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    In Figure 1, the best fit regression line has parameters of  = -4.0389 and  = 0.1681.  y = -4.0389 +0.1681x  0  50  100  150  200  250  0  200  400  600  800  1000  1200  **Estimated Proxy Size**  **Actual Development Hours** |

Figure 1

Continued on next page

Regression, Continued

|  |  |
| --- | --- |
| Using regression in the PSP | Looking at Figure 1, how many hours do you think it would take to develop a program with an estimated proxy size of 500?  Using PROBE method A for time, the estimate would be  or an estimate of 80.011 hours.  The PSP PROBE method uses regression parameters to make better predictions of size and time based on your historical data.  PROBE methods A and B differ only in the historical data (*x* values) used to calculate the regression parameters. In PROBE method A, **estimated** **proxy** size are used as the *x* values. In PROBE method B, **plan** **added and modified** size are used as the *x* values.  PROBE methods for size and time differ only in the historical data (*y* values) used to calculate the regression parameters. To predict improved size estimates, **actual added and modified LOC** are used as the *y* values. To predict time estimates, **actual development times** are used as the *y* values. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Historical Data Used** | | ***x* values** | ***y* values** |
| Size Estimating | PROBE A | Estimated Proxy Size | Actual Added and Modified Size |
| PROBE B | Plan Added and Modified Size | Actual Added and Modified Size |
| Time Estimating | PROBE A | Estimated Proxy Size | Actual Development Time |
| PROBE B | Plan Added and Modified Size | Actual Development Time |

Correlation

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Using correlation in the PSP | Correlation is used in the PSP to judge the quality of the linear relation in various historical process data that are used for planning. For example, the relationships between estimated proxy size and actual time or plan added and modified size and actual time.  For this purpose, we examine the value of the relation *rxy* squared, or . | | | |
| If  is | the relationship is |
| .9 ≤ | predictive; use it with high confidence |
| .7 ≤ *< .9* | strong and can be used for planning |
| .5 ≤ < .7 | adequate for planning but use with caution |
| *< .5* | not reliable for planning purposes |

|  |  |
| --- | --- |
| Limitations of correlation | Correlation doesn’t imply cause and effect.  A strong correlation may be coincidental.  From 1840 to 1960, no U.S. president elected in a year ending in 0 survived his presidency. Coincidence or Correlation?  Many coincidental correlations may be found in historical process data.  To use a correlation, you must understand the cause-and-effect relationship in the process. |

Calculating regression and correlation

|  |  |
| --- | --- |
| Calculating regression and correlation | The formulas for calculating the regression parameters  and  are      The formulas for calculating the correlation coefficient  and  are      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*  •  is the average of the *x* values  •  is the average of the *y* values |

An example

|  |  |
| --- | --- |
| An example | In this example, we will calculate the regression parameters ( and  values) and correlation coefficients  and  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



|  |  |
| --- | --- |
|  | 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* | *x2* | *x\*y* | *y2* |
| 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 |  |  |  |  |  |
|  |  |  |  |  |  |

Continued on next page

An example, Continued

|  |  |
| --- | --- |
| An example, cont. | 1. We can then substitute the values into the formulas                1. We can then substitute the values in the  formula        1. We now find  from the formula |