

Exploring Statistics Using Fathom

Correlation Coefficient and Least-Squares

Understanding the Correlation Coefficient with Fathom

The correlation coefficient is used to quantify the strength of a linear relationship between x and y values in a data set of (x, y) pairs. Recall when we first discussed scatter plots of bivariate data, we could visually determine if x and y had a positive or negative relationship (we say x and y have a positive relationship if increases in x seem to correspond to increases in y , etc.). Now we will discover a way to quantify the intensity of that relationship.

1. First, download the Fathom data set “BoneToHeight.ftm” from

<http://paws.wcu.edu/emcnelis/StatsExamples.html>.

2. We will create two new attributes corresponding to the z-scores of the metacarpal length and stature. To do this, we must remember that a z-score of an observation, x_i is given by:

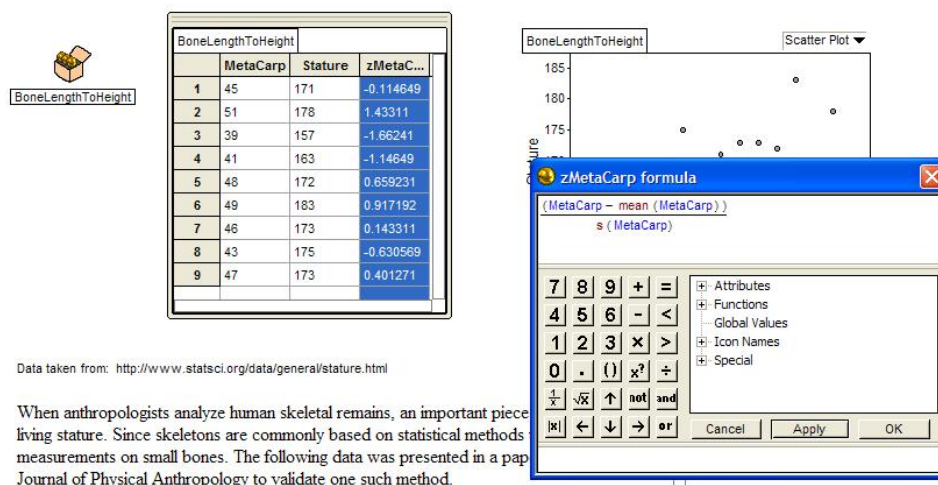
$$z_i = \frac{x_i - \bar{x}}{s}$$

where \bar{x} is the sample mean and s is the sample standard deviation.

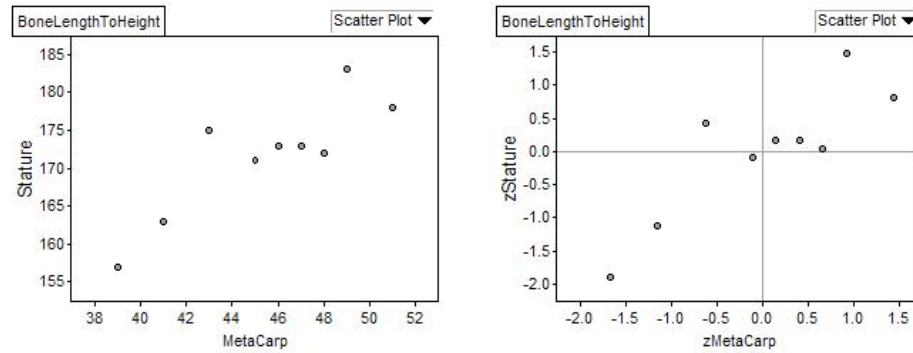
- (a) Title a new attribute “zMetaCarp”.
- (b) With your cursor over the new attribute name, right click and select **Edit Formula**. This will generate a pop-up box in which we’ll specify our z-score formula:

$$\frac{(\text{MetaCarp} - \text{mean}(\text{MetaCarp}))}{s(\text{MetaCarp})}$$

You may type this in, or build it using the functions and attribute names listed in the box to the lower right. **MetaCarp** can be found under the **Attribute** menu, and **mean** and **s** functions can be found under the **Functions** → **Statistical** → **One Attribute** menu.



- (c) Repeat the above steps to generate a new attribute called “zStature”.
3. Now, make a new scatter plot of “zStature” versus “zMetaCarp”.



Food for thought:

- What can you say about the appearance of the scatterplot of the z-scores in comparison to the scatterplot of the “Stature” versus “MetaCarp”?
- What can you say about the values of the z-scores in each quadrant of our new graph?
- What can you say about the values of the products of the z-scores in each quadrant of our new graph?
- What would you guess would be true of the sum of the products of the z-scores, judging from this scatter plot?

Definition 1 (Pearson Correlation Coefficient)

Pearson's sample correlation coefficient, r , is given by

$$r = \frac{\sum z_x * z_y}{n - 1}$$

We can now easily calculate this value in Fathom.

1. Create yet another new attribute, called “zProduct”, and define it to have value

$$z\text{MetaCarp} * z\text{Stature}$$

2. Bring down a summary table from the menu bar and place it on the page.
3. Drag the “zProduct” to the summary table. This will automatically bring the `mean()` for the attribute up. Delete this by right clicking and selecting **Delete Formula**.
4. Add a new formula (right click in the summary table area and select **Add Formula**) with value

$$\frac{\text{sum}()}{(\text{count}()-1)}$$

by typing this in by hand or using the lists under the **Statistical** → **One Attribute** submenu. Your result is the value of your (Pearson correlation) coefficient.

BoneLengthToHeight	Summary Table
↓ →	
zProduct	0.85596828
$S1 = \left(\frac{\text{sum}(z\text{Product})}{(\text{count}(z\text{Product}) - 1)} \right)$	

Can you guess the range of values the correlation coefficient, r , can assume? What about interpreting these values? What values of r indicate a strong linear relationship between x and y ? a weak relationship?

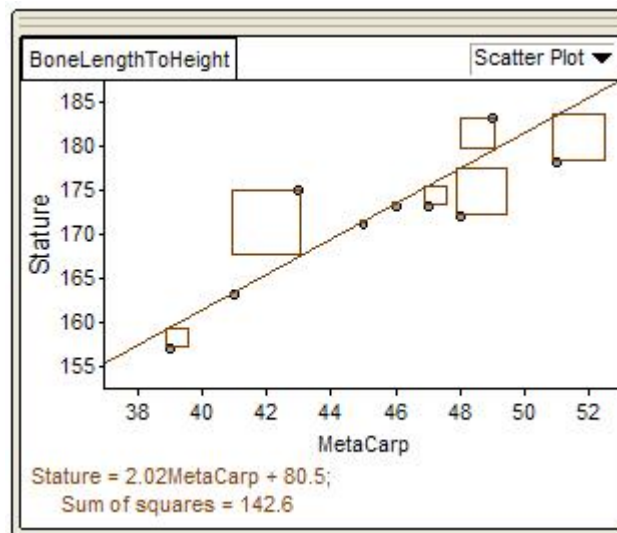
Least-Squares Lines Exploration with Fathom

By the time you get to college, many of you have been exposed to “linear regression” (perhaps using your TI calculator), or finding the best fitting line for a data set. The question is,

How do we quantify “best” in best fitting line? Any ideas?

Let's see how well you can "eyeball" the best fitting line, i.e. the one that has the smallest error in terms of area of the square whose side is the length of the difference between true y -values and predicted y -values on our line.

1. Return to the scatterplot of "Stature" versus "MetaCarp".
2. Right click while on the graph and select the option of adding a **Moveable Line**. This adds a line to the graph going from the lower left corner to the upper left corner. The equation of the line is below the graph.
3. Note, you can move this line using your mouse. If you're to the left or right of center and over the line, you can swivel the line about its center. If you're over the center of the line, you can shift it up or down using the left click mouse. Try to find a line that "fits" the data well.
4. Right click again on the graph and select **Show Squares**. In addition to visualizing the actual squares whose area is added together to quantify total error. This total is now displayed in the window below.



5. When you feel you've done the best you can, compare your result (the **Sum of Squares** values in particular) with the people around you. How did you do?
6. To see the truly best fitting line through the data, right click and select **Least-Squares Line**.