How to find the coefficient of determination (R-Squared) in R?

Authored by

stats writer

December 19, 2025

The calculation and interpretation of the coefficient of determination, often abbreviated as **R-Squared** (R^2), is a fundamental step in evaluating the performance of statistical models, particularly in the context of linear regression. In the R statistical environment, this metric is readily available through built-in functions designed for fitting and summarizing linear models. The standard workflow involves utilizing the powerful lm() function to construct the statistical model, followed by the `summary()` function to display the comprehensive diagnostic results, which include the R-Squared value. The higher the R-Squared value, the better the model fits the data, signifying that a larger proportion of the outcome's variability is explained by the predictors.

The **coefficient of determination** (commonly denoted R2) is the proportion of the variance in the response variable that can be explained by the explanatory variables in a regression model. This crucial metric quantifies the overall goodness-of-fit, reflecting how well the independent variables collectively account for the dispersion observed in the dependent variable.

This tutorial provides a comprehensive example detailing how to find and accurately interpret R2 within a regression model constructed using the R statistical environment. We will walk through data preparation, model fitting, summary extraction, and interpretation of the results.

For detailed context on evaluating model fit: What is a Good R-squared Value?

The Role of R-Squared in Model Assessment

The coefficient of determination (R^2) serves as a vital diagnostic tool for researchers and analysts employing regression techniques. It provides a standardized measure that allows for the straightforward assessment of a model's efficacy in explaining the underlying relationship between variables. Specifically, R^2 reveals the fraction of the total variance observed in the dependent variable that is predictable from the independent variables included in the model. This measure is crucial because raw estimates or coefficients alone do not convey the overall explanatory strength of the model, making R^2 indispensable for comparing different models fitted to the same dataset.

When constructing a predictive model, the goal is often to minimize the unexplained variation, or residual sum of squares. R-Squared encapsulates this minimization objective concisely, ranging strictly between 0 and 1. If R^2 equals 1, the model perfectly predicts the response variable, meaning all data points fall precisely on the regression line, and there is zero unexplained variation. Conversely, an R^2 of 0 suggests that the independent variables have absolutely no linear relationship with the dependent variable, and the model is no better at predicting the outcome than simply using the average response value. Therefore, R-Squared provides immediate feedback on the practical utility and robustness of the chosen statistical framework.

It is important to differentiate the role of R-Squared from measures focused solely on parameter

significance, such as p-values. While p-values inform us whether an individual predictor variable is statistically significant (i.e., its coefficient is likely not zero), R-Squared addresses the collective explanatory power of all predictors combined. A model might contain individually significant predictors but still possess a low R-Squared if the overall magnitude of the variance they explain is small relative to the total variance. For serious modeling efforts, assessing both the significance of individual coefficients and the overall fit provided by R-Squared is mandatory for a complete evaluation.

Prerequisites and Tooling in R

To accurately determine the R-Squared value within the R statistical environment, the primary prerequisite is the existence of a fitted linear regression model object. This model object is generated by the core statistical function, lm() function. The `lm()` function is used to estimate parameters of linear models and is one of the most frequently utilized commands in R for statistical analysis. It requires a formula specifying the relationship between the response variable and the explanatory variables, along with the data frame containing the variables.

Before fitting the model, analysts must ensure that the data is clean, appropriately structured, and meets the basic assumptions of linear regression. These assumptions include linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of error terms. Although R-Squared can be calculated even if assumptions are violated, its reliability and the model's overall validity are compromised. Therefore, careful exploratory data analysis (EDA) and checks for outliers or multicollinearity are essential preprocessing steps that precede model construction using the `lm()` function.

Once the model object is successfully created using `lm()`, the next step involves calling the `summary()` function on that object. The `summary()` function extracts and presents a comprehensive set of statistical results from the fitted model. These results typically include the estimated coefficients, standard errors, t-values, p-values for individual coefficients, residual diagnostics, and, crucially, both the **Multiple R-squared** and the **Adjusted R-squared** values. Accessing these metrics is instantaneous after the initial model fitting, showcasing the efficiency of R for statistical reporting.

Example: Find & Interpret R-Squared in R

Suppose we have the following dataset that contains data for the number of hours studied, preparatory exams taken, and the final exam score received for 15 students. We intend to use a multiple linear regression approach to predict the score based on the other two factors.

The dataset initialization below demonstrates the creation of the data frame, naming it `df`, and assigning the corresponding vectors for the three variables. Viewing the initial rows of the data

frame confirms successful data entry and structure, a necessary verification before proceeding to model fitting. The `score` variable represents the response variable, while `hours` and `prep_exams` act as the explanatory variables.

#create data frame
df <- data.frame(hours=c(1, 2, 2, 4, 2, 1, 5, 4, 2, 4, 4, 3, 6, 5, 3),
prep_exams=c(1, 3, 3, 5, 2, 2, 1, 1, 0, 3, 4, 3, 2, 4, 4),
score=c(76, 78, 85, 88, 72, 69, 94, 94, 88, 92, 90, 75, 96, 90, 82))

#view first six rows of data frame
head(df)

hours prep_exams score
1 1 1 76
2 2 3 78
3 2 3 85
4 4 5 88
5 2 2 72
6 1 2 69

Building and Summarizing the Regression Model

The following code shows how to fit a **multiple linear regression model** to this dataset. We specify the relationship using the formula notation: `score ~ hours + prep_exams`, indicating that score is modeled as a linear combination of the two predictors. The resulting object is stored as `model`, which is then passed to the `summary()` function to generate the comprehensive statistical report.

#fit regression model
model <- lm(score~hours+prep_exams, data=df)

#view model summary
summary(model)

Call:
lm(formula = score ~ hours + prep_exams, data = df)

Residuals:
Min 1Q Median 3Q Max
-7.9896 -2.5514 0.3079 3.3370 7.0352

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.8078 3.5222 20.387 1.12e-10 ***
hours 5.0247 0.8964 5.606 0.000115 ***
prep_exams -1.2975 0.9689 -1.339 0.205339

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.944 on 12 degrees of freedom
Multiple R-squared: 0.7237, Adjusted R-squared: 0.6776
F-statistic: 15.71 on 2 and 12 DF, p-value: 0.0004454

Extracting and Interpreting the R-Squared Value

The R-squared of the model, labeled "Multiple R-squared" and shown near the very bottom of the output, turns out to be **0.7237**. This is the calculated proportion of the total sum of squares that is explained by the regression. This value is critical for assessing the overall quality of the linear fit achieved by the model.

This result means that **72.37%** of the total variation in the exam scores across the 15 students can be statistically accounted for or explained by the combined effect of the number of hours studied and the number of preparatory exams taken. The remaining 27.63% of the variation is considered residual error, stemming from unobserved variables or inherent randomness not captured by the current model structure. A value exceeding 70% typically suggests a reasonably strong model fit in many social science and educational research contexts.

Note that you can also access this value directly and programmatically by using the following syntax, which is highly beneficial for automated data pipelines or conditional scripting:

summary(model)$r.squared

0.7236545

How to Interpret the R-Squared Value

An R-squared value will always range strictly between 0 and 1, representing the proportion of variance explained. A value of 1 indicates that the explanatory variables can perfectly explain the variance in the response variable, implying all observed data points lie exactly on the regression plane. This is rare in real-world data, especially in fields dealing with complex, human-generated variables.

Conversely, a value of 0 indicates that the explanatory variables have no ability to explain the variance in the response variable; the model is no better than predicting the outcome using the mean of the response variable alone. In general, the larger the R-squared value of a regression model, the better the explanatory variables are able to predict the value of the response variable. Our value of 0.7237 demonstrates substantial explanatory power.

However, analysts must remain cautious. A high R-Squared does not guarantee that the model is correctly specified or that the assumptions of linear regression have been met. It merely confirms a strong correlation between the predictors and the response. Therefore, R-Squared should always be evaluated alongside residual plots, p-values, and practical knowledge of the domain to ensure the model is both statistically sound and substantively meaningful.

Considering the Adjusted R-Squared

While the Multiple R-Squared is a useful measure, it possesses a notable flaw: it increases monotonically with the addition of any new predictor, regardless of its statistical relevance. To address this, the **Adjusted R-squared** (0.6776 in our example) is often the preferred metric in multiple regression. This metric incorporates a penalty for increasing the number of predictors (model complexity), adjusting the R-Squared downwards unless the new predictor significantly contributes to the model's explanatory power.

In our analysis, the Adjusted R-squared is only slightly lower than the Multiple R-squared (0.7237 vs. 0.6776). This small deviation suggests that the predictors included (hours and prep exams) are both contributing valuable and non-redundant information to the prediction of the score. If the difference were much larger, it would signal potential overfitting or the inclusion of superfluous variables that should be removed to simplify the model.

Check out this article for details on how to determine whether or not a given R-squared value is considered "good" for a given regression model, based on the specific application and field of study. Understanding the relationship between the two R-squared metrics is paramount for rigorous model selection.

Further reading on a related metric: How to Calculate Adjusted R-Squared in R

Understanding Explanatory and Response Variables