Chapter 12: Linear Regression and Correlation

Introduction

Of all the scenarios that you have studied thus far, you have yet to study the relationships between two quantitative variables. This scenario is the subject of the final unit, which will again progress from graphical displays to numerical summaries and then proceed to mathematical models and statistical inference.

We will use three tools to describe, picture, and quantify the relationship between two quantitative variables:

- **Scatterplot** - a two-dimensional graph of data values
- **Correlation** - a statistic that measures the *strength and direction* of a linear relationship between two quantitative variables
- **Regression equation** - an equation that describes the average relationship between a quantitative response variable and an explanatory variable

12.1: Linear Equations

- In the equatoin $y = ax + b$
  - a represents the **slope** and
  - b represents the y **- intercept**.
- The **slope** of a line is defined by the change in y divided by the change in x.
- The y **- intercept** is the value of y when $x = 0$.

12.2: Scatterplots

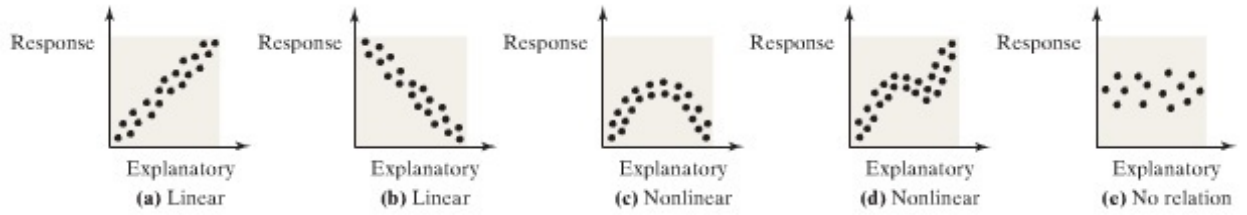
- **Scatterplot** - a two-dimensional graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram.
- **Explanatory variable** or **predictor variable** - plotted on the horizontal (x) axis, used as input and can help explain another variable.
- **Response variable** - plotted on the vertical (y) axis whose value can be explained by the explanatory variable.

Two variables that are linearly related are said to be **positively associated** if, whenever the value of one variable increases, the value of the other variable also increases.

Two variables that are linearly related are said to be **negatively associated** if, whenever the value of one variable increases, the value of the other variable also decreases.

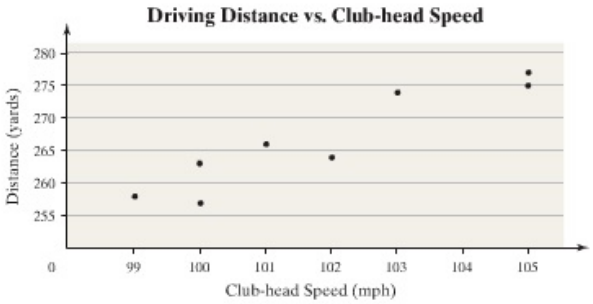
Two variables have a **linear relationship** when the pattern of their relationship resembles a straight line.

If a scatterplot shows a curved line instead of a straight line, the relationship is called **nonlinear** or **curvilinear**.



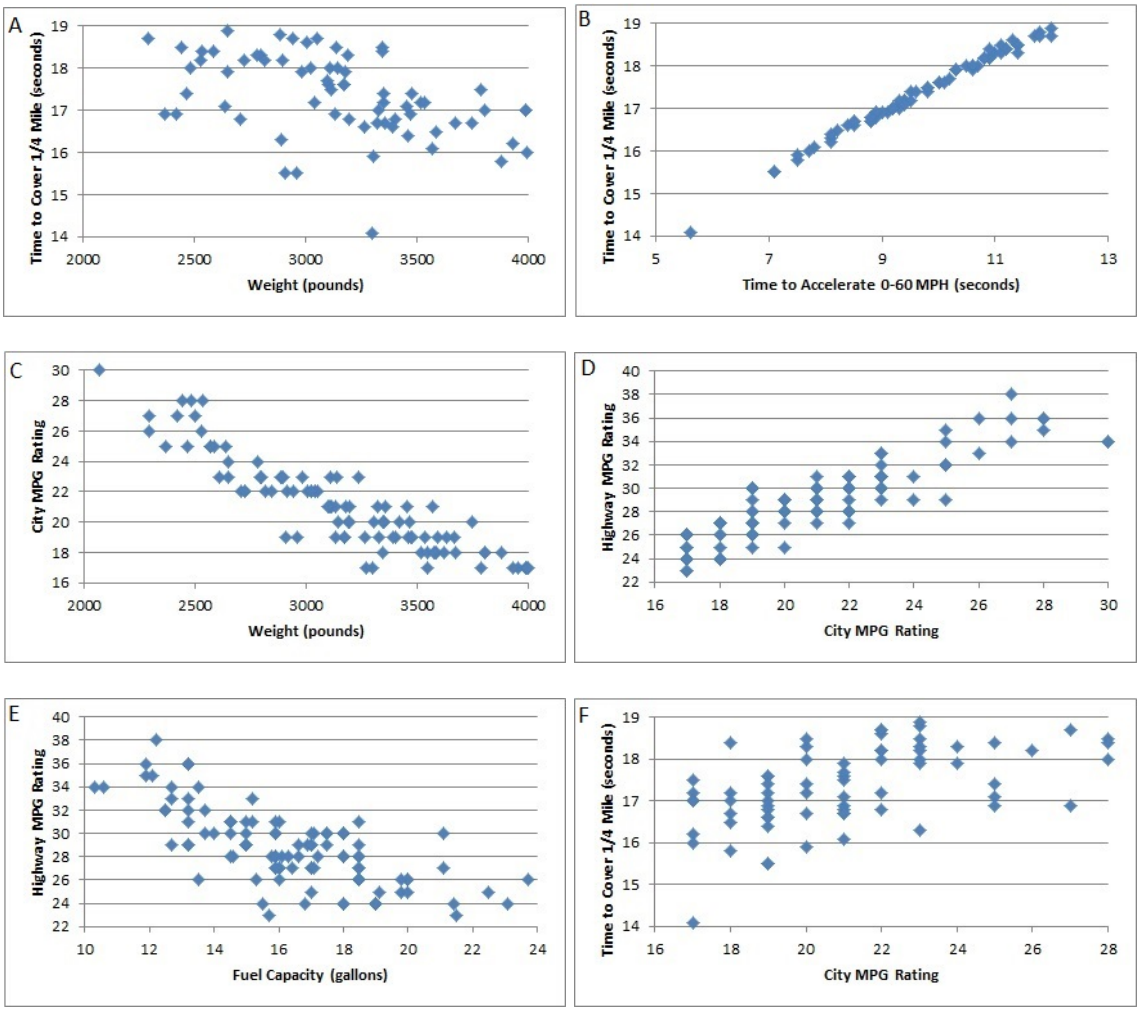
Example 1. A golf pro wanted to investigate the relation between the club-head speed of a golf club (measured in miles per hour) and the distance (in yards) that the ball will travel. He realized that there are other variables besides club-head speed that determine the distance a ball will travel (such as club type, ball type, golfer, and weather conditions). To eliminate the variability due to these variables, the pro used a single model of club and ball. One golfer was chosen to swing the club on a clear, 70-degree day with no wind. The pro recorded the club-head speed, measured the distance that the ball traveled, and collected the data below. Draw a scatter diagram of the data.

Club-Head Speed (mph)	100	102	103	101	105	100	99	105
Distance (yards)	257	264	274	266	277	263	258	275



Example 2. Evaluate the strength and direction of the association between the variables in each graph (on the next page). Fill in the table below:

	Negative			Positive		
	Strongest		Weakest	Weakest		Strongest
Letter of Scatterplot						



Correlation Coefficient

The statistical **correlation** between two quantitative variables is a number that *indicates the strength and the direction of a straight-line relationship.*

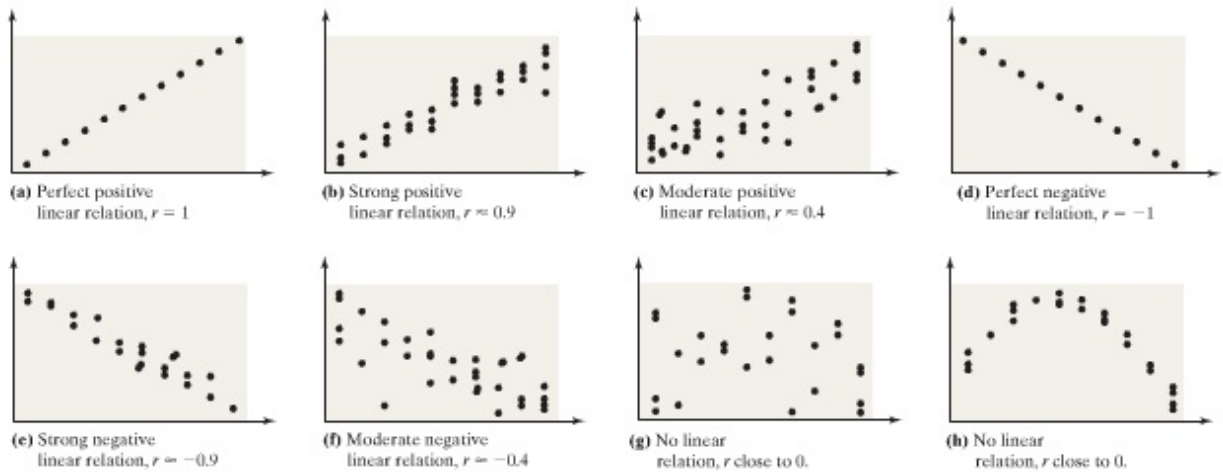
- The *strength* of the relationship is determined by the *closeness of the points to a straight line.*
- The *direction* is determined by whether one variable generally increases or generally decreases when the other variable increases.

The **linear correlation coefficient** or **Pearson product moment correlation coefficient** is a measure of the strength and direction of the 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.

Some Notes about Correlation:

- Correlation measures the strength of the linear relationship between variables.
- Correlation coefficients are always between -1 and $+1$.
- The *magnitude* of the correlation indicates the strength of the relationship, which is the overall closeness of the points to a straight line. The *sign* of the correlation does not matter when assessing the strength of the linear relationship.
- A correlation of either -1 or $+1$ indicates that there is a perfect linear relationship and all data points fall on the same straight line.
- The *sign* of the correlation indicates the direction of the relationship. A *positive* correlation indicates that the two variables tend to increase together. A *negative* correlation indicates that when one variable increases, the other is likely to decrease.
- A correlation of 0 indicates that the best straight line through the data is exactly horizontal, so knowing the value of x does not change the predicted value of y.
- There is more than one way to measure correlation - the *Pearson product moment correlation coefficient* (our r) is the most widely used.
- CORRELATION DOES NOT IMPLY CAUSATION!



Common Errors Involving Correlation:

- Correlation is very sensitive to outliers, so look at the scatterplot.
- Positive correlations imply positive slopes but large correlations do not imply large slopes.
- The lack of a linear relationship does not imply a lack of a relationship in general.
- CORRELATION DOES NOT IMPLY CAUSATION!

Ridiculous Example

- Ice cream consumption has a positive correlation with drowning rates. Does that mean eating ice cream causes drowning?

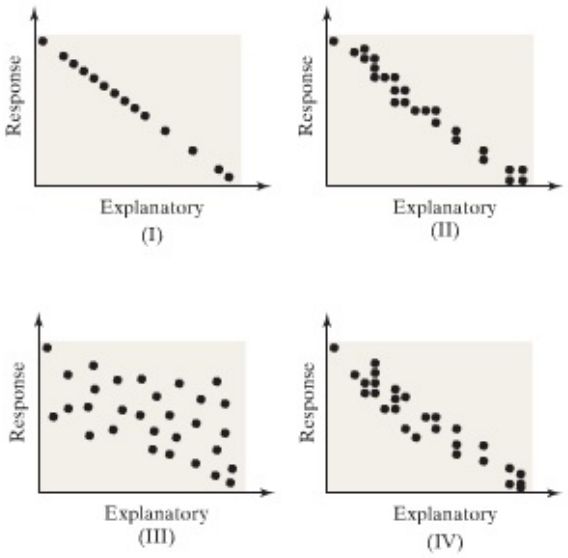
Not-so Ridiculous Example

- The number of hours a person spends working per day appear to be negatively correlated with the number of years a person is expected to live.

Rule of thumb for interpreting values of r:

- If $|r| < 0.25$, conclude no linear correlation
- $0.25 < |r| < 0.50$, conclude weak (positive/negative) linear correlation
- $0.50 < |r| < 0.75$, conclude moderate (positive/negative) linear correlation
- $|r| > 0.75$, conclude strong (positive/negative) linear correlation

Example 3. Match the linear correlation coefficient to the scatter diagram. The scales on the x- and y-axes are the same for each scatter diagram.



(a) $r = -0.969$

(b) $r = -0.049$

(c) $r = -1$

(d) $r = -0.992$

12.3: The Regression Equation

Introduction: Sir Francis Galton (1822 - 1911) developed the technique of regression in his studies on inheritance. He wrote about "the law of universal regression" saying "Each peculiarity in a man is shared by his kinsman, but *on the average* in a less degree". In 1903 Karl Pearson (1857-1936) surveyed 1,078 families and obtained data on fathers' heights (x, in inches) and sons' heights (y, in inches), producing the regression equation $\hat{y} = 33.73 + 0.516x$. Thus, though sons of tall fathers do tend to be tall, they are, on average, not as tall as their fathers. For example, 6' (72") fathers will have, on average, approximately 5'11" (70.882") sons. So, the sons' heights *regressed* toward the average of all fathers' heights, as Galton said.

- **Regression analysis** is the area of statistics that is used to examine the relationship between a quantitative response variable and one or more explanatory variables.

- **Linear regression equation** - an algebraic description of the linear relationship between two variables, defined by
$$\hat{y} = b_0 + b_1 x$$
where x is the explanatory variable, $\hat{y}$ is the PREDICTED response variable, b_0 is the y-intercept, and b_1 is the slope.
  - The equation can be used in the future to **predict** values of a response variable when we only know the values for the explanatory variable.

- All of the points will likely not fall on the line in a regression equation, which means that the actual y will often differ from the predicted y ($\hat{y}$).

- **Residual** - the difference between the observed sample value, y, and the predicted value of y, $\hat{y}$, from the regression equation, computed as $y - \hat{y}$ (obesrved$_y$ - predicted$_y$).

- **Least-squares regression line** - the line that minimizes the sum of the squared errors (or residuals), or rather the sum of the squared vertical distance between the observed values of y and those predicted by the line, $\hat{y}$.

- Guidelines for using the regression equation for predictions:
  - Use the regression equation for predictions only if the graph of the regression line fits the points reasonably well.
  - Use the regression equation for predictions only if r indicates there is a linear correlation present.
  - Use the regression equation for predictions only if the data do not go much beyond the scope of the available sample data.

Determining the Least-Squares Regression Line Using Excel

1. Enter the explanatory variable in column A.
2. Enter the response variable in column B.
3. Select the Data Tab and select Data Analysis.
4. Select the Regression option.
5. With the cursor in the Input Y-Range, select the column that contains the response variable (column B).
6. With the cursor in the Input X-Range, select the column that contains the explanatory variable (column A).
7. Select Output Range, and click on an empty cell, then select OK.

Example 4. Run Example 1 using Excel.

E	F	G	H	I	J	K	L	M
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.938695838							
R Square	0.881149876							
Adjusted R Square	0.861341522							
Standard Error	2.882638465							
Observations	8							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	369.6423729	369.64237	44.4837503	0.000549825			
Residual	6	49.85762712	8.3096045					
Total	7	419.5						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-55.7966102	48.37134953	-1.153505	0.29257431	-174.1570386	62.563818	-174.1570386	62.56381826
X Variable 1	3.166101695	0.47470539	6.669614	0.00054983	2.00453945	4.3276639	2.00453945	4.32766394

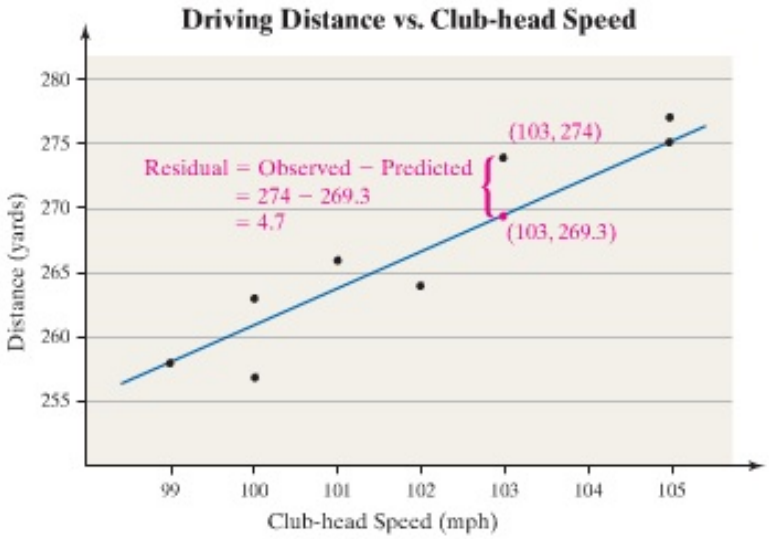
Prediction Errors and Residuals

Before using Excel, a regression line was created by hand. The figure below shows this line, and the regression line was found to be

$$\hat{y} = -22.4967 + 2.833x$$

We found a line that fits the data, but is this the *best* line?

We need a criterion for determining best. Consider the figure below.



Each y-coordinate on the line coresponds to a predicted distance for a given club-head speed. For example, if club-head speed is 103 miles per hour, the predicted distance the ball will travel is $2.833(103) - 22.4967 = 269.3$ yards. The observed distance for this club-head speed is 274 yards. The difference between the observed value of y and the predicted value of y is the error, or **residual**. For a club-head speed of 103 miles per hour, the residual is

$$\begin{aligned} \text{Residual} &= \text{observed } y - \text{ predicted } y \\ &= y - \hat{y} \\ &= 274 - 269.3 \\ &= 4.7 \text{ yards} \end{aligned}$$

The **residual** or **prediction error**, for a club-head speed of 103 miles per hour is 4.7 yards. This is shown in the figure above.

Interpret the Slope and y-Intercept of the Least-Squares Regression Line

The definition of the slope of a line is $\frac{\text{Rise}}{\text{Run}}$ or $\frac{\text{Change in } y}{\text{Change in } x}$. For a line whose slope is $\frac{2}{3}$, if x increases by 3, y will *increase* by 2. Of course, if the slope of a line is $-4 = \frac{-4}{1}$, if x increases by 1, y will *decrease* by 4.

The y-intercept of any line is the point where the graph intersects the vertical axis. It is found by letting $x = 0$ in an equation and solving for y.

We found the regression equation in Example 4 to be $\hat{y} = 3.1661x - 55.7966$.

Interpretation of Slope:

The slope of the regression line is 3.1661. If the club-head speed increases by 1 mile per hour, the distance the ball travels increases by 3.1661 yards, on average. We use the phrase "on average" because we cannot say an increase in swing speed of 1 mile per hour guarantees the distance will increase 3.1661 yards. The slope represents what we expect the response variable (distance) to change by for a 1-unit change in the explanatory variable (swing speed).

For a one unit increase in [the explanatory variable], [the response variable] will [increase/decrease] by [the value of the slope], on average.

Interpretation of the y-intercept:

The y-intercept of the regression line is -55.7966. To interpret the y-intercept, we must first ask two questions:

1. Is 0 a reasonable value for the explanatory variable?

2. Do any observations near $x = 0$ exists in the data set?

If the answer to either of these questions is no, we do not give an interpretation to the y-intercept. In the regression equation in Example 4, a swing speed of 0 miles per hour does not make sense, so an interpretation of the y-intercept is unreasonable.

In general, to interpret a y-intercept, we would say that it is the expected value of the response variable when the value of the explanatory variable is 0.

Interpreting the Squared Correlation, r^2

Consider the club-head speed versus distance data (Example 1). If we were asked to predict the distance of a randomly selected shot, what would be a good guess? Our best guess might be the average distance of all shots taken. Since we don't know this value, we use the average distance from the sample data given, on page 2 of these notes, $\bar{y} = 266.75$ yards.

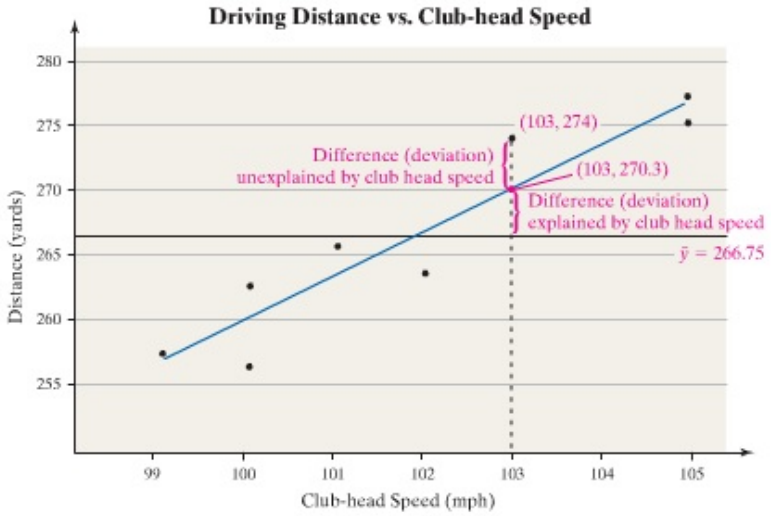
Now suppose we were told this particular shot resulted from a swing with a club-head of 103 mph. Knowing that a linear relation exists between club-head speed and distance allows us to improve our estimate of the distance of the shot. In statistical terms, we say that some of the variation in distance is explained by the linear relation between club-head speed and distance.

The **coefficient of determination,** r^2, measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

Notes about the Coefficient of Determination

- R^2 is not the only way to measure explained variation, but it is the most common.
- R^2 is always between 0 and 1. That is, $0 \leq r^2 \leq 1$.
- A coefficient of determination of 1 indicates that the least squares regression line explains 100% of the variation in the response variable (y).
- A coefficient of determination of 0 indicates that the least squares regression line explains none of the variation in the response variable (y).

Consider the figure below, where a horizontal line is drawn at $\bar{y} = 266.75$. This value represents the predicted distance of a shot without any knowledge of club-head speed. Armed with the additional information that the club-head speed is 103 mph, we increased our guess to 270.3 yards. The difference between the predicted distance of 266.75 yards and the predicted distance of 270.3 yards is due to the fact that the club-head speed is 103 mph. In other words, the difference between the prediction of $\hat{y} = 270.3$ and $\bar{y} = 266.75$ is explained by the linear relation between club-head speed and distance. The observed distance when club-head speed is 103 mph is 274 yards. The difference between our predicted value, $\hat{y} = 270.3$, and the actual value $y = 274$, is due to the factors (variables) other than the club-head speed and also is due to random error. The differences just discussed are called **deviations**.



The deviation between the observed value of the response variable, y, and the mean value of the response variable, $\bar{y}$, is called the **total deviation**, so total deviation $= y - \bar{y}$.

The deviation between the predicted value of the response variable $\hat{y}$, and the mean value of the response variable, $\bar{y}$, is called the **explained deviation**, so explained deviation $= \hat{y} - \bar{y}$.

Finally, the deviation between the observed value and the response variable, y, and the predicted value of the response variable, $\hat{y}$, is called the **unexplained deviation**, so unexplained $= y - \hat{y}$.

Determining the Coefficient of Determination Using Technology

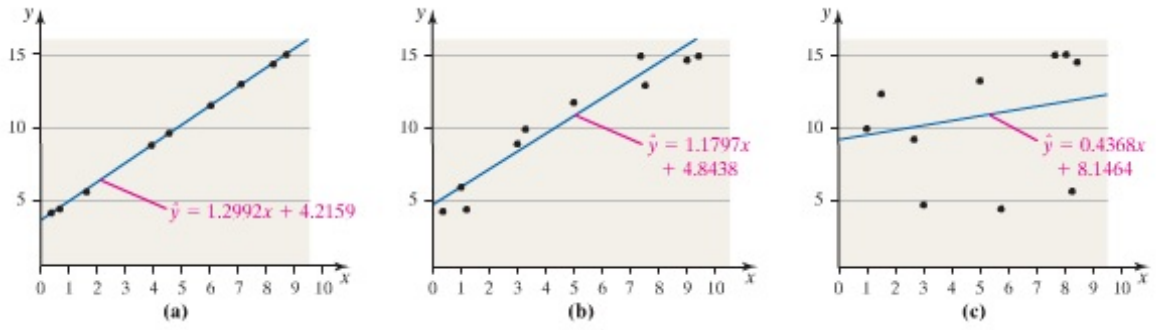
The coefficient of determination, r^2, is given using the same Excel steps give at the beginning of this chapter.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.938695838
R Square	0.881149876
Adjusted R Square	0.861341522
Standard Error	2.882638465
Observations	8

Consider the three data sets given below.

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

The figure below represent the scatter diagrams for each data set.



Data	r^2	Interpretation
A	99.99%	99.99% of the variability in y is explained by the least-squares regression line.
B	94.7%	94.7% of the variability in y is explained by the least-squares regression line.
C	9.4%	9.4% of the variability in y is explained by the least-squares regression line.

Notice that as the explanatory ability of the line decreases, the coefficient of determination, r^2, also decreases.

Example 5. A student at Joliet Junior College conducted a survey of 20 randomly selected full-time students to determine the relation between the number of hours of video game playing each week, x, and grade-point average, y. She found that a linear relation exists between the two variables. The least-squares regression line that describes this relation is $\hat{y} = -0.0526x + 2.9342$. Note that R^2 is equal to 0.372.

(a) Find and interpret the linear correlation coefficient.

(b) Predict the grade-point average of a student who plays video games 8 hours per week.

(c) Interpret the slope.

(d) If appropriate, interpret the y-intercept.

(e) A student who plays video games 7 hours per week has a grade-point average of 2.68. What is the residual for a student who plays video games 7 hours per week?

Example 6. The American black bear is one of eight bear species in the world. It is the smallest North American bear and the most common bear species on the planet. In 1969, Dr. Michael R. Pelton of the University of Tennessee initiated a long-term study of the population in the Great Smoky Mountains National Park. One aspect of the study was to develop a model that could be used to predict a bear's weight (since it is not practical to weigh bears in the field). One variable thought to be related to weight is the length of the bear. The following data represent the lengths and weights of 12 American black bears. Use the output on the next page.

Total Length (cm)	139.0	138.0	139.0	120.5	149.0	141.0	141.0	150.0	166.0	151.5	129.5	150.0
Weight (kg)	110	60	90	60	85	110	95	85	155	140	105	110

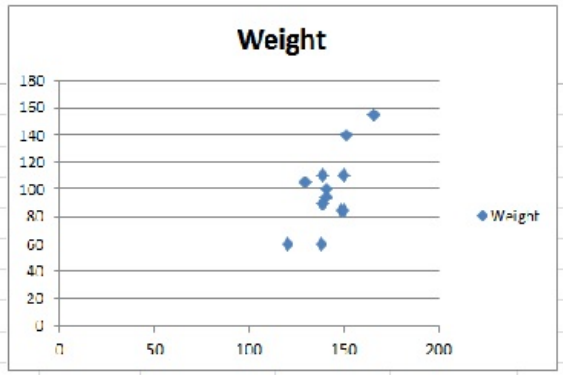
(a) Which variable is the explanatory variable based on the goals of the research?

(b) Determine the regression line between weight and height.

(c) Predict the weight of a bear if the height is 155.5 cm.

(d) Find and interpret r.

(e) Compute and interpret the coefficient of determination.



SUMMARY OUTPUT

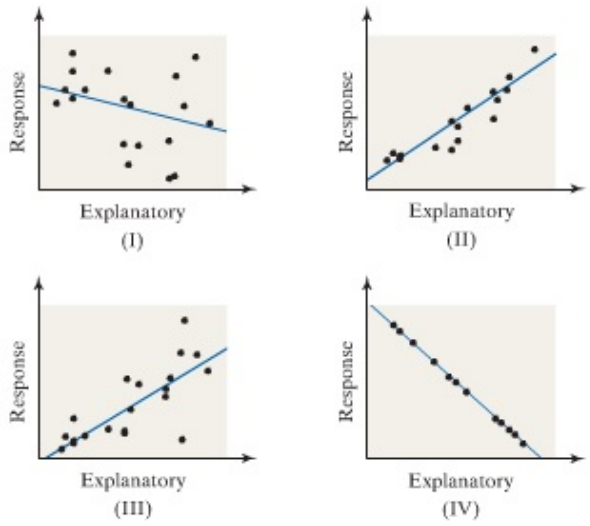
Regression Statistics	
Multiple R	0.70390318
R Square	0.49547968
Adjusted R Square	0.44502765
Standard Error	20.8577004
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4272.480004	4272.48	9.820807	0.010621299
Residual	10	4350.436663	435.0437		
Total	11	8622.916667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-142.470924	77.47377069	-1.83896	0.095762	-315.0932426	30.1513944	-315.093243	30.15139436
Length	1.69416803	0.540608523	3.133817	0.010621	0.489617178	2.89871888	0.489617178	2.898718884

Example 7. Match the coefficient of determination to the scatter diagram. The scales on the horizontal and vertical axis are the same for each scatter diagram.



(a) $r^2 = 0.58$

(b) $r^2 = 0.90$

(c) $r^2 = 1$

(d) $r^2 = 0.12$

Example 8. A school cafeteria has a salad bar that is priced based on weight, with salads costing 30 cents an ounce. Students fill a container that weighs 8 ounces when it is empty. Define x to be the weight of the filled container (in ounces) and y to be the price the student is charged (in dollars). The relationship is $\hat{y} = -2.40 + 0.30x$, and the linear correlation is given as 1.

(a) Find and interpret r.

(b) Calculate R^2 and write a sentence that interprets this value in the context of this situation.

(c) Explain (in words) what it would mean if $x = 8$.

(d) If appropriate, interpret the y-intercept.

(e) Interpret the slope.

(f) If the container plus ingredients weighs 20 ounces, how much does the salad cost the student?