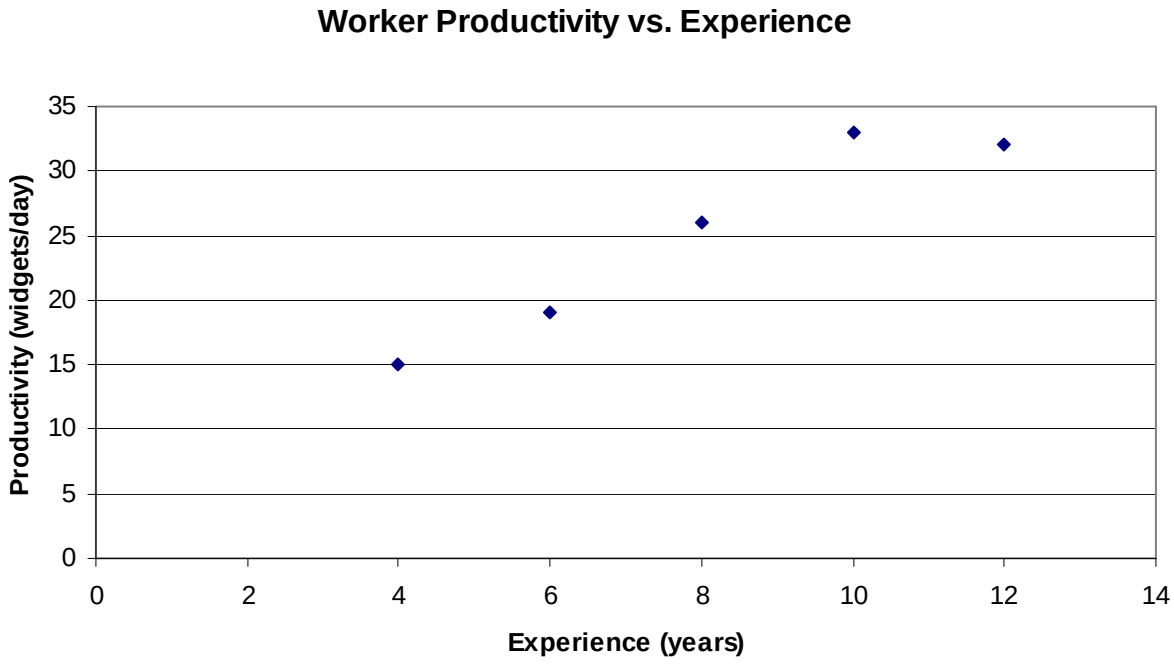
4.1 Scatter Diagrams and Correlation

Definition: The **response (dependent) variable** is the variable whose value can be explained by, or is determined by, the value of the **predictor (independent) variable**.

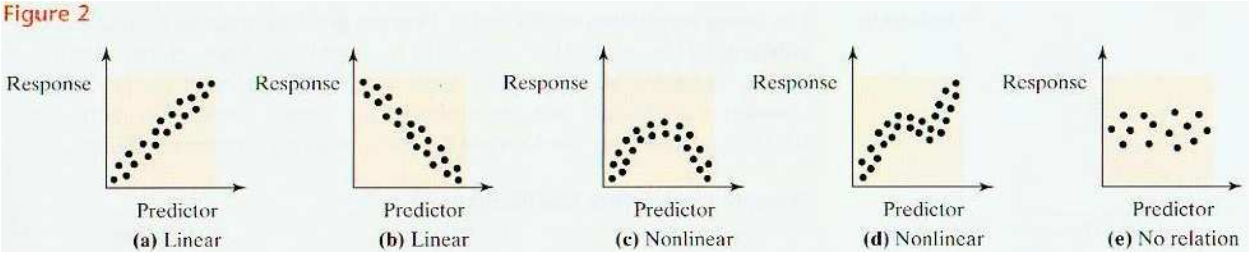
Definition: A **scatter diagram** is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The predictor variable is plotted on the horizontal axis and the response variable is plotted on the vertical axis. Do not connect the points when drawing a scatter diagram.

The data set below represents a random sample of 5 workers in a particular industry. The productivity of each worker was measured at one point in time, and the worker was asked the number of years of job experience. The response variable is productivity, measured in widgets per day, and the predictor variable is experience, measured in years.

Worker	y=Productivity (widgets/day)	x=Experience (years)
1	33	10
2	19	6
3	32	12
4	26	8
5	15	4



Interpreting Scatter Diagrams:



Definition: **Positively associated** variables—when above-average values of one variable are associated with above-average values of the corresponding variable. That is, two variables are positively associated if, whenever the values of the predictor variable increase, the values of the response variable also increase.

Negative associated variables—when above-average values of one variable are associated with below-average values of the corresponding variable. That is, two variables are negatively associated if, whenever the values of the predictor variable increase, the values of the response variable decrease.

Correlation:

Definition: The **linear correlation coefficient** is a measure of the strength of linear relation between two quantitative variables. We use the Greek letter ρ (rho) to represent the population correlation coefficient and r to represent the sample correlation coefficient. The formula for the sample correlation is presented below:

Sample Correlation Coefficient

$$r = \frac{\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}\right)}\sqrt{\left(\sum y_i^{\,2} - \frac{\left(\sum y_i\right)^2}{n}\right)}}$$

Where: $\bar{x}$ is the sample mean of the predictor variable
s_x is the sample standard deviation of the predictor variable
$\bar{y}$ is the sample mean of the response variable
s_y is the sample standard deviation of the response variable
n is the number of individuals in the sample

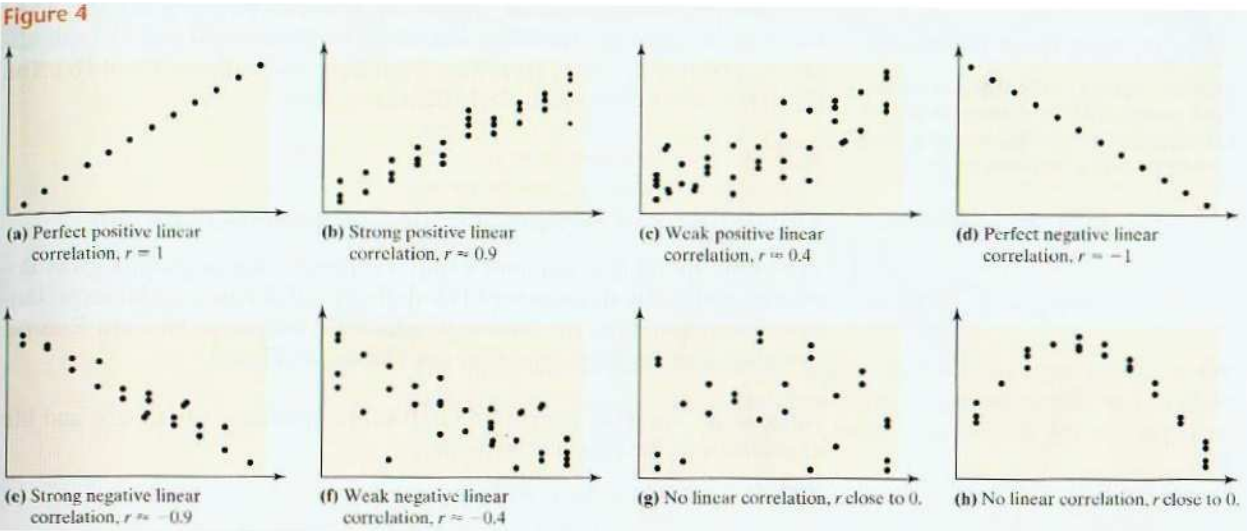
Calculate r using the Productivity-Experience example:

Worker	y=Productivity (widgets/day)	x=Experience (years)	y^2	x^2	x*y
1	33	10			
2	19	6			
3	32	12			
4	26	8			
5	15	4			
	$\sum y = 125$	$\sum x = 40$	$\sum y^2 =$	$\sum x^2 =$	$\sum xy =$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}\right)}\sqrt{\left(\sum y_i^{\,2} - \frac{\left(\sum y_i\right)^2}{n}\right)}}$$

Properties of the Linear Correlation Coefficient:

1. The linear correlation coefficient is always between -1 and +1, inclusive. That is, $-1 \le r \le +1$
2. If r=+1, there is perfect positive linear relation between the two variables.
3. If r=-1, there is perfect negative linear relation between the two variables.
4. The closer r is to +1, the stronger the positive association between the two variables.
5. The closer r is to -1, the stronger the negative association between the two variables.
6. If r is close to 0, there is evidence of no linear relation between the two variables. Because r is a measure of linear relation, a correlation coefficient close to 0 does not imply no relation, just no linear relation.
7. The linear correlation coefficient is a unitless measure of association. So the units of measure for x and y play no role in the interpretation of r.



How would you interpret the correlation coefficient for the **Productivity-Experience** example of r=0.96?

Excel can be used to draw a scatter diagram and calculate a linear correlation coefficient.

Excel—Drawing a scatter diagram

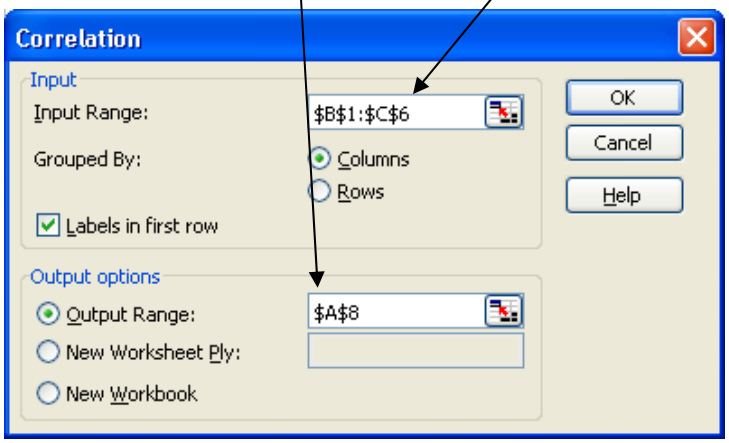
Step 1: Enter the response (dependent) variable in column **B** and the predictor (independent) variable in column **C** in an Excel spreadsheet.
Step 2: Select the **Chart Wizard** icon. Select **Chart Type** of "XY(Scatter)" and follow the instructions.

	A	B	C
1	Worker	y=Productivity (widgets/day)	x=Experience (years)
2	1	33	10
3	2	19	6
4	3	32	12
5	4	26	8
6	5	15	4

Excel—Correlation Coefficient

Step 1: Enter raw data in columns **B** and **C** (the Excel worksheet page is shown below).
Step 2: Select **Tools** from the Windows menu and highlight **Data Analysis.** In the **Analysis Tools** box, highlight "Correlation" and click **OK**. This brings up the **Correlation** box.
Step 3: With the cursor in the "Input Range" box, highlight the data in columns **B** and **C** (include the column headings and check the "Labels in first row" box). In the "Output Range" box, specify a cell for the output (upper left corner of the output range). Click **OK**. The correlation coefficient, r, will appear as the lower off-diagonal element in a 2x2 matrix.

	A	B	C
1	Worker	y=Productivity (widgets/day)	x=Experience (years)
2	1	33	10
3	2	19	6
4	3	32	12
5	4	26	8
6	5	15	4
7			
8		*y=Productivity (widgets/day)*	*x=Experience (years)*
9	y=Productivity (widgets/day)	1	
10	x=Experience (years)	0.96	1



4.2 Least-Squares Regression

In this section we will learn how to fit a regression model to a scatter diagram of data. Our focus is on regression models that produce straight (or linear) lines. Fitting a regression line involves determining the equation for the line that "best" represents the data. Remember, a straight line has two characteristics—(1) y-intercept and (2) slope. Thus, our goal will be to find values for the y-intercept and the slope of the regression line.

In inferential statistics, we are concerned with estimating population parameters based on statistics from a sample. Likewise, in regression analysis we have a **population model**, which represents the true linear relation between y and x for the entire population of data, and a **statistical model**, which represents the linear relation between y and x in the sample data. By applying statistical inference, the linear relation from the sample is used as an estimate of the unknown (true) population relation.

Population Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

where y_i = dependent (response) variable
x_i = independent (predictor) variable
β_0 and β_1 are the intercept and slope parameters, respectively
ε_i = random error term with mean 0 and variance σ_ε^2
Note that ε_i accounts for factors, other than x, that affect y.

Statistical Model: $y_i = b_0 + b_1 x_i + u_i$

where y_i = dependent (response) variable
x_i = independent (predictor) variable
b_0 and b_1 are the intercept and slope statistics, respectively
u_i = estimated error (or disturbance) term that accounts for factors, other than x, that affect y

Coefficient	Statistic		Parameter
Intercept	b_0	$\longrightarrow$	β_0
Slope	b_1	$\longrightarrow$	β_1

The statistics, b_0 and b_1, serve as proxies (estimates) for the unknown population parameters, β_0 and β_1.

The least-squares regression coefficients (b_0 and b_1) are calculated based on the Least-Squares Regression Criterion.

Definition: **Least-Squares Regression Criterion**

The **least-squares regression line** is the one that minimizes the sum of the squared errors. It is the line that minimizes the square of the vertical distance between observed values of y and those predicted by the line, $\hat{y}$ (read "y-hat"). We represent this as

Minimize Σ (errors^2)

Application of the Least-Squares Regression Criterion produces the **Normal Equations** shown below:

(1) $b_0 n + b_1 \Sigma x = \Sigma y$

(2) $b_0 \Sigma x + b_1 \Sigma x^2 = \Sigma xy$

These equations can be solved to find formulas for b_0 and b_1 (the solution is done in reverse order).

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

The final least-squares regression equation is written as: $y = b_0 + b_1 x + u$

Example—Application of Least-Squares Regression for Productivity-Experience Data

Worker	y=Productivity (widgets/day)	x=Experience (years)	x^2	$x*y$
1	33	10	100	330
2	19	6	36	114
3	32	12	144	384
4	26	8	64	208
5	15	4	16	60
	$\Sigma y=125$	$\Sigma x=40$	$\Sigma x^2=360$	$\Sigma xy=1,096$
	$\bar{y}=25$	$\bar{x}=8$		

Calculate the slope coefficient (b_1) and the intercept coefficient (b_0) using the least-squares formulas:

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} =$$

$$b_0 = \bar{y} - b_1\bar{x} =$$

Alternative way to calculate b_1: If you have already calculated r, b_1 can be found using the formula $b_1 = r * \frac{s_y}{s_x}$, where s_y and s_x are, respectively, the standard deviations of the observations in the y-sample and the x-sample (see pp. 198, text).

The estimated least-squares regression equation is written as: $y =$ ___+ ___ $x + u$

Interpretation of the estimated regression coefficients:

b_0=____: _____

_____

Scope of the Model
The y-intercept, or b_0, will have meaning only if the following two conditions are met:
(1) A value of 0 for the independent (predictor) variable makes sense;
(2) There are observed values of the independent variable near 0.
The second condition is particularly important because statisticians do not use the regression model to make predictions **outside the scope of the model** (more on this later when we discuss using the regression model to predict).

b_1=___ (Remember the slope equals the change in y divided by the change in x, i.e., $b_1 = \frac{\Delta y}{\Delta x}$.)

_____

_____

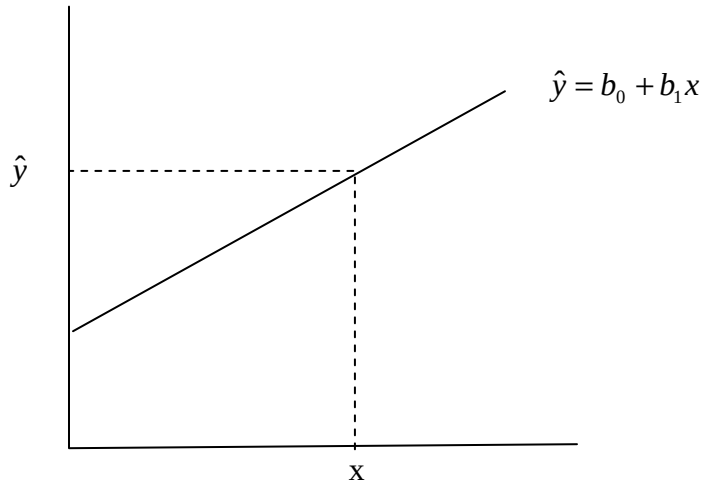
Prediction: The estimated least-squares regression equation can be used for prediction by setting u=0 in the *Statistical Model*:

$y = b_0 + b_1 x + u$

$\hat{y} = b_0 + b_1 x + 0$ Note: When u is set to 0, a hat is put on the y-variable.

$\hat{y} = b_0 + b_1 x$

where $\hat{y}$ is the predicted value of y.



Productivity-Experience Example: $\hat{y} = 5.8 + 2.4x$

(1) Predict worker productivity for a worker with 7 years of experience.

(2) Predict worker productivity for a worker with 20 years of experience.

Scope of the Regression Model

Statisticians do not use the regression model to make predictions **outside the scope of the model**, i.e., they do not use the regression model to make predictions for values of the independent (predictor) variable that are much larger or smaller than those observed in the dataset. This rule is followed because we are not certain of the behavior of the line outside the range of the dataset used to estimate the regression model.

For example, it is not appropriate to use the line to predict the productivity of a worker with 20 years of experience. The highest value of the independent variable (in the dataset used to estimate the regression model) is 12 years of experience and we cannot be certain that the linear regression line will continue out to 20 years of experience. It is likely at high levels of experience that productivity will level off as shown in the figure below.

Accuracy of Predictions from the Least-Squares Regression Model:

Are the predictions from a regression model perfectly accurate?—The answer is NO. Let's examine how we predict with a regression model and why the predictions are usually not perfectly accurate. First, the prediction from a regression model is represented by the equation

$$\hat{y} = b_0 + b_1 x$$

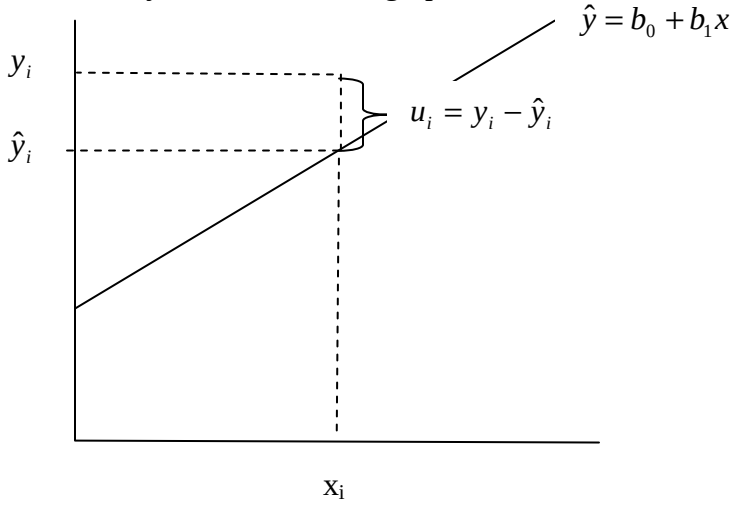
The statistical regression model is written as

$$y = b_0 + b_1 x + u$$

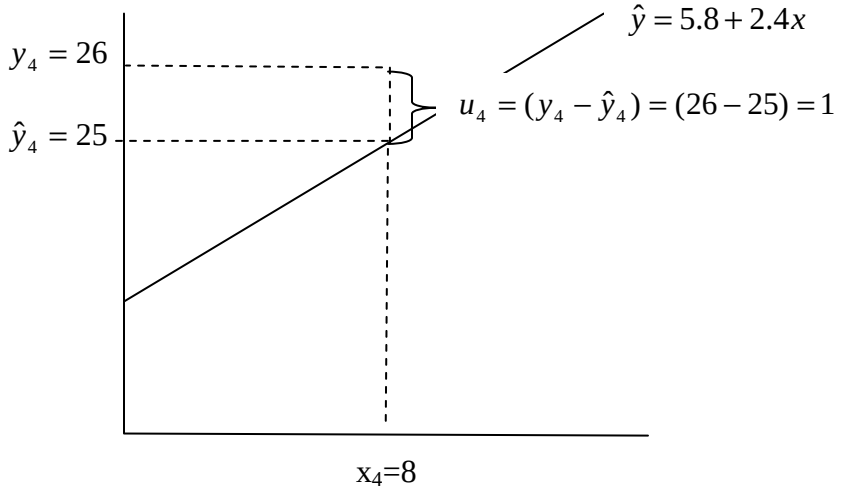
$$y = \hat{y} + u$$

.

The inaccuracy (or error) of a prediction is represented by $u = y - \hat{y}$. The term "u" is referred to as an error term that represents all the factors, other than x, that account for (or explain) y. The error term, u, is represented as the vertical distance between the observed value of y and the predicted value of y, as shown in the graph below.



To get a better understanding of the error term, let's calculate an error for the **Productivity-Experience** example. Take the 4th worker with x=8 years of experience, and calculate the worker's predicted productivity: $\hat{y}_4 = 5.8 + 2.4(8) = 25\, widgets/day$. The actual productivity of the 4th worker was 26 widgets/day. The error in this case is u_4=26-25=1 widget/day. The error is represented in the graph below.



Why do errors occur when making predictions?

Let's think about why our prediction of productivity for the 4th worker was inaccurate. That is, why did the 4th worker produce 26 widgets/day when the regression line predicted that he would produce 25 widgets/day? Remember that $\hat{y} = 25$ represents the productivity of the "average worker" with 8 years experience. The u, or the difference between 26 and 25, is due to factors, other than x, that account for y. For example, the 4th worker could work harder than the average worker. Or, the 4th worker could have more innate talent, or have many other characteristics—that are different from those of the average worker—that account for the higher productivity,

In short, the 4th worker's productivity includes two components: (1) his predicted productivity based on years of experience, and (2) an adjustment in productivity (which can be positive or negative) that is due to the particular "other" characteristics of the 4th worker such as work ethic, ability, etc. The sum total of the two components account for the 4th worker's total productivity, i.e.,

$$y_4 = \hat{y}_4 + u_4$$
$$26 = 25 + 1$$

In the application of the regression model to real-world, scientific data, there will always be positive and negative u's that account for unknown factors, other than x, that explain the y variable. This is true for any type of scientific relation which involves a complicated process. Only when scientists know everything, will the u's go away…and that is a long way, or infinitely, out into the future.

How well does the regression line fit the data?

Two measures are widely used to measure the fit of a regression line—(1) coefficient of determination, and (2) standard error of the estimate. We will discuss each in turn.

Coefficient of Determination:

The coefficient of determination provides a quantitative measure of how well the regression line fits the scatter plot. The coefficient of determination measures the strength of relation that exists between two variables.

Definition: The **coefficient of determination, R^2**, measures the proportion of the total variation in the dependent (response) variable that is explained by the least-squares regression line.
R^2 is a number between 0 and 1, inclusive, i.e., $0 \le R^2 \le 1$. If $R^2 = 0$, the regression line has no explanatory power; if $R^2 = 1$, the regression line explains 100% of the variation in the dependent variable.

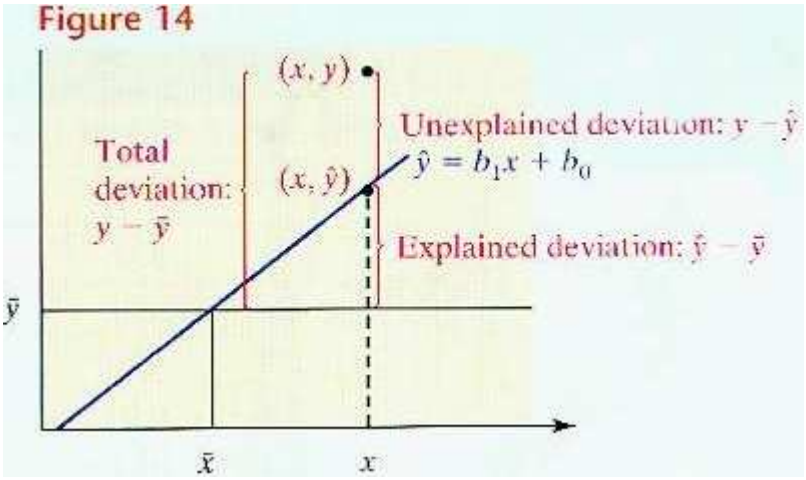
To understand the coefficient of determination, we need to understand three types of deviation.

(1) Total deviation = $y - \bar{y}$

(2) Explained deviation = $\hat{y} - \bar{y}$

(3) Unexplained deviation = $y - \hat{y}$

Figure 14



From the figure above notice that total deviation is the sum of the explained deviation and the unexplained deviation:

Total deviation = Explained deviation + Unexplained deviation

$y - \bar{y}$ $\qquad$ $\hat{y} - \bar{y}$ $\qquad$ $y - \hat{y}$

In statistics we are generally concerned with squared deviations…Remember that variance in Ch. 3 was defined as the average of the squared deviations from the mean. Likewise, here we are concerned with squared deviations, which are referred to as *variation*. Although beyond the scope of this course, it can be shown that there are three sources of variation:

(1) total variation, $\sum (y - \bar{y})^2$

(2) explained variation, $\sum (\hat{y} - \bar{y})^2$

(3) unexplained variation, $\sum (y - \hat{y})^2$

Furthermore, total variation is equal to the sum of explained variation and unexplained variation, i.e.,

Total variation = Explained variation + Unexplained variation

If we divide this equation by total variation, we get

$$1 = \frac{\text{explained variation}}{\text{total variation}} + \frac{\text{unexplained variation}}{\text{total variation}}$$

Using algebra we can solve for R^2:

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

Calculate R^2 based on the correlation coefficient, r: This approach involves first calculating the correlation coefficient and then squaring r to find R^2. For the **Productivity-Experience** example, we found r=0.96, and therefore, $R^2=(0.96)^2=0.92$.

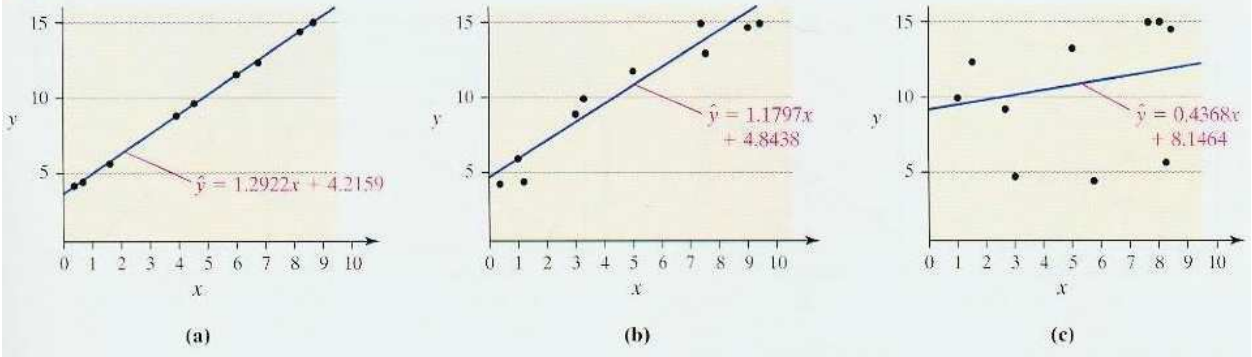
An R^2 value of 0.92 says that 92% of the variation in productivity (y) is explained by the linear relation with experience (x). This leaves 8% of the variation in worker productivity left to be explained by other factors (not x) such as work ethic, ability, health condition, etc.

Relating R^2 to Scatter Diagrams:

TABLE 5

DATA SET A		DATA SET B		DATA SET C	
x	y	x	y	x	y
3.6	8.9	3.1	8.9	2.8	8.9
8.3	15.0	9.4	15.0	8.1	15.0
0.5	4.8	1.2	4.8	3.0	4.8
1.4	6.0	1.0	6.0	8.3	6.0
8.2	14.9	9.0	14.9	8.2	14.9
5.9	11.9	5.0	11.9	1.4	11.9
4.3	9.8	3.4	9.8	1.0	9.8
8.3	15.0	7.4	15.0	7.9	15.0
0.3	4.7	0.1	4.7	5.9	4.7
6.8	13.0	7.5	13.0	5.0	13.0

Figure 15



The coefficients of determination (R^2's) for the three datasets are: 1.0, 0.947, and 0.094, respectively.

Standard Error of the Estimate:

The variance of the error terms (ε_i) from the **population regression model** is represented by σ_ε^2, which is a population parameter. Since a population parameter is unknown to a practicing statistician, it must be estimated using a sample statistic. In this case, the standard error of the estimate, s_e, serves as the statistic used to estimate σ_ε. The **standard error of the estimate** is found using the formula:

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum u_i^2}{n-2}} = \sqrt{\frac{\sum error^2}{n-2}}$$

Remember that y_i = observed value of y for the ith individual

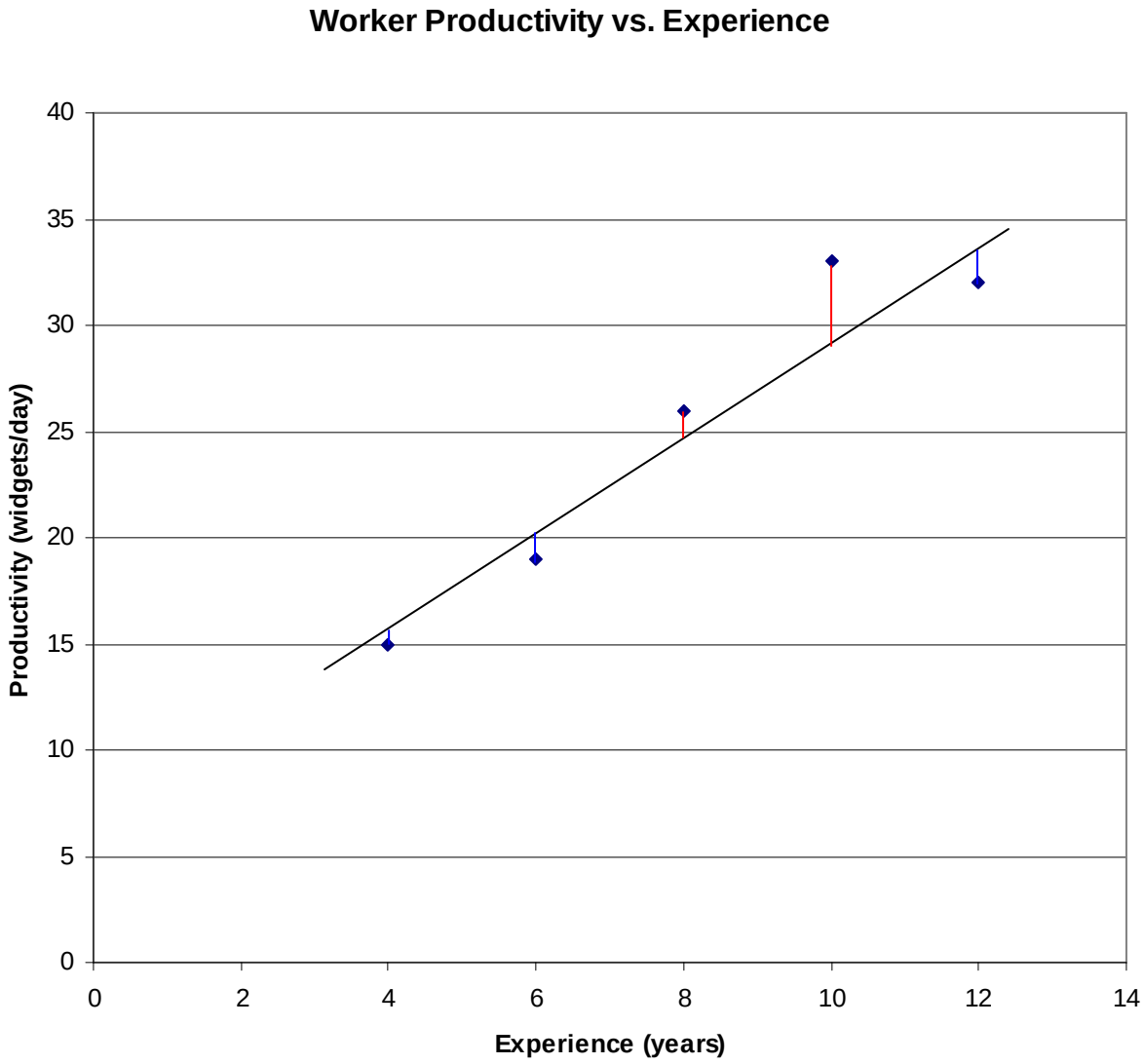
$\hat{y}_i$ = predicted value of y for the ith individual

$u_i = y_i - \hat{y}_i$ or prediction error for the ith individual

The standard error of the estimate measures the standard deviation of the scatter points about the least-squares regression line.

To develop a better understanding of the standard error of the estimate, let's calculate s_e for the **Productivity-Experience** example. Carry out the calculations below.

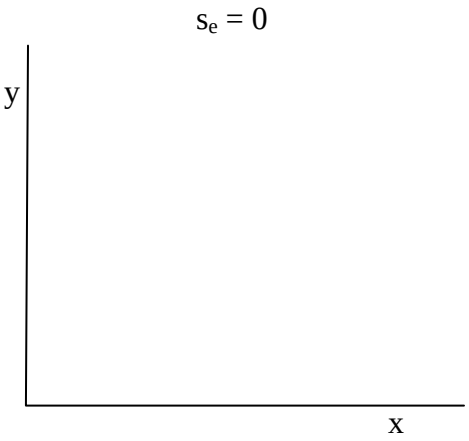
Worker	Y=Productivity (widgets/day)	X=Experience (years)	$\hat{y}$	$u_i = y_i - \hat{y}_i$	u_i^2
1	33	10			
2	19	6			
3	32	12			
4	26	8			
5	15	4			
				Σu_i	Σu_i^2=



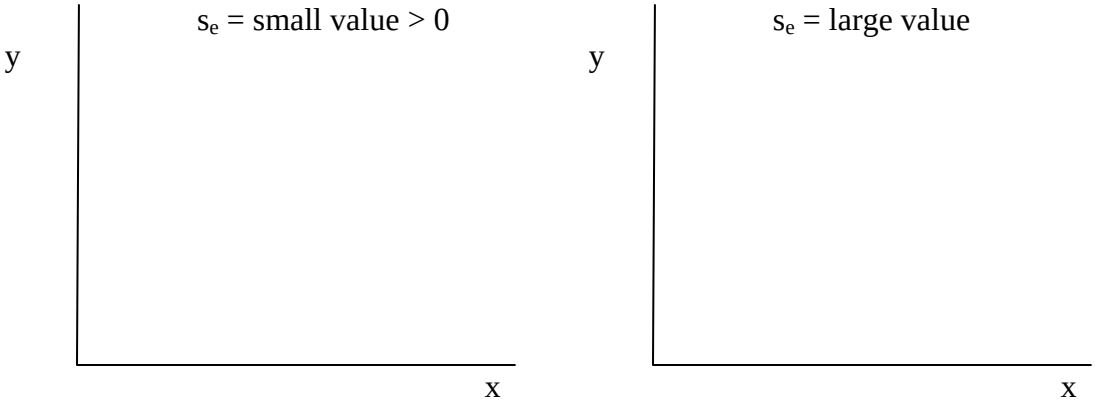
$\Sigma u_i = 0$: The sum of the vertical distance the points are off of the regression line is 0. That is, the sum of the 2 red segments (positive errors) is exactly equal to the sum of the 3 blue segments (i.e., the absolute value of the blue errors).

Σu_i^2 is minimized by Least-Squares Regression Criterion: The least-squares regression line produces the minimum sum of squared deviations of any possible linear line. That is, if you were to draw any other line (with a different b_0 and b_1 than that of the least-squares line) the sum of the squared deviations from the line would be greater than from the least-squares regression line. See "Least-Squares Criterion Revisited," p. 198.

How do you interpret the standard error of the estimate? At one extreme is the case where $s_e=0$. In this situation, all the scatter points in the scatter diagram fall on a linear line and there is no error in any prediction.



In real-world situations, there generally are errors in the predictions and the scatter points in the scatter diagram fall off of the least-squares regression line. The more scattered the points are about the regression line, the higher the s_e. For example, the scatter diagram to the right has the highest s_e.



Hypothesis Test Regarding the Slope Coefficient, β_1.

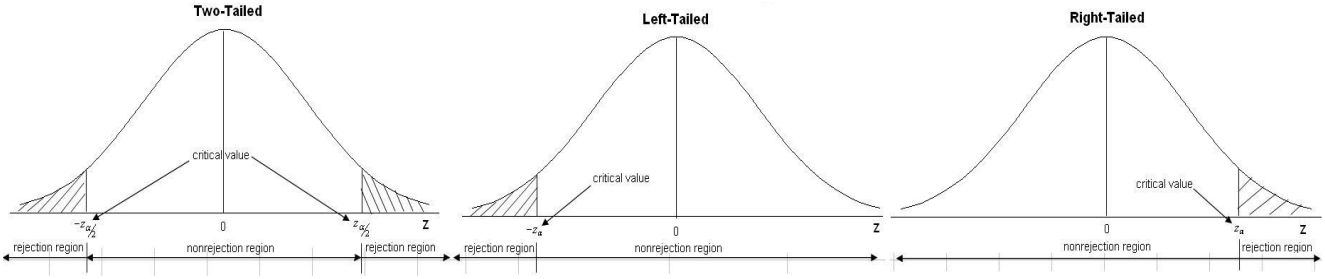
Assumptions:

- A random sample is drawn.
- The error terms are normally distributed with mean 0 and constant variance σ_ε^2.

Step 1: A claim is made regarding the linear relation between a response (dependent) variable, y, and a predictor (independent) variable, x. The null hypothesis is most often specified as no relation between y and x, i.e., $\beta_1=0$. The hypotheses can be structured in one of three ways.

Two-Tailed	Left-Tailed	Right-Tailed
H_0: $\beta_1 = 0$	H_0: $\beta_1 = 0$	H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$	H_1: $\beta_1 < 0$	H_1: $\beta_1 > 0$
$\beta_1 < 0$ or $\beta_1 > 0$		

Step 2: Choose a significance level, α. The level of significance is used to determine the **critical t-value**, with (n-2) degrees of freedom. The rejection (critical) region is the set of all values of the test statistic (from step 3 below) such that the null hypothesis is rejected.



Step 3: Compute the test statistic: $t = \dfrac{b_1 - \beta_1}{s_{b_1}}$, which follows Student's t-distribution with df=n-2. The estimated standard deviation of b_1 is $s_{b_1} = \dfrac{s_e}{\sqrt{\sum \left(x_i - \bar{x}\right)^2}}$ where s_e is the standard error of the estimate.

Step 4: Draw a conclusion:

- Compare the calculated t-value (or test statistic) to the critical t-value and state whether or not H_0 is rejected at the specified α.

Two-Tailed	Left-Tailed	Right-Tailed
If $t < -t_{\alpha/2}$ *or* $t > t_{\alpha/2}$ *reject the null hypothesis*	*If* $t < -t_\alpha$ *reject the null hypothesis.*	*If* $t > t_\alpha$ *reject the null hypothesis.*

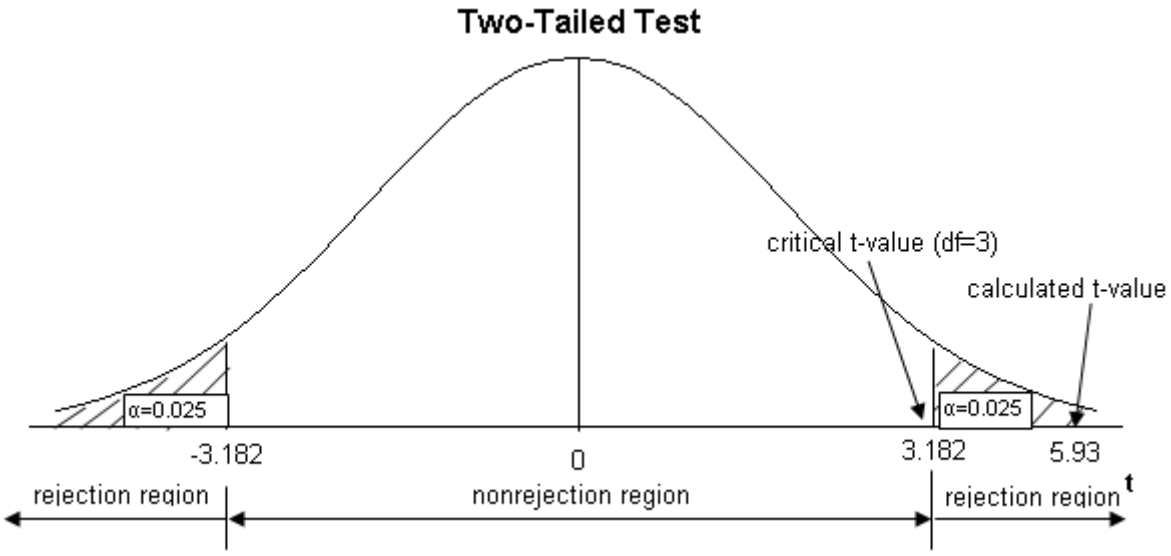
- Interpret the conclusion in the context of the problem

Hypothesis Test Regarding the Slope Coefficient, β_1.

Problem: A labor economist would like to measure the relation between the productivity of workers and the number of years of experience of the workers. A random sample of 5 workers is chosen. A random work day is chosen and the productivity of each worker is measured and recorded. Work experience (in years) is taken from each worker. Carry out a hypothesis test of the relevant null hypothesis at the 5% significance level. Show all of your calculations and justify your conclusion.

Step 1: H_0: $\beta_1 = 0$ (No relation exists between worker productivity and experience.)
H_1: $\beta_1 \neq 0$ (A positive or negative relation exists between productivity and experience.)

Step 2: Select $\alpha = 0.05$ and find the critical value of t (df=5-2=3).



Step 3: Draw a random sample of n=5 workers and collect data on each worker's productivity and experience. Calculate the standard deviation of b_1, s_{b_1}, and the test statistic, *t*.

Worker	y=Productivity (widgets/day)	x=Experience (years)
1	33	10
2	19	6
3	32	12
4	26	8
5	15	4

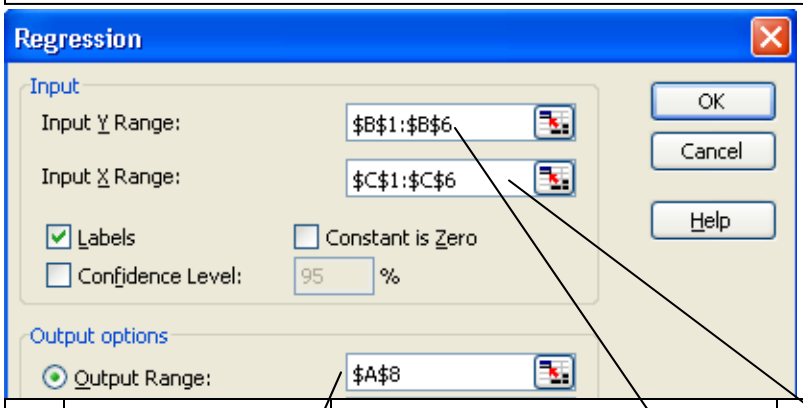
$$t = \frac{(b_1 - \beta_1)}{s_{b_1}} = \frac{(2.4 - 0)}{0.405} = \frac{2.4}{0.405} = 5.93$$

where $s_{b_1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{2.56}{\sqrt{40}} = 0.405$

Step 4: Conclusion—Because the calculated t = 5.93 is greater than the critical t = 3.182 (and in the rejection region), reject H_0 at the 0.05 significance level. A significant positive relation exists between productivity and experience.

Excel—Estimating a linear regression model

Step 1: Enter raw data in columns B and C (the Excel worksheet page is shown below).
Step 2: Select **Tools** from the Windows menu and highlight **Data Analysis.** In the **Analysis Tools** box, highlight "Regression" and click **OK**. This brings up the **Regression** box.
Step 3: With the cursor in the "Input Y Range" box, highlight the data in column **B** (including the column heading at the top of the column). Move the cursor to the "Input X Range" box and highlight the data in column **C**. Check the "Labels" box.
Step 4: Under Output Options, move the cursor to the "Output Range" box and specify a cell for the output (upper left corner of the output range). Click **OK** and the regression output results will appear as shown below.



	A	B	C	D	E
1	**Worker**	**y=Productivity (widgets/day)**	**x=Experience (years)**		
2	1	33	10		
3	2	19	6		
4	3	32	12		
5	4	26	8		
6	5	15	4		
7					
8	**SUMMARY OUTPUT**				
9					
10	***Regression Statistics***				
11	**Multiple R**	0.96			
12	**R Square**	0.9216			
13	**Adjusted R Square**	0.895466667			
14	**Standard Error**	2.556038602			
15	**Observations**	5			
16					
17	**ANOVA**				
18		***df***	***SS***	***MS***	***F***	***Signif F***
19	**Regression**	1	230.4	230.4	35.26531	0.009546
20	**Residual**	3	19.6	6.533333		
21	**Total**	4	250			
22						
23		***Coefficients***	***Standard Error***	***t Stat***	***P-value***	
24	**Intercept (b_0)**	5.8	3.42928564	1.691314	0.189355	
25	**x=Experience (b_1)**	2.4	s_{b1}=0.40414518	5.93846	0.009546	

CHAPTER 4 Describing the Relation between Two Variables

- Correlation Coefficient: $r = \dfrac{\sum\left(\dfrac{x_i - \bar{x}}{s_x}\right)\left(\dfrac{y_i - \bar{y}}{s_y}\right)}{n-1}$

$$r = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sqrt{\left(\sum x_i^{\,2} - \dfrac{\left(\sum x_i\right)^2}{n}\right)}\sqrt{\left(\sum y_i^{\,2} - \dfrac{\left(\sum y_i\right)^2}{n}\right)}}$$

- The equation of the least-squares regression line is

  $\hat{y} = b_0 + b_1 x$ where $\hat{y}$ is the predicted value, $b_1 = r\dfrac{s_y}{s_x}$

  is the slope, and $b_0 = \bar{y} - b_1\bar{x}$ is the intercept.

- Residual = observed y – predicted y = $y - \hat{y}$.

- Coefficient of Determination: R^2 = the percent of total variation in the response variable that is explained by the least-squares regression line.

- $R^2 = r^2$ for the least-squares regression model.

- Least-Squares Regression Normal Equations:
  (1) $b_0 n + b_1\Sigma x = \Sigma y$

  (2) $b_0\Sigma x + b_1\Sigma x^2 = \Sigma xy$

- Solution of normal equations for b_0 and b_1:

  $b_1 = \dfrac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$

  $b_0 = \bar{y} - b_1\bar{x}$

- The final least-squares regression equation is written as:
  $y = b_0 + b_1 x + u$