

DM Assignment 02

17.1.2016

Adithya Abraham Philip

1PI13CS008

Questions and Answers

1. What conclusion, if any, can be obtained from the covariance and correlation values of age and income?

Ans: Covariance: 139829.44286941574

Correlation: 0.7527261606278606

The closer the correlation value is to +1 or -1, the more linearly related the two attributes are.

Since we have a positive correlation, we know that as age increases, income tends to increase and vice versa.

Also, the correlation value (0.75) is close to 1, which shows that there is a large degree of **linear** correlation between the two attributes.

2. Suppose the income were measured in thousands instead of the actual value (for example, and income value of 17,456 now becomes 17.456), how would the results of task 4 above change (if at all)? Explain.

Ans: The **covariance** will be **multiplied by the same factor** (0.001). This is because covariance = $E(A.B) - E(A).E(B)$

which becomes

$0.001 * \text{covariance} = 0.001 * E(A.B) - 0.001 * E(A).E(B)$.

In other words, the mean and average of product of attribute values both get multiplied by the same factor, when a single attribute is multiplied, and consequently covariance is also multiplied by the same factor.

The **correlation coefficient** remains **unaffected** (no change) as both the numerator (covariance) and the denominator (product of std. dev. of both attributes) are multiplied by the same factor, and therefore the factors cancel each other out. In short, it is dimensionless. This is in line with the semantic meaning of correlation, which should not depend on the unit of measurement of the attributes.