# Explained variation

From Wikipedia, the free encyclopedia

In statistics, explained variation measures the proportion to which a mathematical model accounts for the variation (dispersion) of a given data set. Often, variation is quantified as variance; then, the more specific term explained variance can be used.

The complementary part of the total variation is called unexplained or residual.

## Contents

## Definition in terms of information gain

### Information gain by better modelling

Following Kent (1983),[1] we use the Fraser information (Fraser 1965)[2]

$$F(\theta) = \int \mathrm{d}r\, g(r)\, \ln f(r;\theta)$$

where $g(r)$ is the probability density of a random variable $R$, and $f(r;\theta)$ with $\theta \in \Theta_i$ $(i = 0, 1)$ are two families of parametric models. Model family 0 is the simpler one, with a restricted parameter space $\Theta_0 \subset \Theta_1$.

Parameters are determined by maximum likelihood estimation,

$$\theta_i = \arg\max_{\theta \in \Theta_i} F(\theta).$$

The information gain of model 1 over model 0 is written as

$$\Gamma(\theta_1 : \theta_0) = 2[F(\theta_1) - F(\theta_0)]$$

where a factor of 2 is included for convenience. Γ is always nonnegative; it measures the extent to which the best model of family 1 is better than the best model of family 0 in explaining g(r).

### Information gain by a conditional model

Assume a two-dimensional random variable $R = (X, Y)$ where X shall be considered as an explanatory variable, and Y as a dependent variable. Models of family 1 "explain" Y in terms of X,

$$f(y|x;\theta),$$

whereas in family 0, X and Y are assumed to be independent. We define the randomness of Y by $D(Y) = \exp[-2F(\theta_0)]$, and the randomness of Y, given X, by $D(Y|X) = \exp[-2F(\theta_1)]$. Then,

$$\rho_C^2 = 1 - D(Y|X)/D(Y)$$

can be interpreted as proportion of the data dispersion which is "explained" by X.

## Special cases and generalized usage

For special models, the above definition yields particularly appealing results. Regrettably, these simplified definitions of explained variance are used even in situations where the underlying assumptions do not hold.

### Linear regression

The fraction of variance unexplained is an established concept in the context of linear regression. The usual definition of the coefficient of determination is based on the fundamental concept of explained variance.

### Correlation coefficient as measure of explained variance

Let X be a random vector, and Y a random variable that is modeled by a normal distribution with centre $\mu + \Psi^{\mathrm{T}} X$. In this case, the above-derived proportion of randomness $\rho_C^2$ equals the squared correlation coefficient $R^2$.

Note the strong model assumptions: the centre of the Y distribution must be a linear function of X, and for any given x, the Y distribution must be normal. In other situations, it is generally not justified to interpret $R^2$ as proportion of explained variance.

### Explained variance in principal component analysis

"Explained variance" is routinely used in principal component analysis. The relation to the Fraser-Kent information gain remains to be clarified.

# Criticism

As the fraction of "explained variance" equals the correlation coefficient $R^2$, it shares all the disadvantages of the latter: it reflects not only the quality of the regression, but also the distribution of the independent (conditioning) variables.

In the words of one critic: "Thus $R^2$ gives the 'percentage of variance explained' by the regression, an expression that, for most social scientists, is of doubtful meaning but great rhetorical value. If this number is large, the regression gives a good fit, and there is little point in searching for additional variables. Other regression equations on different data sets are said to be less satisfactory or less powerful if their $R^2$ is lower. Nothing about $R^2$ supports these claims".[3]:58 And, after constructing an example where $R^2$ is enhanced just by jointly considering data from two different populations: "'Explained variance' explains nothing."[3][4]:183

# See also

- Variance reduction

# References

1. ^ Kent, J. T. (1983). "Information gain and a general measure of correlation". Biometrika 70 (1): 163–173. doi:10.1093/biomet/70.1.163 (http://dx.doi.org/10.1093%2Fbiomet%2F70.1.163). JSTOR 2335954 (https://www.jstor.org/stable/2335954).
2. ^ Fraser, D. A. S. (1965). "On Information in Statistics". Ann. Math. Statist. 36 (3): 890–896. doi:10.1214/aoms/1177700061 (http://dx.doi.org/10.1214%2Faoms%2F1177700061).
3. ^ a b Achen, C. H. (1982). Interpreting and Using Regression. Beverly Hills: Sage. ISBN 0-8039-1915-8.
4. ^ Achen, C. H. (1990). "'What Does "Explained Variance" Explain?: Reply". Political Analysis 2 (1): 173–184. doi:10.1093/pan/2.1.173 (http://dx.doi.org/10.1093%2Fpan%2F2.1.173).

# External links

- Variance, explained and unexplained (http://www.documentingexcellence.com/stat_tool/variance.htm)
- Explained variance (http://spirxpert.com/statistical7.htm)
- Explained and Unexplained Variance on a graph (http://darwin.cwru.edu/~witte/statistics/explained_variance.htm)

Categories: Data analysis | Regression analysis