Looking at Data - Relationships
Least-Squares Regression

IPS Chapter 2.3

Objectives (IPS Chapter 2.3)

Least-squares regression

- Regression lines
- Prediction and Extrapolation
- Correlation and $r^2$
- Transforming relationships (Skip)
Introduction

Correlation tells us about strength (scatter) and direction of the linear relationship between two quantitative variables. In addition, we would like to have a numerical description (mathematical model) of how both variables vary together. For instance, is one variable increasing faster than the other one? And we would like to make predictions based on that numerical description.

Regression lines

- A regression line is a straight line that describes how a response variable $y$ changes as an explanatory variable $x$ changes.

- We often use a regression line to predict the value of $y$ for a given value of $x$.

- In regression, the distinction between explanatory and response variables is important. Changing the roles of $x$ & $y$ gives a different regression line even though the correlation coefficient “$r$” stays the same.
The least-squares regression line

The least-squares regression line is the unique line such that the sum of the squared vertical (y) distances between the data points and the line is as small as possible.

Distances between the points and line are squared so all are positive values. This is done so that distances can be properly added (Pythagoras).

Properties

The least-squares regression line can be shown to have this equation:

\[ \hat{y} = a + bx \]

\( \hat{y} \) is the predicted y value (y hat)

b is the slope

a is the y-intercept
Finding the equation of the least-squares regression line

1. First we calculate the slope of the line, \( b \).
   \[
   \text{Slope} = r \frac{s_y}{s_x}
   \]
   where \( r \) is the correlation,
   \( s_y \) is the standard deviation of the response variable \( y \),
   \( s_x \) is the standard deviation of the explanatory variable \( x \).

2. Once we know \( b \), the slope, we can calculate \( a \), the \textit{y-intercept}.
   \[
   \text{Intec.} = \bar{y} - b \bar{x}
   \]
   where \( \bar{x} \) and \( \bar{y} \) are the sample means of the \( x \) and \( y \) variables.

 Typically, we use a \textit{2-var stats calculator} or \textit{stats software}.

The equation completely describes the regression line.

To plot the regression line you only need to plug two \( x \) values into the equation, get \( y \), and draw the line that goes through those points.

The points you use for drawing the regression line are derived from the equation.

They are \textit{NOT} points from your sample data (except by pure coincidence).

**Hint:** The regression line always passes through the mean of \( x \) and \( y \).
Adding the least-squares regression equation to the scatterplot in Excel 2007

1. Click on the graph first to activate the Design ribbon.
2. For more information check this demo: http://www.marin.edu/~npsomas/Demos/How2Create_ScatterPlot/how2create_scatterplot_viewlet.swf.html
3. Also available from the class Web page ~ http://www.marin.edu/~npsomas/Math115/

Calculating the least-squares regression equation

- Using TI – 83/84

**Example 2.5 More Brain Activity and Stress.** The data from Example 2.1 are repeated below. (a) What is the equation of the least-squares regression line for predicting brain activity from a social distress score? Make a scatterplot with this line drawn on it. (b) Use the equation of the regression line to get the predicted brain activity level for a distress score of 2. (c) What percent of the variation in brain activity among these subjects is explained by the straight-line relationship with social distress score?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Social distress</th>
<th>Brain activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.26</td>
<td>-0.055</td>
</tr>
<tr>
<td>2</td>
<td>1.85</td>
<td>-0.040</td>
</tr>
<tr>
<td>3</td>
<td>1.10</td>
<td>-0.026</td>
</tr>
<tr>
<td>4</td>
<td>2.50</td>
<td>-0.017</td>
</tr>
<tr>
<td>5</td>
<td>2.17</td>
<td>-0.017</td>
</tr>
<tr>
<td>6</td>
<td>2.67</td>
<td>0.017</td>
</tr>
<tr>
<td>7</td>
<td>2.01</td>
<td>0.021</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>Social distress</th>
<th>Brain activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>2.18</td>
<td>0.025</td>
</tr>
<tr>
<td>9</td>
<td>2.58</td>
<td>0.027</td>
</tr>
<tr>
<td>10</td>
<td>2.75</td>
<td>0.033</td>
</tr>
<tr>
<td>11</td>
<td>2.75</td>
<td>0.064</td>
</tr>
<tr>
<td>12</td>
<td>3.33</td>
<td>0.077</td>
</tr>
<tr>
<td>13</td>
<td>3.65</td>
<td>0.124</td>
</tr>
</tbody>
</table>
Calculating the least-squares regression equation

- Using TI – 83/84

**Solution.** (a) We obtain the linear regression line using the same **LinReg(a+bx)** command that computes the correlation. After entering data into lists, say L1 and L2, we would enter the command **LinReg(a+bx)** L1,L2. However, in order to store the equation and be able to use it for graphing or prediction, we will add a new parameter to the LinReg command. Add a comma after L2 and press **VARS**, press **5** to move the highlight to **Y-VARS**, press **ENTER** to select option 1:Function and **ENTER** to select Y1. Finally, press **ENTER** to perform the calculations.

![Linear Regression Calculation](image1)

Calculating the least-squares regression equation

- Using TI – 83/84

You can see the equation of the line stored by pressing **Y=**. Our equation is

\[
\text{Brain-Activity} = -0.126 + 0.061 \times \text{Distress}
\]

To graph the data along with the regression line, press **ZOOM 9** (**F5** for **ZoomData** from the **Plot Setup** Screen on a TI-89).

![Graph of Linear Regression](image2)
BEWARE!!!

Not all calculators and software use the same convention. Some use:

\[ \hat{y} = a + bx \]

And some use:

\[ \hat{y} = ax + b \]

Make sure you know what YOUR calculator gives you for a and b before you answer homework or exam questions.

The distinction between explanatory and response variables is crucial in regression. If you exchange \( y \) for \( x \) in calculating the regression line, you will get the wrong line.

Regression examines the distance of all points from the line in the \( y \) direction only.

Hubble telescope data about galaxies moving away from earth:

These two lines are the two regression lines calculated either correctly (\( x = \) distance, \( y = \) velocity, solid line) or incorrectly (\( x = \) velocity, \( y = \) distance, dotted line).
Making predictions

The equation of the least-squares regression allows you to predict $y$ for any $x$ within the range studied.

Nobody in the study drank 6.5 beers, but by finding the value of $\hat{y}$ from the regression line for $x = 6.5$ we would expect a blood alcohol content of 0.094 mg/ml.

\[ \hat{y} = 0.0144 \times 6.5 + 0.0008 \]
\[ \hat{y} = 0.936 + 0.0008 = 0.9444 \text{mg/ml} \]

There is a positive linear relationship between the number of powerboats registered and the number of manatee deaths. (Florida, US)

The least squares regression line has the equation: \[ \hat{y} = 0.125x - 41.4 \]

Thus if we were to limit the number of powerboat registrations to 500,000, what could we expect for the number of manatee deaths?

\[ \hat{y} = 0.125(500) - 41.4 \Rightarrow \hat{y} = 62.5 - 41.4 = 21.1 \]

Roughly 21 manatees.
Making predictions using the regression equation

- Using TI – 83/84 (Back to Example 2.5)

(b) With the equation of the regression line computed and displayed on the graph as above, we can use the [Y=] on a TI-89 to evaluate the function for a specific \(x\) value. Press [TRACE]. At the upper right, the active plot is displayed, which indicates we are tracing the scatterplot. Press [2] to switch to the line. The top line will change to display the equation of the line. Type in 2 (the distress value we want to find a brain activity for) and press [ENTER]. We see that a distress score of 2 gives a predicted brain activity of –0.0045.

Making predictions using the regression equation

- Using TI – 83/84 (Optional)

Alternately, we can access the \(Y1\) function from the \text{Home} screen. To do so, press [VARS] arrow right to \text{Y-VARS}, enter [1] for \text{Function}, enter [1] for \(Y1\), then enter the command \(Y1(2)\). Press [ENTER] to perform the calculation.
Verifying \((x \text{-bar}, y \text{-bar})\) lies on the regression line

- Using TI – 83/84 (Optional)

We can also verify that the point \((\bar{x}, \bar{y})\) is on the regression line; but first we must compute the statistics. We can do so simultaneously with the **2-Var Stats** command from the STAT CALC menu, because the two data sets are the same size. Enter the command **2-Var Stats L1,L2**. Then enter **Y1(\bar{x})** by recalling \(\bar{x}\) from the **VARS** Statistics menu.

![Image of 2-Var Stats calculations](image1)

**Coefficient of determination, \(r^2\)**

\(r^2\), the coefficient of determination, is the square of the correlation coefficient.

\(r^2\) represents the percentage of the variance in \(y\) (vertical scatter from the regression line) that can be explained by changes in \(x\).
Explained variation in the response attributed to the explanatory

- Using TI – 83/84 (Back to Example 2.5)

(c) With the calculator’s diagnostics turned on, the LinReg(a+bx) command also displays the values of $r$ and $r^2$. In this case, $r^2 = 0.7713$. Thus, 77.13% of the variation in brain activity among these subjects is explained by the straight-line relationship with social distress score.
Extrapolation

Extrapolation is the use of a regression line for predictions \textit{outside the range of x values} used to obtain the line.

Extrapolation DOES NOT always make sense as seen in this example. (Why?)

The $y$-intercept

The $y$-intercept is the value where the line crosses the $y$-axis ($x = 0$). Sometimes the $y$-intercept is not biologically possible. Here we have negative blood alcohol content, which makes no sense…

But the negative value is appropriate for the equation of the regression line.

There is a lot of scatter in the data, and the line is just an estimate.