**Regression (Continued)**

**Recap**

* How is regression different from correlation
* What is that line which generalises well the relationship between two variables called?
* What are the two parts of a line or a regression equation?
* What is slope?
* What is intercept?

**Topics**

* **Computing the regression equation**
* **Measures of Variation**

**Computing the regression equation**

**Recall the Regression Line Equation**

Yc = B0 + B1X

B1 = [ n(∑(XY)) – ∑X∑Y ] / [ n∑X^2 – (∑X)^2 ]

B0= Y̅ - B1X̅

**Example**

A researcher wants to find out if there is a relationship between the heights of the sons and the heights of their father. In other words, do tall fathers have tall sons? He took a random sample of 6 fathers and their 6 sons. Their heights in inches are given below.

|  |  |
| --- | --- |
| **Father (X)** | **Son (Y)** |
| 63 | 66 |
| 65 | 68 |
| 66 | 65 |
| 67 | 67 |
| 67 | 69 |
| 68 | 70 |

1. Compute the regression line
2. Based upon the relationship between the heights, what would be the estimate of the height of the son, if the father’s height is 70 inches.

**Steps**

1. Draw Scatter diagram
2. Compute Beta coefficients

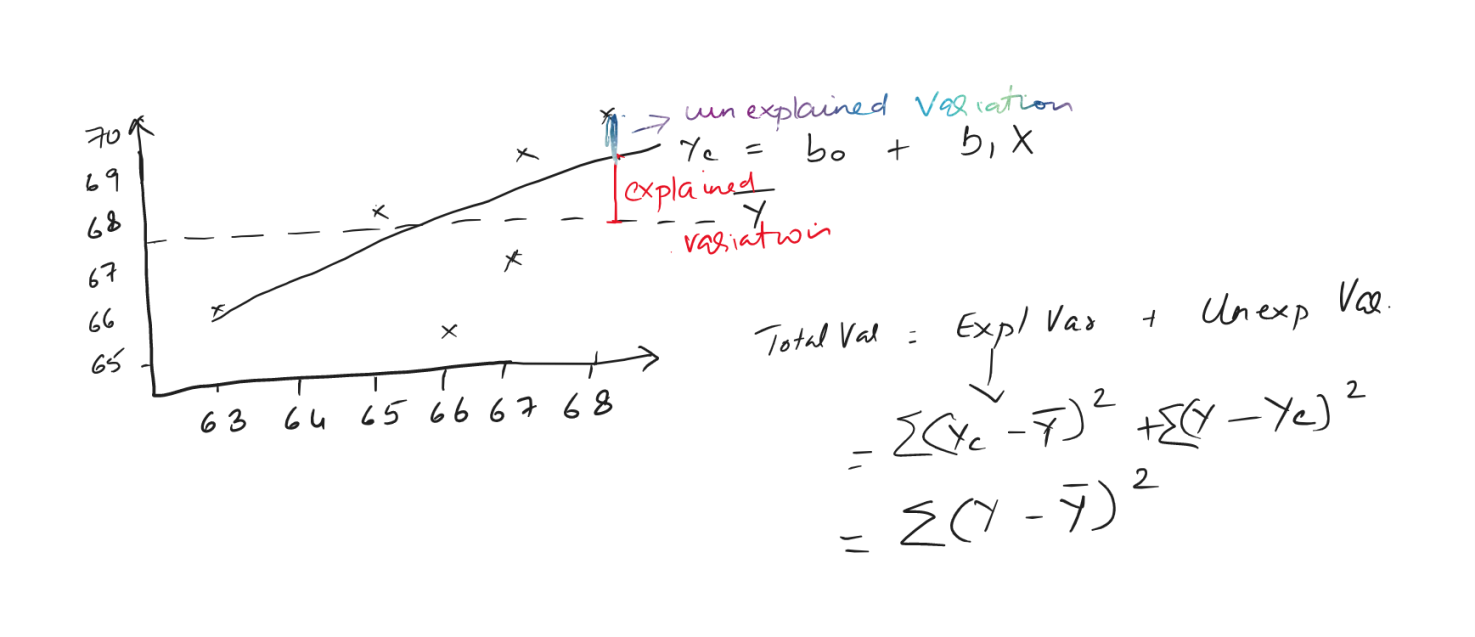
B1 = 0.625, B0 =26.25

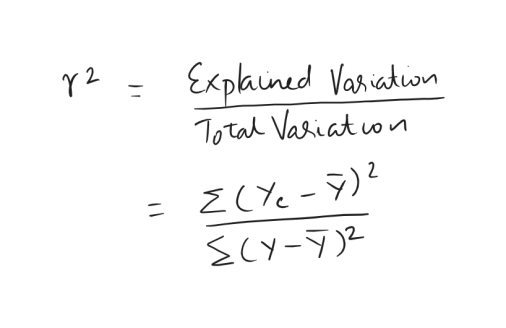
Yc = 26.25 + .625(70) = 70

**Measures of Variation**

We have found a line through the scatter points which best fits the data.

* But how good is this fit
* How reliable is the estimated value of Yc?
* How close are the values of Yc to the observed values of Y?
* The closer these values are to each other, the better the fit.
* This means that if the points in the scatter diagram are closely spaced around the regression line, then the estimated value Yc will be close to the observed value of Y and hence this estimate can be considered as highly reliable.
* Accordingly, a measure of variability of scatter around the regression line would determine the reliability of this estimate Yc.
* The smaller this estimate, the more dependable the prediction will be.
* The idea is similar to standard deviation which is also a measure of scatter of data around the mean.
* This measure is known as standard error of the estimate and is used to determine the dispersion of observed values of Y about the regression line.





Example

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | intercept | 26.25 | slope | 0.625 |
| X | Y | Yc | (Yc-Y̅)^2 | (Y- Y̅)^2 |  |  |
| 63 | 66 | 65.625 | 3.515625 | 2.25 |  |  |
| 65 | 68 | 66.875 | 0.390625 | 0.25 |  |  |
| 66 | 65 | 67.5 | 0 | 6.25 |  |  |
| 67 | 67 | 68.125 | 0.390625 | 0.25 |  |  |
| 67 | 69 | 68.125 | 0.390625 | 2.25 |  |  |
| 68 | 70 | 68.75 | 1.5625 | 6.25 |  |  |
|  |  |  | **6.25** | **17.5** |  |  |
| Y̅ | 67.5 |  |  |  |  |  |
|  |  | R^2 | 0.357143 |  |  |  |

**Conclusions**

* The value of r-squared tells us that 35.7% of the variation in Y is explained by the variation in X.
* This indicates a weak relationship
* R-squared ranges from 0 to 100%.