Linear Regression and Correlation

Chapter 13

McGraw-Hill/Irwin

GOALS

1. Understand and interpret the terms *dependent* and *independent variable*.
2. Calculate and interpret the *coefficient of correlation*, the *coefficient of determination*, and the *standard error of estimate*.
3. Conduct a test of hypothesis to determine whether the coefficient of correlation in the population is zero.
4. Calculate the least squares regression line.
5. Construct and interpret confidence and prediction intervals for the dependent variable.

Regression Analysis - Introduction

- Recall in Chapter 4 the idea of showing the relationship between *two* variables with a scatter diagram was introduced.
- In that case we showed that, as the age of the buyer increased, the amount spent for the vehicle also increased.
- In this chapter we carry this idea further. Numerical measures to express the strength of relationship between two variables are developed.
- In addition, an equation is used to express the relationship between variables, allowing us to estimate one variable on the basis of another.

EXAMPLES

1. Is there a relationship between the amount Healthtex spends per month on advertising and its sales in the month?
2. Can we base an estimate of the cost to heat a home in January on the number of square feet in the home?
3. Is there a relationship between the miles per gallon achieved by large pickup trucks and the size of the engine?
4. Is there a relationship between the number of hours that students studied for an exam and the score earned?

Correlation Analysis

Correlation Analysis is the study of the relationship between variables. It is also defined as group of techniques to measure the association between two variables.

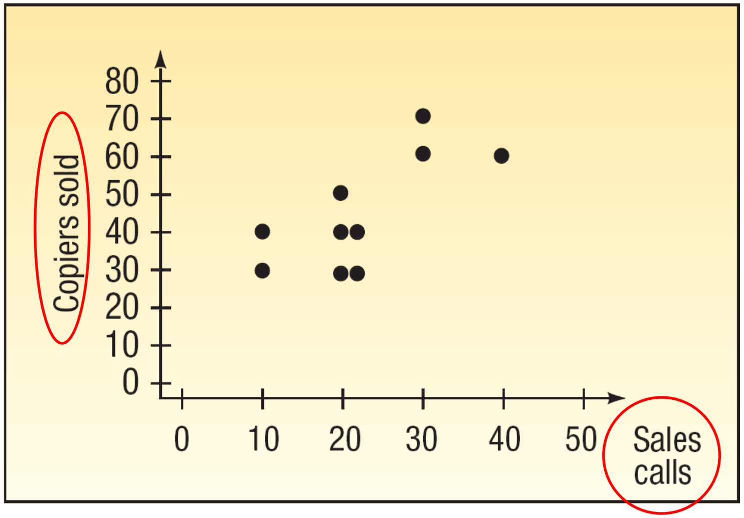
Scatter Diagram is a chart that portrays the relationship between the two variables. It is the usual first step in correlations analysis

The Dependent Variable is the variable being predicted or estimated.
The Independent Variable provides the basis for estimation. It is the predictor variable.

Scatter Diagram Example

The sales manager of Copier Sales of America, which has a large sales force throughout the United States and Canada, wants to determine whether there is a relationship between the number of sales calls made in a month and the number of copiers sold that month. The manager selects a random sample of 10 representatives and determines the number of sales calls each representative made last month and the number of copiers sold.

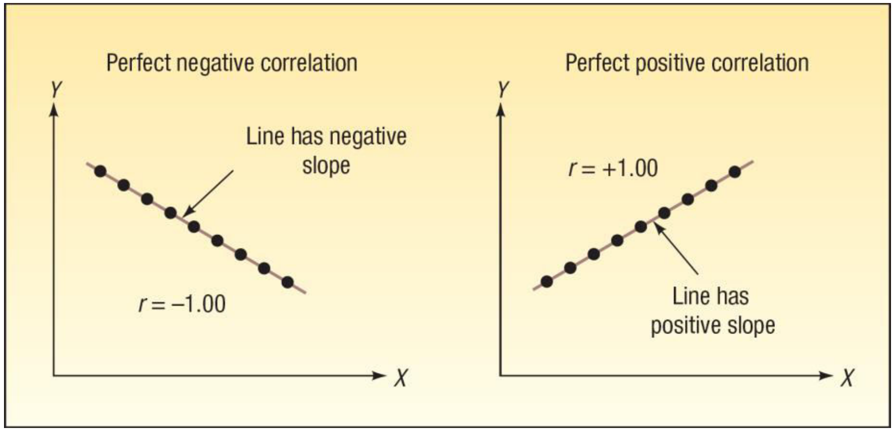
Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70



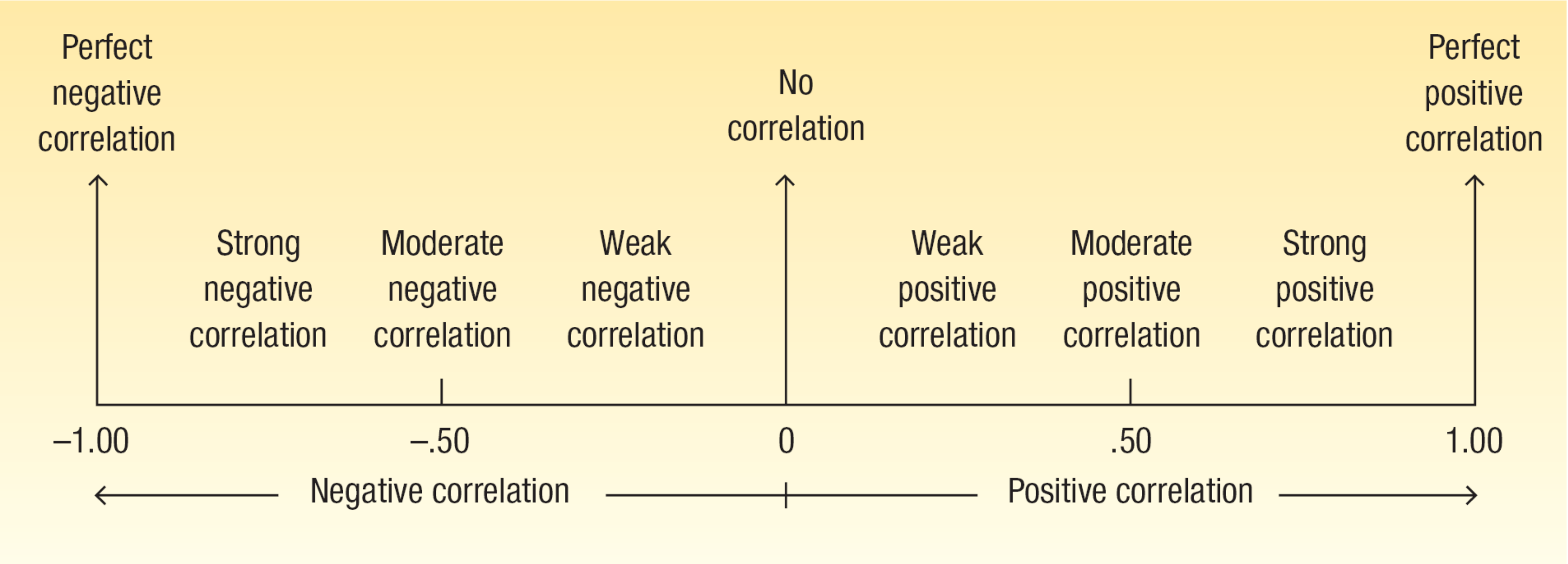
The Coefficient of Correlation, *r*

The Coefficient of Correlation (*r*) is a measure of the strength of the relationship between two variables.

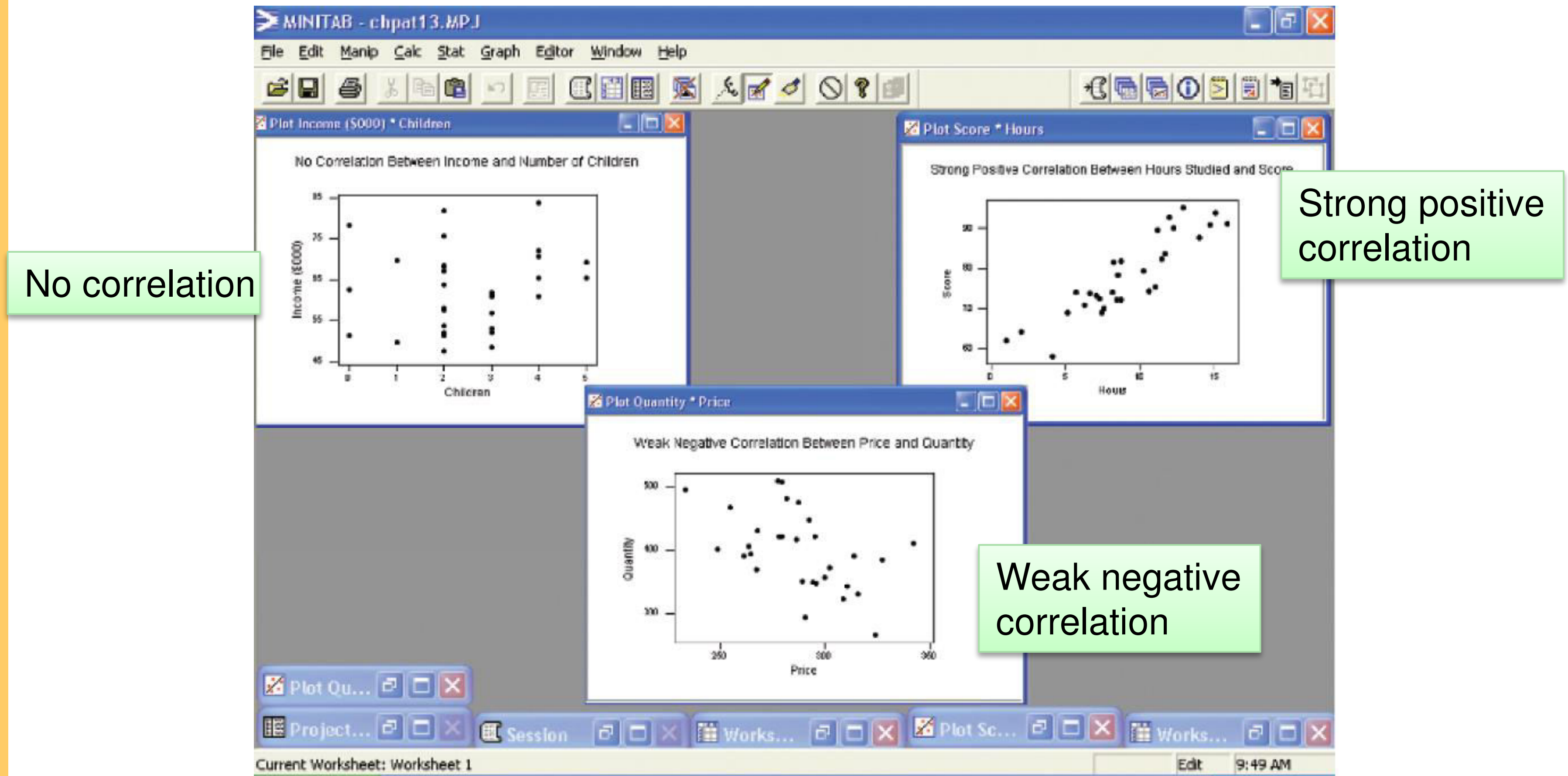
- It shows the direction and strength of the linear relationship between two interval or ratio-scale variables
- It can range from -1.00 to +1.00.
- Values of -1.00 or +1.00 indicate perfect and strong correlation.
- Values close to 0.0 indicate weak correlation.
- Negative values indicate an inverse relationship and positive values indicate a direct relationship.



Correlation Coefficient - Interpretation



Minitab Scatter Plots



Coefficient of Determination

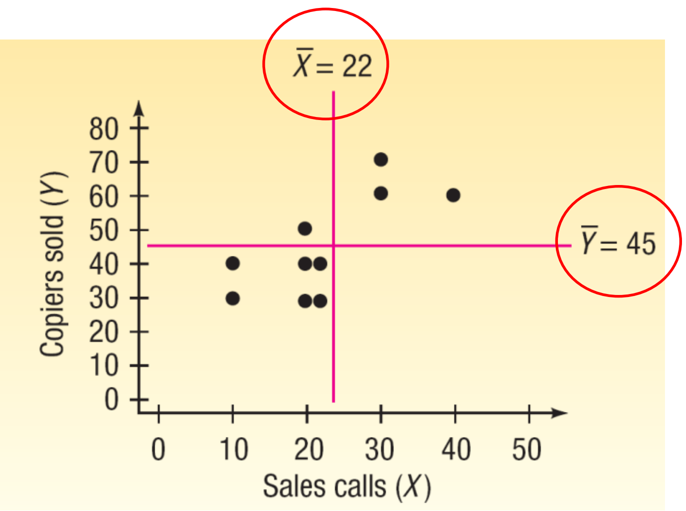
The coefficient of determination (r^2) is the proportion of the total variation in the dependent variable (Y) that is explained or accounted for by the variation in the independent variable (X). It is the square of the coefficient of correlation.

- It ranges from 0 to 1.
- It does not give any information on the direction of the relationship between the variables.

Correlation Coefficient - Example

Using the Copier Sales of America data which a scatterplot is shown below, compute the correlation coefficient and coefficient of determination.

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70



Using the formula:

CORRELATION COEFFICIENT $$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n-1)s_x s_y}$$

Correlation Coefficient - Example

Sales Representative	Calls, Y	Sales, X	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
Tom Keller	20	30	−2	−15	30
Jeff Hall	40	60	18	15	270
Brian Virost	20	40	−2	−5	10
Greg Fish	30	60	8	15	120
Susan Welch	10	30	−12	−15	180
Carlos Ramirez	10	40	−12	−5	60
Rich Niles	20	40	−2	−5	10
Mike Kiel	20	50	−2	5	−10
Mark Reynolds	20	30	−2	−15	30
Soni Jones	30	70	8	25	200
					900

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y} = \frac{900}{(10 - 1)(9.189)(14.337)} = 0.759$$

How do we interpret a correlation of 0.759?
First, it is positive, so we see there is a direct relationship between the number of sales calls and the number of copiers sold. The value of 0.759 is fairly close to 1.00, so we conclude that the association is strong.

However, does this mean that more sales calls ***cause*** more sales?
No, we have not demonstrated cause and effect here, only that the two variables—sales calls and copiers sold—are related.

Coefficient of Determination (r^2) – Copier Sales Example

- The coefficient of determination, r^2 ,is 0.576, found by $(0.759)^2$

- This is a proportion or a percent; we can say that 57.6 percent of the variation in the number of copiers sold is explained, or accounted for, by the variation in the number of sales calls.

Testing the Significance of the Correlation Coefficient

H_0: $\rho = 0$ (the correlation in the population is 0)
H_1: $\rho \neq 0$ (the correlation in the population is not 0)

Reject H_0 if:

$t > t_{\alpha/2,n-2}$ or $t < -t_{\alpha/2,n-2}$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$ with $n - 2$ degrees of freedom

Testing the Significance of the Correlation Coefficient – Copier Sales Example

H_0: $\rho = 0$ (the correlation in the population is 0)

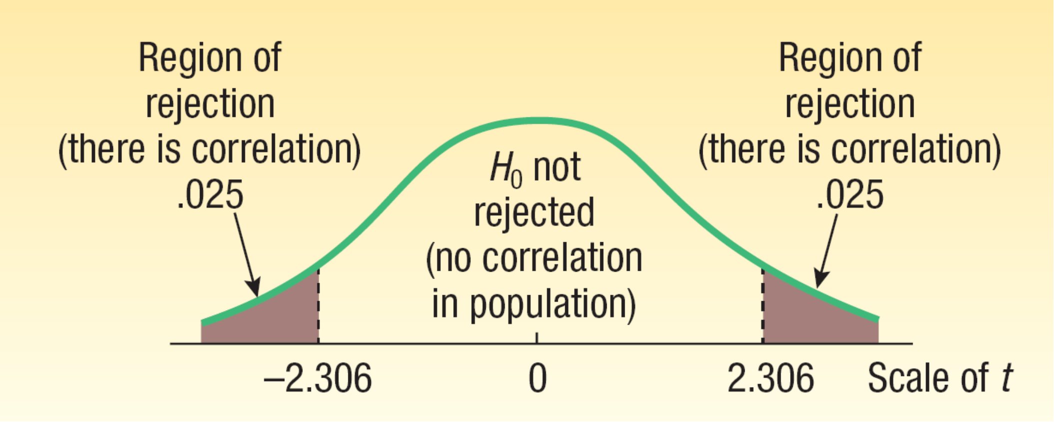
H_1: $\rho \neq 0$ (the correlation in the population is not 0)

Reject H_0 if:

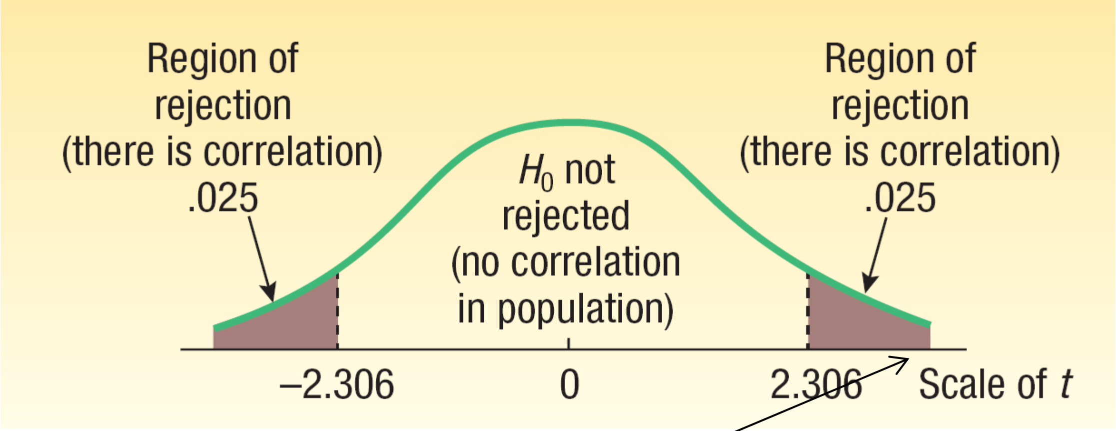
$t > t_{\alpha/2,n-2}$ or $t < -t_{\alpha/2,n-2}$

$t > t_{0.025,8}$ or $t < -t_{0.025,8}$

$t > 2.306$ or $t < -2.306$



Testing the Significance of the Correlation Coefficient – Copier Sales Example



Computing t, we get

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.759\sqrt{10-2}}{\sqrt{1-.759^2}} = 3.297$$

The computed t (3.297) is within the rejection region, therefore, we will reject H_0. This means the correlation in the population is not zero. From a practical standpoint, it indicates to the sales manager that there is correlation with respect to the number of sales calls made and the number of copiers sold in the population of salespeople.

Linear Regression Model

GENERAL FORM OF LINEAR REGRESSION EQUATION $$\hat{Y} = a + bX$$

where

$\hat{Y}$ read Y hat, is the estimated value of the Y variable for a selected X value.

a is the Y-intercept. It is the estimated value of Y when $X = 0$. Another way to put it is: a is the estimated value of Y where the regression line crosses the Y-axis when X is zero.

b is the slope of the line, or the average change in $\hat{Y}$ for each change of one unit (either increase or decrease) in the independent variable X.

X is any value of the independent variable that is selected.

Computing the Slope of the Line and the Y-intercept

SLOPE OF THE REGRESSION LINE

$$b = r\frac{s_y}{s_x}$$

where
r is the correlation coefficient.
s_y is the standard deviation of Y (the dependent variable).
s_x is the standard deviation of X (the independent variable).

***Y*-INTERCEPT**

$$a = \overline{Y} - b\overline{X}$$

where
$\overline{Y}$ is the mean of Y (the dependent variable).
$\overline{X}$ is the mean of X (the independent variable).

Regression Analysis

In regression analysis we use the independent variable (X) to estimate the dependent variable (Y).

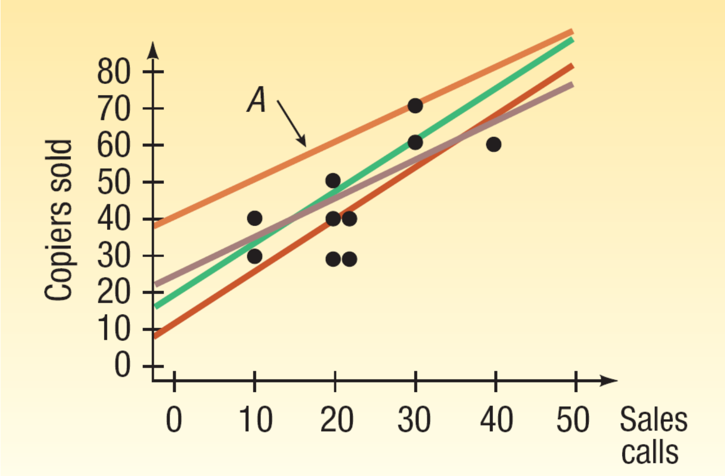
- The relationship between the variables is linear.
- Both variables must be at least interval scale.
- The least squares criterion is used to determine the equation.

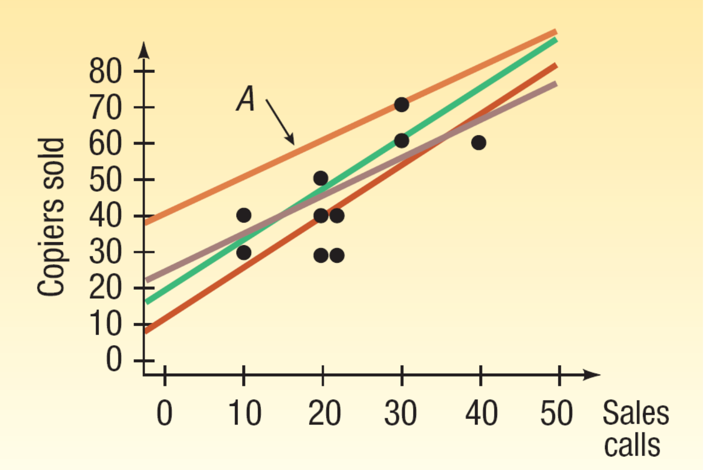
REGRESSION EQUATION An equation that expresses the linear relationship between two variables.

LEAST SQUARES PRINCIPLE Determining a regression equation by minimizing the sum of the squares of the vertical distances between the actual *Y values and* the predicted values of *Y.*

Regression Analysis – Least Squares Principle

- The least squares principle is used to obtain *a* and *b*.





- The equations to determine *a* and *b* are:

$$b = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$$

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n}$$

Regression Equation - Example

Recall the example involving Copier Sales of America. The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 10 sales representatives. Use the least squares method to determine a linear equation to express the relationship between the two variables.

What is the expected number of copiers sold by a representative who made 20 calls?

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

Finding the Regression Equation - Example

Step 1 – Find the slope (*b*) of the line

$$b = r\left(\frac{s_y}{s_x}\right) = .759\left(\frac{14.337}{9.189}\right) = 1.1842$$

Step 2 – Find the *y*-intercept (*a*)

$$a = \overline{Y} - b\overline{X} = 45 - 1.1842(22) = 18.9476$$

The regression equation is :

$$\hat{Y} = a + bX$$

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(20)$$

$$\hat{Y} = 42.6316$$

Computing the Estimates of Y

Step 1 – Using the regression equation, substitute the value of each X to solve for the estimated sales

Sales Representative	Sales Calls (X)	Estimated Sales ($\hat{Y}$)	Sales Representative	Sales Calls (X)	Estimated Sales ($\hat{Y}$)
Tom Keller	20	42.6316	Carlos Ramirez	10	30.7896
Jeff Hall	40	66.3156	Rich Niles	20	42.6316
Brian Virost	20	42.6316	Mike Kiel	20	42.6316
Greg Fish	30	54.4736	Mark Reynolds	20	42.6316
Susan Welch	10	30.7896	Soni Jones	30	54.4736

Tom Keller

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(20)$$

$$\hat{Y} = 42.6316$$

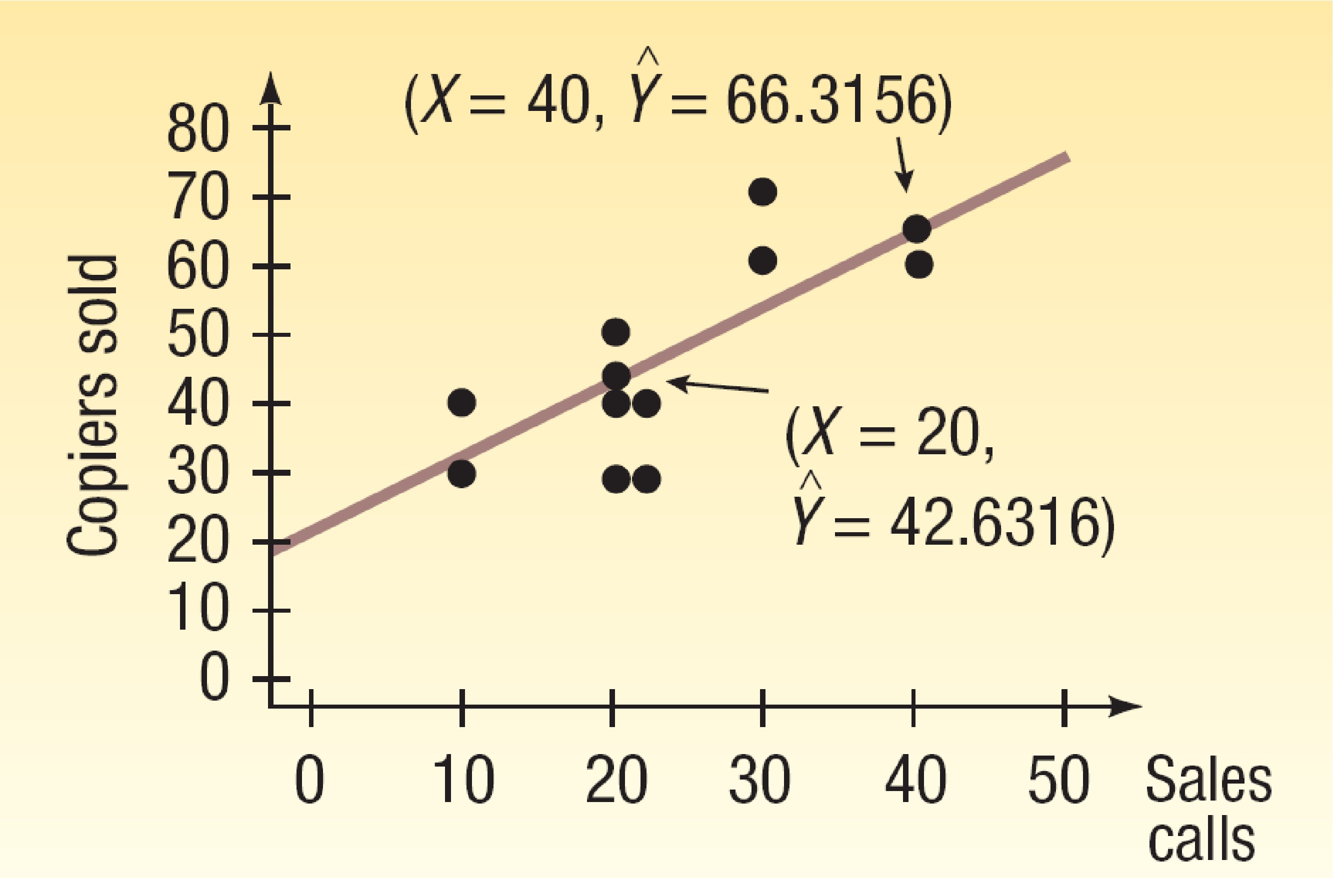
Soni Jones

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(30)$$

$$\hat{Y} = 54.4736$$

Plotting the Estimated and the Actual Y's



The Standard Error of Estimate

- The standard error of estimate measures the scatter, or dispersion, of the observed values around the line of regression
- Formulas used to compute the standard error:

$$s_{y.x} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n-2}}$$

$$s_{y.x} = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{n-2}}$$

Standard Error of the Estimate - Example

Recall the example involving Copier Sales of America. The sales manager determined the least squares regression equation is given below.

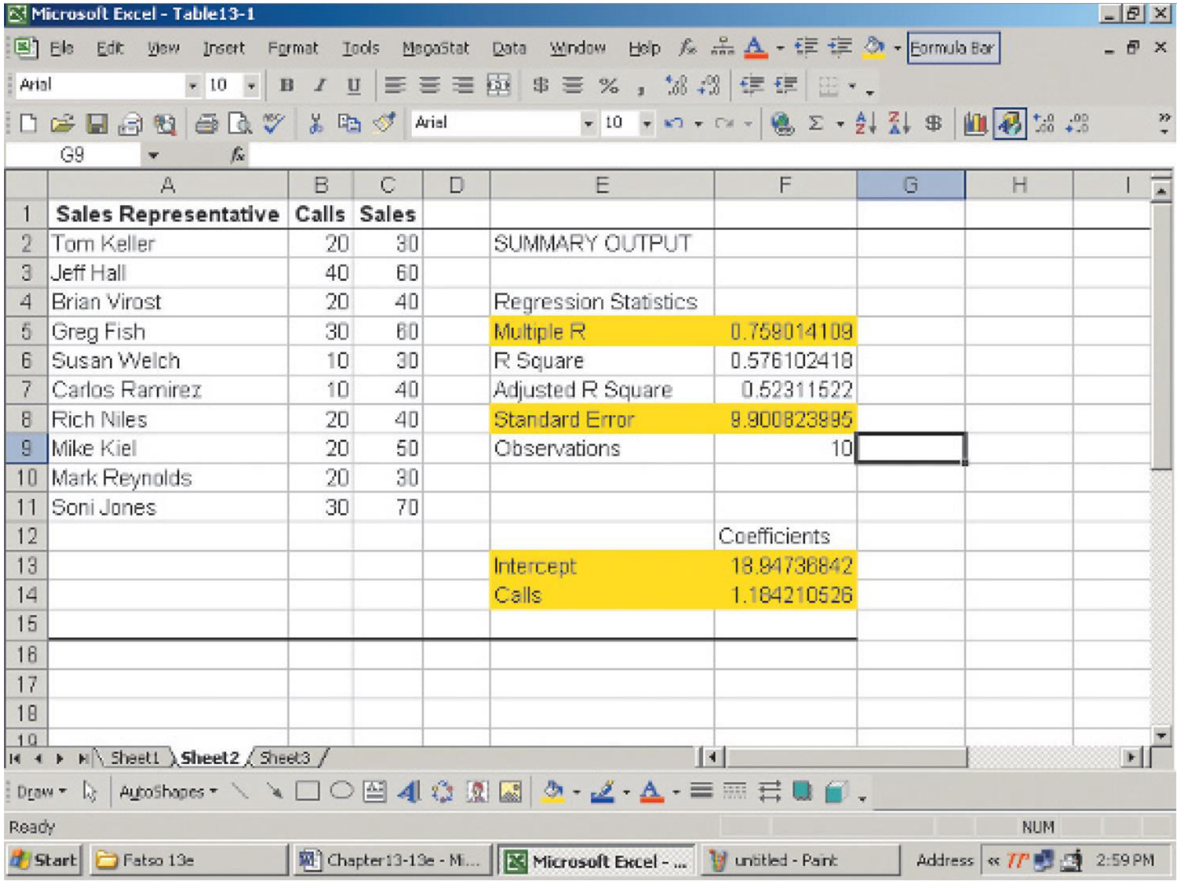
Determine the standard error of estimate as a measure of how well the values fit the regression line.

$$\hat{Y} = 18.9476 + 1.1842X$$

Sales Representative	Actual Sales, (Y)	Estimated Sales, $(\hat{Y})$	Deviation, $(Y - \hat{Y})$	Deviation Squared, $(Y - \hat{Y})^2$
Tom Keller	30	42.6316	−12.6316	159.557
Jeff Hall	60	66.3156	−6.3156	39.887
Brian Virost	40	42.6316	−2.6316	6.925
Greg Fish	60	54.4736	5.5264	30.541
Susan Welch	30	30.7896	−0.7896	0.623
Carlos Ramirez	40	30.7896	9.2104	84.831
Rich Niles	40	42.6316	−2.6316	6.925
Mike Kiel	50	42.6316	7.3684	54.293
Mark Reynolds	30	42.6316	−12.6316	159.557
Soni Jones	70	54.4736	15.5264	241.069
			0.0000	784.211

$$s_{y.x} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{784.211}{10-2}} = 9.901$$

Standard Error of the Estimate - Excel



Sales Representative	Calls	Sales
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

SUMMARY OUTPUT

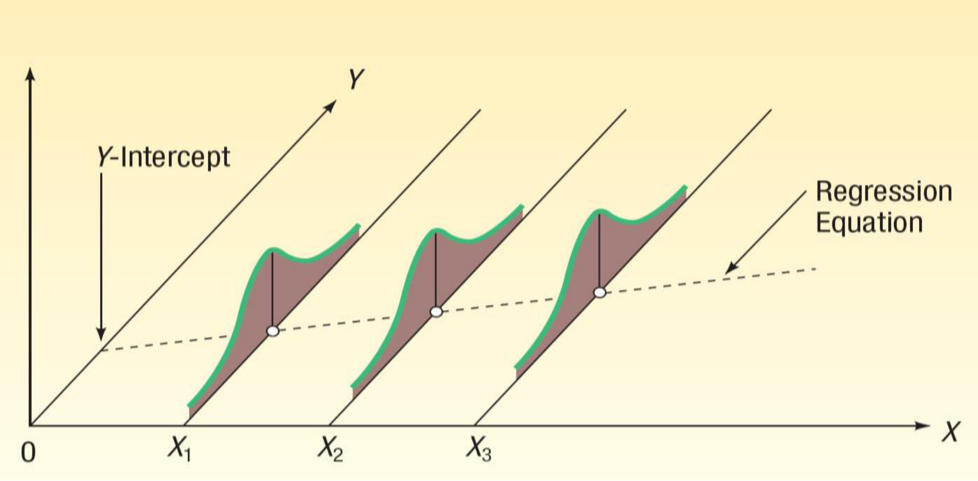
Regression Statistics	
Multiple R	0.759014109
R Square	0.576102418
Adjusted R Square	0.52311522
Standard Error	9.900823995
Observations	10

	Coefficients
Intercept	18.94736842
Calls	1.184210526

Assumptions Underlying Linear Regression

For each value of *X*, there is a group of *Y* values, and these

- *Y* values are *normally distributed*. The *means* of these normal distributions of *Y* values all lie on the straight line of regression.
- The *standard deviations* of these normal distributions are equal.
- The *Y* values are statistically independent. This means that in the selection of a sample, the *Y* values chosen for a particular *X* value do not depend on the *Y* values for any other *X* values.



Confidence Interval and Prediction Interval Estimates of Y

- A **confidence interval** reports the *mean* value of *Y* for a given *X.*
- *A* **prediction interval** reports the *range of values* of *Y* for a *particular* value of *X.*

CONFIDENCE INTERVAL FOR THE MEAN OF *Y*, GIVEN *X*

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad [13\text{–}7]$$

PREDICTION INTERVAL FOR *Y*, GIVEN *X*

$$\hat{Y} \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad [13\text{–}8]$$

Confidence Interval Estimate - Example

We return to the Copier Sales of America illustration. Determine a 95 percent confidence interval for all sales representatives who make 25 calls.

CONFIDENCE INTERVAL FOR THE MEAN OF *Y*, GIVEN *X*

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad [13\text{–}7]$$

where

$\hat{Y}$ is the predicted value for any selected X value.
X is any selected value of X.
$\bar{X}$ is the mean of the Xs, found by $\Sigma X/n$.
n is the number of observations.
$s_{y \cdot x}$ is the standard error of estimate.
t is the value of t from Appendix B.2 with $n - 2$ degrees of freedom.

Confidence Interval Estimate - Example

CONFIDENCE INTERVAL FOR THE MEAN OF *Y*, GIVEN *X*

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad [13\text{–}7]$$

Step 1 – Compute the point estimate of Y

In other words, determine the number of copiers we expect a sales representative to sell if he or she makes 25 calls.

The regression equation is :

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(25)$$

$$\hat{Y} = 48.5526$$

Confidence Interval Estimate - Example

CONFIDENCE INTERVAL FOR THE MEAN OF *Y*, GIVEN *X*

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad \textbf{[13–7]}$$

Step 2 – Find the value of *t*

- To find the *t* value, we need to first know the number of degrees of freedom. In this case the degrees of freedom is n - 2 = 10 – 2 = 8.
- We set the confidence level at 95 percent. To find the value of *t*, move down the left-hand column of Appendix B.2 to 8 degrees of freedom, then move across to the column with the 95 percent level of confidence.
- The value of *t* is 2.306.

Confidence Interval Estimate - Example

CONFIDENCE INTERVAL FOR THE MEAN OF Y, GIVEN X

$$\hat{Y} \pm t(s_{y\cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \quad [13\text{–}7]$$

Step 3 – Compute $\left(X - \bar{X}\right)^2$ and $\sum\left(X - \bar{X}\right)^2$

Sales Representative	Sales Calls, (X)	Copier Sales, (Y)	$(X - \bar{X})$	$(X - \bar{X})^2$
Tom Keller	20	30	−2	4
Jeff Hall	40	60	18	324
Brian Virost	20	40	−2	4
Greg Fish	30	60	8	64
Susan Welch	10	30	−12	144
Carlos Ramirez	10	40	−12	144
Rich Niles	20	40	−2	4
Mike Kiel	20	50	−2	4
Mark Reynolds	20	30	−2	4
Soni Jones	30	70	8	64
			0	760

Confidence Interval Estimate - Example

CONFIDENCE INTERVAL FOR THE MEAN OF Y, GIVEN X

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad [13\text{–}7]$$

Step 4 – Use the formula above by substituting the numbers computed in previous slides

$$\text{Confidence Interval} = \hat{Y} \pm ts_{y \cdot x}\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}}$$

$$= 48.5526 \pm 2.306(9.901)\sqrt{\frac{1}{10} + \frac{(25 - 22)^2}{760}}$$

$$= 48.5526 \pm 7.6356$$

Thus, the 95 percent confidence interval for the average sales of all sales representatives who make 25 calls is from 40.9170 up to 56.1882 copiers.

Prediction Interval Estimate - Example

We return to the Copier Sales of America illustration. Determine a 95 percent prediction interval for Sheila Baker, a West Coast sales representative who made 25 calls.

Prediction Interval Estimate - Example

PREDICTION INTERVAL FOR *Y*, GIVEN *X*

$$\hat{Y} \pm t s_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{\Sigma(X - \overline{X})^2}} \quad \textbf{[13–8]}$$

Step 1 – Compute the point estimate of Y

In other words, determine the number of copiers we expect a sales representative to sell if he or she makes 25 calls.

The regression equation is :

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(25)$$

$$\hat{Y} = 48.5526$$

Prediction Interval Estimate - Example

PREDICTION INTERVAL FOR *Y*, GIVEN *X*

$$\hat{Y} \pm t s_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{\Sigma(X - \overline{X})^2}} \qquad [13\text{–}8]$$

Step 2 – Using the information computed earlier in the confidence interval estimation example, use the formula above.

$$\text{Prediction Interval} = \hat{Y} \pm t s_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{\Sigma(X - \overline{X})^2}}$$

$$= 48.5526 \pm 2.306(9.901)\sqrt{1 + \frac{1}{10} + \frac{(25 - 22)^2}{760}}$$

$$= 48.5526 \pm 24.0746$$

If Sheila Baker makes 25 sales calls, the number of copiers she will sell will be between about 24 and 73 copiers.

Confidence and Prediction Intervals – Minitab Illustration