

10.E: Outliers (Optional Exercises)

Use the following information to answer the next four exercises. The scatter plot shows the relationship between hours spent studying and exam scores. The line shown is the calculated line of best fit. The correlation coefficient is 0.69.

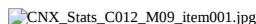
 CNX_Stats_C012_M09_item001.jpg

Figure 10.E. 5.

? Exercise 12.7.4

Do there appear to be any outliers?

Answer

Yes, there appears to be an outlier at (6, 58).

? Exercise 12.7.5

A point is removed, and the line of best fit is recalculated. The new correlation coefficient is 0.98. Does the point appear to have been an outlier? Why?

? Exercise 12.7.6

What effect did the potential outlier have on the line of best fit?

Answer

The potential outlier flattened the slope of the line of best fit because it was below the data set. It made the line of best fit less accurate as a predictor for the data.

? Exercise 12.7.7

Are you more or less confident in the predictive ability of the new line of best fit?

? Exercise 12.7.8

The Sum of Squared Errors for a data set of 18 numbers is 49. What is the standard deviation?

Answer

$s = 1.75$

? Exercise 12.7.9

The Standard Deviation for the Sum of Squared Errors for a data set is 9.8. What is the cutoff for the vertical distance that a point can be from the line of best fit to be considered an outlier?

Bring It Together

? Exercise 12.7.10

The average number of people in a family that received welfare for various years is given in Table.

Year	Welfare family size
1969	4.0
1973	3.6
1975	3.2

Year	Welfare family size
1979	3.0
1983	3.0
1988	3.0
1991	2.9

- Using “year” as the independent variable and “welfare family size” as the dependent variable, draw a scatter plot of the data.
- Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
- Find the correlation coefficient. Is it significant?
- Pick two years between 1969 and 1991 and find the estimated welfare family sizes.
- Based on the data in Table, is there a linear relationship between the year and the average number of people in a welfare family?
- Using the least-squares line, estimate the welfare family sizes for 1960 and 1995. Does the least-squares line give an accurate estimate for those years? Explain why or why not.
- Are there any outliers in the data?
- What is the estimated average welfare family size for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.
- What is the slope of the least squares (best-fit) line? Interpret the slope.

? Exercise 12.7.11

The percent of female wage and salary workers who are paid hourly rates is given in Table for the years 1979 to 1992.

Year	Percent of workers paid hourly rates
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

- Using “year” as the independent variable and “percent” as the dependent variable, draw a scatter plot of the data.
- Does it appear from inspection that there is a relationship between the variables? Why or why not?
- Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
- Find the correlation coefficient. Is it significant?
- Find the estimated percents for 1991 and 1988.
- Based on the data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
- Are there any outliers in the data?

- h. What is the estimated percent for the year 2050? Does the least-squares line give an accurate estimate for that year? Explain why or why not.
- i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Answer

- a. Check student's solution.
- b. yes
- c. $\hat{y} = -266.8863 + 0.1656x$
- d. 0.9448; Yes
- e. 62.8233; 62.3265
- f. yes
- g. yes; (1987, 62.7)
- h. 72.5937; no
- i. $slope = 0.1656$.

As the year increases by one, the percent of workers paid hourly rates tends to increase by 0.1656.

Use the following information to answer the next two exercises. The cost of a leading liquid laundry detergent in different sizes is given in Table.

Size (ounces)	Cost (\$)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

? Exercise 12.7.12

- a. Using “size” as the independent variable and “cost” as the dependent variable, draw a scatter plot.
- b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
- c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
- d. Find the correlation coefficient. Is it significant?
- e. If the laundry detergent were sold in a 40-ounce size, find the estimated cost.
- f. If the laundry detergent were sold in a 90-ounce size, find the estimated cost.
- g. Does it appear that a line is the best way to fit the data? Why or why not?
- h. Are there any outliers in the given data?
- i. Is the least-squares line valid for predicting what a 300-ounce size of the laundry detergent would you cost? Why or why not?
- j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

? Exercise 12.7.13

- a. Complete Table for the cost per ounce of the different sizes.
- b. Using “size” as the independent variable and “cost per ounce” as the dependent variable, draw a scatter plot of the data.
- c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
- d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
- e. Find the correlation coefficient. Is it significant?
- f. If the laundry detergent were sold in a 40-ounce size, find the estimated cost per ounce.
- g. If the laundry detergent were sold in a 90-ounce size, find the estimated cost per ounce.
- h. Does it appear that a line is the best way to fit the data? Why or why not?
- i. Are there any outliers in the the data?

- j. Is the least-squares line valid for predicting what a 300-ounce size of the laundry detergent would cost per ounce? Why or why not?
- k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Answer

a.

Size (ounces)	Cost (\$)	cents/oz
16	3.99	24.94
32	4.99	15.59
64	5.99	9.36
200	10.99	5.50

- b. Check student's solution.
- c. There is a linear relationship for the sizes 16 through 64, but that linear trend does not continue to the 200-oz size.
- d. $\hat{y} = 20.2368 - 0.0819x$
- e. $r = -0.8086$
- f. 40-oz: 16.96 cents/oz
- g. 90-oz: 12.87 cents/oz
- h. The relationship is not linear; the least squares line is not appropriate.
- i. no outliers
- j. No, you would be extrapolating. The 300-oz size is outside the range of x .
- k. $slope = -0.08194$; for each additional ounce in size, the cost per ounce decreases by 0.082 cents.

? Exercise 12.7.14

According to a flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate (\$)	Approximate Probate Fees and Taxes (\$)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

- a. Decide which variable should be the independent variable and which should be the dependent variable.
- b. Draw a scatter plot of the data.
- c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
- d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
- e. Find the correlation coefficient. Is it significant?
- f. Find the estimated total cost for a next taxable estate of \$1,000,000. Find the cost for \$2,500,000.
- g. Does it appear that a line is the best way to fit the data? Why or why not?
- h. Are there any outliers in the data?
- i. Based on these results, what would be the probate fees and taxes for an estate that does not have any assets?
- j. What is the slope of the least-squares (best-fit) line? Interpret the slope.

? Exercise 12.7.15

The following are advertised sale prices of color televisions at Anderson's.

Size (inches)	Sale Price (\$)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the data.
- Does it appear from inspection that there is a relationship between the variables? Why or why not?
- Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
- Find the correlation coefficient. Is it significant?
- Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.
- Does it appear that a line is the best way to fit the data? Why or why not?
- Are there any outliers in the data?
- What is the slope of the least-squares (best-fit) line? Interpret the slope.

Answer

- Size is x , the independent variable, price is y , the dependent variable.
- Check student's solution.
- The relationship does not appear to be linear.
- $\hat{y} = -745.252 + 54.75569x$
- $r = 0.8944$, yes it is significant
- 32-inch: \$1006.93, 50-inch: \$1992.53
- No, the relationship does not appear to be linear. However, r is significant.
- yes, the 60-inch TV
- For each additional inch, the price increases by \$54.76

? Exercise 12.7.16

Table shows the average heights for American boys in 1990.

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

- Decide which variable should be the independent variable and which should be the dependent variable.
- Draw a scatter plot of the data.
- Does it appear from inspection that there is a relationship between the variables? Why or why not?
- Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
- Find the correlation coefficient. Is it significant?
- Find the estimated average height for a one-year-old. Find the estimated average height for an eleven-year-old.
- Does it appear that a line is the best way to fit the data? Why or why not?
- Are there any outliers in the data?
- Use the least squares line to estimate the average height for a sixty-two-year-old man. Do you think that your answer is reasonable? Why or why not?
- What is the slope of the least-squares (best-fit) line? Interpret the slope.

? Exercise 12.7.17

State	# letters in name	Year entered the Union	Ranks for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

- What are the independent and dependent variables?
- What do you think the scatter plot will look like? Make a scatter plot of the data.
- Does it appear from inspection that there is a relationship between the variables? Why or why not?
- Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
- Find the correlation coefficient. What does it imply about the significance of the relationship?
- Find the estimated areas for Alabama and for Colorado. Are they close to the actual areas?
- Use the two points in part f to plot the least-squares line on your graph from part b.
- Does it appear that a line is the best way to fit the data? Why or why not?
- Are there any outliers?
- Use the least squares line to estimate the area of a new state that enters the Union. Can the least-squares line be used to predict it? Why or why not?
- Delete "Hawaii" and substitute "Alaska" for it. Alaska is the forty-ninth, state with an area of 656,424 square miles.
- Calculate the new least-squares line.
- Find the estimated area for Alabama. Is it closer to the actual area with this new least-squares line or with the previous one that included Hawaii? Why do you think that's the case?
- Do you think that, in general, newer states are larger than the original states?

Answer

- a. Let rank be the independent variable and area be the dependent variable.
- b. Check student's solution.
- c. There appears to be a linear relationship, with one outlier.
- d. $\hat{y}(\text{area}) = 24177.06 + 1010.478x$
- e. $r = 0.50047$, r is not significant so there is no relationship between the variables.
- f. Alabama: 46407.576 Colorado: 62575.224
- g. Alabama estimate is closer than Colorado estimate.
- h. If the outlier is removed, there is a linear relationship.
- i. There is one outlier (Hawaii).
- j. rank 51: 75711.4; no

k.

Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Alaska	6	1959	51	656,424
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

- l. $\hat{y} = -87065.3 + 7828.532x$
- m. Alabama: 85,162.404; the prior estimate was closer. Alaska is an outlier.
- n. yes, with the exception of Hawaii

10.E: Outliers (Optional Exercises) is shared under a [CC BY](https://creativecommons.org/licenses/by/4.0/) license and was authored, remixed, and/or curated by LibreTexts.