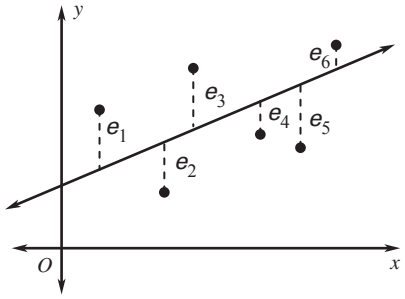
LESSON
31 Goodness of Fit

GOAL Use indicators of goodness of fit to analyze and compare linear models.

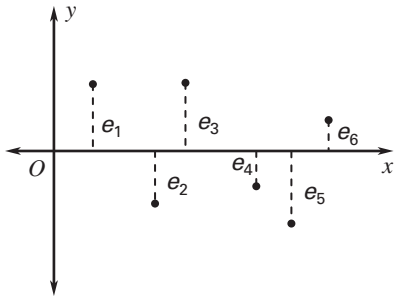
You have already learned how to fit a least-squares line to a set of paired data and how to fit a median-median line to a set of paired data. Unless the data form a set of collinear points, neither of these linear models will pass through all the data points. In this lesson, you will consider different ways to measure how well a linear model fits a set of data.

Given a set of data and a linear model, the difference between an actual value of the dependent variable y and the value that is predicted by the linear model (denoted $\hat{y}$ in this lesson) is called a **residual**.

The top figure shows a scatter plot for a set of data and a linear model. The residuals, e_1 through e_6, are the "signed" vertical distances between the data points and the line. That is, the residual may be positive (when the data point lies above the line) or negative (when the data point lies below the line). Note that the residual is 0 when a data point lies on the line. The letter e (for "error") is often used for residuals.



The bottom figure shows a *residual plot*, which is a scatter plot of points whose x-values are those from the data set and whose y-values are the corresponding residuals. If a line is a good fit for a set of data, the absolute values of the residuals are relatively small and more or less evenly distributed above and below the x-axis in a residual plot. This gives an informal idea of what is meant by the goodness of fit of a linear model.



One numerical indicator of the goodness of fit of a linear model is the **sum of the squared errors**, which is obtained by finding the residuals, squaring these values, and then finding their sum. The closer the sum of the squared errors is to 0, the better the linear model fits the data.

Sum of the Squared Errors

Use these steps to find the sum of the squared errors for a linear model.

1. Find the residuals (i.e., the difference between each y-value in the data set and the y-value predicted by the model).
2. Square the residuals.
3. Find the sum of the squared residuals.

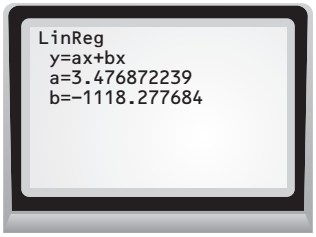
EXAMPLE 1 Calculating the Sum of the Squared Errors

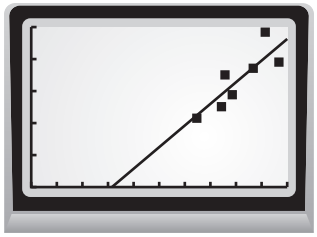
The table shows the lengths (in feet) and passenger capacities of various cruise ships. Use a graphing calculator to find the least-squares line for the data. Then calculate the sum of the squared errors.

Length, *x* (ft)	644	720	754	781	866	915	965
Capacity, *y*	1090	1266	1748	1440	1870	2435	1950

SOLUTION

First find the least-squares line using a graphing calculator. Rounding the values of a and b to three significant digits, you obtain the equation $y = 3.48x - 1120$.





Calculate the sum of the squared errors. Use the graphing calculator's *table* feature to find the y-values predicted by the model.

x	644	720	754	781	866	915	965
y	1090	1266	1748	1440	1870	2435	1950
$\hat{y}$	1121	1386	1504	1598	1894	2064	2238
$y - \hat{y}$	−31	−120	244	−158	−24	371	−288
$(y - \hat{y})^2$	961	14,400	59,536	24,964	576	137,641	82,944

The sum of the squared errors is

$961 + 14{,}400 + 59{,}536 + 24{,}964 + 576 + 137{,}641 + 82{,}944 = 321{,}022.$

Another goodness-of-fit indicator is the **mean absolute deviation**, which is obtained by first finding the absolute value of the difference between each y-value in the data set and the y-value predicted by the model and then finding the mean of the absolute values. The closer the mean absolute deviation is to 0, the better the linear model fits the data.

Mean Absolute Deviation

Use these steps to find the mean absolute deviation for a linear model.

1. Find the residuals (i.e., the difference between each y-value in the data set and the y-value predicted by the model).
2. Find the absolute value of the residuals.
3. Find the mean of the absolute values.

EXAMPLE 2 Calculating the Mean Absolute Deviation

Use the data and the least-squares line from Example 1 to calculate the mean absolute deviation.

Length, *x* (ft)	644	720	754	781	866	915	965
Capacity, *y*	1090	1266	1748	1440	1870	2435	1950

SOLUTION

The least-squares line for the data from Example 1 was determined to be $y = 3.48x - 1120$.

Modify the last row of the table from Example 1 to show the absolute value of the difference between each y-value and the predicted y-value.

x	644	720	754	781	866	915	965
y	1090	1266	1748	1440	1870	2435	1950
$\hat{y}$	1121	1386	1504	1598	1894	2064	2238
$y - \hat{y}$	–31	–120	244	–158	–24	371	–288
$\lvert y - \hat{y} \rvert$	31	120	244	158	24	371	288

The mean absolute deviation is

$$\frac{31 + 120 + 244 + 158 + 24 + 371 + 288}{7} \approx 177.$$

CHECK Examples 1 and 2

The table shows the number of games won and the number of points scored by seven teams during the National Football League's 2008 season.

Games won, *x*	2	4	5	7	8	11	12
Points scored, *y*	232	294	263	336	370	385	427

1. Use a graphing calculator to find the least-squares line for the data. Then calculate the sum of the squared errors.
2. Calculate the mean absolute deviation.

Notice that the errors, or residuals, are squared when calculating the sum of the squared errors. The absolute value of the errors is used when calculating the mean absolute deviation. These steps prevent positive and negative errors from "canceling" each other. Thus, the sum of the squared errors and the mean absolute deviation are both always greater than or equal to zero.

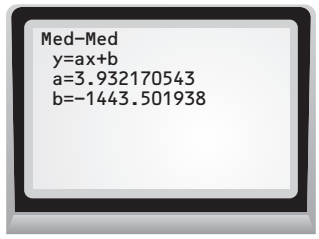
In Examples 1 and 2, you calculated the sum of the squared errors and the mean absolute deviation of a least-squares line. You can also calculate these indicators for a median-median line.

EXAMPLE 3 Analyzing Goodness of Fit of a Median-Median Line

Use a graphing calculator to find the median-median line for the data in Example 1. Then calculate the sum of the squared errors and the mean absolute deviation.

SOLUTION

With the data already entered into the calculator, press STAT, choose the CALC menu, and select Med-Med.



Rounding the values of a and b, you obtain the equation $y = 3.93x - 1440$.

The calculations for the sum of the squared errors and the mean absolute deviation are shown below.

x	644	720	754	781	866	915	965
y	1090	1266	1748	1440	1870	2435	1950
$\hat{y}$	1091	1390	1523	1629	1963	2156	2352
$y - \hat{y}$	−1	−124	225	−189	−93	279	−402
$(y - \hat{y})^2$	1	15,376	50,625	35,721	8649	77,841	161,604
$\lvert y - \hat{y} \rvert$	1	124	225	189	93	279	402

The sum of the squared errors for the median-median line is

$1 + 15{,}376 + 50{,}625 + 35{,}721 + 8649 + 77{,}841 + 161{,}604 = 349{,}817.$

The mean absolute deviation for the median-median line is

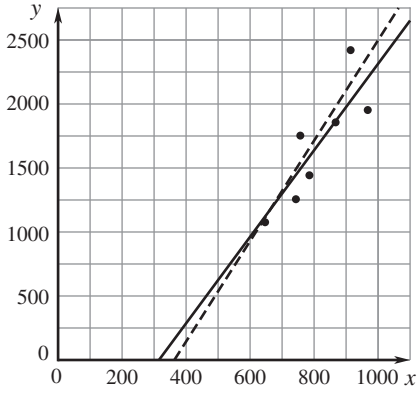
$$\frac{1 + 124 + 225 + 189 + 93 + 279 + 402}{7} \approx 188.$$

EXAMPLE 4 Comparing Linear Models Visually

Compare the linear models in Examples 1 and 3 visually.

SOLUTION

Graph the least-squares line (shown with a solid line) and the median-median line (shown with a dashed line) on the same coordinate plane, along with the data. Notice that both lines very nearly pass through the data points (644, 1090) and (866, 1870). Also, while the lines are close together for x-values between 600 and 800, they diverge for x-values between 800 and 1000.



EXAMPLE 5 Comparing Linear Models Using Goodness of Fit

Compare the linear models in Examples 1 and 3 using the sum of the squared errors and the mean absolute deviation.

SOLUTION

The goodness-of-fit indicators for the least-squares line are:

Sum of the squared errors: 321,022 (Example 1)

Mean absolute deviation: 177 (Example 2)

The goodness-of-fit indicators for the median-median line are:

Sum of the squared errors: 349,817 (Example 3)

Mean absolute deviation: 188 (Example 3)

The sum of the squared errors and the mean absolute deviation are both lower for the least-squares fit line. By both measures of goodness of fit, the least-squares line fits the data better than the median-median line.

CHECK Examples 3, 4, and 5

In Exercises 3–5, use the football data from Exercises 1 and 2.

Games won, *x*	2	4	5	7	8	11	12
Points scored, *y*	232	294	263	336	370	385	427

3. Use a graphing calculator to find the median-median line for the data. Then calculate the sum of the squared errors and the mean absolute deviation.

4. Compare the least-squares line and the median-median line visually.

5. Compare the least-squares line and the median-median line using the sum of the squared errors and the mean absolute deviation.

EXERCISES

1. The table gives the engine displacement (in liters) and horsepower for five cars from the same manufacturer for the same model year.

Engine displacement, *x* (L)	1.6	2.2	2.4	3.5	6.2
Horsepower, *y*	106	155	169	211	430

a. Find the least-squares line for the data.

b. Find the sum of the squared errors for the least-squares line.

c. Find the mean absolute deviation for the least-squares line.

2. Repeat Exercise 1, but this time use the median-median line.

3. Compare the linear models from Exercises 1 and 2 visually.

4. Compare the linear models from Exercises 1 and 2 using the sum of the squared errors and the mean absolute deviation.

5. The table gives the number of floors and the heights (in meters) of eight tall office buildings in London.

Number of floors, *x*	50	47	45	41	35	33	32	26
Heights, *y* (m)	235.1	183.0	199.5	179.8	164.3	153.0	151.0	124.9

a. Find the least-squares line for the data.

b. Find the sum of the squared errors for the least-squares line.

c. Find the mean absolute deviation for the least-squares line.

6. Repeat Exercise 5, but this time use the median-median line.

7. Compare the linear models from Exercises 5 and 6 visually.

8. Compare the linear models from Exercises 5 and 6 using the sum of the squared errors and the mean absolute deviation.

9. The table gives the sepal length and width (both in centimeters) of seven irises from the same species.

Length, *x* (cm)	5.1	4.9	4.7	4.6	5.0	5.4	4.6
Width, *y* (cm)	3.5	3.0	3.2	3.1	3.6	3.9	3.4

a. Find the least-squares line for the data.

b. Find the sum of the squared errors for the least-squares line.

c. Find the mean absolute deviation for the least-squares line.

10. Repeat Exercise 9, but this time use the median-median line.

11. Compare the linear models from Exercises 9 and 10 visually.

12. Compare the linear models from Exercises 9 and 10 using the sum of the squared errors and the mean absolute deviation.

13. The table gives Hank Aaron's number of at bats and home runs during his last eight seasons as a Major League Baseball player.

At bats, *x*	547	516	495	449	392	340	465	271
Home runs, *y*	44	38	47	34	40	20	12	10

a. Find the least-squares line for the data.

b. Find the sum of the squared errors for the least-squares line.

c. Find the mean absolute deviation for the least-squares line.

14. Repeat Exercise 13, but this time use the median-median line.

15. Compare the linear models from Exercises 13 and 14 visually.

16. Compare the linear models from Exercises 13 and 14 using the sum of the squared errors and the mean absolute deviation.

17. What must be true about a set of data if the sum of the squared errors or the mean absolute deviation for a line fit to the data is 0? Explain.