

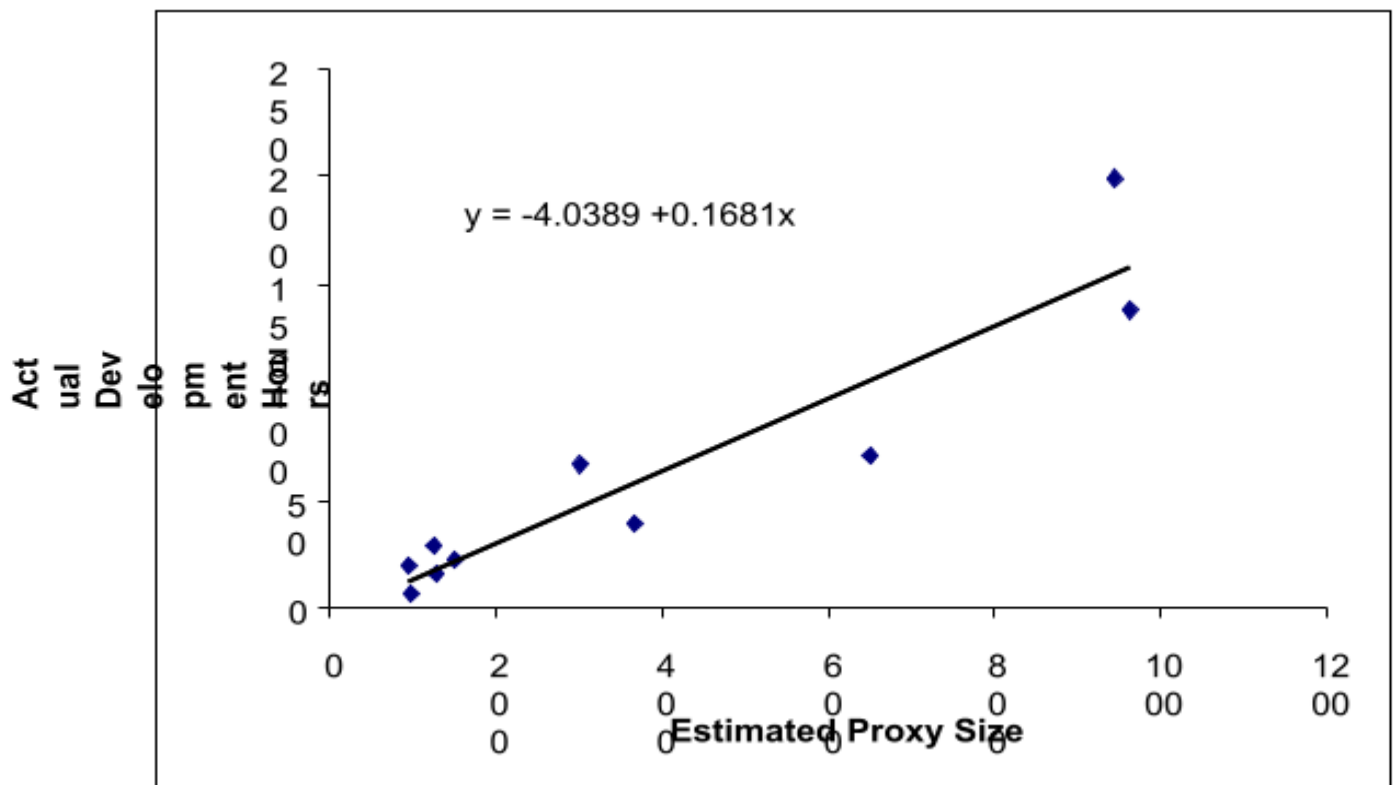
Regression

Overview

Linear regression is a way of optimally fitting a line to a set of data. The linear regression line is the line where the distance from all points to that line is minimized. The equation of a line can be written as

$$y = \beta_0 + \beta_1 x$$

In Figure 1, the best fit regression line has parameters of $\beta_0 = -4.0389$ and $\beta_1 = 0.1681$.



Correlation

Overview

The correlation calculation determines the relationship between two sets of numerical data.

The correlation $r_{x,y}$ can range from +1 to -1.

- Results near +1 imply a strong positive relationship; when x increases, so does y .
 - Results near -1 imply a strong negative relationship; when x increases, y decreases.
 - Results near 0 imply no relationship.
-

For this purpose, we examine the value of the relation r_{xy} squared, or r^2 .

| If r^2 is | the relationship is |
|--------------------|--|
| $.9 \leq r^2$ | predictive; use it with high confidence |
| $.7 \leq r^2 < .9$ | strong and can be used for planning |
| $.5 \leq r^2 < .7$ | adequate for planning but use with caution |
| $r^2 < .5$ | not reliable for planning purposes |

Calculating regression and correlation

Calculating regression and correlation

The formulas for calculating the regression parameters β_0 and β_1 are

$$\beta_1 = \frac{\left(\sum_{i=1}^n x_i y_i \right) - (n x_{avg} y_{avg})}{\left(\sum_{i=1}^n x_i^2 \right) - (n x_{avg}^2)}$$

$$\beta_0 = y_{avg} - \beta_1 x_{avg}$$

The formulas for calculating the correlation coefficient $r_{x,y}$ and r^2 are

$$r_{x,y} = \frac{n\left(\sum_{i=1}^n x_i y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{\sqrt{\left[n\left(\sum_{i=1}^n x_i^2\right) - \left(\sum_{i=1}^n x_i\right)^2\right]\left[n\left(\sum_{i=1}^n y_i^2\right) - \left(\sum_{i=1}^n y_i\right)^2\right]}}$$

$$r^2 = r * r$$

where

- Σ is the symbol for summation
 - i is an index to the n numbers
 - x and y are the two paired sets of data
 - n is the number of items in each set x and y
 - x_{avg} is the average of the x values
 - y_{avg} is the average of the y values
-

An example

An example

In this example, we will calculate the regression parameters (β_0 and β_1 values) and correlation coefficients $r_{x,y}$ and r^2 of the data in the Table 3.

| n | x | y |
|-----|-----|------|
| 1 | 130 | 186 |
| 2 | 650 | 699 |
| 3 | 99 | 132 |
| 4 | 150 | 272 |
| 5 | 128 | 291 |
| 6 | 302 | 331 |
| 7 | 95 | 199 |
| 8 | 945 | 1890 |
| 9 | 368 | 788 |
| 10 | 961 | 1601 |

Table 3

$$\beta_1 = \frac{\left(\sum_{i=1}^n x_i y_i \right) - (n x_{avg} y_{avg})}{\left(\sum_{i=1}^n x_i^2 \right) - (n x_{avg}^2)}$$

1. In this example there are 10 items in each dataset and therefore we set $n = 10$.
2. We can now solve the summation items in the formulas.

| n | x | y | x^2 | $x*y$ | y^2 |
|-------|-------------------------------------|-------------------------------------|-----------------------------------|-------------------------------------|-----------------------------------|
| 1 | 130 | 186 | 16900 | 24180 | 34596 |
| 2 | 650 | 699 | 422500 | 454350 | 488601 |
| 3 | 99 | 132 | 9801 | 13068 | 17424 |
| 4 | 150 | 272 | 22500 | 40800 | 73984 |
| 5 | 128 | 291 | 16384 | 37248 | 84681 |
| 6 | 302 | 331 | 91204 | 99962 | 109561 |
| 7 | 95 | 199 | 9025 | 18905 | 39601 |
| 8 | 945 | 1890 | 893025 | 1786050 | 3572100 |
| 9 | 368 | 788 | 135424 | 289984 | 620944 |
| 10 | 961 | 1601 | 923521 | 1538561 | 2563201 |
| Total | $\sum_{i=1}^{10} x_i = 3828$ | $\sum_{i=1}^{10} y_i = 6389$ | $\sum_{i=1}^{10} x_i^2 = 2540284$ | $\sum_{i=1}^{10} x_i y_i = 4303108$ | $\sum_{i=1}^{10} y_i^2 = 7604693$ |
| | $x_{avg} = \frac{3828}{10} = 382.8$ | $y_{avg} = \frac{6389}{10} = 638.9$ | | | |

Continued on next page

An example, Continued

An example, cont. 3. We can then substitute the values into the formulas

$$\beta_1 = \frac{(4303108) - (10 * 382.8 * 638.9)}{(2540284) - (10 * 382.8^2)}$$

$$\beta_1 = \frac{1857399}{1074926} = 1.727932$$

$$r_{x,y} = \frac{10(4303108) - (3828)(6389)}{\sqrt{[10(2540284) - (3828)^2][10(7604693) - (6389)^2]}}$$

$$r_{x,y} = \frac{18573988}{\sqrt{[10749256][35227609]}} \quad r_{x,y} = \frac{18573988}{19459460.1}$$

$$r_{x,y} = 0.9545$$

$$r^2 = 0.9111$$

4. We can then substitute the values in the β_0 formula

$$\beta_0 = y_{avg} - \beta_1 x_{avg}$$

$$\beta_0 = 638.9 - 1.727932 * 382.8 = -22.5525$$

5. We now find y_k from the formula $y_k = \beta_0 + \beta_1 x_k$

$$= -22.5525 + 1.727932 * 386 = 644.4294$$

Assignment instructions

- Calculate the linear regression parameters β_0 and β_1 and correlation coefficients $r_{x,y}$ and r^2 for a set of n pairs of data,
- given an estimate, x_k calculate an improved prediction, y_k where

$$y_k = \beta_0 + \beta_1 x_k$$

Table 1 contains historical estimated and actual data for 10 programs. For program 11, the developer has estimated a proxy size of 386 LOC.

Thoroughly test the program. At a minimum, run the following four test cases.

- **Test 1:** Calculate the regression parameters and correlation coefficients between **estimated proxy size** and **actual added and modified size** in Table 1. Calculate plan added and modified size given an estimated proxy size of $x_k = 386$.
- **Test 2:** Calculate the regression parameters and correlation coefficients between **estimated proxy size** and **actual development time** in Table 1. Calculate time estimate given an estimated proxy size of $x_k = 386$.
- **Test 3:** Calculate the regression parameters and correlation coefficients between **plan added and modified size** and **actual added and modified size** in Table 1. Calculate plan added and modified size given an estimated proxy size of $x_k = 386$.
- **Test 4:** Calculate the regression parameters and correlation coefficients between **plan added and modified size** and **actual development time** in Table 1. Calculate time estimate given an estimated proxy size of $x_k = 386$.

| Program Number | Estimated Proxy Size | Plan Added and Modified size | Actual Added and Modified Size | Actual Development Hours |
|----------------|----------------------|------------------------------|--------------------------------|--------------------------|
| 1 | 130 | 163 | 186 | 15.0 |
| 2 | 650 | 765 | 699 | 69.9 |
| 3 | 99 | 141 | 132 | 6.5 |
| 4 | 150 | 166 | 272 | 22.4 |
| 5 | 128 | 137 | 291 | 28.4 |
| 6 | 302 | 355 | 331 | 65.9 |
| 7 | 95 | 136 | 199 | 19.4 |
| 8 | 945 | 1206 | 1890 | 198.7 |
| 9 | 368 | 433 | 788 | 38.8 |
| 10 | 961 | 1130 | 1601 | 138.2 |

Table 1

Expected results are provided in Table 2.

Expected results

| Test | Expected Values | | | | | Actual Values | | | | |
|---------------|-----------------|-----------|-----------|--------|----------|---------------|-----------|-----------|--------|----------|
| | β_0 | β_1 | $r_{x,y}$ | r^2 | y_k | β_0 | β_1 | $r_{x,y}$ | r^2 | y_k |
| Test 1 | -22.55 | 1.7279 | 0.9545 | 0.9111 | 644.429 | -22.5525 | 1.7279 | 0.9544 | 0.9110 | 644.4293 |
| Test 2 | -4.039 | 0.1681 | 0.9333 | .8711 | 60.858 | -4.0388 | 0.1681 | 0.9333 | 0.8710 | 60.8580 |
| Test 3 | -23.92 | 1.43097 | .9631 | .9276 | 528.4294 | -23.9238 | 1.4309 | 0.9631 | 0.9275 | 528.4294 |
| Test 4 | -4.604 | 0.140164 | .9480 | .8988 | 49.4994 | -4.6037 | 0.1401 | 0.9480 | 0.9897 | 49.4993 |

Table 2