Chapter 15

Describing Relationships:

Regression, Prediction, and Causation

Lecture Slides

Case Study: Describing Relationships—Regression, Prediction, and Causation 1

Predicting the future course of the stock market could make you rich. No wonder lots of people and lots of computers pore over market data looking for patterns. There are some surprising methods.

The Super Bowl indicator says that the football Super Bowl, played in January or early February, predicts how stocks will behave each year.

Case Study: Describing Relationships—Regression, Prediction, and Causation 2

The current National Football League (NFL) was formed by merging the original NFL with the American Football League (AFL) and consists of two conferences:

- National Football Conference (NFC)
- American Football Conference (AFC)

The indicator claims that stocks go up in years when a team from the NFC (or from the old NFL) wins and down when an AFC team wins.

Case Study: Describing Relationships—Regression, Prediction, and Causation 3

The indicator was right in 38 of 48 years between the first Super Bowl in 1967 and 2014.

For purposes of the legend, we will regard the Baltimore Ravens as an old NFL team because they were the Cleveland Browns before the franchise moved to Baltimore in 1996. We will also regard the Tampa Bay Buccaneers as an NFC team, although they were neither a premerger team nor an old NFL team and started out as an AFC team.

The indicator is right over 75% of the time, which seems impressive.

Case Study: Describing Relationships—Regression, Prediction, and Causation 4

In 2015, the Patriots, an AFC team, won the Super Bowl. The Super Bowl indicator predicted stocks would fall in 2015. Should I have avoided investing in 2015?

In this chapter, we will study statistical methods to predict one variable from others that go well beyond just counting ups and downs. We will also distinguish between the ability to predict one variable from others and the issue of whether changes in one variable are caused by changes in others. By the end of this chapter, you will be able to critically evaluate the Super Bowl indicator.

Regression Lines 1

If a scatterplot shows a straight-line relationship between two quantitative variables, we would like to summarize this overall pattern by drawing a line on the graph.

A regression line summarizes the relationship between two variables, but only in a specific setting: one of the variables helps explain or predict the other. That is, regression describes a relationship between an explanatory and a response variable.

Regression Lines 2

A **regression line** is a straight line that describes how a response variable *y* changes as explanatory variable *x* changes. We often use a regression line to predict the value of *y* for a given value of *x*.

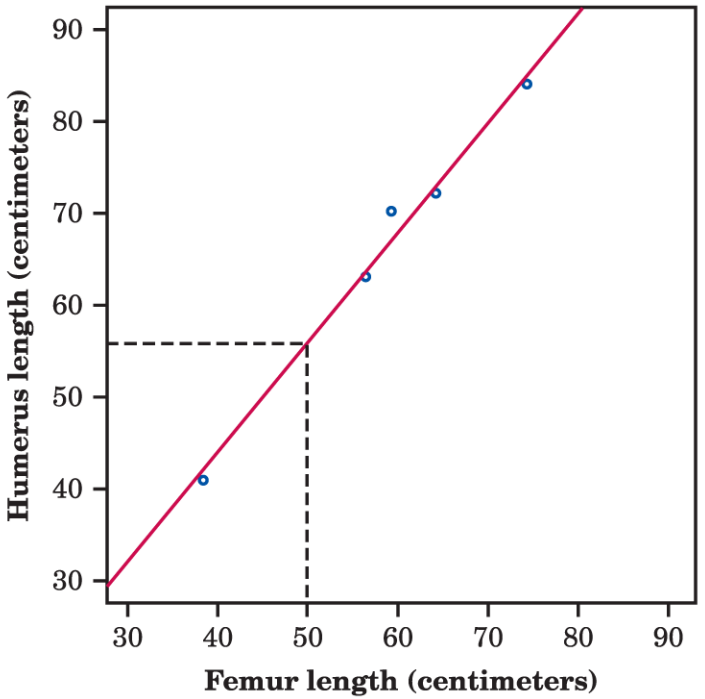
General regression: find a function h such that $\sum_{i=1}^{n} (y_i - h(x_i))^2$ is minimal.

Linear regression: find two constants a and b such that $\sum_{i=1}^{n} (y_i - (a + bx_i))^2$ is minimal.

Example: Fossil Bones 1

The lengths of two bones in fossils of the extinct archaeopteryx closely follow a straight-line pattern.

Figure 15.1 plots the lengths for the five available fossils with a regression line on the plot.

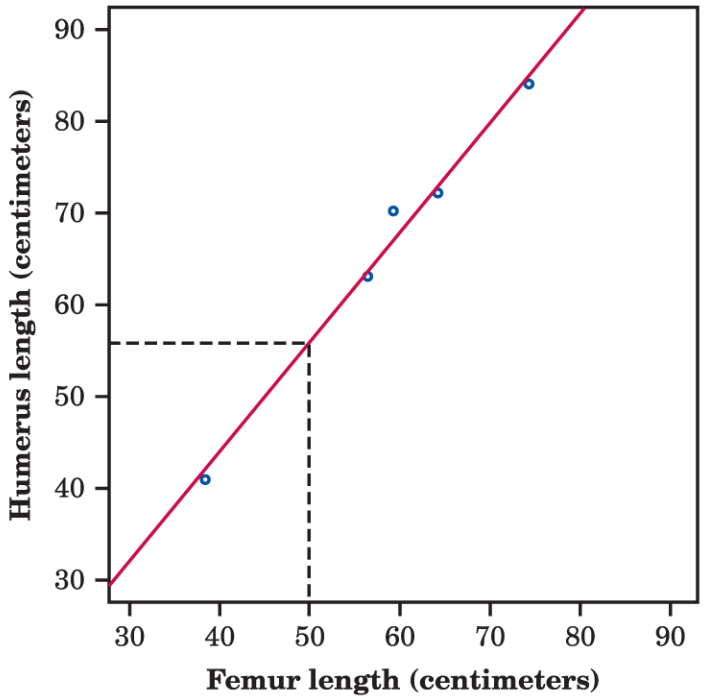


Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Fossil Bones 2

Another archaeopteryx fossil is incomplete.

Its femur is 50 cm long, but the humerus is missing. Can we predict how long the humerus is?

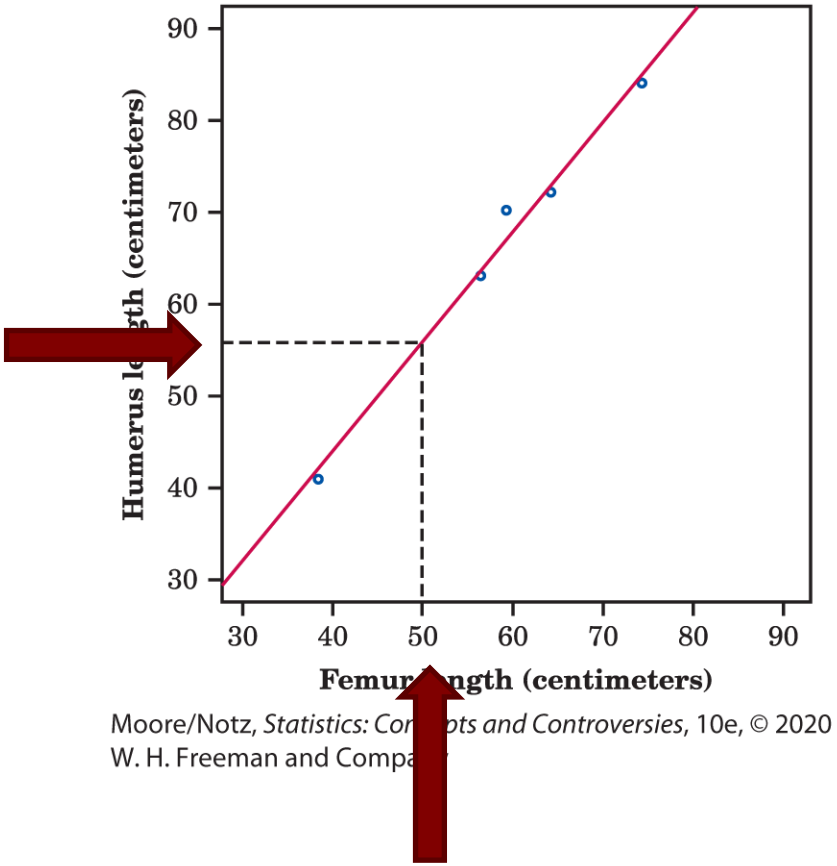


Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Fossil Bones 3

For a femur length of 50 cm we predict a length of about 56 cm.

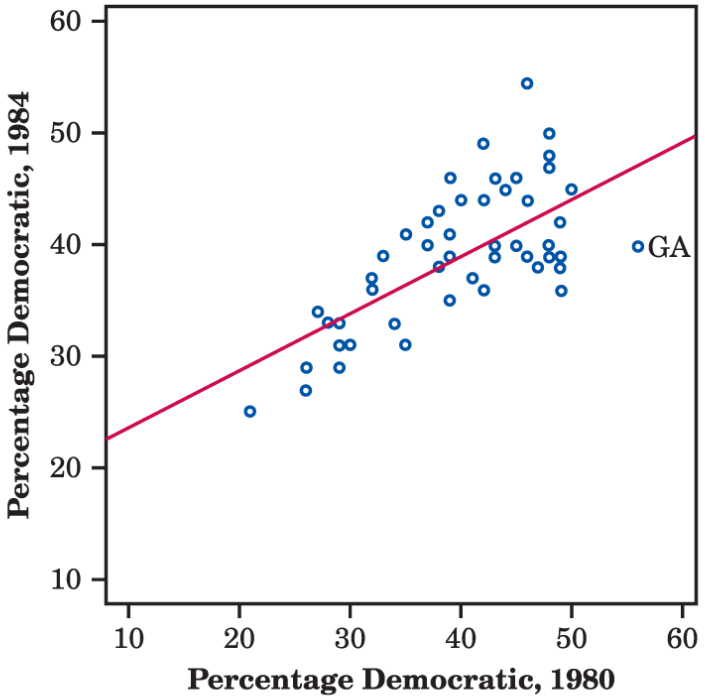
This is the length the humerus would have if this fossil's point lay exactly on the line. All the other points are close to the line, so we think the missing point will also be close to the line. That is, we think this prediction will be quite accurate.



Moore/Notz, *Statistics: Co ots and Controversies*, 10e, © 2020 W. H. Freeman and Compa

Example: Presidential Elections, the Reagan Years 1

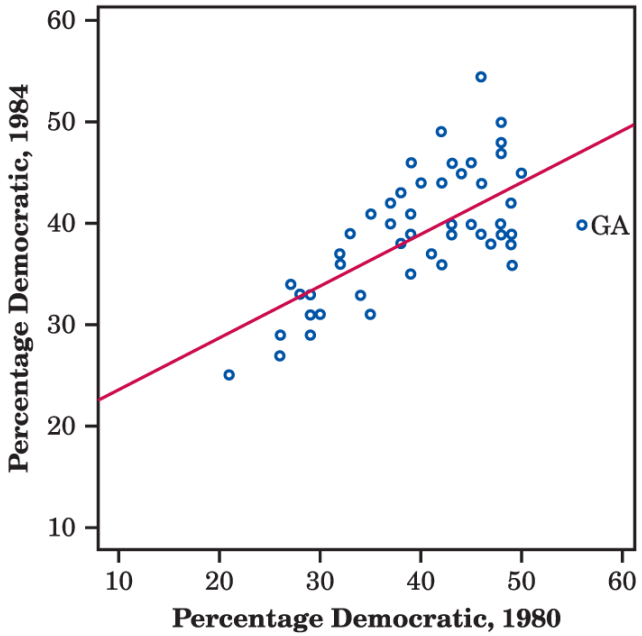
Republican Ronald Reagan was elected president twic in 1980 and in 1984. His economic policy of tax cut to stimulate the economy, eventually leading to increases in tax revenue, was still advocated by some Republican presidential candidates in 2015.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e,

Example: Presidential Elections, the Reagan Years 2

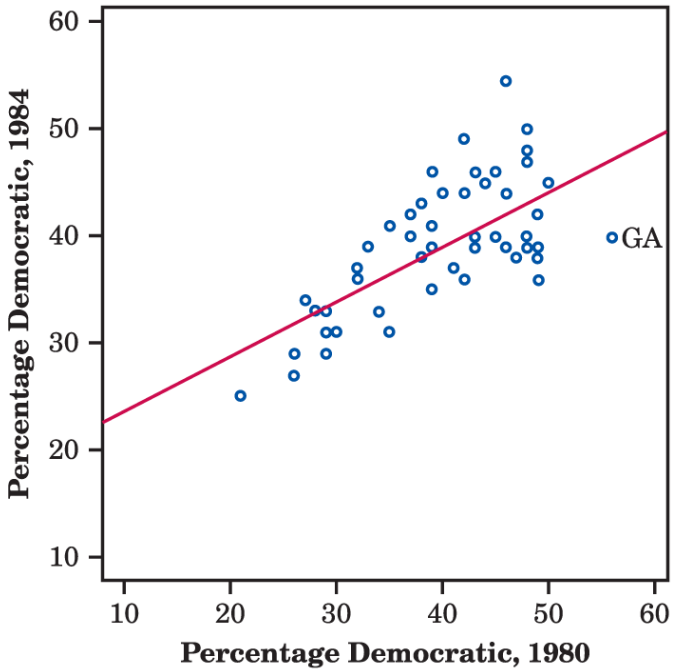
Figure 15.2 plots the percentage of voters in each state who voted for Reagan's Democratic opponents: Jimmy Carter in 1980 and Walter Mondale in 1984. The plot shows a positive straight-line relationship.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e,

Example: Presidential Elections, the Reagan Years 3

There is one outlier: Georgia, President Carter's home state, voted 56% for the Democrat Carter in 1980 but only 40% Democratic in 1984. We could use the regression line drawn in Figure 15.2 to predict a state's 1984 vote from its 1980 vote.

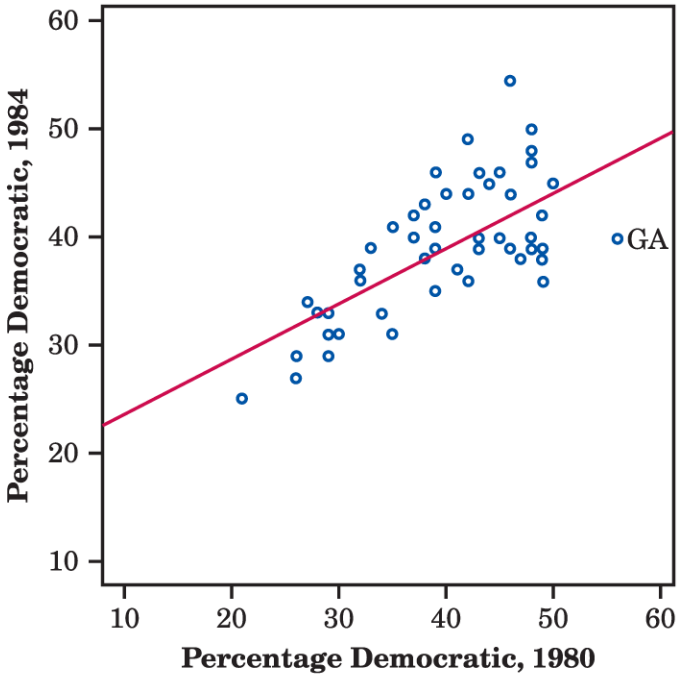


Moore/Notz, *Statistics: Concepts and Controversies*, 10e,
© 2020 W. H. Freeman and Company

Example: Presidential Elections, the Reagan Years 4

The points in this figure are more widely scattered about the line than are the points in the fossil bone plot in Figure 15.1.

The scatter of the points makes it clear that predictions of voting will be generally less accurate than predictions of bone length.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e,

Regression Equations 1

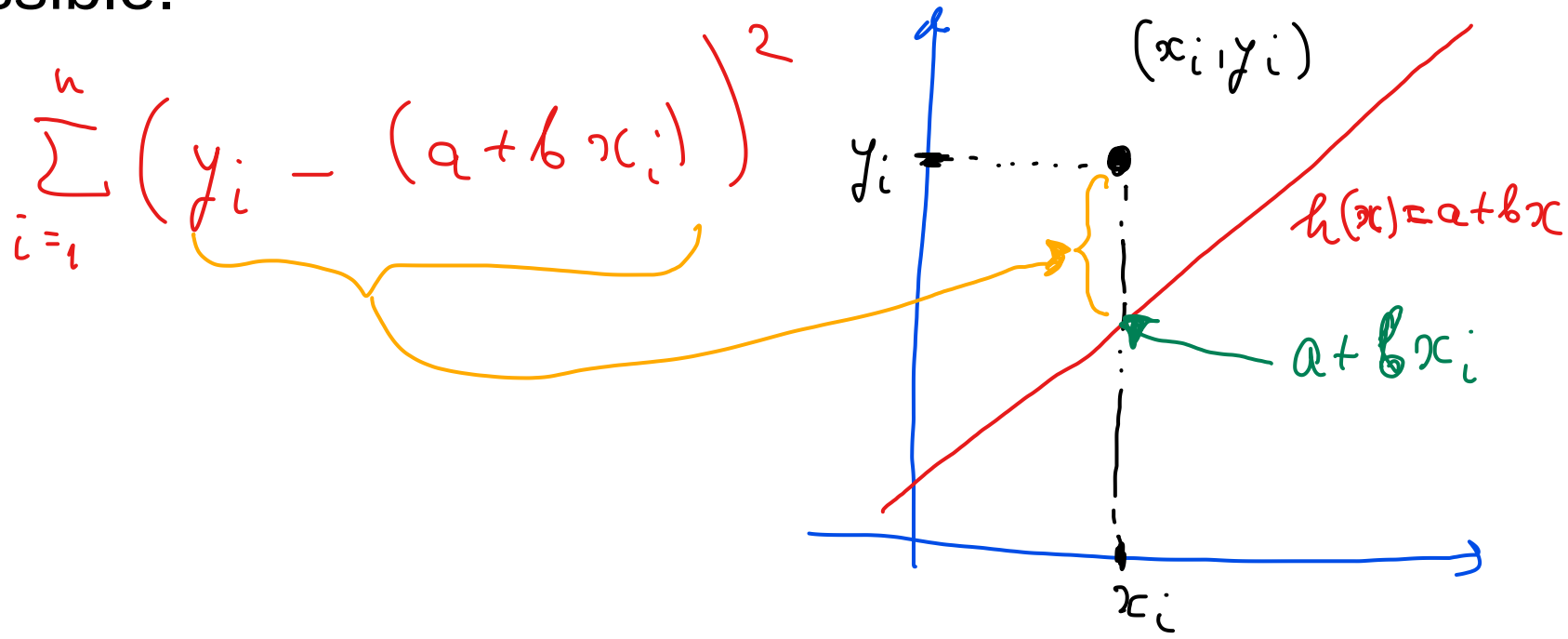
In regression, we want to predict *y* from *x*, so we want a line that is close to the points in the vertical (*y*) direction.

It is hard to concentrate on just the vertical distances when drawing a line by eye, so we use computer software to find the line.

There are many ways to make the collection of vertical distances "as small as possible." The most common is the least-squares method.

Regression Lines 3

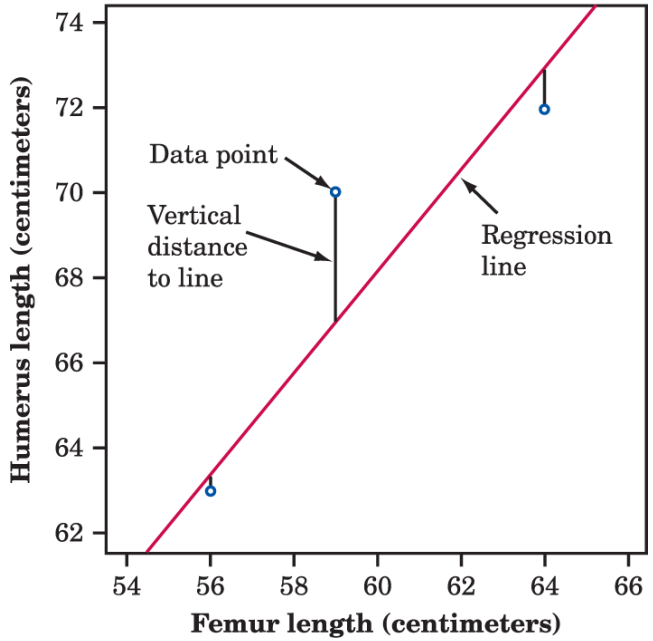
The **least-squares regression line** of *y* on *x* is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.



Regression Equations 2

Figure 15.3 illustrates the least-squares idea. This figure magnifies the center part of Figure 15.1 to focus on three of the points.

We see the vertical distances of these three points from the regression line. To find the least-squares line, look at these vertical distances (all five for the fossil data), square them, and move the line until the sum of the squares is the smallest it can be for any line.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e,

Regression Equations 3

In writing the equation of a line, *x* stands, as usual, for the explanatory variable and *y* for the response variable.

The equation of a line has the form $y = a + bx$.

The number *b* is the slope of the line, the amount by which *y* changes when *x* increases by 1 unit. The number *a* is the intercept, the value of *y* when $x = 0$.

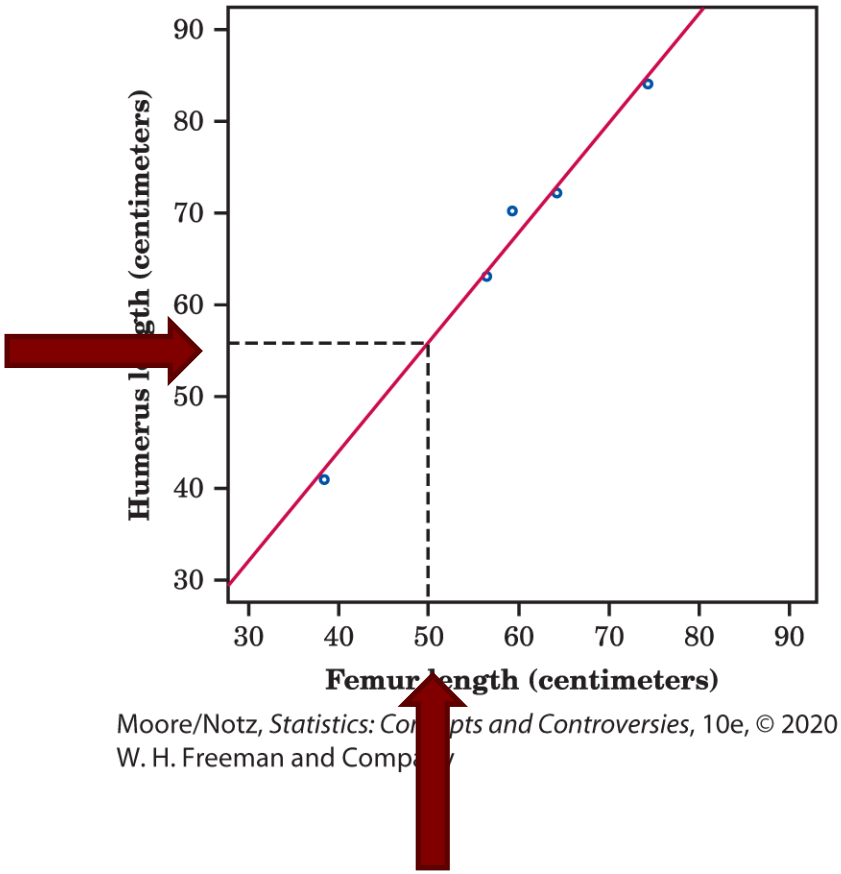
To make a prediction, substitute your *x*-value into the equation and calculate the resulting *y*-value.

Example: Fossil Bones 4

In the fossil example, we used the “up-and-over” method in Figure 15.1 to predict the humerus length for a fossil whose femur length is 50 cm.

It is roughly 56 cm.

The equation of the least-squares line is humerus length = −3.66 + (1.197 × femur length)



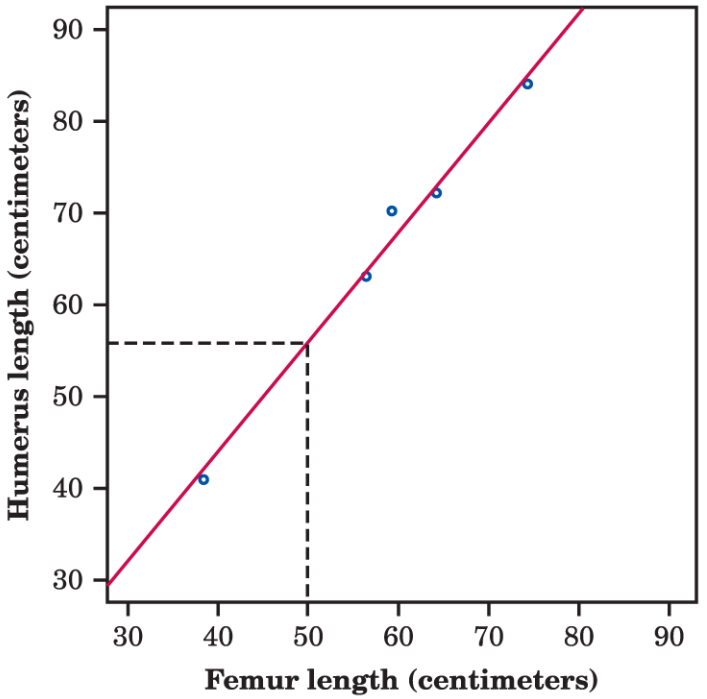
Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Example: Fossil Bones 5

Use the equation to predict by substituting the value of x and calculate y.

The predicted humerus length for a fossil with a femur 50 cm long is humerus length = −3.66 + (1.197)(50) = 56.2 cm.

More precise than the “up and over” method!



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Understanding Prediction 1

Computers make prediction easy and automatic, even from very large sets of data.

Issue 1:

However, regression software will happily fit a straight line to a curved relationship.

Issue 2:

Also, the computer cannot decide which is the explanatory variable and which is the response variable. This is important because the same data give two different lines, depending on which is the explanatory variable.

Understanding Prediction 2

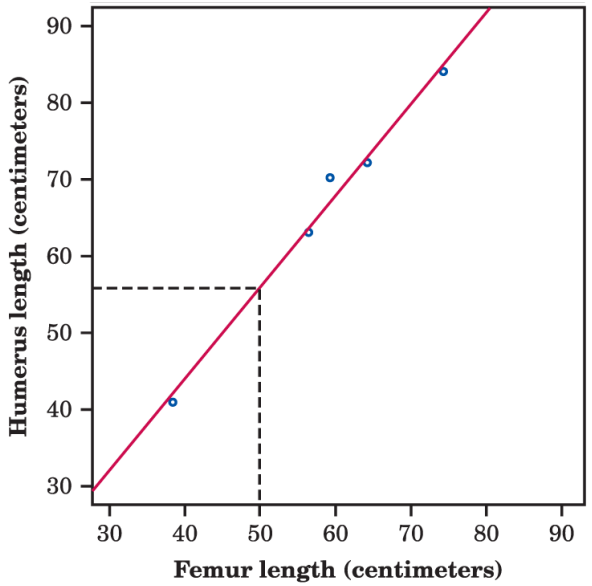
In practice, we often use several explanatory variables to predict a response.

As part of its admissions process, a college might use SAT Math and Verbal scores and high school grades in English, math, and science (five explanatory variables) to predict first-year college grades.

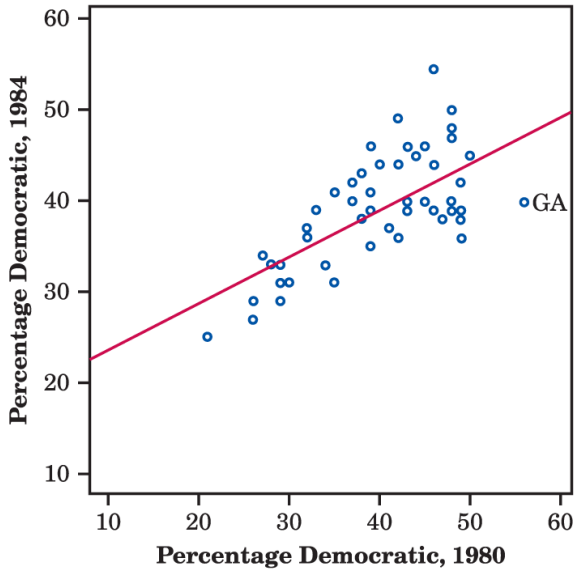
Although the details are messy, all statistical methods of predicting a response share some basic properties of least-squares regression lines.

Understanding Prediction 3

Prediction is based on fitting some "model" to a set of data. In Figures 15.1 and 15.2, our model is a straight line that we draw through the points in a scatterplot. Other prediction methods use more elaborate models.



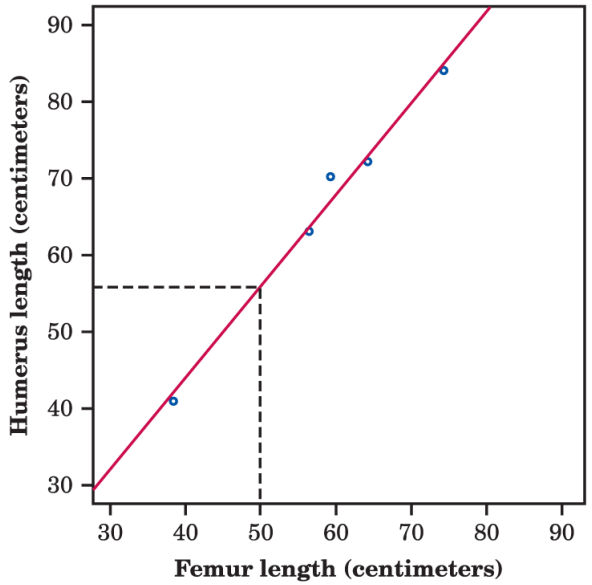
Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company



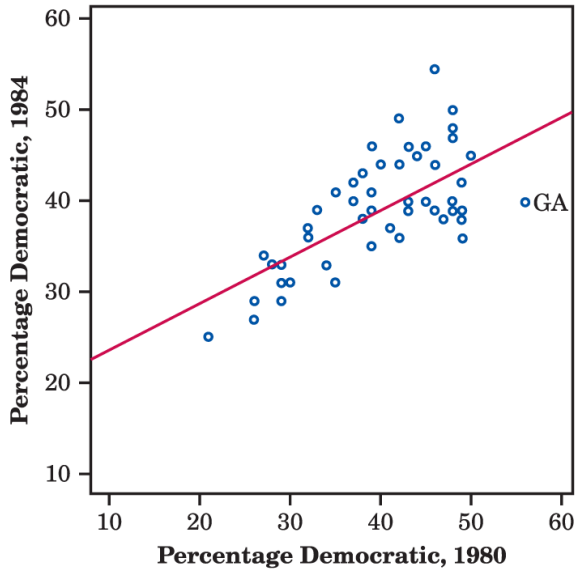
Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Understanding Prediction 4

Prediction works best when the model fits the data closely. Compare, again, Figure 15.1, where the data closely follow a line, with Figure 15.2, where they do not. Prediction is more trustworthy in Figure 15.1. Also, it is not so easy to see patterns when there are many variables, but if the data do not have strong patterns, prediction may be very inaccurate.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Understanding Prediction 5

Prediction outside the range of the available data is risky. Suppose that you have data on a child's growth between 3 and 8 years of age. You find a strong straight-line relationship between age *x* and height *y*. If you fit a regression line to these data and use it to predict height at age 25 years, you will predict that the child will be 8 feet tall. Growth slows down and stops at maturity, so extending the straight line to adult ages is foolish. No wonder economic predictions are often wrong. Prediction outside the range of available data is referred to as **extrapolation**. Beware of **extrapolation**!

Example: Predicting the National Deficit 1

The Congressional Budget Office is required to submit annual reports that predict the federal budget and its deficit or surplus for the next 5 years. These forecasts depend on future economic trends (unknown) and on what Congress will decide about taxes and spending (also unknown).

Even the prediction of the state of the budget, if existing policies remain unchanged, has been wildly inaccurate. The forecast made in January 2008 for 2012, for example, underestimated the deficit by nearly $1000 billion!

Example: Predicting the National Deficit 2

Senator Everett Dirksen once said, “A billion here and a billion there and pretty soon you are talking real money.”

In 1999, the Budget Office predicted a surplus (ignoring Social Security) of $996 billion over the following 10 years. Politicians debated what to do with the money, but no one else believed the prediction (correctly, as it turned out).

Example: Predicting the National Deficit 3

In 2012, there was a $1087 billion deficit, in 2013 a $680 billion deficit, and in 2014 a $483 billion deficit.

The forecast in January 2015 is for a $652 billion deficit in 2019. Time will tell how accurate this forecast is.

Correlation and Regression 1

Correlation measures the direction and strength of a straight-line relationship.

Regression draws a line to describe the relationship.

Regression requires choosing an explanatory variable and correlation does not.

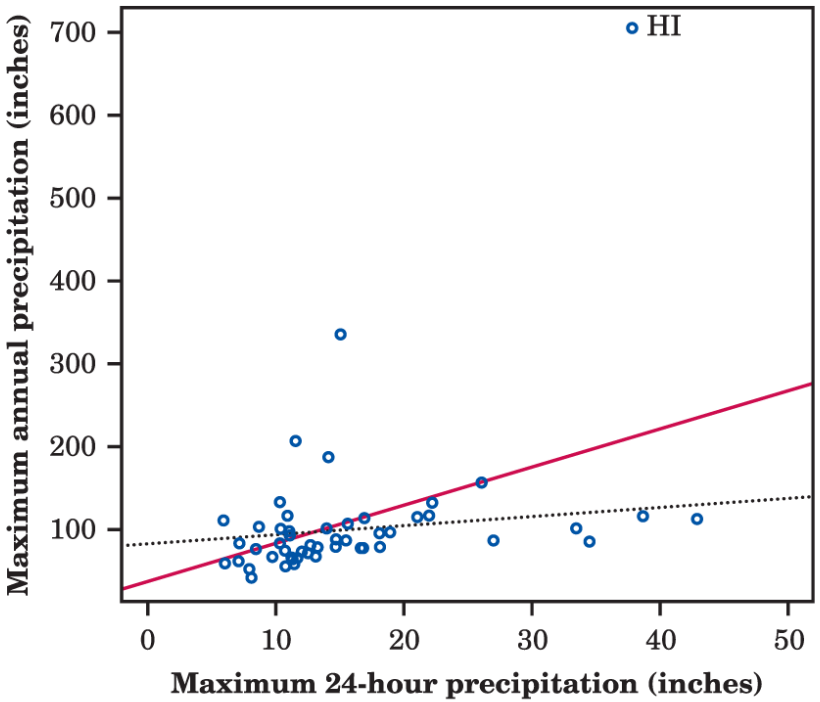
Both correlation and regression are strongly affected by outliers. Be wary if your scatterplot shows strong outliers.

Correlation and Regression 2

Figure 15.4 plots the record-high yearly precipitation in each state against that state's record-high 24-hour precipitation.

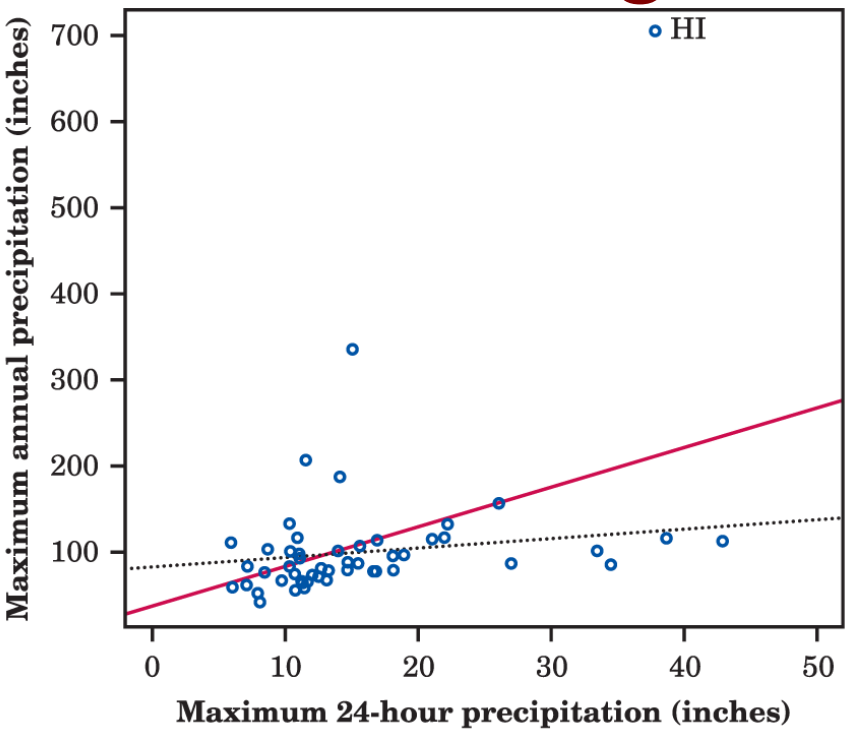
Hawaii is a high outlier, with a yearly record of 704.83 inches of rain recorded at Kukui in 1982. The correlation for all 50 states in Figure 15.4 is 0.510.

If we leave out Hawaii, the correlation drops to $r = 0.248$.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Correlation and Regression 3



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

The solid line in the figure is the least-squares line for predicting the annual record from the 24-hour record. If we leave out Hawaii, the least-squares line drops down to the dotted line. This line is nearly flat: There is little relation between yearly and 24-hour record precipitation once we decide to ignore Hawaii.

Correlation and Regression 4

The usefulness of the regression line for prediction depends on the strength of the association, the correlation. The square of the correlation is the right measure.

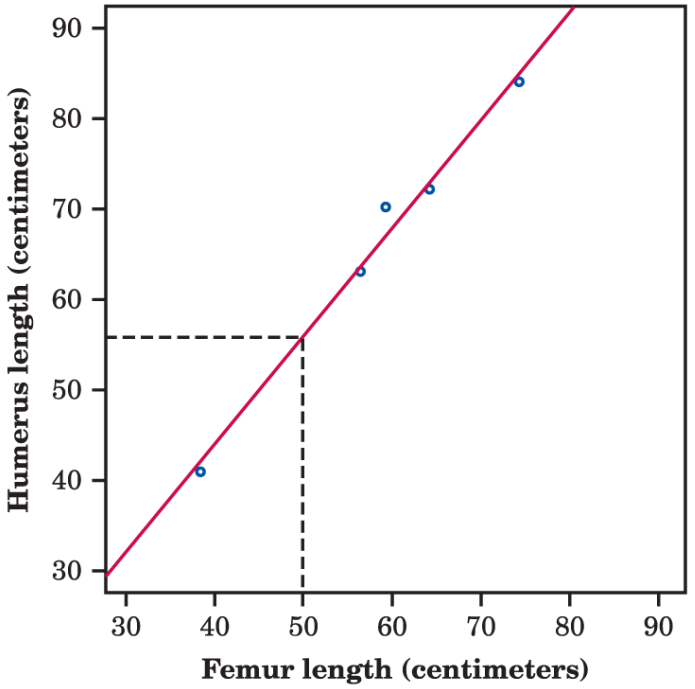
The **square of the correlation**, r^2, is the proportion of the variation in the values of *y* that is explained by the least-squares regression of *y* on *x*.

The idea is that when there is a straight-line relationship, some of the variation in *y* is accounted for by the fact that as *x* changes, it pulls *y* along with it.

Example: Fossil Bones 6

Look again at Figure 15.1. There is a lot of variation in the humerus lengths of these five fossils, from a low of 41 cm to a high of 84 cm.

The scatterplot shows that we can explain almost all of this variation by looking at femur length and at the regression line. Because $r = 0.994$ for these data, $r^2 = (0.994)^2 = 0.988$. So the variation "along the line" as femur length pulls humerus length with it accounts for 98.8% of all the variation in humerus length.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

The Question of Causation 1

There is a strong relationship between cigarette smoking and death rate from lung cancer.

Does smoking cigarettes cause lung cancer?

There is a strong association between the availability of handguns in a nation and that nation's homicide rate from guns.

Does easy access to handguns cause more murders?

The Question of Causation 2

It says right on the pack that cigarettes cause cancer.

Whether more guns cause more murders is hotly debated.

Why is the evidence for cigarettes and cancer better than the evidence for guns and homicide?

We already know three big facts about statistical evidence for cause and effect.

The Question of Causation 3

Statistics and Causation

1. A strong relationship between two variables does not always mean that changes in one variable cause changes in the other.
2. The relationship between two variables is often influenced by other variables lurking in the background.
3. The best evidence for causation comes from randomized comparative experiments.

The Question of Causation 4

Statistics and Causation

4. The observed relationship between two variables may be due to **direct causation, common response, or confounding**. Two or more of these factors may be present together. **Common response** is when a lurking variable influences both *x* and *y* and creates a high correlation, even if *x* and *y* have nothing to do with each other.

5. An observed relationship can, however, be used for prediction without worrying about causation as long as the patterns found in past data continue to hold true.

Evidence for Causation 1

Although difficult, it is sometimes possible to build a strong case for causation in the absence of experiments.

The evidence that smoking causes lung cancer is about as strong as nonexperimental evidence can be. Doctors had long observed that most lung cancer patients were smokers. Observational studies comparing smokers and “similar” (in the sense of characteristics such as age, gender, and overall health) nonsmokers showed a strong association between smoking and death from lung cancer.

Evidence for Causation 2

Could the association be explained by lurking variables that the studies could not measure?

Maybe there is a genetic factor that predisposes people both to nicotine addiction and to lung cancer? Smoking and lung cancer would then be positively associated even if smoking had no direct effect on the lungs.

How were these objections overcome? What are the criteria for establishing causation when we cannot do an experiment?

Evidence for Causation 3

The association is strong. The association between smoking and lung cancer is very strong.

The association is consistent. Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.

Higher doses are associated with stronger responses. People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.

Evidence for Causation 4

The alleged cause precedes the effect in time. Lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer kills more men than any other form of cancer. Lung cancer in women rose along with smoking, again with a lag of about 30 years, and has now passed breast cancer as the leading cause of cancer death among women.

The alleged cause is plausible. Experiments with animals show that tars from cigarette smoke do cause cancer.

Correlation, Prediction, and Big Data 1

In 2008, Google researchers were able to track influenza's spread across the United States much faster than the Centers for Disease Control and Prevention (CDC) could.

Google used computer algorithms to explore millions of online Internet searches and discovered a correlation between what people searched for online and whether they had flu symptoms.

The Google researchers used this correlation to make their surprisingly accurate predictions.

Correlation, Prediction, and Big Data 2

Massive databases, or big data, that are collected by Google, Facebook, credit card companies, and others contain petabytes, or 1015 bytes of data, and continue to grow in size.

Big data allow researchers, businesses, and industry to search for correlations and patterns in data that will enable them to make accurate predictions about public health, economic trends, or consumer behavior. Using big data to make predictions is increasingly common. Big data explored with clever algorithms opens exciting possibilities.

Correlation, Prediction, and Big Data 3

Proponents for big data often make the following claims:

- Big data include all members of a population, eliminating the need for statistical sampling.
- There is no need to worry about causation because correlations are all we need to know for making accurate predictions.
- Scientific and statistical theory is unnecessary because, with enough data, the numbers speak for themselves.

Correlation, Prediction, and Big Data 4

Are these claims correct?

Big data are often enormous convenience samples, the result of recording huge numbers of web searches, credit card purchases, or mobile phones pinging the nearest phone tower. This is not equivalent to having information about the entire population of interest.

It is possible to record every message on Twitter and use these data to draw conclusions about public opinion. However, Twitter users are not representative of the population as a whole.

Correlation, Prediction, and Big Data 5

Are these claims correct?

It is true that correlation can be exploited for purposes of prediction even if there is no causal relation between explanatory and response variables.

If you have no idea what is behind a correlation you have no idea what might cause prediction to fail, especially when the correlation is exploited to extrapolate to new situations.

Correlation, Prediction, and Big Data 6

Google Flu Trends continued to accurately track the spread of influenza until the 2012–2013 flu season, when Google's estimate of the spread of flu-like illnesses was overstated by a factor of almost two.

A possible explanation was that the news was full of stories about the flu and this provoked Internet searches by people who were otherwise healthy.

The failure to understand why search terms were correlated with the spread of flu resulted in incorrectly assuming previous correlations extrapolated into the future.

Correlation, Prediction, and Big Data 7

Adding to the perception of the infallibility of big data are news reports touting successes, with few reports of the failures.

The claim that theory is unnecessary because the numbers speak for themselves is misleading when all the numbers concerning successes and failures of big data are not reported.

Statistical theory has much to say that can prevent data analysts from making serious errors.

Correlation, Prediction, and Big Data 8

Providing examples of where mistakes have been made and explaining how, with proper statistical understanding and tools, those mistakes could have been avoided is an important contribution.

The era of big data is exciting and challenging and has opened incredible opportunities for researchers, businesses, and industry.

But big data are not exempt from statistical pitfalls such as bias and extrapolation.

Statistics in Summary 1

- **Regression** is the name for statistical methods that fit some model to data in order to predict a response variable from one or more explanatory variables.

- The simplest kind of regression fits a straight line on a scatterplot for use in predicting *y* from *x*. The most common way to fit a line is the **least-squares** method, which finds the line that makes the sum of the squared vertical distances of the data points from the line as small as possible.

Statistics in Summary 2

- The **squared correlation r^2** tells us what fraction of the variation in the responses is explained by the straight-line relationship between *y* and *x*.

- **Extrapolation**, or prediction outside the range of the data, is risky because the pattern may be different there. Beware of extrapolation!

Statistics in Summary 3

- A strong relationship between two variables is not always evidence that changes in one variable **cause** changes in the other. Lurking variables can create relationships through **common response** or **confounding**.

- If we cannot do experiments, it is often difficult to get convincing evidence for causation.