CHAPTER 3
Describing Relationships

3.2
Least-Squares Regression

Regression Line

Linear (straight-line) relationships between two quantitative variables are common and easy to understand. A **regression line** summarizes the relationship between two variables, but only in settings where one of the variables helps explain or predict the other.

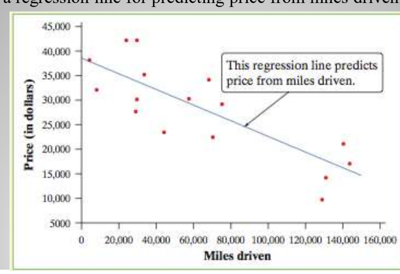
A **regression line** is a line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x.

Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

Example 1: Everyone knows that cars and trucks lose value the more they are driven. Can we predict the price of a used Ford F-150 SuperCrew 4 × 4 if we know how many miles it has on the odometer? A random sample of 16 used Ford F-150 SuperCrew 4 × 4s was selected from among those listed for sale at autotrader.com. The number of miles driven and price (in dollars) were recorded for each of the trucks. Here are the data:

Miles driven	70,583	129,484	29,932	29,953	24,495	75,678	8359	4447
Price (in dollars)	21,994	9500	29,875	41,995	41,995	28,986	31,891	37,991
Miles driven	34,077	58,023	44,447	68,474	144,162	140,776	29,397	131,385
Price (in dollars)	34,995	29,988	22,896	33,961	16,883	20,897	27,495	13,997

The figure below is a scatterplot of these data. The plot shows a moderately strong, negative linear association between miles driven and price with no outliers. The correlation is $r = -0.815$. The line on the plot is a regression line for predicting price from miles driven.



A regression line is a *model* for the data, much like density curves. The equation of a regression line gives a compact mathematical description of what this model tells us about the response variable y and the explanatory variable x.

Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis).
A **regression line** relating y to x has an equation of the form

$$\hat{y} = a + bx$$

In this equation,

*$\hat{y}$ (read "y hat") is the **predicted value** of the response variable y for a given value of the explanatory variable x.

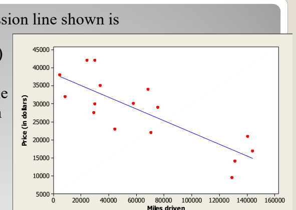
*b is the **slope**, the amount by which y is predicted to change when x increases by one unit.

*a is the **y intercept**, the predicted value of y when $x = 0$.

Example 2: The equation of the regression line shown is

$\widehat{\text{price}} = 38257 - 0.1629(\text{miles driven})$

Identify the slope and y intercept of the regression line. Interpret each value in context.



The slope $b = -0.1629$ tells us that the price of a used Ford F-150 is predicted to go down by 0.1629 dollars (16.29 cents) for each additional mile that the truck has been driven.

The y intercept $a = 38,257$ is the predicted price of a Ford F-150 that has been driven 0 miles.

Prediction

We can use a regression line to predict the response $\hat{y}$ for a specific value of the explanatory variable x.

Extrapolation is the use of a regression line for prediction far outside the interval of values of the explanatory variable x used to obtain the line. Such predictions are often not accurate.

Don't make predictions using values of x that are much larger or much smaller than those that actually appear in your data.

Residuals
In most cases, no line will pass exactly through all the points in a scatterplot. A good regression line makes the vertical distances of the points from the line as small as possible.

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

residual = observed y – predicted y

residual = $y - \hat{y}$

Example 3: Find and interpret the residual for the Ford F-150 that had 70,583 miles driven and a price of $21,994.

$$\widehat{\text{price}} = 38257 - 0.1629(\text{miles driven})$$

$$\widehat{\text{price}} = 38,257 - 0.1629(70,583)$$

$$\widehat{\text{price}} = 26,759 \text{ dollars}$$

residual = observed y – predicted y

$= y - \hat{y} = 21,994 - 26759 = -4765$

That is, the actual price of this truck is $4765 lower than expected, based on its mileage. The actual price might be lower than predicted as a result of other factors. For example, the truck may have been in an accident or may need a new paintjob.

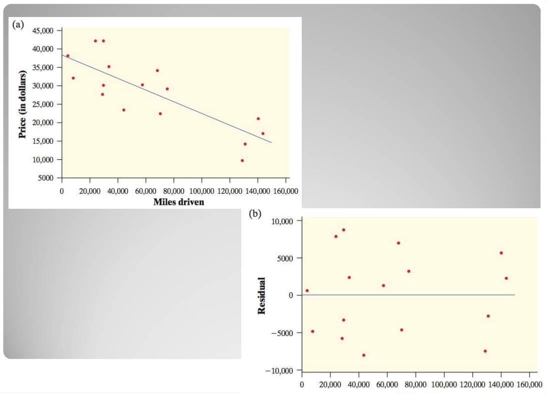
AP EXAM TIP: There's no firm rule for how many decimal places to show for answers on the AP exam. *My advice:* Give your answer correct to two or three nonzero decimal places. *Exception:* If you're using one of the tables in the back of the book, give the value shown in the table.

Least-Squares Regression Line

The **least-squares regression line** of y on x is the line that makes the sum of the squared residuals as small as possible.

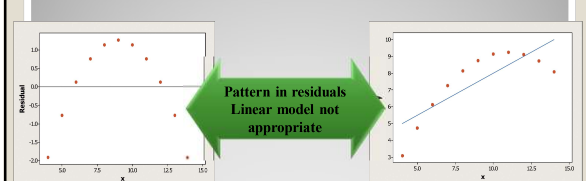
Residual Plots
One of the first principles of data analysis is to look for an overall pattern and for striking departures from the pattern. A regression line describes the overall pattern of a linear relationship between two variables. We see departures from this pattern by looking at the residuals.

A **residual plot** is a scatterplot of the residuals against the explanatory variable. Residual plots help us assess how well a regression line fits the data.



Examining Residual Plots

A residual plot magnifies the deviations of the points from the line, making it easier to see unusual observations and patterns. The residual plot should show no obvious patterns. The residuals should be relatively small in size.



How Well the Line Fits the Data: The Role of r^2 in Regression

If we use a least-squares regression line to predict the values of a response variable y from an explanatory variable x, **the standard deviation of the residuals (s)** is given by

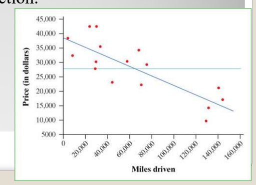
$$s = \sqrt{\frac{\sum residuals^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

This value gives the approximate size of a "typical" prediction error (residual).

The standard deviation of the residuals gives us a numerical estimate of the average size of our prediction errors. There is another numerical quantity that tells us how well the least-squares regression line predicts values of the response

The Coefficient of Determination

Example 4: Suppose that we randomly selected an additional used Ford F-150 that was on sale. What should we predict for its price? The figure below shows a scatterplot of the truck data that we have studied throughout this section, including the least-squares regression line. Another horizontal line has been added at the mean y-value, $\bar{y} = \$27{,}834$. If we don't know the number of miles driven for the additional truck, we can't use the regression line to make a prediction. What should we do? Our best strategy is to use the mean price of the other 16 trucks as our prediction.



The **coefficient of determination** r^2 is the fraction of the variation in the values of y that is accounted for by the least-squares regression line of y on x. We can calculate r^2 using the following formula:

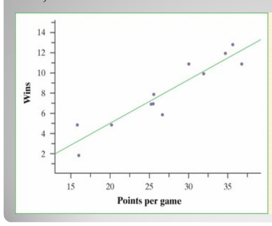
$$r^2 = 1 - \frac{\sum \text{residuals}^2}{\sum (y_i - \bar{y})^2}$$

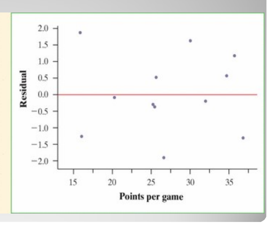
r^2 tells us how much better the LSRL does at predicting values of y than simply guessing the mean y for each value in the dataset.

AP EXAM TIP: Students often have a hard time interpreting the value of r^2 on AP exam questions. They frequently leave out key words in the definition. Our advice: Treat this as a fill-in-the-blank exercise. Write "____% of the variation in [response variable name] is accounted for by the LSRL of y **(context)** on x **(context)**."

STOP

Example 5: In Section 3.1, we looked at the relationship between the average number of points scored per game x and the number of wins y for the 12 college football teams in the Southeastern Conference. A scatterplot with the least-squares regression line and a residual plot are shown. The equation of the least-squares regression line is $\hat{y} = -3.75 + 0.437x$. Also, $s = 1.24$ and $r^2 = 0.88$.





a) Calculate and interpret the residual for South Carolina, which scored 30.1 points per game and had 11 wins.

The predicted amount of wins for South Carolina is

$$\hat{y} = -3.75 + 0.437(30.1) = 9.40 \text{ wins}$$

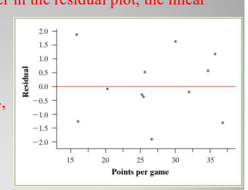
The residual for South Carolina is

$$\text{residual} = y - \hat{y} = 11 - 9.40 = 1.60 \text{ wins}$$

South Carolina won 1.60 more games than expected, based on the number of points they scored per game.

b) Is a linear model appropriate for these data? Explain.

Because there is no obvious pattern left over in the residual plot, the linear model is appropriate.



c) Interpret the value $s = 1.24$.

The standard deviation of the residual, 1.24, measures the typical distance between the actual number of wins and their predicted number of wins.

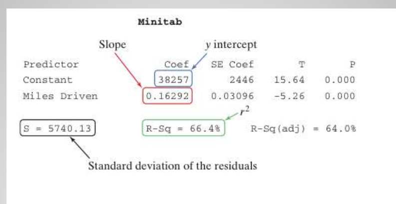
d) Interpret the value $r^2 = 0.88$.

About 88% of the variation in wins is accounted for by the LSRL of wins on points per game.

Interpreting Computer Regression Output

A number of statistical software packages produce similar regression output. Be sure you can locate

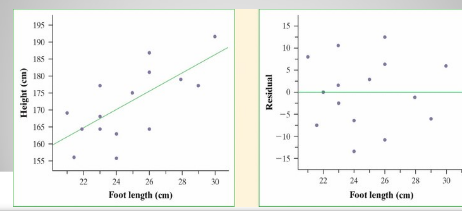
the slope b, the y intercept a, the values of s and r^2.



Example 6: A random sample of 15 high school students was selected from the U.S.CensusAtSchool database. The foot length (in centimeters) and height (in centimeters) of each student in the sample were recorded. Least-squares regression was performed on the data. A scatterplot with the regression line added, a residual plot, and some computer output from the regression are shown below.

Predictor	Coef	SE Coef	T	P
Constant	103.41	19.50	5.30	0.000
Foot length	2.7469	0.7833	3.51	0.004

S = 7.95126 R-Sq = 48.6% R-Sq(adj) = 44.7%



a) What is the equation of the least-squares regression line that describes the relationship between foot length and height? Define any variables you use.

Predictor	Coef	SE Coef	T	P
Constant	103.41	19.50	5.30	0.000
Foot length	2.7469	0.7833	3.51	0.004

S = 7.95126 R-Sq = 48.6% R-Sq(adj) = 44.7%

The equation is $\hat{y} = 103.41 + 2.7469\,x$, where $\hat{y}$ = predicted height (in centimeters) and x is foot length (in centimeters).

We could also write:
predicted height = 103.41 + 2.7469 (foot length)

b) Interpret the slope of the regression line in context.

For each additional centimeter of foot length, the least-squares regression line predicts an increase of 2.7469 cm in height.

c) Find the correlation.

To find the correlation, we take the square root of r^2:

$$r = \pm\sqrt{0.486} = \pm 0.697.$$

Because the scatterplot shows a positive association, $r = 0.697$.

d) Is a line an appropriate model to use for these data? Explain how you know.

Because the scatterplot shows a linear association and the residual plot has no obvious leftover patterns, a line is an appropriate model to use for these data.

Calculating the Equation of the Least-Squares Line

We can use technology to find the equation of the least-squares regression line. We can also write it in terms of the means and standard deviations of the two variables and their correlation.

How to Calculate the Least-Squares Regression Line

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations of the two variables and their correlation r.

The least-squares regression line is the line $\hat{y} = a + bx$ with **slope**

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

And ***y* intercept**

$$a = \bar{y} - b\bar{x}$$

AP EXAM TIP: The formula sheet for the AP exam uses different notation for these equations:

$$b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1\bar{x}$$

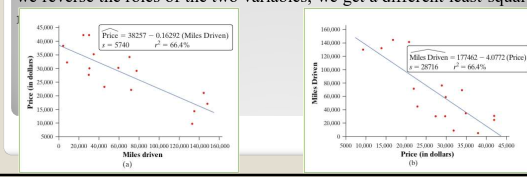
That's because the least-squares line is written as $\hat{y} = b_0 + b_1 x$.
We prefer our simpler versions without the subscripts!

Example 7: In the previous example, we used data from a random sample of 15 high school students to investigate the relationship between foot length (in centimeters) and height (in centimeters). The mean and standard deviation of the foot lengths are $\overline{x} = 24.76$ cm and $s_x = 2.71$ cm. The mean and standard deviation of the heights are $\overline{y} = 171.43$ cm and $s_y = 10.69$ cm. The correlation between foot length and height is $r = 0.697$. Find the equation of the least-squares regression line for predicting height from foot length. Show your work.

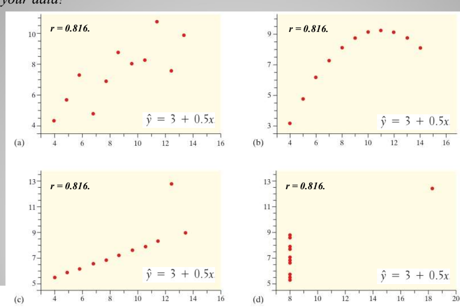
Correlation and Regression Wisdom

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, you should be aware of their limitations.

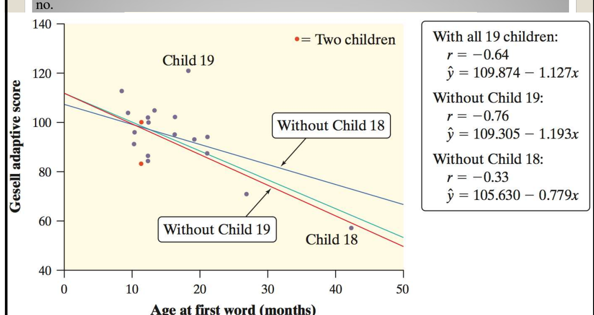
1. *The distinction between explanatory and response variables is important in regression.* This isn't true for correlation: switching x and y doesn't affect the value of r. Least-squares regression makes the distances of the data points from the line small only in the y direction. If we reverse the roles of the two variables, we get a different least-squares l



2. *Correlation and regression lines describe only linear relationships.* You can calculate the correlation and the least-squares line for any relationship between two quantitative variables, but the results are useful only if the scatterplot shows a linear pattern. *Always plot your data!*



3. *Correlation and least-squares regression lines are not resistant.* You already know that the correlation r is not resistant. One unusual point in a scatterplot can greatly change the value of r. Is the least-squares line resistant? Not surprisingly, the answer is no.



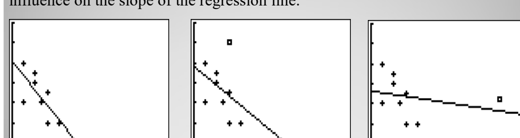
Least-squares lines make the sum of the squares of the vertical distances to the points as small as possible. A point that is extreme in the x direction with no other points near it pulls the line toward itself. We call such points **influential**.

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction but not the x direction of a scatterplot have large residuals. Other outliers may not have large residuals.

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line.

The best way to verify that a point is influential is to find the regression line both with and without the unusual point. If the line moves more than a small amount when the point is deleted, the point is influential.

Graphs I, II, and III are the same except for the point symbolized by the box in graphs II and III. Graph I below has no outliers or influential points. Graph II has an outlier that has an effect on the correlation. Graph III has an outlier that is an influential point that has an effect on the regression slope. Compare the correlation coefficient and regression lines for each graph. Note that the outlier in Graph II has some effect on the slope and a significant effect on the correlation coefficient. The influential point in Graph III has about the same effect on the correlation coefficient as the outlier in Graph II, but a major influence on the slope of the regression line.



Graph I: $\hat{y} = -0.73x + 4.06$
$r = -0.69$

Graph II: $\hat{y} = -0.49x + 3.83$
$r = -0.36$

Graph III: $\hat{y} = -0.09x + 2.64$
$r = -0.26$

4. *Association does not imply causation*. When we study the relationship between two variables, we often hope to show that changes in the explanatory variable *cause* changes in the response variable. *A strong association between two variables is not enough to draw conclusions about cause and effect*. Sometimes an observed association really does reflect cause and effect. A household that heats with natural gas uses more gas in colder months because cold weather requires burning more gas to stay warm. In other cases, an association is explained by lurking variables, and the conclusion that x causes y is not valid.

Association does not Imply Causation
An association between an explanatory variable x and a response variable y, even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y.

A serious study once found that people with two cars live longer than people who only own one car. Owning three cars is even better, and so on. There is a substantial positive correlation between number of cars x and length of life y.

Correlations such as those in the previous example are sometimes called "nonsense correlations." The correlation is real. What is nonsense is the conclusion that changing one of the variables causes changes in the other. A "lurking variable"—such as personal wealth in this example—that influences both x and y can create a high correlation even though there is no direct connection between x and y.