Linear Regression and Correlation

Chapter 13

GOALS

- Understand and interpret the terms dependent and independent variable.
- Calculate and interpret the coefficient of correlation, the coefficient of determination, and the standard error of estimate.
- Conduct a test of hypothesis to determine whether the coefficient of correlation in the population is zero.
- Calculate the least squares regression line.
- 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.

Regression Analysis - Uses

Some examples.

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

Correlation Analysis

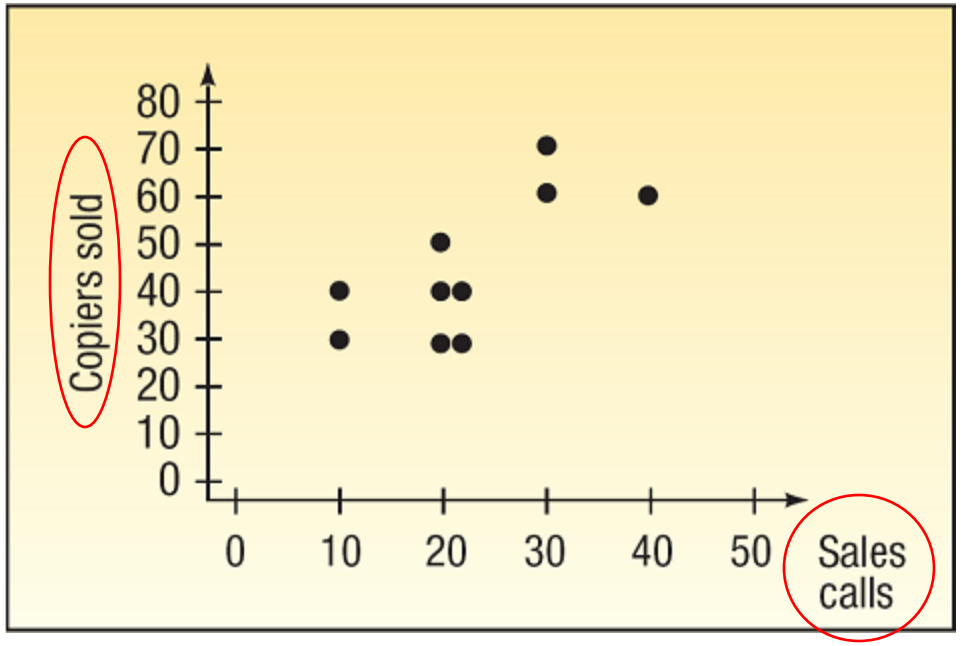
- Correlation Analysis is the study of the relationship between variables. It is also defined as group of techniques to measure the association between two variables.
- A Scatter Diagram is a chart that portrays the relationship between the two variables. It is the usual first step in correlations analysis
  - The Dependent Variable is the variable being predicted or estimated.
  - The Independent Variable provides the basis for estimation. It is the predictor variable.

Regression 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

Scatter Diagram

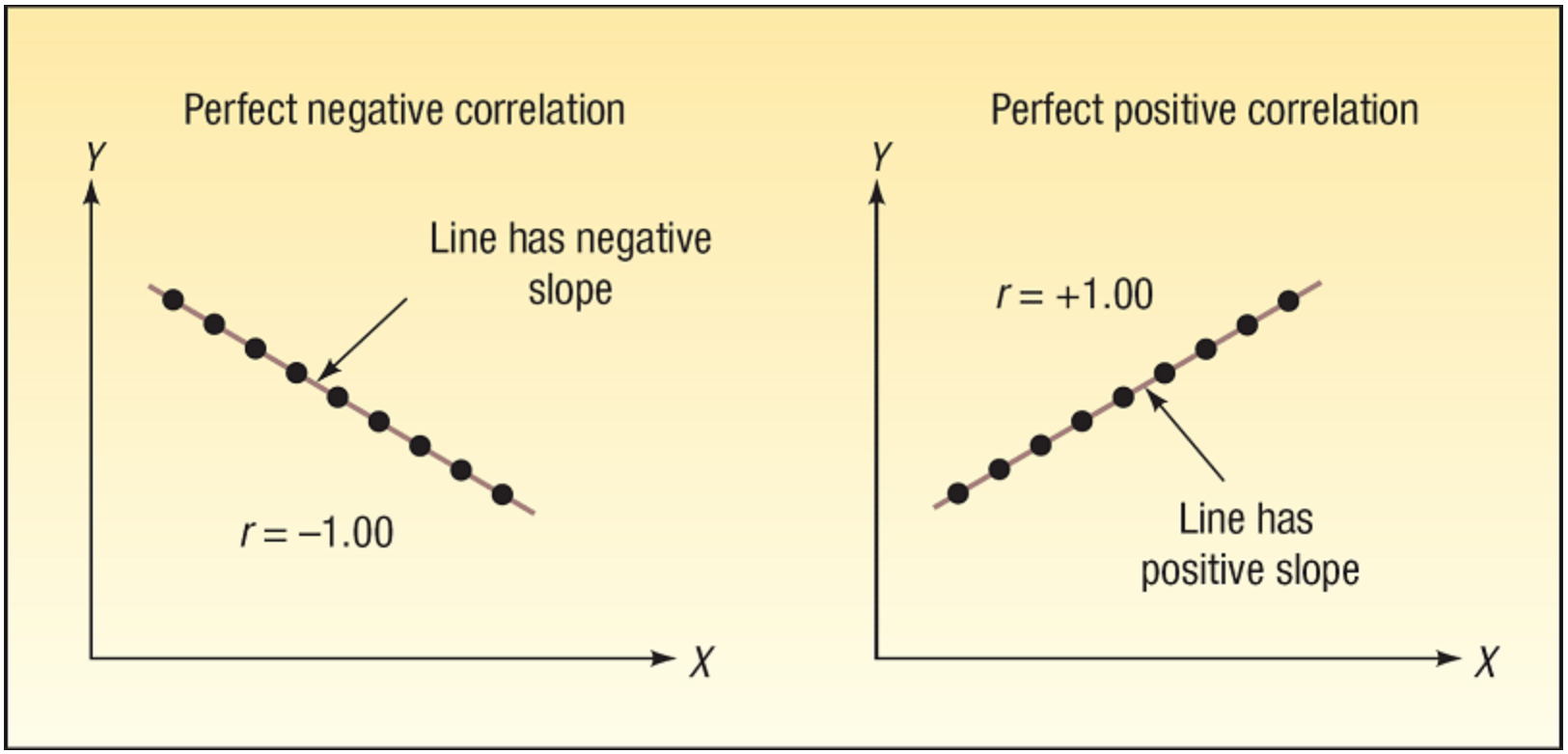


The Coefficient of Correlation, *r*

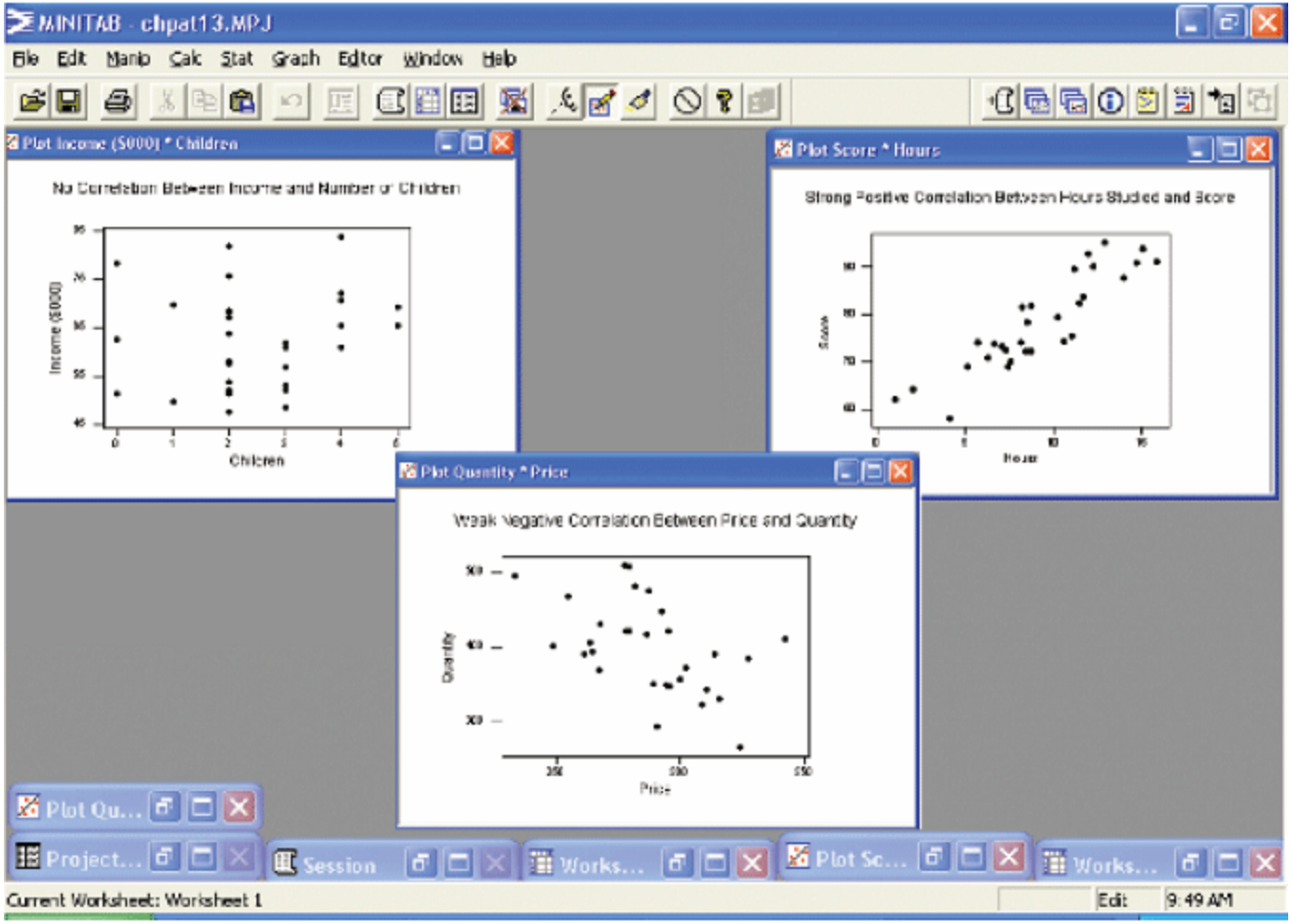
The Coefficient of Correlation (*r*) is a measure of the strength of the relationship between two variables. It requires interval or ratio-scaled data.

- 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.

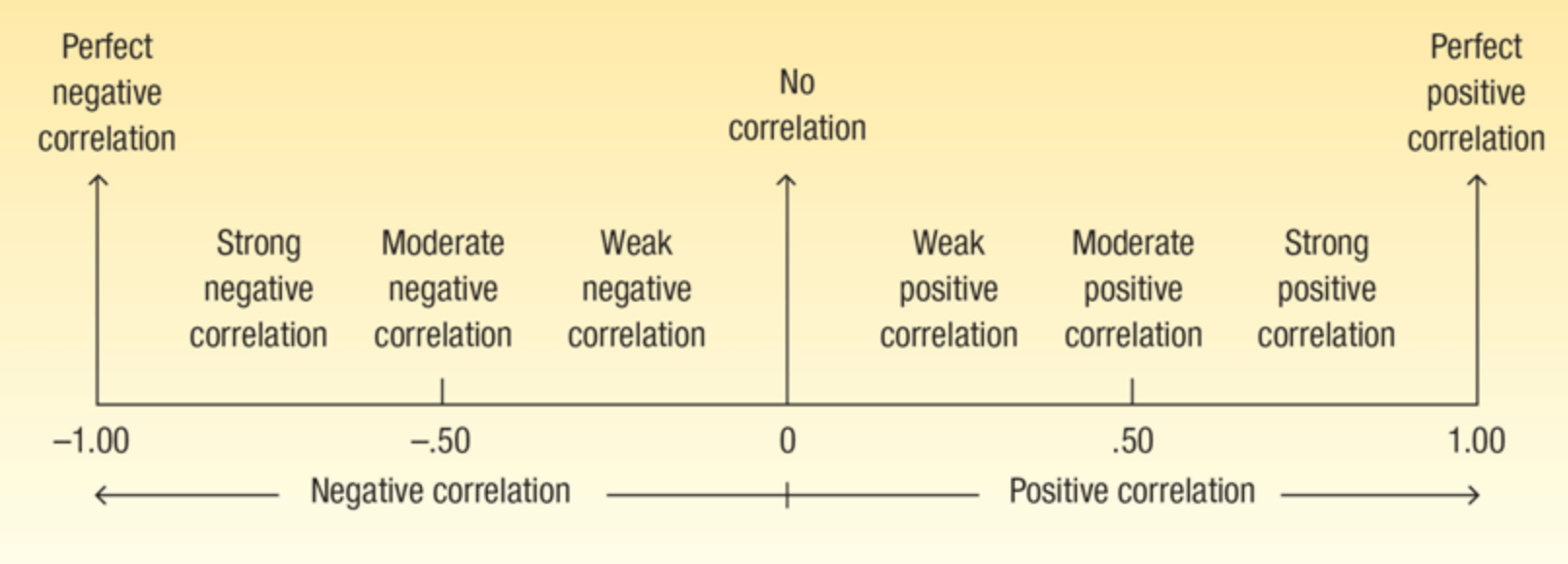
Perfect Correlation



Minitab Scatter Plots



Correlation Coefficient - Interpretation



Correlation Coefficient - Formula

CORRELATION COEFFICIENT

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

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 was developed earlier, compute the correlation coefficient and coefficient of determination.

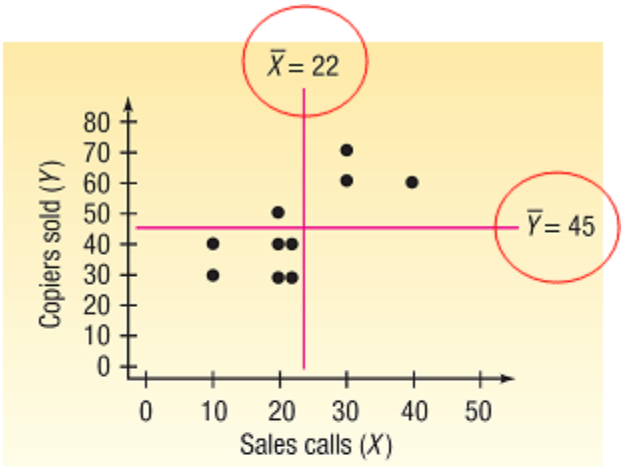
Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

Correlation Coefficient - Example

Using the formula:

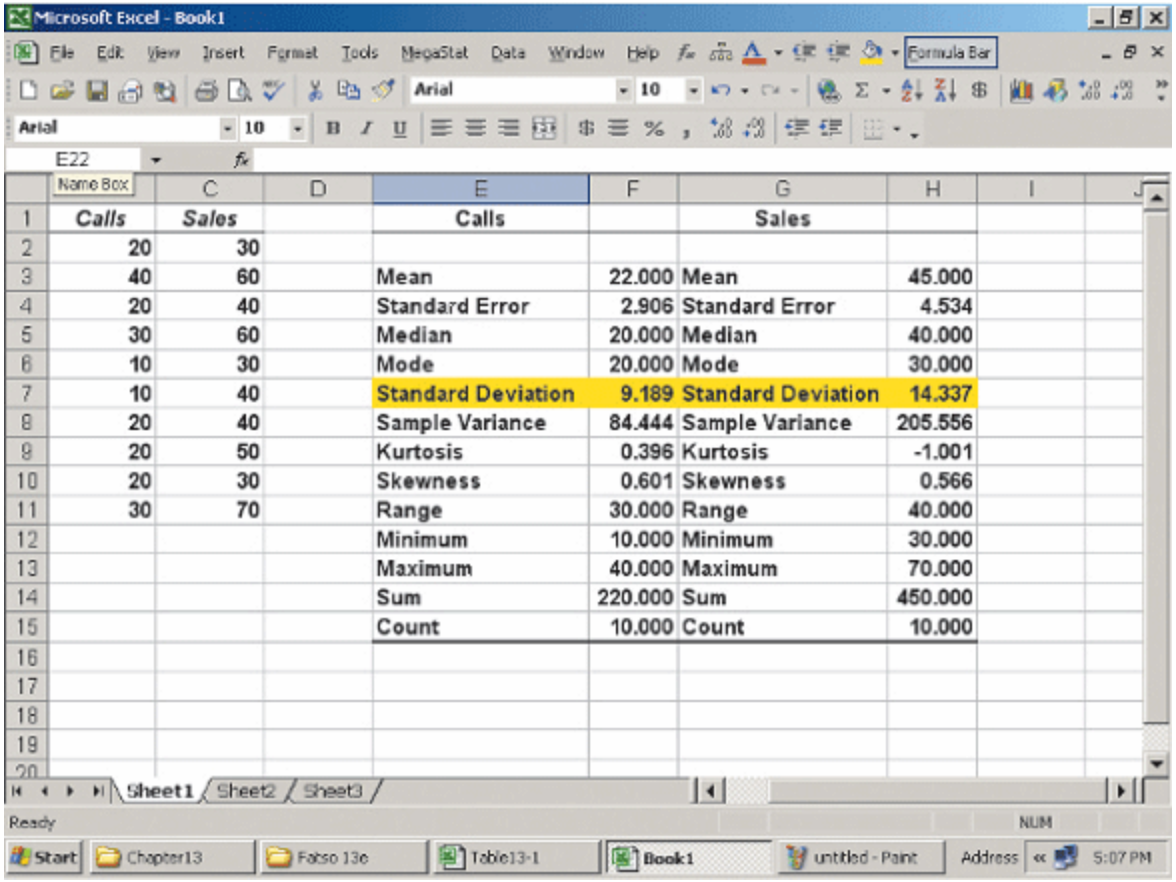
CORRELATION COEFFICIENT

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



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

Correlation Coefficient – Excel Example



	C	D	E	F	G	H
1	*Calls*	*Sales*	Calls		Sales	
2	20	30				
3	40	60	Mean	22.000	Mean	45.000
4	20	40	Standard Error	2.906	Standard Error	4.534
5	30	60	Median	20.000	Median	40.000
6	10	30	Mode	20.000	Mode	30.000
7	10	40	Standard Deviation	9.189	Standard Deviation	14.337
8	20	40	Sample Variance	84.444	Sample Variance	205.556
9	20	50	Kurtosis	0.396	Kurtosis	-1.001
10	20	30	Skewness	0.601	Skewness	0.566
11	30	70	Range	30.000	Range	40.000
12			Minimum	10.000	Minimum	30.000
13			Maximum	40.000	Maximum	70.000
14			Sum	220.000	Sum	450.000
15			Count	10.000	Count	10.000

Correlation Coefficient - Example

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

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

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

Testing the Significance of the Correlation Coefficient - Example

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

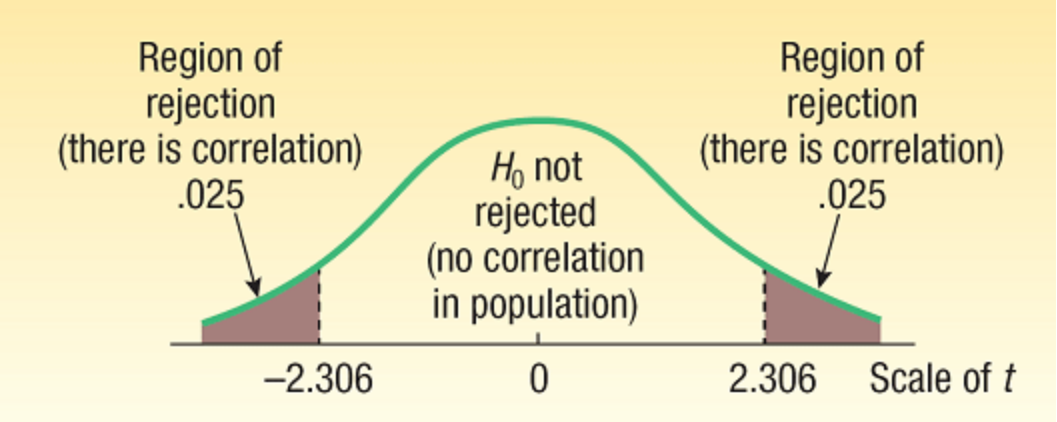
H_1: $\rho \neq 0$ (the correlation in the population is not 0)

Reject H_0 if:

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

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

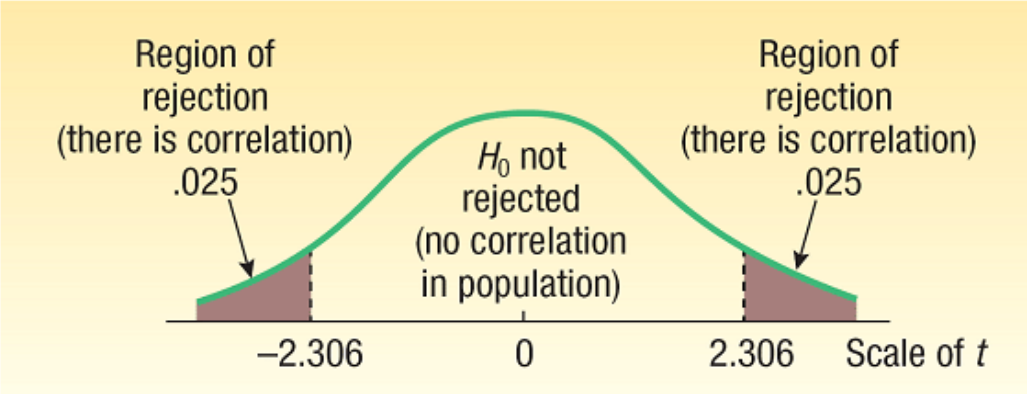
$t > 2.306$ or $t < -2.306$



Testing the Significance of the Correlation Coefficient - Example

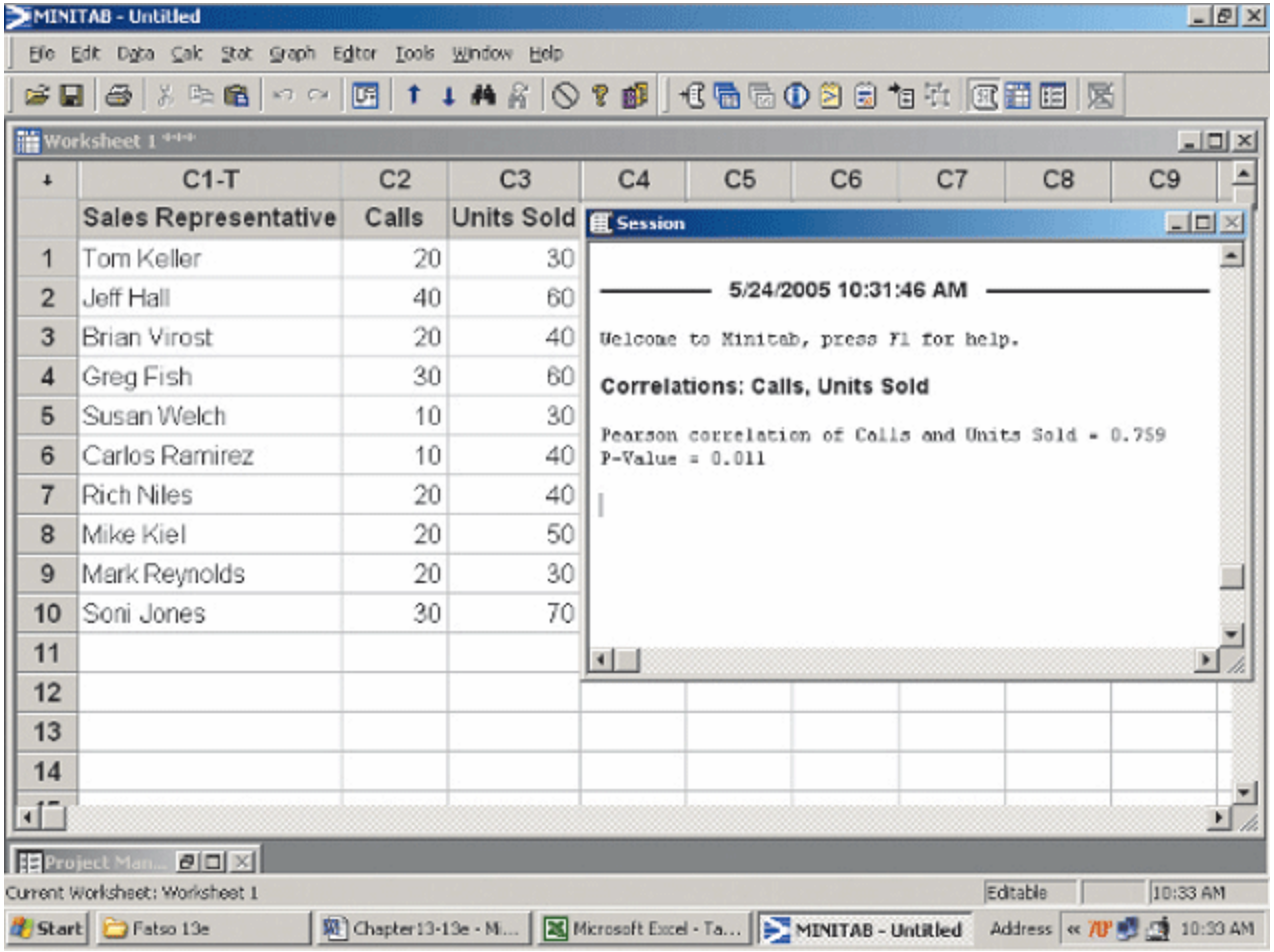
Computing t, we get

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



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.

Minitab


MINITAB - Untitled
File Edit Data Calc Stat Graph Editor Tools Window Help
Worksheet 1
C1-T C2 C3 C4 C5 C6 C7 C8 C9
Sales Representative Calls Units Sold
1 Tom Keller 20 30
2 Jeff Hall 40 60
3 Brian Virost 20 40
4 Greg Fish 30 60
5 Susan Welch 10 30
6 Carlos Ramirez 10 40
7 Rich Niles 20 40
8 Mike Kiel 20 50
9 Mark Reynolds 20 30
10 Soni Jones 30 70
11
12
13
14
Session
5/24/2005 10:31:46 AM
Welcome to Minitab, press F1 for help.
Correlations: Calls, Units Sold
Pearson correlation of Calls and Units Sold = 0.759
P-Value = 0.011
Project Man...
Current Worksheet: Worksheet 1
Editable
10:33 AM
Start
Fatso 13e
Chapter13-13e - Mi...
Microsoft Excel - Ta...
MINITAB - Untitled
Address
10:33 AM

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

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).

Computing the Y-Intercept

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.

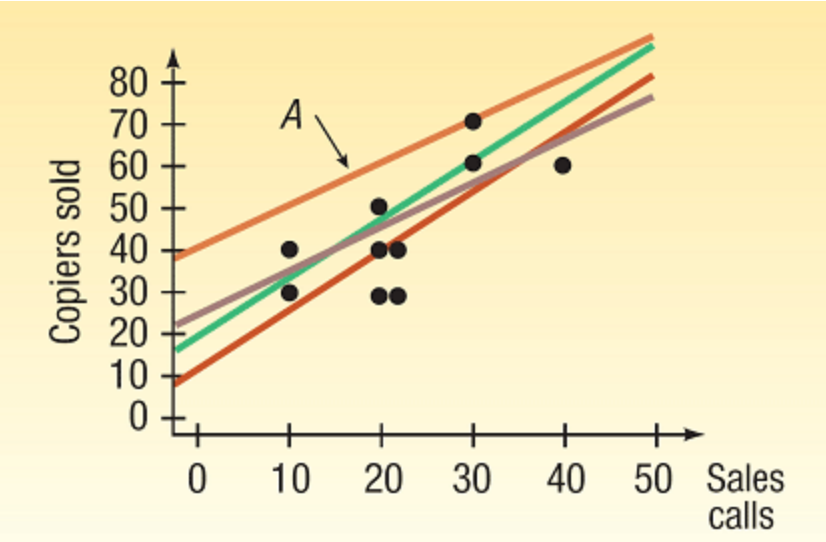
Regression Analysis – Least Squares Principle

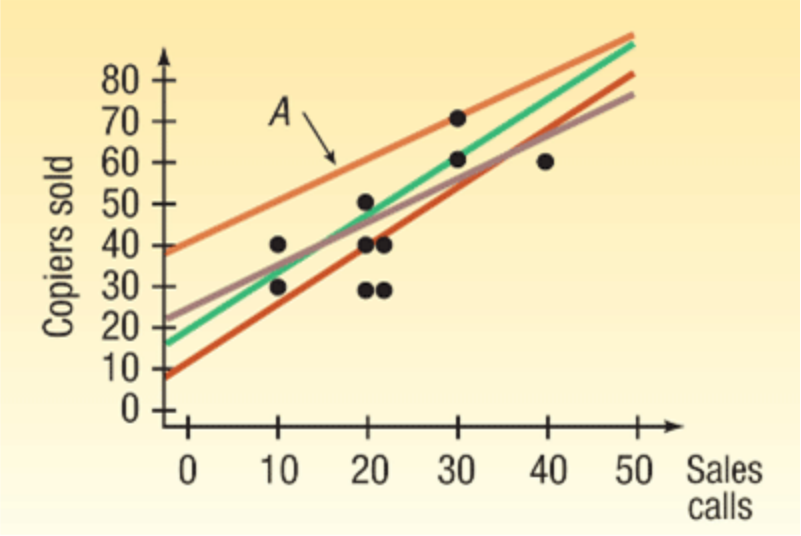
- The least squares principle is used to obtain *a* and *b*.
- The equations to determine *a* and *b* are:

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

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

Illustration of the Least Squares Regression Principle





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)

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

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

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

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

$$\hat{Y} = 42.6316$$

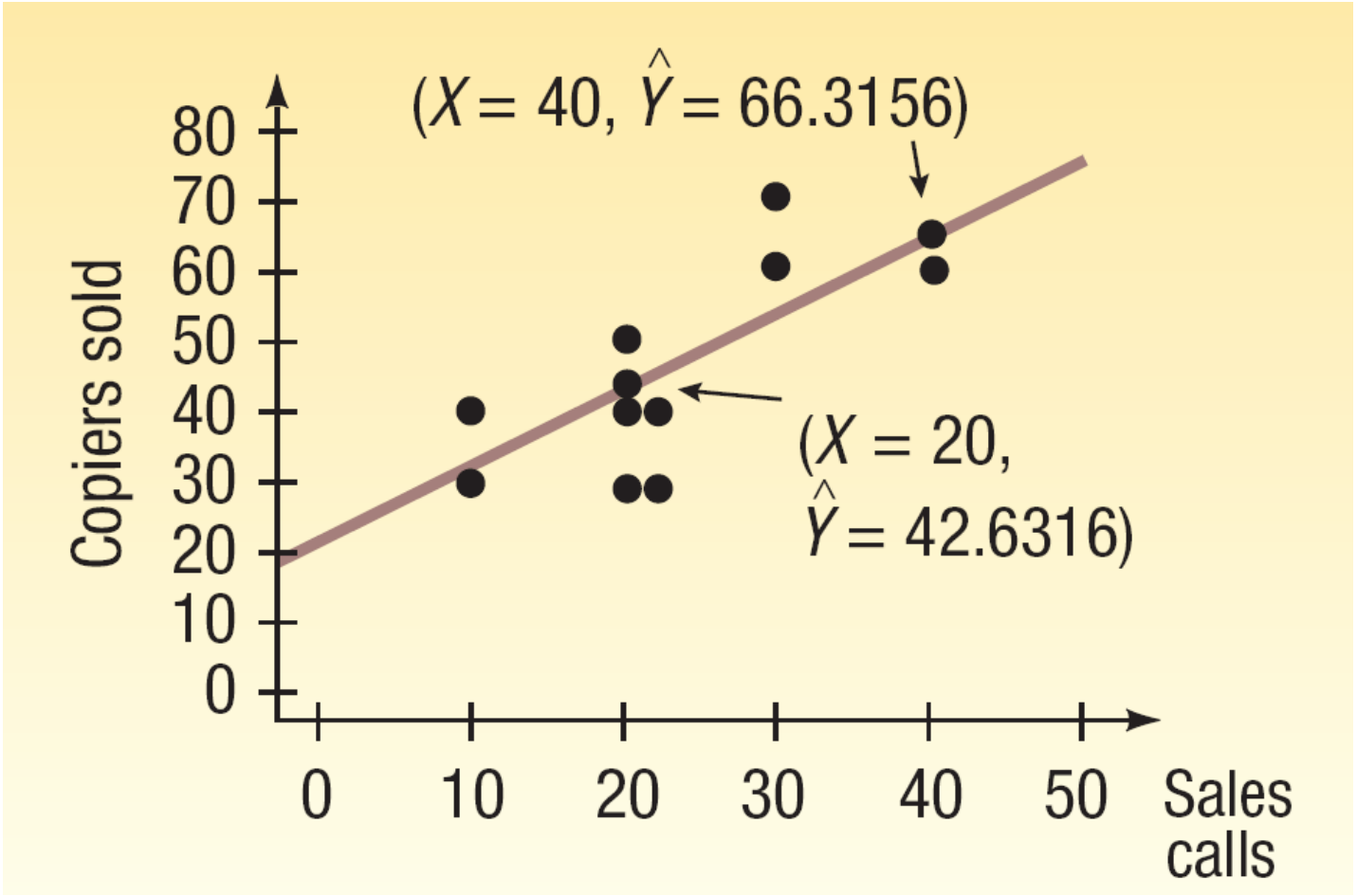
Soni Jones

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

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

$$\hat{Y} = 54.4736$$

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
- The formulas that are 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.

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

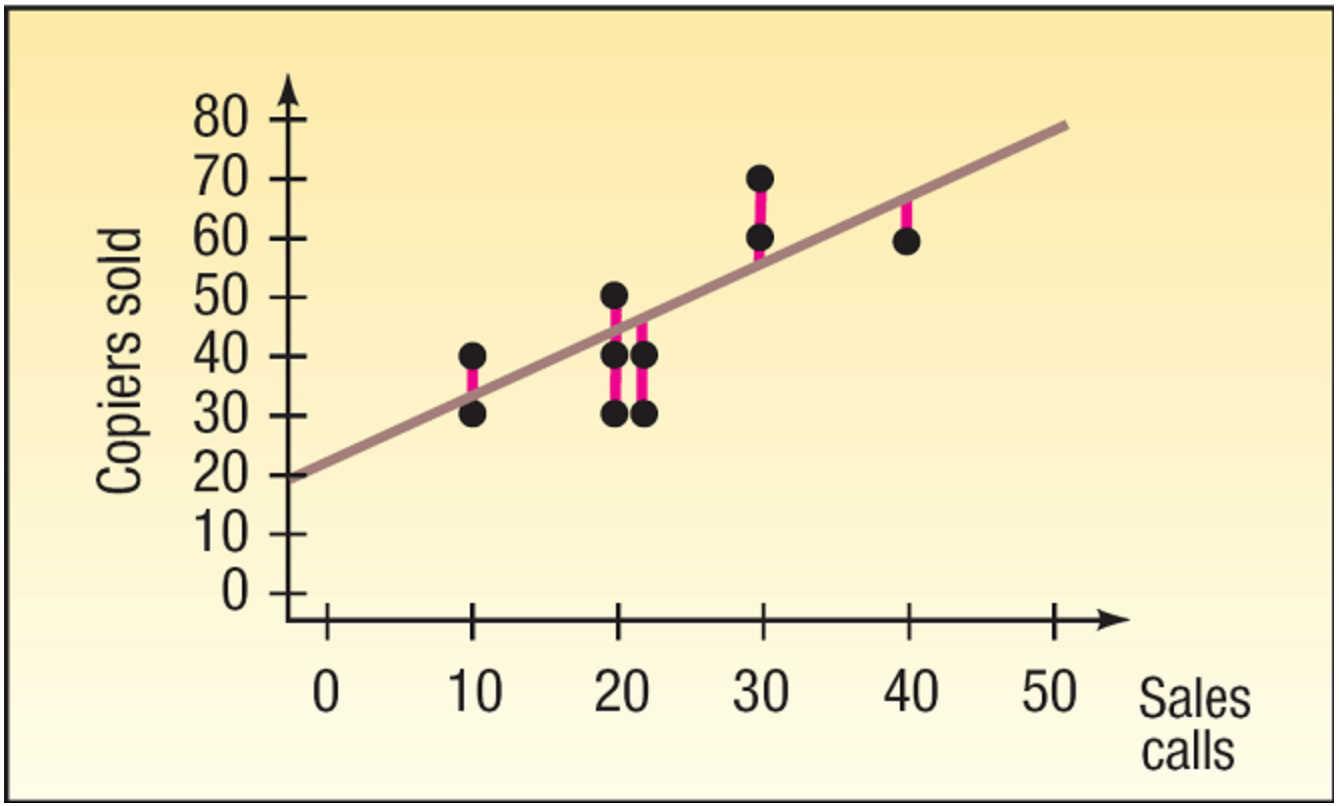
Sales Representative	Actual Sales, (Y)	Estimated Sales, ($\hat{Y}$)	Deviation, ($Y - \hat{Y}$)	Deviation Squared, ($Y - \hat{Y}$)^2
Tom Keller	30	42.6316	−12.6316	159.557
Jeff Hall	60	66.3156	−6.3156	39.887
Brian Virost	40	42.6316	−2.6316	6.925
Greg Fish	60	54.4736	5.5264	30.541
Susan Welch	30	30.7896	−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

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

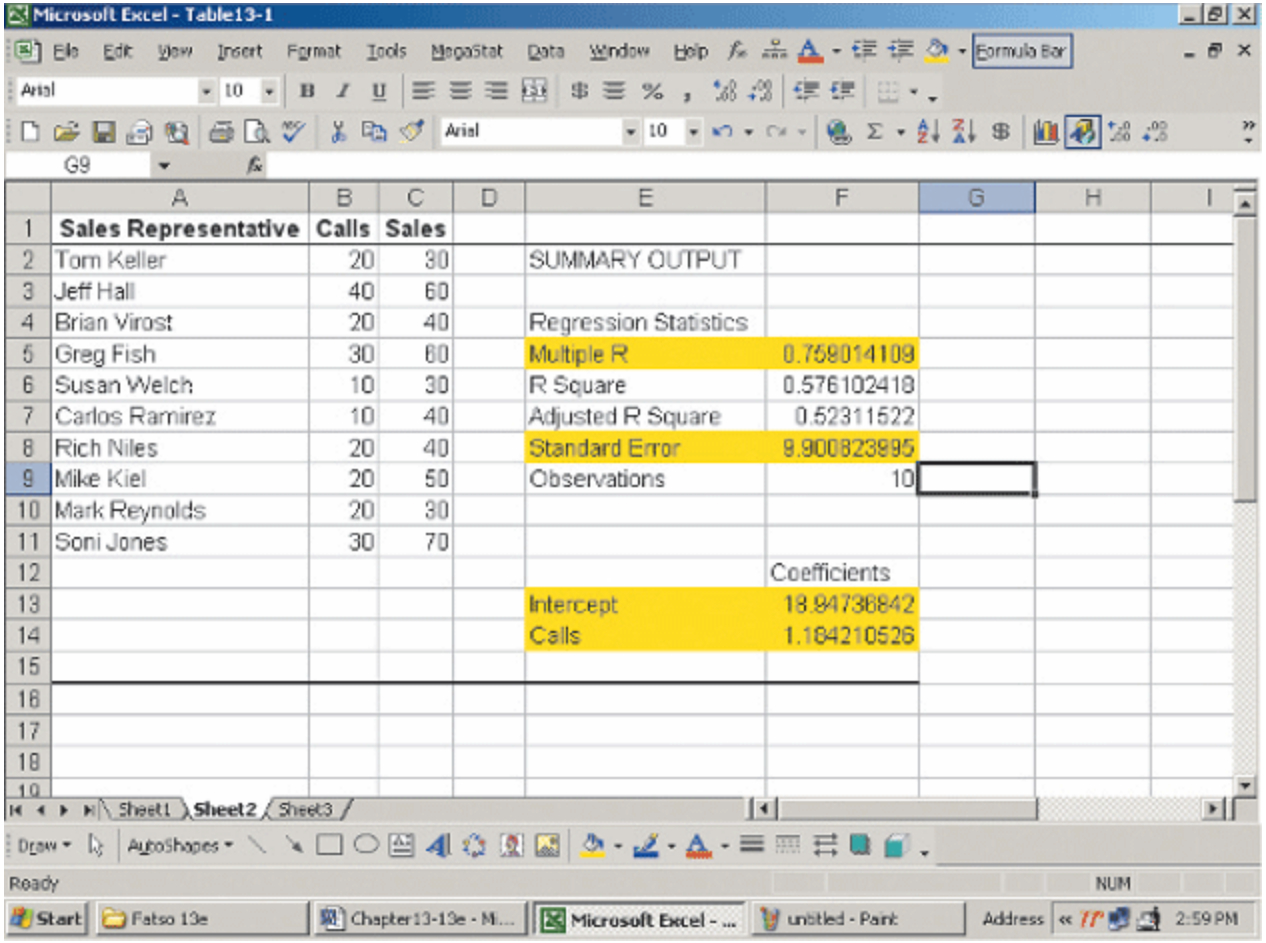
$$= \sqrt{\frac{784.211}{10-2}} = 9.901$$

Graphical Illustration of the Differences between Actual *Y* – Estimated *Y*

$$(Y - \hat{Y})$$



Standard Error of the Estimate - Excel

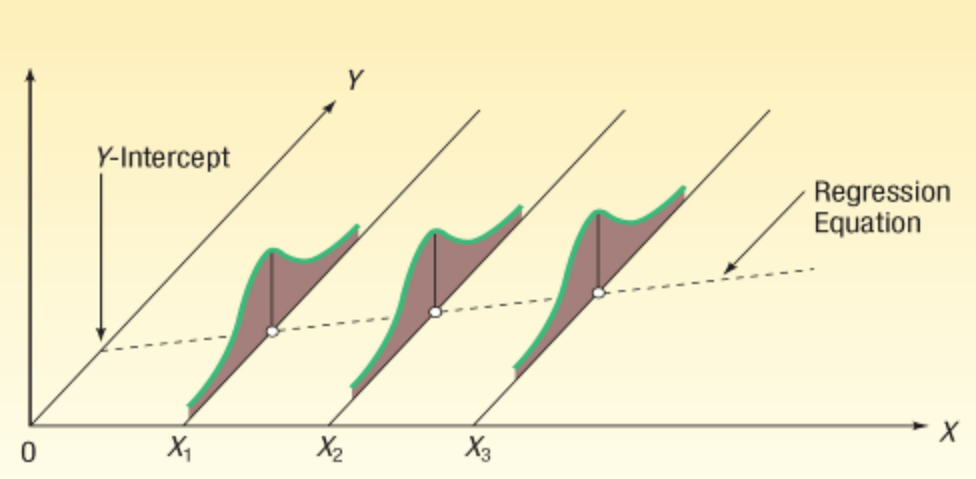


	A	B	C	D	E	F	G	H	I
1	**Sales Representative**	**Calls**	**Sales**						
2	Tom Keller	20	30		SUMMARY OUTPUT				
3	Jeff Hall	40	60						
4	Brian Virost	20	40		Regression Statistics				
5	Greg Fish	30	60		Multiple R	0.759014109			
6	Susan Welch	10	30		R Square	0.576102418			
7	Carlos Ramirez	10	40		Adjusted R Square	0.52311522			
8	Rich Niles	20	40		Standard Error	9.900823995			
9	Mike Kiel	20	50		Observations	10			
10	Mark Reynolds	20	30						
11	Soni Jones	30	70						
12						Coefficients			
13					Intercept	18.84736842			
14					Calls	1.184210526			
15									
16									
17									
18									
19									

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*

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

where

$\hat{Y}$ is the predicted value for any selected X value.
X is any selected value of X.
$\overline{X}$ 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*

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

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*

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

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

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*

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

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*

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

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

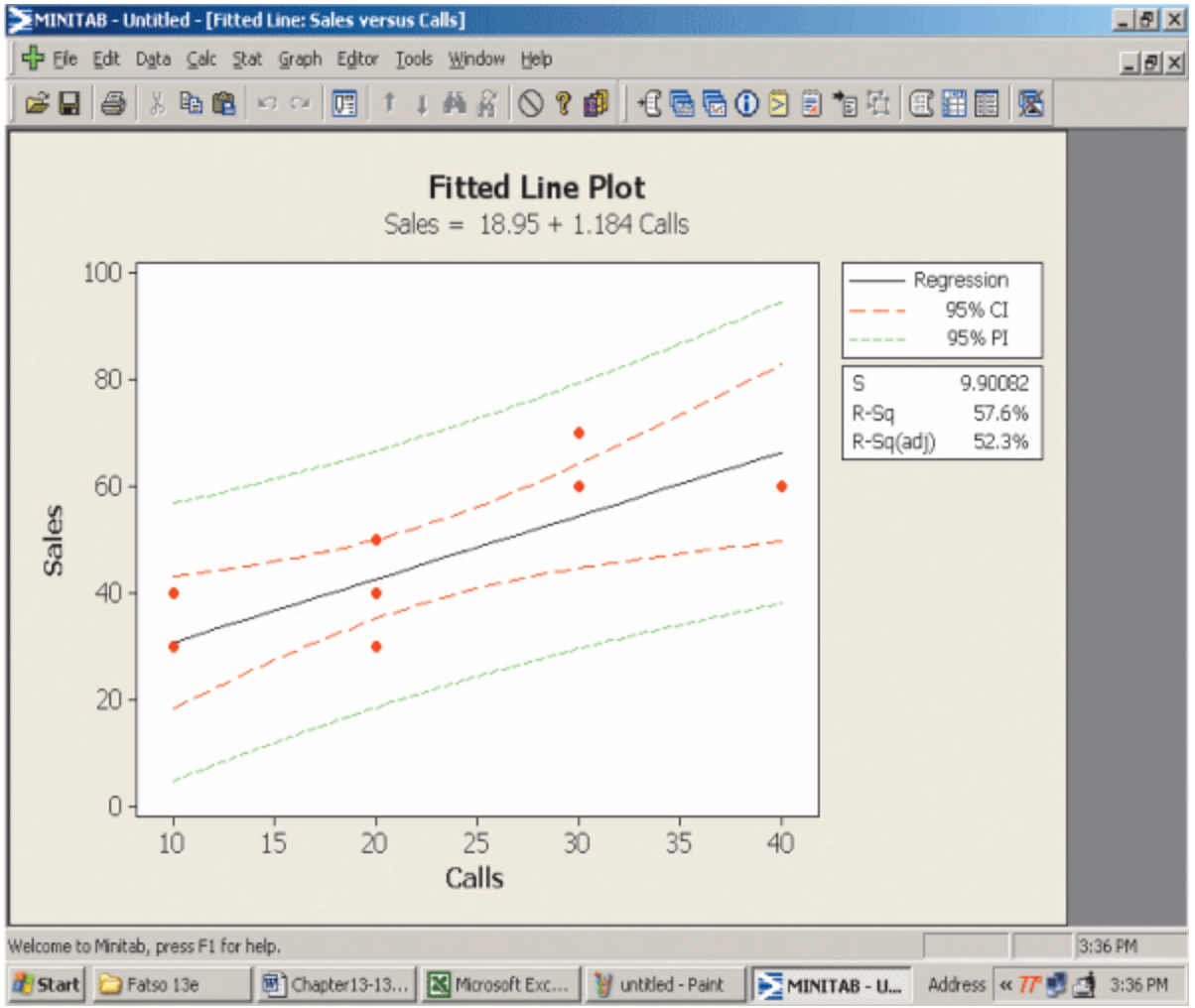
$$\text{Prediction Interval} = \hat{Y} \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{\Sigma(X - \overline{X})^2}}$$

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

$$= 48.5526 \pm 24.0746$$

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



Transforming Data

- The coefficient of correlation describes the strength of the *linear* relationship between two variables. It could be that two variables are closely related, but there relationship is not linear.
- Be cautious when you are interpreting the coefficient of correlation. A value of *r* may indicate there is no linear relationship, but it could be there is a relationship of some other nonlinear or curvilinear form.

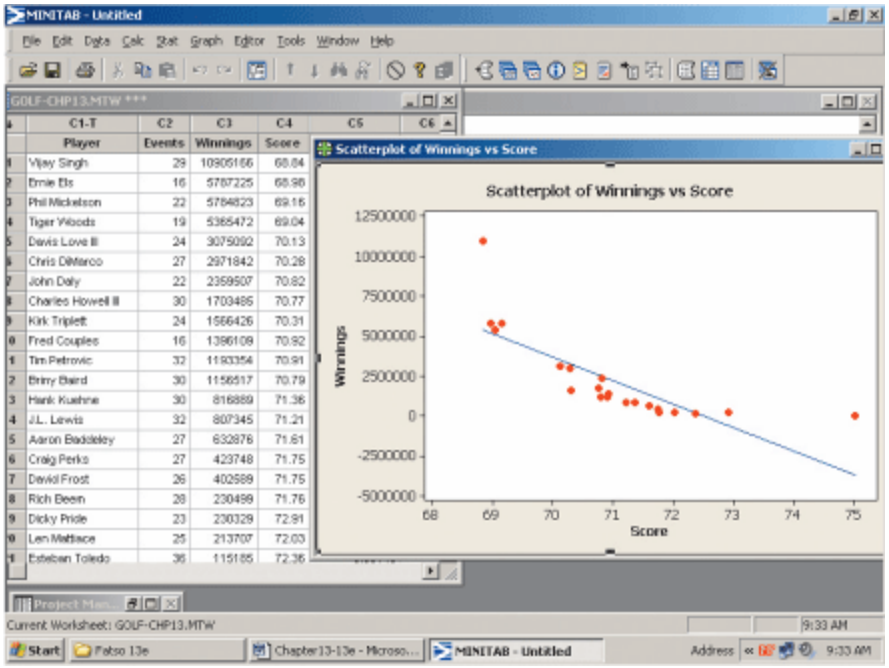
Transforming Data - Example

On the right is a listing of 22 professional golfers, the number of events in which they participated, the amount of their winnings, and their mean score for the 2004 season. In golf, the objective is to play 18 holes in the least number of strokes. So, we would expect that those golfers with the lower mean scores would have the larger winnings. To put it another way, score and winnings should be inversely related. In 2004 Tiger Woods played in 19 events, earned $5,365,472, and had a mean score per round of 69.04. Fred Couples played in 16 events, earned $1,396,109, and had a mean score per round of 70.92. The data for the 22 golfers follows.

Player	Events	Winnings	Score
Vijay Singh	29	$10,905,166	68.84
Ernie Els	16	5,787,225	68.98
Phil Mickelson	22	5,784,823	69.16
Tiger Woods	19	5,365,472	69.04
Davis Love III	24	3,075,092	70.13
Chris DiMarco	27	2,971,842	70.28
John Daly	22	2,359,507	70.82
Charles Howell III	30	1,703,485	70.77
Kirk Triplett	24	1,566,426	70.31
Fred Couples	16	1,396,109	70.92
Tim Petrovic	32	1,193,354	70.91
Briny Baird	30	$ 1,156,517	70.79
Hank Kuehne	30	816,889	71.36
J. L. Lewis	32	807,345	71.21
Aaron Baddeley	27	632,876	71.61
Craig Perks	27	423,748	71.75
David Frost	26	402,589	71.75
Rich Beem	28	230,499	71.76
Dicky Pride	23	230,329	72.91
Len Mattiace	25	213,707	72.03
Esteban Toledo	36	115,185	72.36
David Gossett	25	21,250	75.01

Scatterplot of Golf Data

- The correlation between the variables Winnings and Score is 0.782. This is a fairly strong inverse relationship.
- However, when we plot the data on a scatter diagram the relationship does not appear to be linear; it does not seem to follow a straight line.



What can we do to explore other (nonlinear) relationships?

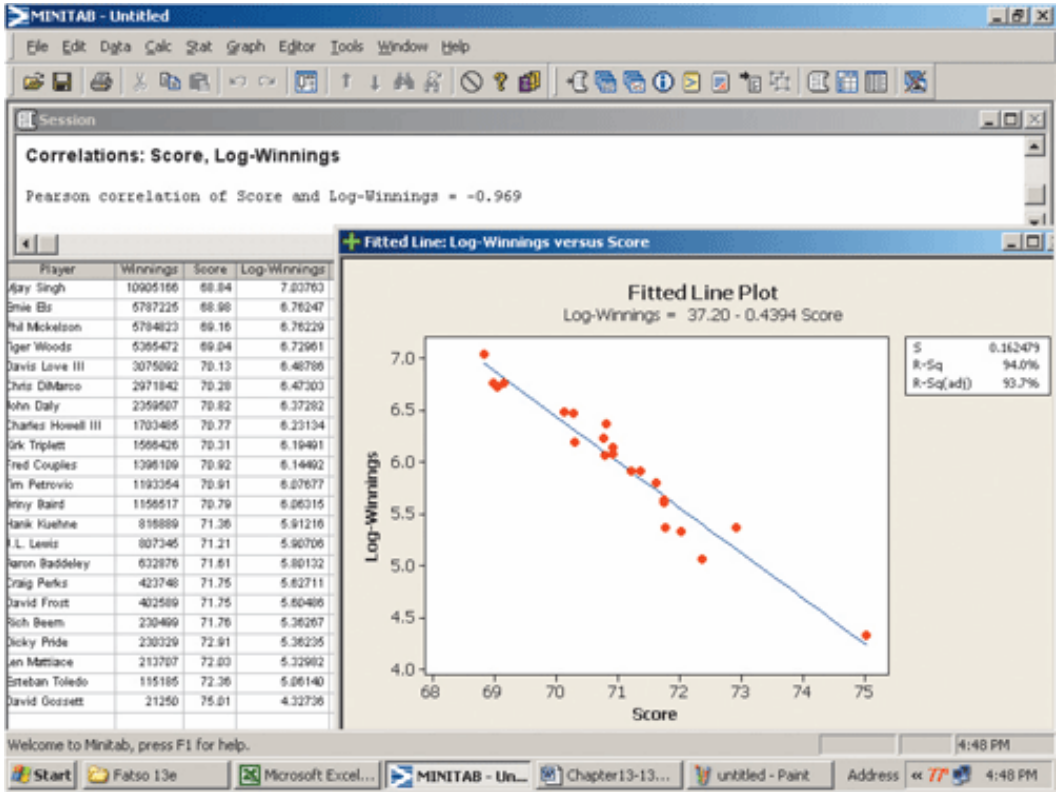
One possibility is to transform one of the variables. For example, instead of using *Y* as the dependent variable, we might use its log, reciprocal, square, or square root. Another possibility is to transform the independent variable in the same way. There are other transformations, but these are the most common.

Transforming Data - Example

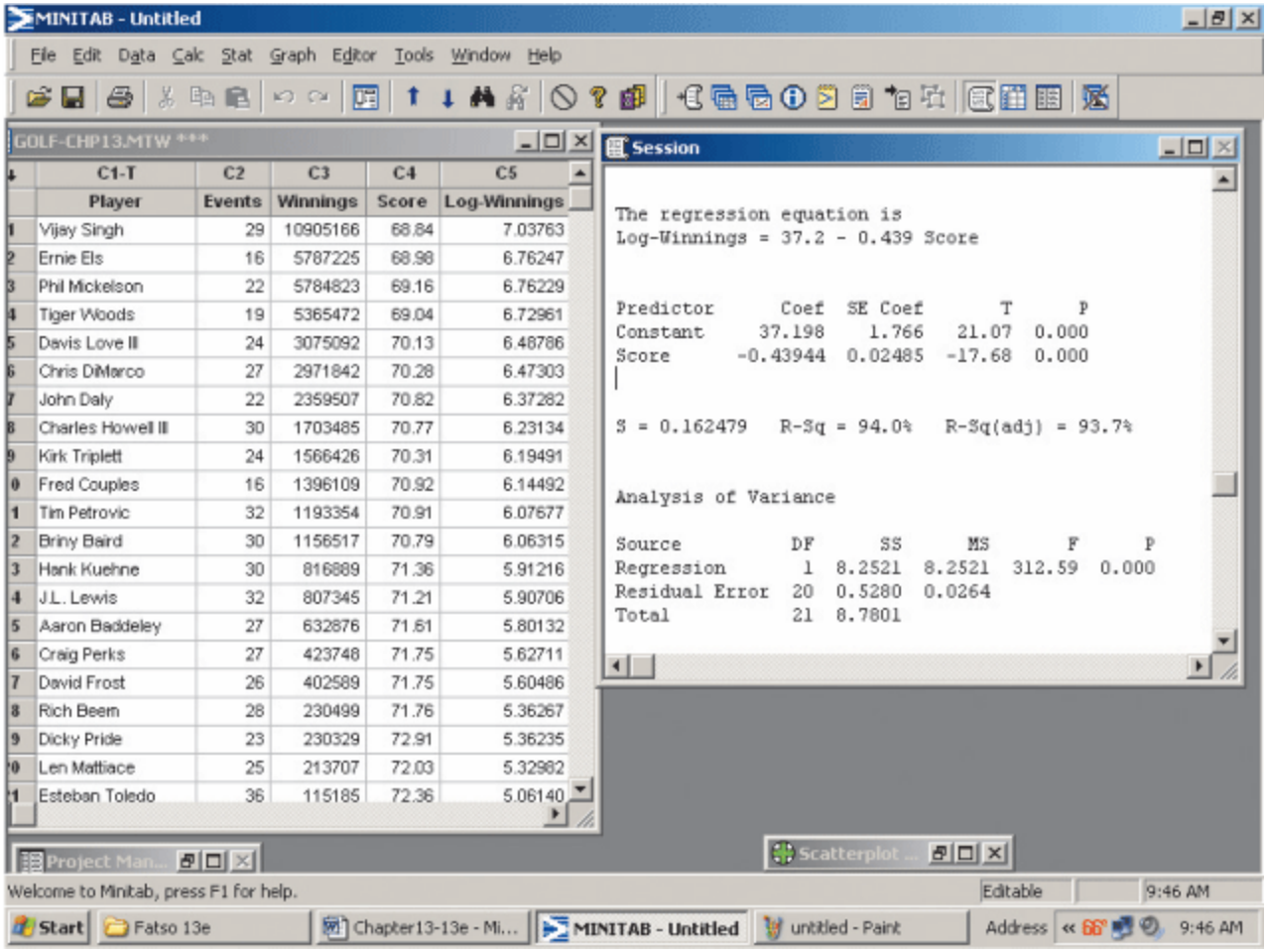
In the golf winnings example, changing the scale of the dependent variable is effective. We determine the log of each golfer's winnings and then find the correlation between the log of winnings and score. That is, we find the log to the base 10 of Tiger Woods' earnings of $5,365,472, which is 6.72961.

Player	Winnings	Score	Log-Winnings
Vijay Singh	10905166	68.84	7.03763
Ernie Els	5787225	68.98	6.76247
Phil Mickelson	5784823	69.16	6.76229
Tiger Woods	5365472	69.04	6.72961
Davis Love III	3075092	70.13	6.48786
Chris DiMarco	2971842	70.28	6.47303
John Daly	2359507	70.82	6.37282
Charles Howell III	1703485	70.77	6.23134
Kirk Triplett	1566426	70.31	6.19491
Fred Couples	1396109	70.92	6.14492
Tim Petrovic	1193354	70.91	6.07677
Briny Baird	1156517	70.79	6.06315
Hank Kuehne	816889	71.36	5.91216
J.L. Lewis	807345	71.21	5.90706
Aaron Baddeley	632876	71.61	5.80132
Craig Perks	423748	71.75	5.62711
David Frost	402589	71.75	5.60486
Rich Beem	230499	71.76	5.36267
Dicky Pride	230329	72.91	5.36235
Len Mattiace	213707	72.03	5.32982
Esteban Toledo	115185	72.36	5.06140
David Gossett	21250	75.01	4.32736

Scatter Plot of Transformed Y



Linear Regression Using the Transformed Y



	C1-T	C2	C3	C4	C5
	Player	Events	Winnings	Score	Log-Winnings
1	Vijay Singh	29	10905166	68.84	7.03763
2	Ernie Els	16	5787225	68.98	6.76247
3	Phil Mickelson	22	5784823	69.16	6.76229
4	Tiger Woods	19	5365472	69.04	6.72961
5	Davis Love III	24	3075092	70.13	6.48786
6	Chris DiMarco	27	2971842	70.28	6.47303
7	John Daly	22	2359507	70.82	6.37282
8	Charles Howell III	30	1703485	70.77	6.23134
9	Kirk Triplett	24	1566426	70.31	6.19491
10	Fred Couples	16	1396109	70.92	6.14492
11	Tim Petrovic	32	1193354	70.91	6.07677
12	Briny Baird	30	1156517	70.79	6.06315
13	Hank Kuehne	30	816889	71.36	5.91216
14	J.L. Lewis	32	807345	71.21	5.90706
15	Aaron Baddeley	27	632876	71.61	5.80132
16	Craig Perks	27	423748	71.75	5.62711
17	David Frost	26	402589	71.75	5.60486
18	Rich Beem	28	230499	71.76	5.36267
19	Dicky Pride	23	230329	72.91	5.36235
20	Len Mattiace	25	213707	72.03	5.32982
21	Esteban Toledo	36	115185	72.36	5.06140

Using the Transformed Equation for Estimation

Based on the regression equation, a golfer with a mean score of 70 could expect to earn:

$$\hat{Y} = 37.198 - .43944X$$
$$= 37.198 - .43944(70) = 6.4372$$

- The value 6.4372 is the log to the base 10 of winnings.
- The antilog of 6.4372 is 2.736
- So a golfer that had a mean score of 70 could expect to earn $2,736,528.

End of Chapter 13