Lesson 15: Interpreting Residuals from a Line

Classwork

Example 1: Residuals

One way to think about how useful a line is for describing a relationship between two variables is to use the line to predict the values for the points in the scatter plot. These predicted values could then be compared to the actual

 values. For example, the first data point in the table from Lesson 14 represents a man with a shoe length of inches and height of inches. If you use the line to predict this man’s height, you would get:

Because his actual height was inches, you can calculate the prediction error by subtracting the predicted value from the actual value. This prediction error is called a *residual*. For the first data point, the residual is calculated as follows:

1. For the line , calculate the missing values and add them to complete the table.

|  |  |  |  |
| --- | --- | --- | --- |
| = Shoe Length |  = Height | Predicted *-*value | Residual |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

2. Why is the residual in the table’s first row positive, and the residual in the second row negative?

3. What is the sum of the residuals? Why did you get a number close to zero for this sum? Does this mean that all of the residuals were close to 0?

When you use a line to describe the relationship between two numerical variables, **the *best* line is the line that makes the residuals as small as possible overall.**

4. If the residuals tend to be small, what does that say about the fit of the line to the data?

**The most common choice for the *best* line is the line that makes the sum of the *squared* residuals as small as possible.**

Add a column on the right of the table in Exercise 8. Calculate the square of each residual and place the answer in the column**.**

1. Why do we use the sum of the squared residuals instead of just the sum of the residuals (without squaring)? Hint: Think about whether the sum of the residuals for a line can be small even if the prediction errors are large. Can this happen for squared residuals?

1. What is the sum of the squared residuals for the line and the data of Exercise 1?

The line that has a smaller sum of squared residuals for this data set than any other line is called the ***least-squares line*.** This line can also be called the ***best-fit line* or the *line of best fit* (or regression line).** For the shoe-length and height data for the sample of 10 men, the line is the least-squares line. No other line would have a smaller sum of squared residuals for this data set than this line. Once you have found the equation of the least-squares line, the values of the slope and -intercept of the line often reveals something interesting about the relationship you are modeling. The slope of the least-squares line is the change in the predicted value of the variable associated with an increase of one in the value of the -variable.

1. Give an interpretation of the slope of the least-squares line *x* for predicting height from shoe size for adult men.
2. Explain why it does not make sense to interpret the -intercept of 25.3 as the predicted height for an adult male whose shoe length is zero.

Example 2: Calculating Prediction Errors

The gestation time for an animal is the typical duration between conception and birth. The longevity of an animal is the typical lifespan for that animal. The gestation times (in days) and longevities (in years) for 13 types of animals are shown in the table below along with the scatter plot.

|  |  |  |
| --- | --- | --- |
| **Animal** | **Gestation Time (days)** | **Longevity (years)** |
| Baboon | 187 | 20 |
| Black Bear | 219 | 18 |
| Beaver | 105 | 5 |
| Bison | 285 | 15 |
| Cat | 63 | 12 |
| Chimpanzee | 230 | 20 |
| Cow | 284 | 15 |
| Dog | 61 | 12 |
| Fox (Red) | 52 | 7 |
| Goat | 151 | 8 |
| Lion | 100 | 15 |
| Sheep | 154 | 12 |
| Wolf | 63 | 5 |

Exercises 1–4

Finding the equation of the least-squares line relating longevity to gestation time for these types of animal provides the equation to predict longevity. How good is the line? In other words, if you were given the gestation time for another type of animal not included in the original list, how accurate would the least-squares line be at predicting the longevity of that type of animal?

1. Using a graphing calculator, verify that the equation of the least-squares line is:

, where represents the gestation time (in days) and represents longevity in years.

The least-squares line has been added to the scatter plot.



2. Suppose a particular type of animal has a gestation

time of 200 days. Approximately what value does the line

predict for the longevity of that type of animal?

3. Would the value you predicted in question (2) necessarily be

the exact value for the longevity of that type of animal?

Could the actual longevity of that type of animal be longer

than predicted? Could it be shorter?

You can investigate further by looking at the types of animal included in the original data set. Take the lion, for example. Its gestation time is days. You also know that its longevity is years, but what does the least-squares line *predict* for the lion’s longevity? Substituting days into the equation, you get: or approximately . The least-squares line predicts the lion’s longevity to be approximately years.

4. How close is this to being correct? More precisely, how much do you have to add to 10.6 to get the lion’s true longevity of 15?

You can show the prediction error on the graph above:

Exercises 5–6

5. Let’s continue to think about the gestation times and longevities of animals. Let’s specifically investigate how accurately the least-squares line predicted the longevity of the black bear.

* 1. What is the gestation time for the black bear?
	2. Use the gestation time from (a) and the least-squares line to predict the black bear’s longevity. Round your answer to the nearest tenth.
	3. What is the actual longevity of the black bear?
	4. How much do you have to add to the predicted value to get the actual longevity of the black bear?
	5. Show your answer to part (e) on the graph as a vertical line segment.

6. Repeat this activity for the sheep.

* 1. Substitute the sheep’s gestation time for *x* into the equation to find the predicted value for the sheep’s longevity. Round your answer to the nearest tenth.
	2. What do you have to add to the predicted value in order to get the actual value of the sheep’s longevity? (Hint: Your answer should be negative.)
	3. Show your answer to part (b) on the graph as a vertical line segment. Write a sentence describing points in the graph for which a negative number would need to be added to the predicted value in order to get the actual value.

In each example above, you found out how much needs to be added to the predicted value in order to find the true value of the animal’s longevity. In order to find this you have been calculating:

actual value – predicted value

This quantity is referred to as a residual. It is summarized as:

residual = actual *y* value – predicted *y* value

You can now work out the residuals for all of the points in our animal longevity example. The values of the residuals are shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Animal** | **Gestation Time (days)** | **Longevity (years)** | **Residual** |
| Baboon | 187 | 20 | 5.9 |
| Black Bear | 219 | 18 | 2.7 |
| Beaver | 105 | 5 | −5.8 |
| Bison | 285 | 15 | −3.0 |
| Cat | 63 | 12 | 2.9 |
| Chimpanzee | 230 | 20 | 4.2 |
| Cow | 284 | 15 | −2.9 |
| Dog | 61 | 12 | 2.9 |
| Fox (Red) | 52 | 7 | −1.7 |
| Goat | 151 | 8 | −4.6 |
| Lion | 100 | 15 | 4.4 |
| Sheep | 154 | 12 | −0.8 |
| Wolf | 63 | 5 | −4.1 |

These residuals show that the actual longevity of an animal should be within six years of the longevity predicted by the least-squares line.

Suppose you selected a type of animal that is not included in the original data set, and the gestation time for this type of animal is 270 days. Substituting *x* = 270 into the equation of the least-squares line you get:

years.

Exercises 7–8

Think about what the *actual* longevity of this type of animal might be.

7. Could it be 30 years? How about 5 years?

8. Judging by the size of the residuals in our table, what kind of values do you think would be reasonable for the longevity of this type of animal?

Exercises 9–10

Continue to think about the gestation times and longevities of animals. The gestation time for the type of animal called the ocelot is known to be 85 days. The least-squares line predicts the longevity of the ocelot to be:

 years

9. Based on the residuals in Example 2, would you be surprised to find that the longevity of the ocelot was 2 years? Why, or why not? What might be a sensible range of values for the actual longevity of the ocelot?

10. We know that the actual longevity of the ocelot is 9 years. What is the residual for the ocelot?

Lesson Summary

* When a least-squares line is used to calculate a predicted value, the prediction error can be measured by:

residual = actual -value – predicted -value

* On the graph, the residuals are the vertical distances of the points from the least-squares line.
* The residuals give us an idea how close a prediction might be when the least-squares line is used to make a prediction for a value that is not included in the data set.

Problem Set

The time spent in surgery and the cost of surgery was recorded for six patients. The results and scatter plot are shown below.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |  |
| --- | --- |
| **Time (minutes)** | **Cost ($)** |
| 14 | 1,510 |
| 80 | 6,178 |
| 84 | 5,912 |
| 118 | 9,184 |
| 149 | 8,855 |
| 192 | 11,023 |

 |  |

1. Below is the equation of the least-squares line relating cost to time. (Indicate slope to the nearest tenth and -intercept to the nearest whole number.)

, where = time and = cost

1. Draw the least-squares line on the graph above. (Hint: Substitute = 30 into your equation to find the predicted *-* value. Plot the point (30, your answer) on the graph. Then substitute = 180 into the equation and plot the point. Join the two points with a straightedge.)
2. What does the least-squares line predict for the cost of a surgery that lasts 118 minutes? (Calculate the cost to the nearest cent.)
3. How much do you have to add to your answer to question (3) to get the actual cost of surgery for a surgery lasting 118 minutes? (This is the residual.)
4. Show your answer to question (4) as a vertical line between the point for that person in the scatter plot and the least-squares line.
5. Remember that the residual is the actual -value minus the predicted -value. Calculate the residual for the surgery that took 149 minutes and cost $8,855.
6. Calculate the other residuals, and write all the residuals in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Time (minutes) | Cost ($) | Predicted y-value ($) | Residual |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

1. Suppose that a surgery took 100 minutes.
	1. What does the least-squares line predict for the cost of this surgery?
	2. Would you be surprised if the actual cost of this surgery were $9000? Why or why not?
	3. Interpret the slope of the least-squares line.