Simple Linear Regression III – Least Squares Line

Example 1.

If we want to denote the trend describing the relation between quiz #1 and quiz #2 scores in example given on the previous lecture then we might want to “fit” a line into scatter diagram.

Apparently we made a choice of line that fits the data but how?
The idea applied in the plot above is so called method of the least squares. The line given by that method is called the least-squares line or simple regression line.

Using slope-intercept form we need the equation to relate $X$ and $Y$ such as

$$Y = a + b \cdot X$$  \hspace{1cm} (1.1)

Let’s denote quiz #1 score as $X$ and quiz #2 score as $Y$. We are looking for the $b$, the slope and $a$, the intercept of the line so that the sum

$$\sum \left( y - (a + bx) \right)^2$$  \hspace{1cm} (1.2)

is the smallest possible. We are minimizing the sum of squares of “vertical distance” from points to the line. (Attend the lecture for more explanation!)

Using calculus methods that are beyond the scope of our course it is easy to show that the slope must be

$$b = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{n \cdot \sum x^2 - \left( \sum x \right)^2}$$  \hspace{1cm} (1.3)

For the intercept we can use the fact that $\bar{y} = a + b \bar{x}$ (this means that the “point of both means is on the least squares line) which leads to

$$a = \bar{y} - b \cdot \bar{x}$$  \hspace{1cm} (1.4)

or if we prefer we can use straightforward calculation for the intercept through

$$a = \frac{\sum x^2 \cdot \sum y - \sum x \cdot \sum xy}{n \cdot \sum x^2 - \left( \sum x \right)^2}.$$  \hspace{1cm} (1.5)
We can also express the slope and intercept in terms of Pearson Correlation Coefficient and variances. Reminder:

It helps to get organized as I did in the table 1 below.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$xy$</th>
<th>$x^2$</th>
<th>$y^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>15</td>
<td>255</td>
<td>289</td>
<td>225</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>300</td>
<td>225</td>
<td>400</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>15</td>
<td>150</td>
<td>100</td>
<td>225</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>225</td>
<td>225</td>
<td>225</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>80</td>
<td>100</td>
<td>64</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>400</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>15</td>
<td>17</td>
<td>255</td>
<td>225</td>
<td>289</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>80</td>
<td>64</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>50</td>
<td>25</td>
<td>100</td>
</tr>
</tbody>
</table>

$\sum x = 115 \quad \sum y = 133 \quad \sum xy = 1795 \quad \sum x^2 = 1653 \quad \sum y^2 = 2037$

Using (1.3) and (1.4) we get

\[
b = \frac{n \cdot \sum xy - \sum x \cdot \sum y}{n \cdot \sum x^2 - (\sum x)^2} = \frac{10 \cdot 1795 - 133 \cdot 115}{10 \cdot 1653 - 115^2} \approx 0.80332829
\]

\[
a = \bar{y} - b \cdot \bar{x} \approx 13.3 - 0.8033 \cdot 11.5 \approx 4.06172467.
\]

Therefore the line itself has the slope-intercept equation

\[Y = 4.0617 + 0.8033 \cdot X.\]
We can use this “trend” line to make a prediction. Assume that a student has taken the quiz #1 and scored 10. What can we expect from him on the quiz #2?

\[ Y = 4.0617 + 0.8033 \cdot 10 = 12.0947. \]

Our expectation/estimate is approximately 12 points.

***

Example 2.

Let’s revisit the Boston Park example from the last lecture.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>xy</th>
<th>( x^2 )</th>
<th>( y^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>5</td>
<td>50</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>30</td>
<td>225</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>16</td>
<td>256</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>9</td>
<td>1</td>
<td>81</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>28</td>
<td>16</td>
<td>49</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>48</td>
<td>36</td>
<td>64</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>18</td>
<td>324</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>60</td>
<td>144</td>
<td>25</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>42</td>
<td>196</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>42</td>
<td>49</td>
<td>36</td>
</tr>
<tr>
<td>103</td>
<td>47</td>
<td>343</td>
<td>1347</td>
<td>295</td>
</tr>
</tbody>
</table>

Using (1.3) and (1.4) we get

\[
\begin{align*}
 b &= \frac{n \cdot \sum xy - \sum x \cdot \sum y}{n \cdot \sum x^2 - (\sum x)^2} = \frac{10 \cdot 343 - 103 \cdot 47}{10 \cdot 1347 - 103^2} \approx -0.4932 \\
 a &= \bar{y} - b \cdot \bar{x} \approx 4.7 - (-0.4932) \cdot 10.3 = 9.77996.
\end{align*}
\]

Therefore the least-squares line itself has the slope-intercept equation

\[ Y = 9.77996 - 0.4932 \cdot X. \]

The graph of the line is on the figure 2.
We can use this “trend” line to make a prediction. How many muggings we should expect if there are 8 police officers on duty?

\[ Y = 9.77996 - 0.4932 \cdot 8 = 5.83436 \, . \]

Our expectation/estimate is approximately 6 muggings (rounded to the nearest integer).

---

**Few Important Notes**

1. What happens when there is no significant correlation, for example in the case of “shoe size and quiz” example from the previous lecture? Since there is no fitting line as there is linear regression (correlation) we should not attempt to calculate least-squares regression line. Even though the formulas would lead to

\[ Y = 13.0036 + 0.0314 \cdot X \]
there is no significant impact of $X$ on $Y$. That means that shoe size has no bearing on the quiz scores. If you are required to make a prediction about second quiz score for any student based on the shoe size you might as well state the average second quiz score, i.e., 13.3 as the predicted value.

2. It can be shown that slope and Pearson correlation coefficient are in the following relation:

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

Obviously negative $r$ (Pearson Correlation Coefficient) means the slope of least-squares line will be negative. Positive $r$ means positive slope of least-squares line. The value of $r$ close to 0 means that the slope of least-squares line is also close to 0 i.e., the line is close to horizontal.

3. When correlation is significant then prediction using LSRL is possible but only within the range of the data. For example in Boston Park case we would be able to estimate the number of muggings only when we have between 1 and 18 police officers on duty since this is the range of sample we have our data from.

**Homework:**

Section 10-2: Problems 13, 14, 15, 16. (In the problem 16 read the note #1 above!)

[Note about the textbook notation: the author uses the following notation

$$y' = a + bx$$

for the least-squares line.]