How to Interpret R Squared and Goodness of Fit in Regression Analysis

by KnowledgeHut (https://www.knowledgehut.com/blog/author/knowledgehut-editor) 12th Sep, 2020 Last updated on 17th Mar, 2021 10 mins read

Regression Analysis is a set of statistical processes that are at the core of data science. In the field of numerical simulation, it represents the most well-understood models and helps in interpreting machine learning algorithms. Their real-life applications can be seen in a wide range of domains, ranging from advertising and medical research to agricultural science and even different sports.

In linear regression models, R-squared is a goodness-fit-measure. It takes into account the strength of the relationship between the model and the dependent variable. Its convenience is measured on a scale of 0 – 100%.

Once you have a fit linear regression model, there are a few considerations that you need to address:

- How well does the model fit the data?
- How well does it explain the changes in the dependent variable?

In this article, we will learn about R-squared (R2), its interpretation, limitations, and a few miscellaneous insights about it.

Let us first understand the fundamentals of Regression Analysis and its necessity.

What is Regression Analysis?

Regression Analysis is a well-known statistical learning technique that allows you to examine the relationship between the independent variables (or explanatory variables) and the dependent variables (or response variables). It requires you to formulate a mathematical model that can be used to determine an estimated value which is nearly close to the actual value.

The two terms essential to understanding Regression Analysis:

- Dependent variables - The factors that you want to understand or predict.
- Independent variables - The factors that influence the dependent variable.

Consider a situation where you are given data about a group of students on certain factors: number of hours of study per day, attendance, and scores in a particular exam. The Regression technique allows you to identify the most essential factors, the factors that can be ignored and the dependence of one factor on others.

There are mainly two objectives of a Regression Analysis technique:

- Explanatory analysis - This analysis understands and identifies the influence of the explanatory variable on the response variable concerning a certain model.
- Predictive analysis - This analysis is used to predict the value assumed by the dependent variable.

Why use Regression Analysis?

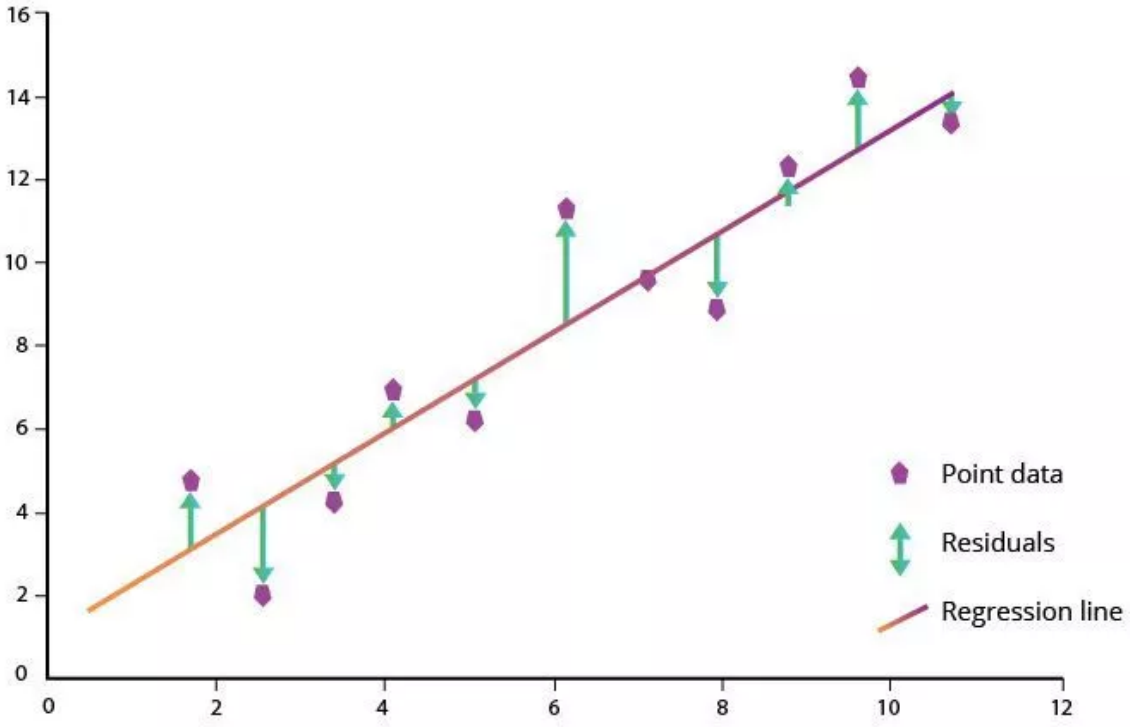
The technique generates a regression equation where the relationship between the explanatory variable and the response variable is represented by the parameters of the technique.

You can use the Regression Analysis to perform the following:

- To model different independent variables.
- To add continuous and categorical variables having numerous distinct groups based on a characteristic.
- To model the curvature using polynomial terms.
- To determine the effect of a certain independent variable on another variable by assessing the interaction terms.

What are Residuals?

Residuals identify the deviation of observed values from the expected values. They are also referred to as error or noise terms. A residual gives an insight into how good our model is against the actual value but there are no real-life representations of residual values.



Source: hatarilabs.com (http://hatarilabs.com)

Regression Line and residual plots

The calculation of the real values of intercept, slope, and residual terms can be a complicated task. However, the Ordinary Least Square (OLS) regression technique can help us to speculate on an efficient model. The technique minimizes the sum of the squared residuals. With the help of the residual plots (https://en.wikipedia.org/wiki/Partial_residual_plot), you can check whether the observed error is consistent with the stochastic error (differences between the expected and observed values must be random and unpredictable).

What is Goodness-of-Fit?

The Regression Analysis is a part of the linear regression technique. It examines an equation that reduces the distance between the fitted line and all of the data points. Determining how well the model fits the data is crucial in a linear model.

A general idea is that if the deviations between the observed values and the predicted values of the linear model are small and unbiased, the model has a well-fit data.

In technical terms, "Goodness-of-fit" is a mathematical model that describes the differences between the observed values and the expected values or how well the model fits a set of observations. This measure can be used in statistical hypothesis testing.

How to assess Goodness-of-fit in a regression model?

According to statisticians, if the differences between the observations and the predicted values tend to be small and unbiased, we can say that the model fits the data well. The meaning of unbiasedness in this context is that the fitted values do not reach the extremes, i.e. too high or too low during observations.

As we have seen earlier, a linear regression model gives you the outlook of the equation which represents the minimal difference between the observed values and the predicted values. In simpler terms, we can say that linear regression identifies the smallest sum of squared residuals probable for the dataset.

Determining the residual plots represents a crucial part of a regression model and it should be performed before evaluating the numerical measures of goodness-of-fit, like R-squared. They help to recognize a biased model by identifying problematic patterns in the residual plots.

However, if you have a biased model, you cannot depend on the results. If the residual plots look good, you can assess the value of R-squared and other numerical outputs.

What is R-squared?

In data science (https://www.knowledgehut.com/data-science/data-science-with-python-certification-training), R-squared (R2) is referred to as the coefficient of determination or the coefficient of multiple determination in case of multiple regression.

In the linear regression model, R-squared acts as an evaluation metric to evaluate the scatter of the data points around the fitted regression line. It recognizes the percentage of variation of the dependent variable.

R-squared and the Goodness-of-fit

R-squared is the proportion of variance in the dependent variable that can be explained by the independent variable.

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total Variance}}$$

The value of R-squared stays between 0 and 100%:

- 0% corresponds to a model that does not explain the variability of the response data around its mean. The mean of the dependent variable helps to predict the dependent variable and also the regression model.
- On the other hand, 100% corresponds to a model that explains the variability of the response variable around its mean.

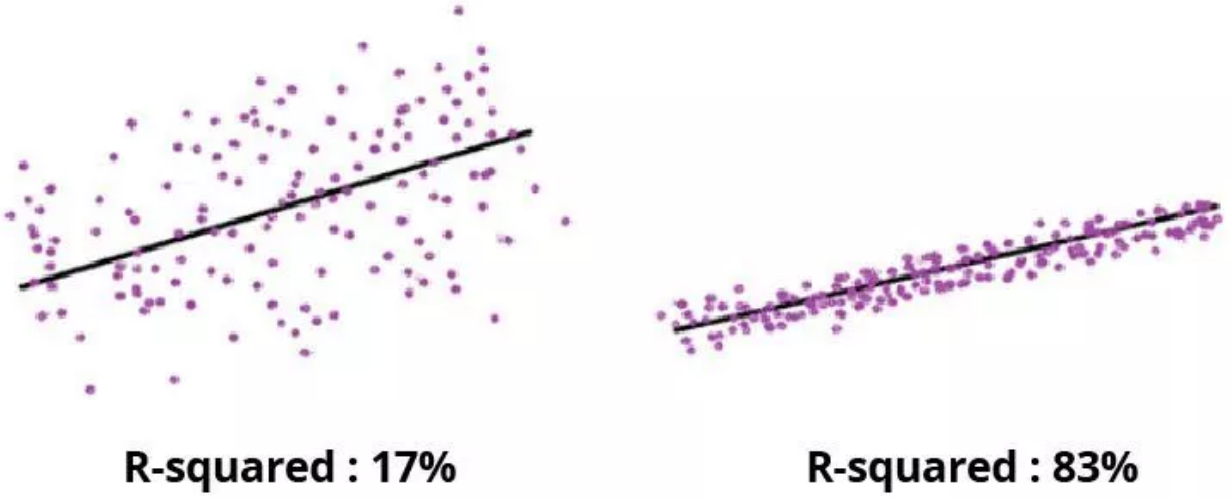
If your value of R2 is large, you have a better chance of your regression model fitting the observations.

Although you can get essential insights about the regression model in this statistical measure, you should not depend on it for the complete assessment of the model. It does not give information about the relationship between the dependent and the independent variables.

It also does not inform about the quality of the regression model. Hence, as a user, you should always analyze R2 along with other variables and then derive conclusions about the regression model.

Visual Representation of R-squared

You can have a visual demonstration of the plots of fitted values by observed values in a graphical manner. It illustrates how R-squared values represent the scatter around the regression line.



As observed in the pictures above, the value of R-squared for the regression model on the left side is 17%, and for the model on the right is 83%. In a regression model, when the variance accounts to be high, the data points tend to fall closer to the fitted regression line.

However, a regression model with an R2 of 100% is an ideal scenario which is actually not possible. In such a case, the predicted values equal the observed values and it causes all the data points to fall exactly on the regression line.

Interpretation of R-squared

The simplest interpretation of R-squared is how well the regression model fits the observed data values. Let us take an example to understand this.

Consider a model where the R2 value is 70%. This would mean that the model explains 70% of the fitted data in the regression model. Usually, when the R2 value is high, it suggests a better fit for the model.

The correctness of the statistical measure does not only depend on R2 but can depend on other several factors like the nature of the variables, the units on which the variables are measured, etc. So, a high R-squared value is not always likely for the regression model and can indicate problems too.

A low R-squared value is a negative indicator for a model in general. However, if we consider the other factors, a low R2 value can also end up in a good predictive model.

Calculation of R-squared

R- squared can be evaluated using the following formula:

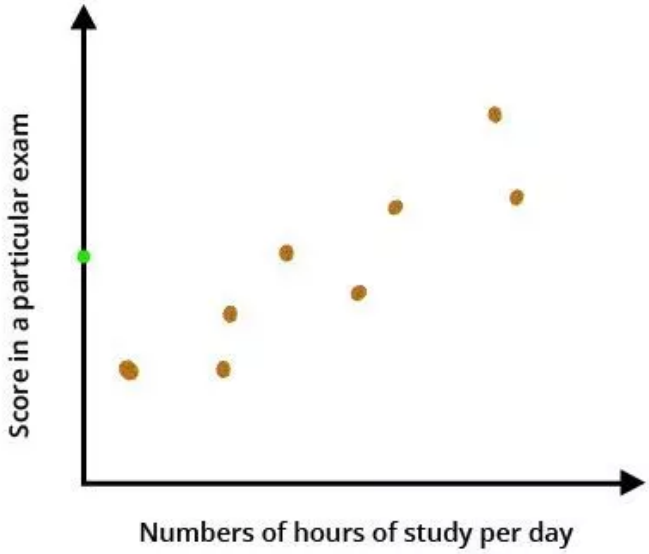
$$\text{R-squared} = \frac{SS_{regresion}}{SS_{total}}$$

Where:

- SSregression – Explained sum of squares due to the regression model.
- SStotal – The total sum of squares.

The sum of squares due to regression assesses how well the model represents the fitted data and the total sum of squares measures the variability in the data used in the regression model.

Now let us come back to the earlier situation where we have two factors: number of hours of study per day and the score in a particular exam to understand the calculation of R-squared more effectively. Here, the target variable is represented by the score and the independent variable by the number of hours of study per day.

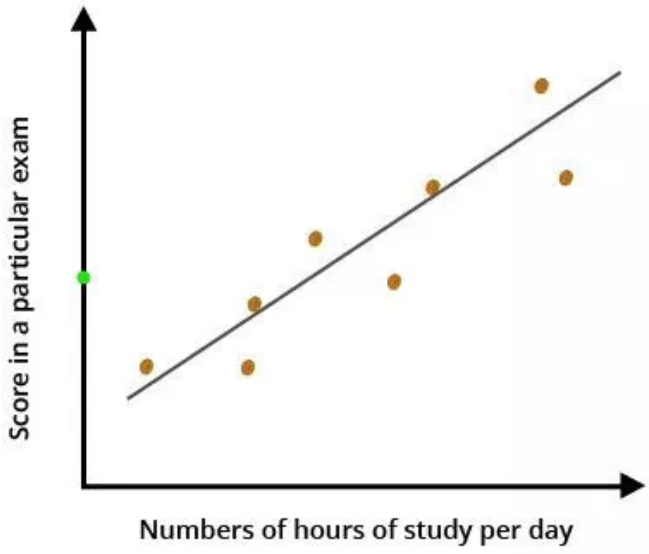


In this case, we will need a simple linear regression model and the equation of the model will be as follows:

$$\hat{y} = w1x1 + b$$

The parameters w1 and b can be calculated by reducing the squared error over all the data points. The following equation is called the least square function:

$$\text{minimize } \sum(yi - w1x1i - b)2$$



Now, to calculate the goodness-of-fit, we need to calculate the variance:

$$var(u) = 1/n\sum(ui - \bar{u})2$$

where, n represents the number of data points.

Now, R-squared calculates the amount of variance of the target variable explained by the model, i.e. function of the independent variable.

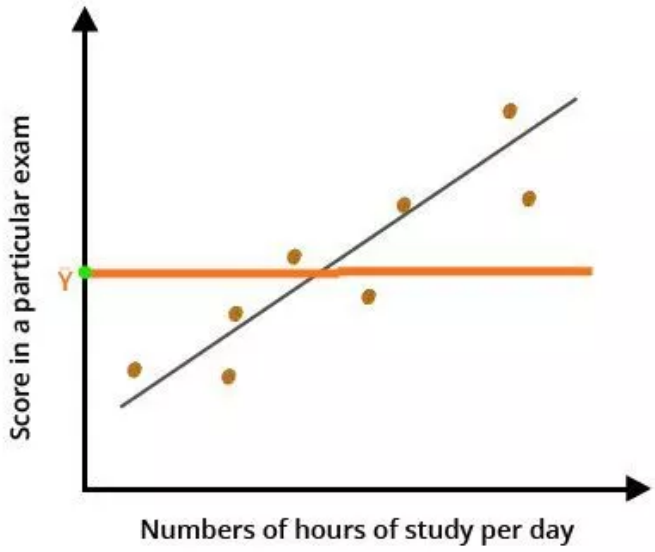
However, in order to achieve that, we need to calculate two things:

- Variance of the target variable:

$var(avg) = \sum(y_i - \bar{Y})^2$

- Variance of the target variable around the best-fit line:

$var(model) = \sum(y_i - \hat{y})^2$



Finally, we can calculate the equation of R-squared as follows:

$R2 = 1 - [var(model)/var(avg)] = 1 - [\sum(y_i - \hat{y})^2 / \sum(y_i - \bar{Y})^2]$

Limitations of R-squared

Some of the limitations of R-squared are:

- R-squared cannot be used to check if the coefficient estimates and predictions are biased or not.
- R-squared does not inform if the regression model has an adequate fit or not.

To determine the biasedness of the model, you need to assess the residuals plots. A good model can have a low R-squared value whereas you can have a high R-squared value for a model that does not have proper goodness-of-fit.

Low R-squared and High R-squared values

Regression models with low R2 do not always pose a problem. There are some areas where you are bound to have low R2 values. One such case is when you study human behavior. They tend to have R2 values less than 50%. The reason behind this is that predicting people is a more difficult task than predicting a physical process.

You can draw essential conclusions about your model having a low R2 value when the independent variables of the model have some statistical significance. They represent the mean change in the dependent variable when the independent variable shifts by one unit.

However, if you are working on a model to generate precise predictions, low R-squared values can cause problems.

Now, let us look at the other side of the coin. A regression model with high R2 value can lead to – as the statisticians call it – specification bias. This type of situation arises when the linear model is underspecified due to missing important independent variables, polynomial terms, and interaction terms.

To overcome this situation, you can produce random residuals by adding the appropriate terms or by fitting a non-linear model.

Model overfitting and data mining techniques can also inflate the value of R2. The model they generate might provide an excellent fit to the data but actually the results tend to be completely deceptive.

Conclusion

Let us summarize what we have covered in this article so far:

- Regression Analysis and its importance
- Residuals and Goodness-of-fit
- R-squared: Representation, Interpretation, Calculation, Limitations
- Low and High R2 values

Although R-squared is a very intuitive measure to determine how well a regression model fits a dataset, it does not narrate the complete story. If you want to get the full picture, you need to have an in-depth knowledge of R2 along with other statistical measures and residual plots.

For gaining more information on the limitations of the R-squared, you can learn about Adjusted R-squared and Predicted R-squared which provide different insights to assess a model's goodness-of-fit. You can also take a look at a different type of goodness-of-fit measure, i.e. Standard Error of the Regression.

KnowledgeHut
Author

KnowledgeHut is an outcome-focused global ed-tech company. We help organizations and professionals unlock excellence through skills development. We offer training solutions under the people and process, data science, full-stack development, cybersecurity, future technologies and digital transformation verticals.
Website : https://www.knowledgehut.com (https://www.knowledgehut.com)

BLOGS (HTTPS://WW...

Will You Surely Get A Job After PMP® Certification? (https://www.knowledgehut.com/blog/project-management/will-you-surely-get-a-job-after-pmpr-certification)

Published 22 May 2018

BLOGS (HTTPS://WW...

Most Recent Updates In PMBOK® Guide And PMP® Classes By KnowledgeHut (https://www.knowledgehut.com/blog/project-management/most-recent-updates-in-pmbokr-guide-and-pmpr-classes-by-knowledgehut)

Published 16 Mar 2018

BLOGS (HTTPS://WW...

PMP or Prince 2 – The Management Certification Suitable For You (https://www.knowledgehut.com/blog/project-management/pmp-or-prince-2-the-management-certification-suitable-for-you)

Published 08 Dec 2017

BLOGS (HTTPS://WW...

How the PMP Certification Impacts Your Salary? (https://www.knowledgehut.com/blog/project-management/how-the-pmp-certification-impacts-your-salary)

Published 13 Oct 2017

BLOGS (HTTPS://WW...

WRITE FOR US (HTTPS://WWW.KNOWLEDGEHUT.COM/BLOG/WRITE-FOR-US)

SUGGESTED BLOGS

(https://www.knowledgehut.com/blog/data-science/regression-analysis-and-its-techniques-in-data-science)

BLOGS (HTTPS://WWW.KNOWLE...

5636

Regression Analysis and Its Techniques in Data Sci... (https://www.knowledgehut.com/blog/data-science/regression-analysis-and-its-techniques-in-data-science)

by Priyankur Sarkar (https://www.knowledgehut.com/blog/author/priyankur-sarkar) | 07 Jan 2022

As a Data Science enthusiast, you might already ... READ MORE (HTTPS://WWW.KNOWLEDGEHUT.COM/BLOG/DATA-SCIENCE/REGRESSION-ANALYSIS-AND-ITS-TECHNIQUES-IN-DATA-SCIENCE)

(https://www.knowledgehut.com/blog/data-science/data-scientist-learning-path)

BLOGS (HTTPS://WWW.KNOWLE...

5364

How to Become a Data Scientist (https://www.knowledgehut.com/blog/data-science/data-scientist-learning-path)

by Priyankur Sarkar (https://www.knowledgehut.com/blog/author/priyankur-sarkar) | 10 Jan 2022

According to a recent Harvard Business Review arti... READ MORE (HTTPS://WWW.KNOWLEDGEHUT.COM/BLOG/DATA-SCIENCE/DATA-SCIENTIST-LEARNING-PATH)

(https://www.knowledgehut.com/blog/data-science/how-to-become-a-data-engineer)

BLOGS (HTTPS://WWW.KNOWLE...

4812

How to Become a Data Engineer (https://www.knowledgehut.com/blog/data-science/how-to-become-a-data-engineer)

by Priyankur Sarkar (https://www.knowledgehut.com/blog/author/priyankur-sarkar) | 10 Jan 2022 | 12 mins read

Data Engineering is typically a software engineeri... READ MORE (HTTPS://WWW.KNOWLEDGEHUT.COM/BLOG/DATA-SCIENCE/HOW-TO-BECOME-A-DATA-ENGINEER)

LOAD MORE

Connect with us

(https://www.linkedin.com/company/knowledgehut) (https://www.facebook.com/KnowledgeHut.Global)
(https://www.instagram.com/knowledgehut.global) (https://www.youtube.com/user/TheKnowledgehut)
(https://twitter.com/KnowledgeHut)

Get Our Weekly Newsletter

Enter Your E-mail SUBSCRIBE

We Accept

USA : +1-469-442-0620 (tel:+1-469-442-0620), +1-832-684-0080 (tel:+1-832-684-0080)

India : +91-84484-45027 (tel:+91-84484-45027)
Toll Free: 1800-121-9232 (tel:1800-121-9232)

UK: +44-2080890434 (tel:+44-2080890434)

Canada: +1-613-707-0763 (tel:+1-613-707-0763)

Singapore: +65-315-83941 (tel:+65-315-83941)

Malaysia: +601548770914 (tel:+601548770914)
Australia: +61-290995641 (tel:+61-290995641)

UAE: Toll Free 8000180860 (tel:8000180860)

New Zealand: +64-36694791 (tel:+64-36694791)

Ireland: +353-14402544 (tel:+353-14402544)