sales were evenly distributed among the five weekdays, Monday through Friday, for a randomly selected non-holiday week. Corporate records released the following sales figures, in billions of dollars.

Day	Mon	Tues	Wed	Thurs	Fri
Sales (billions)	32	36	38	35	40

At the 5% significance level, do the data suggest that sales are evenly distributed throughout the weekdays?

11.2 Chi-Square Independence Test

- Assumptions: 1. All expected frequencies are at least 1.
 2. At most 20% of the expected frequencies are less than 5.

- step 1: State H_0 (not associated / independent) and H_1 (associated / dependent) in words.
 step 2: Enter the data into matrix A. Run the χ^2 Independence Test.
 step 3: Check matrix B to see that the assumptions are met.
 step 4: Decide on significance level α .
 step 5: Check the test statistic and P-value returned by your calculator.
 step 6: Accept/ Reject H_0 .
 step 7: State conclusion in words.

example: The Gallup Organization conducts periodic surveys to gauge the support by U.S. adults for regional primary elections. The question asked is, "It has been proposed that four individual primaries be held in different weeks of June during presidential election years. Does this sound like a good idea or a poor idea?" Here is a contingency table for responses by political affiliation.

At the 5% level of significance, do the data suggest that the feelings of adults on the issue of regional primaries are dependent on political affiliation?

	Good Idea	Poor Idea	No Opinion
Republican	266	266	186
Democrat	308	250	176
Independent	28	27	21

step 1: H_0 :

H_1 :

- step 2: Enter values into calculator. Run program.
 step 3: check assumptions:

- step 4: $\alpha =$ _____
 step 5: critical value and test statistic:

step 6:

step 7:

example: The American Bar Association publishes information on lawyer characteristics. The following contingency table cross classifies 307 randomly selected lawyers by status in practice and size of city practicing in.

Do the data provide sufficient evidence to conclude that the characteristics **size of city** and **status in practice** are statistically dependent? Use $\alpha = .05$.

	< 250,000	250,000 – 499,999	> 500,000
Government	12	4	14
Judicial	8	1	2
Private Practice	122	31	69
Salaried	19	7	18

step 1: Ho:

H₁:

step 2: Enter values into calculator. Run program.

step 3: assumptions:

example: In 1989, roughly 58 million Americans suffered injuries. More males (31.7 million) were injured than females (26.3 million). Those statistics do not tell us whether males and females tend to be injured in similar circumstances. One set of categories commonly used for accident circumstances is "while at work," "home," "motor vehicle," and "other." In order to decide whether there is an association between accident circumstance and sex, a safety official in a large city took a random sample of accident reports. He obtained the following data.

	Male	Female
While at work	18	4
Home	26	28
Motor vehicle	4	6
Other	36	24

Do the data provide sufficient evidence to conclude that in this city, accident circumstances and sex are statistically dependent? Perform the required hypothesis test at the 5% significance level.