What is Explained Variance? (Definition & Example)

Authored by

Mohammed looti

October 29, 2025

The concept of **explained variance**--sometimes referred to as explained variation--is a cornerstone of quantitative statistics and modeling. It provides a precise measure of how much of the inherent variability observed in a response variable (the outcome we are studying) can be reliably attributed to the influence of the predictor variable(s) incorporated into a statistical model. In simpler terms, explained variance quantifies the effectiveness of your model's inputs in accounting for the observed changes in the phenomenon under investigation.

Achieving a high value for explained variance is typically the goal in predictive modeling, as it signifies that the chosen statistical model is highly effective at capturing and describing the underlying patterns and relationships within the dataset. Conversely, a lower value suggests that the majority of the variation in the outcome remains unexplained, implying that crucial explanatory factors are either missing from the model or that the relationship is inherently noisy.

This critical metric is fundamental for evaluating the quality and explanatory power of various statistical frameworks. Explained variance is most prominently featured in the output reports of two major statistical methodologies:

Analysis of Variance (ANOVA): Used predominantly for hypothesis testing, ANOVA assesses whether the means of three or more independent groups are statistically different, quantifying the variation explained by the group membership.
Regression Analysis: Employed to model and quantify the structural relationship between predictors and a response variable, facilitating both prediction and causal inference.

A clear comprehension of how explained variance is calculated and, more importantly, how it should be interpreted within the context of these statistical methods is essential for drawing valid and impactful conclusions from data analysis projects.

It is important to remember that explained variance is always contrasted with its counterpart: **residual variance** (or unexplained variance). Residual variance represents the portion of the variability in the response variable that the current statistical model simply cannot account for, often due to measurement error or missing variables.

Understanding the Foundation: Total Variance

To fully appreciate the significance of explained variance, one must first grasp the foundational statistical concept of **variance** itself. Statistically, variance serves as a measure of dispersion, quantifying how spread out a set of data points are relative to their mean. A high variance value signals substantial dispersion, meaning data points are scattered far from the average, whereas a low variance suggests tight clustering around the mean.

Every dataset inherently contains a specific amount of **total variance**. When statisticians and

researchers develop a model, the overarching objective is often to systematically decompose and identify the factors that contribute meaningfully to this observed variability. For example, if we analyze a dataset of historical stock prices, the total variance represents the overall volatility. A model might then attempt to explain this volatility based on factors like trading volume or interest rate changes.

Conceptually, the total variance present in a response variable is neatly partitioned into two mutually exclusive components: the fraction that our model successfully accounts for (the explained variance) and the remaining fraction that the model fails to capture (the residual variance). This essential decomposition mechanism is central to evaluating a model's descriptive power and its utility in describing the underlying phenomena under investigation.

Defining Explained Variance and Its Explanatory Power

Specifically, **explained variance** quantifies the extent to which the variables chosen for a statistical model successfully reduce the uncertainty associated with the dependent variable. It acts as a powerful measure of the model's explanatory capacity, detailing how much of the observed spread or variation in the data can be logically attributed to the identified relationships between the predictor variables and the outcome. This metric is indispensable for judging the overall goodness of fit.

Imagine attempting to predict the annual yield of a crop. The total variability in yield across different farms is high. If your model, utilizing factors such as fertilizer amount and hours of sunlight, can explain 85% of this variability, it implies these inputs are highly influential and relevant. The remaining 15% would constitute the residual variance, possibly stemming from unmeasured elements like soil quality inconsistencies or localized pest outbreaks.

A high percentage of explained variance suggests a robust and meaningful relationship between the input variables and the outcome, leading to models that offer both more accurate predictions and a deeper, more actionable understanding of the underlying causal processes. Conversely, when explained variance is low, it serves as a critical diagnostic warning that the current set of predictors is insufficient to adequately model the response, signaling the need to search for other, more relevant factors or to refine the model specification.

Explained Variance in Analysis of Variance (ANOVA)

The **Analysis of Variance (ANOVA)** technique is specifically designed to test for statistically significant differences among the means of multiple groups. When an ANOVA model is computed, the output is summarized in a structured ANOVA table, which systematically decomposes the dataset's total variability into components that are either explained by the group differences or remain unexplained.

In the context of ANOVA, the explained variance is formally represented by the **Sum of Squares (SS) Between Groups**, which may also be labeled as SS Model or SS Factor in various software outputs. This specific component quantifies the amount of variability in the response variable that is directly attributable to the distinct categorization (the independent variable). Essentially, it isolates the variation caused solely by the differences in the group treatments or conditions.

To illustrate how this concept appears in practice, we can examine a typical ANOVA table output. The table below clearly segregates the sources of variance and provides the corresponding Sum of Squares values:

ANOVA

	Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Explained Variance	Between Groups	192.2	2	96.1	2.357532	0.113848	3.354131
Unexplained Variance	Within Groups	1100.6	27	40.76296			
	Total	1292.8	29				

Referring to this sample ANOVA output, the crucial metric for explained variance is found under the "SS" (Sum of Squares) column corresponding to the "Between Groups" row. Here, the explained variation, quantified by the SS Between Groups, is precisely **192.2**. This numerical value represents the total amount of variation in the outcome measure that has been successfully accounted for by the categorical grouping factor used in the analysis.

Interpreting Explained Variance and Statistical Significance in ANOVA

While the Sum of Squares Between Groups gives us a raw magnitude of the explained variance (192.2 in our example), this value alone does not confirm whether the explanation is statistically meaningful or simply due to random chance. To test for statistical significance, we rely on the **F-statistic** (or F-ratio), which is calculated as a quotient comparing the explained variance per degree of freedom to the unexplained variance per degree of freedom.

The F-statistic is derived by dividing the **Mean Sum of Squares (MS) Between Groups** by the Mean Sum of Squares Within Groups (often termed MS Error or MS Residual). The Mean Sum of Squares converts the raw Sum of Squares into an average variance by dividing it by its corresponding degrees of freedom. This resulting F-ratio provides the critical test: are the differences between the group means significantly larger than the random variation observed within the groups?

Using the data from our preceding ANOVA table, we perform the calculation for the F-value:

F-statistic = MSbetween / MSwithin
F = 96.1 / 40.76296
F = **2.357**

For this particular ANOVA model, the calculated F-value is 2.357. Statistical software also provides a corresponding **p-value** (0.113848 in this case), which is essential for evaluating the viability of the **null hypothesis**--the default assumption that all group means are equal and that the explained variance is negligible.

To make a decision, we compare the p-value against a predefined significance level (typically $\alpha = 0.05$). Since our observed p-value (0.113848) is greater than 0.05, we lack sufficient statistical evidence to **reject the null hypothesis**. Therefore, although the model explains 192.2 units of variance, we cannot confidently assert that the differences between the group means are statistically significant at the 0.05 level.

Explained Variance in Linear Regression Models

When moving into the domain of **regression analysis**, the concept of explained variance is quantified and standardized through the metric known as **R-squared** (R2), or the coefficient of determination. Regression models are powerful tools used to mathematically describe the form and strength of the relationship between a response variable and one or several predictor variables, serving as the basis for prediction and modeling.

The **R-squared** value provides the exact proportion of the total variance in the outcome measure that is successfully predicted by the independent variables included in the model. This proportion is frequently cited as the primary indicator of the model's overall quality and its "goodness of fit" to the observed data points. Understanding the range of R-squared is crucial for interpretation:

An R-squared value of **0** suggests that the predictors explain absolutely none of the variability in the response. The model has no explanatory power beyond simply using the mean of the data.
A value of **1** (or 100%) signifies a perfect fit, indicating that the predictor variables flawlessly account for all of the variability in the response variable, leaving zero residual error.

To see how these components are structured, we turn to a standard regression output table, which breaks down the Sum of Squares:

Regression Statistics	
Multiple R	0.98294208
R Square	0.96617513
Adjusted R	0.95651089
Standard E	0.91826226
Observatio	10

ANOVA

	df	SS	MS	F	Significance F
Explained Variance — Regression	2	168.5976	84.29878	99.97417	7.11748E-06
Unexplained Variance — Residual	7	5.9024	0.843206		
Total	9	174.5			

	Coefficients	Standard Error	t Stat	P-value
Intercept	17.1158537	1.171711716	14.60756	1.68E-06
x1	1.01829268	0.348522419	2.921742	0.022285
x2	0.39634146	0.325643032	1.217104	0.263001

In this typical regression output summary, the explained variance is explicitly labeled as the Sum of Squares Regression (SS Regression), which amounts to **168.5976**. The overall variability present in the response variable is captured by the Sum of Squares Total (SS Total), listed here as **174.5**.

Calculating and Interpreting the R-squared Value

To derive the final proportion of explained variance--the **R-squared** value--for this regression model, we formalize the relationship by dividing the explained variance (SS Regression) by the total variance (SS Total). This calculation transforms the raw Sum of Squares values into a standardized, easily interpretable measure of model performance.

R-squared Formula: SS Regression / SS Total
R-squared Calculation: 168.5976 / 174.5
Result: R-squared = **0.966**

An R-squared result of **0.966** (or 96.6%) is exceptionally strong in most research contexts. This high figure implies that the predictor variables chosen for this model are capable of explaining 96.6% of the total variation observed in the response variable. Such a high proportion indicates an outstanding fit, meaning the model is highly predictive and the relationship described is very robust.

However, analysts must always interpret a high R-squared with caution. While it confirms predictive power, it does not confirm causality, nor does it guarantee that the model is correctly specified or free from inherent bias. Furthermore, in models involving multiple predictor variables, especially when comparing models with different numbers of inputs, it is best practice to rely on the **adjusted R-squared**. This variation penalizes the inclusion of unnecessary predictors, providing a more reliable and conservative estimate of the true explained variance.

The Significance of Explained Variance in Research and Modeling

The quantitative assessment provided by **explained variance** is indispensable across various fields of statistical analysis and scientific research. Primarily, it offers researchers a direct, quantitative measure of their model's effectiveness and explanatory strength. By calculating this metric, analysts can rigorously validate theoretical assumptions and confidently assess how well their chosen independent variables account for the complexity of the phenomenon being studied, guiding future investigative steps.

Furthermore, explained variance serves as a crucial criterion for sophisticated model comparison and selection. When a researcher constructs several competing models designed to predict the same response variable, comparing their respective explained variance values (such as R-squared) allows for an objective determination of which model achieves the superior fit to the observed data. This process is essential for identifying the most parsimonious and effective set of predictor variables.

Finally, a solid grasp of explained variance dramatically enhances the interpretability and communication of statistical findings. It moves the discussion beyond merely reporting whether relationships are statistically significant to quantifying the practical magnitude of that relationship. By clearly stating the proportion of the outcome's variability that has been explained, researchers can better convey the strength, relevance, and real-world impact of their conclusions to academic colleagues and public audiences alike, fostering a comprehensive understanding of complex systems.