**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*

**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. |

|  |  |  |
| --- | --- | --- |
|  | 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 |

**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,**

|  |  |
| --- | --- |
| **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 |

**Assignment instructions**

|  |  |
| --- | --- |
|  | • Calculate the linear regression parameters  and  and correlation coefficients  and  for a set of *n* pairs of data,  • given an estimate,  calculate an improved prediction,  where    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 = 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 = 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  = 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 = 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** | | | | |
|  |  |  |  |  |  |  |  |  |  |  |
| Test 1 | -22.55 | 1.7279 | 0.9545 | 0.9111 | 644.429 | -22.55 | 1.7279 | 0.9545 | 0.9111 | 644.429 |
| Test 2 | -4.039 | 0.1681 | 0.9333 | .8711 | 60.858 | -4.039 | 0.1681 | 0.9333 | .8711 | 60.858 |
| Test 3 | -23.92 | 1.43097 | .9631 | .9276 | 528.4294 | -23.92 | 1.43097 | .9631 | .9276 | 528.4294 |
| Test 4 | -4.604 | 0.140164 | .9480 | .8988 | 49.4994 | -4.604 | 0.140164 | .9480 | .8988 | 49.4994 |

**Table 2**

Los valores esperados y los valores reales que calculamos indican que los modelos de regresión lineal utilizados son precisos y consistentes. Esto valida la efectividad de los modelos para predecir y explicar las variables dependientes basadas en los datos proporcionados. Los resultados consistentes entre los diferentes tests refuerzan la fiabilidad de estas técnicas de regresión para el análisis de datos y la toma de decisiones basadas en estos modelos.