A line of best fit can model an association between two variables. In this lesson, lines of best fit are estimated by sketching them on a scatterplot.

Once a line of best fit is created, the slope and y-intercept can be determined. In statistical analyses, the slope is often written as the amount of change expected in the dependent variable (\(\Delta y\)), when the independent variable is changed by one unit (\(\Delta x = 1\)). The y-intercept is the predicted value of the dependent variable when the value of the independent variable is zero. In statistical scatterplots, the vertical axis is often not drawn at the origin, so the y-intercept can be someplace other than where the line of best fit crosses the vertical axis of the scatterplot.

Associations between two variables are often described by their form, direction, strength of association, and outliers. Form refers to the shape of the pattern in the scatterplot: linear, some type of curve, or perhaps no pattern at all. Direction refers to whether the pattern is in general positive (increasing from left to right) or negative (decreasing from left to right). For linear relationships, the slope can be used to describe the steepness in addition to the direction. The strength of the relationship indicates how closely the data are to the line of best fit. Outliers are data points far removed from the bulk of the data.

For additional information, see the Math Notes box in Lessons 4.1.2 and review the Checkpoint 7 materials. (Note that the following problems do not use a graphing calculator, while the problems in Checkpoint 7 assume a graphing calculator is available. Graphing calculators are introduced for statistical calculations in Lesson 4.1.4.)

Example

It seems reasonable that there would be a relationship between the amount of time a student spends studying and their GPA. Suppose you were interested in predicting a student’s GPA based on the hours they study per week. You were able to randomly select 12 students and obtain this information from each student.

<table>
<thead>
<tr>
<th>Hours of Study per Week</th>
<th>4</th>
<th>5</th>
<th>11</th>
<th>1</th>
<th>15</th>
<th>2</th>
<th>10</th>
<th>6</th>
<th>7</th>
<th>0</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPA</td>
<td>2.9</td>
<td>3.3</td>
<td>3.9</td>
<td>2.2</td>
<td>4.1</td>
<td>1.8</td>
<td>4.6</td>
<td>2.9</td>
<td>2.2</td>
<td>3</td>
<td>3.3</td>
<td>4.5</td>
</tr>
</tbody>
</table>

a. Without a calculator, make a scatterplot.

*Example continues on next page →*
Example continued from previous page.

b. Estimate a line of best fit, draw it with a ruler, and determine its equation.

If it is reasonable, draw the line of best fit through a point where grid lines intersect or a data point. That way, the equation will be easier to write. In this problem, the data points (1, 2.2) and (11, 3.9) seem to be on a line that “fits” the data.

Using those two points, and techniques from Lesson 2.3.2, the equation for the estimated line of best fit is

\[ y = 2.03 + 0.17x \]

where \( y \) is the predicted GPA and \( x \) is the number of hours per week the student studies. Note that the variables and their units were defined.

c. Interpret the slope and y-intercept in context.

The slope indicates that a student’s GPA is expected to increase by 0.17 points for every additional hour of studying per week. The y-intercept predicts that students who do not study at all will have a GPA of 2.03.

d. Describe the association.

The form is linear; it does not appear to be curved nor simply a collection of randomly scattered points. The direction is positive; in general, students who study more also have higher GPAs. A student’s GPA is expected to increase by 0.17 points for every additional hour of studying per week. The strength is moderate: it is strong enough to easily see its form, but there is scatter about the line. There do not appear to be any outliers.

In summary one could say there is a moderately strong linear relationship between study hours and GPA for students with no outliers in our data.

Problems

1. It seems reasonable that the power of a car is related to its gas mileage. Suppose a random sample of 10 cars of various models is selected and the engine power and city gas mileage is recorded for each one.

<table>
<thead>
<tr>
<th>Power (hp)</th>
<th>197</th>
<th>170</th>
<th>166</th>
<th>230</th>
<th>381</th>
<th>438</th>
<th>170</th>
<th>326</th>
<th>451</th>
<th>290</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mileage (mpg)</td>
<td>16</td>
<td>24</td>
<td>19</td>
<td>15</td>
<td>13</td>
<td>20</td>
<td>21</td>
<td>11</td>
<td>10</td>
<td>15</td>
</tr>
</tbody>
</table>

a. Create a scatterplot of the data.

b. Estimate a line of best fit and determine its equation.

c. Interpret the slope and y-intercept in context.

d. Describe the association.
2. Many people believe that students who are strong in music are also strong in mathematics. The principal at University High School wonders if that same connection exists between music students and English students. The principal went through the records for the past year and found 10 students who were enrolled in both Advanced Placement Music and Advanced Placement English. He compared their final exam scores.

a. Create a scatterplot of the data.

b. Estimate a line of best fit and determine its equation.

c. Interpret the slope and y-intercept in context.

d. Describe the association.

**Answers**

1. a. See scatterplot at right.

b. Answers can vary. If you ignore the possible outlier at (438, 20), a reasonable line could pass through the points (170, 21) and (290, 15). The line of best fit has equation $y = 29.5 - 0.05x$.

c. For every increase of 1 horsepower, the gas mileage is expected to decrease by 0.05 mpg. The y-intercept means that a 0 horsepower car would get 29.5 mpg. This does not make sense because the y-axis is far from the data and thus this is extrapolation. (Prediction models of all types are unreliable when you extrapolate them.)

d. When the outlier at (438, 20) is removed, there appears to be a strong, negative, linear relationship. If the outlier is not removed, the association is more moderate (not as strong). For every increase of 1 horsepower, gas mileage is expected to decrease by about 0.05 mpg.
2. a. See plot at right.

b. Answers can vary. A reasonable line could pass through the points (64, 90) and (97, 68). The line of best fit has equation $y = 132.7 - 0.67x$.

c. An increase of 1 point in the music score results in a predicted decrease of 0.67 points in the English score. The $y$-intercept would mean that a student who scored 0 on the music test is expected to score 133 on the English test. This does not make sense. The $y$-axis is far outside of the data and represents an extrapolation. (Prediction models of all types are unreliable when you extrapolate them.)

d. There is a lot of scatter around the line of best fit. The linear association is weak. An increase of 1 point in the music score results in a predicted decrease of 0.67 points in the English score. There are no apparent outliers.