My Store Glossary Home About Me Contact Me

Statistics By Jim

Making statistics intuitive

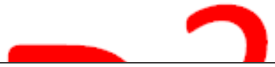
Graphs Basics Hypothesis Testing Regression ANOVA Probability Time Series
Fun

How To Interpret R-squared in Regression Analysis

By Jim Frost — 126 Comments

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

After fitting a linear regression model, you need to determine how well the model fits the data. Does it



X

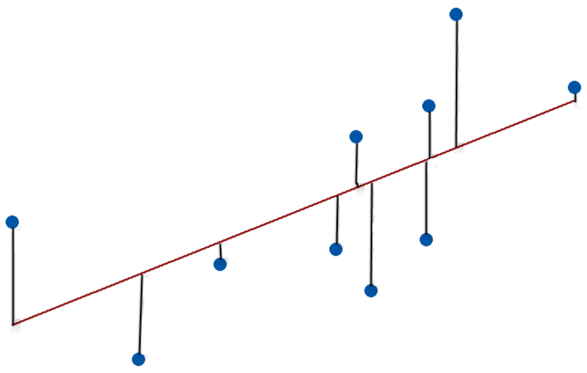
this post, we'll examine R-squared (R^2), highlight some of its limitations, and discover some surprises. For instance, small R-squared values are not always a problem, and high R-squared values are not necessarily good!

Related posts: When Should I Use Regression Analysis? and How to Perform Regression Analysis using Excel

X

observed values and their fitted values. To be precise, linear regression finds the smallest sum of squared residuals that is possible for the dataset.

Statisticians say that a regression model fits the data well if the differences between the observations and the predicted values are small and unbiased. Unbiased in this context means that the fitted values are not systematically too high or too low anywhere in the observation space.



Residuals are the distance between the observed value and the fitted value.

However, before assessing numeric measures of goodness-of-fit, like R-squared, you should evaluate the residual plots. Residual plots can expose a biased model far more effectively than the numeric output by displaying problematic patterns in the residuals. If your model is biased, you cannot trust the results. If your residual plots look good, go ahead and assess your R-squared and other statistics.

Read my post about checking the residual plots.

R-squared and the Goodness-of-Fit

R-squared evaluates the scatter of the data points around the fitted regression line. It is also called the coefficient of determination, or the coefficient of multiple determination for multiple regression. For the same data set, higher R-squared values represent smaller differences

X

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

R-squared is always between 0 and 100%:

- 0% represents a model that does not explain any of the variation in the response variable around its mean. The mean of the dependent variable predicts the dependent variable as well as the regression model.
- 100% represents a model that explains all the variation in the response variable around its mean.

Usually, the larger the R^2, the better the regression model fits your observations. However, this guideline has important caveats that I'll discuss in both this post and the next post.

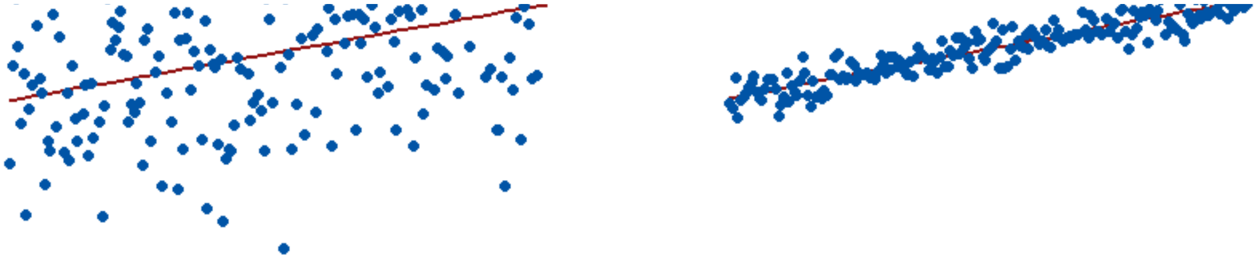
Linear regression uses the sum of squares for your model to find R-squared. Learn more about these calculations in my post, Sum of Squares.

Related post: What are Independent and Dependent Variables?

Visual Representation of R-squared

To visually demonstrate how R-squared values represent the scatter around the regression line, you can plot the fitted values by observed values.

X



The R-squared for the regression model on the left is 15%, and for the model on the right it is 85%. When a regression model accounts for more of the variance, the data points are closer to the regression line. In practice, you’ll never see a regression model with an R^2 of 100%. In that case, the fitted values equal the data values and, consequently, all the observations fall exactly on the regression line.

R-squared has Limitations

You cannot use R-squared to determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots.

R-squared does not indicate if a regression model provides an adequate fit to your data. A good model can have a low R^2 value. On the other hand, a biased model can have a high R^2 value!

Are Low R-squared Values Always a Problem?

No! Regression models with low R-squared values can be perfectly good models for several reasons.

behavior generally have R^2 values less than 50%. People are just harder to predict than things like physical processes.

Fortunately, if you have a low R-squared value but the independent variables are statistically significant, you can still draw important conclusions about the relationships between the variables. Statistically significant coefficients continue to represent the mean change in the dependent variable given a one-unit shift in the independent variable. Clearly, being able to draw conclusions like this is vital.

Related post: How to Interpret Regression Models that have Significant Variables but a Low R-squared

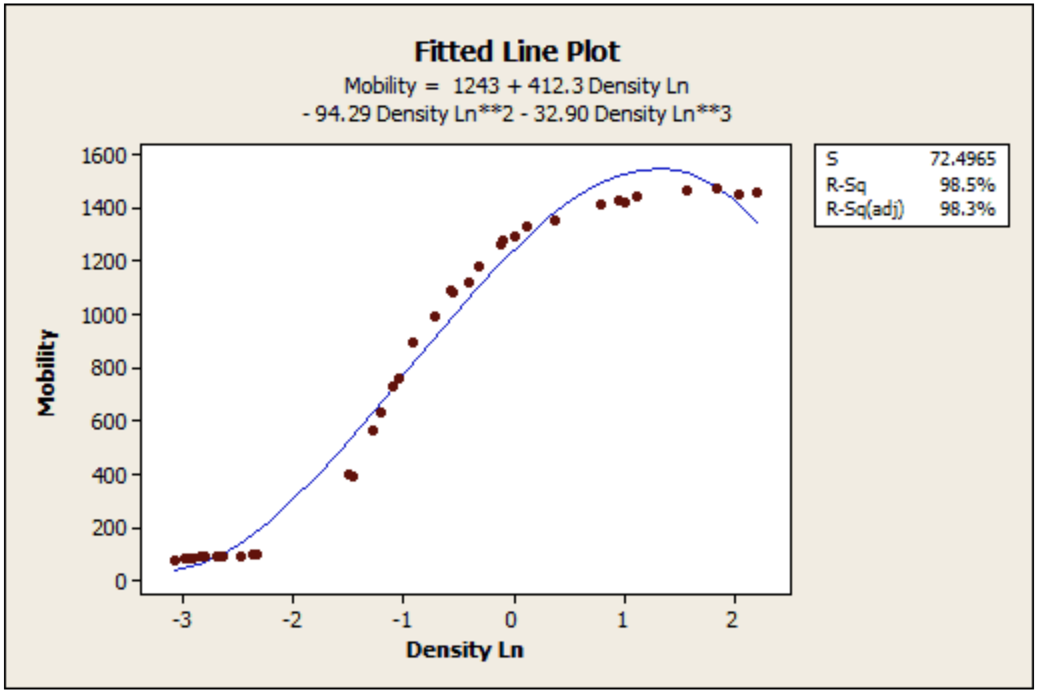
There is a scenario where small R-squared values can cause problems. If you need to generate predictions that are relatively precise (narrow prediction intervals), a low R^2 can be a showstopper.

How high does R-squared need to be for the model to produce useful predictions? That depends on the precision that you require and the amount of variation present in your data. A high R^2 is necessary for precise predictions, but it is not sufficient by itself, as we'll uncover in the next section.

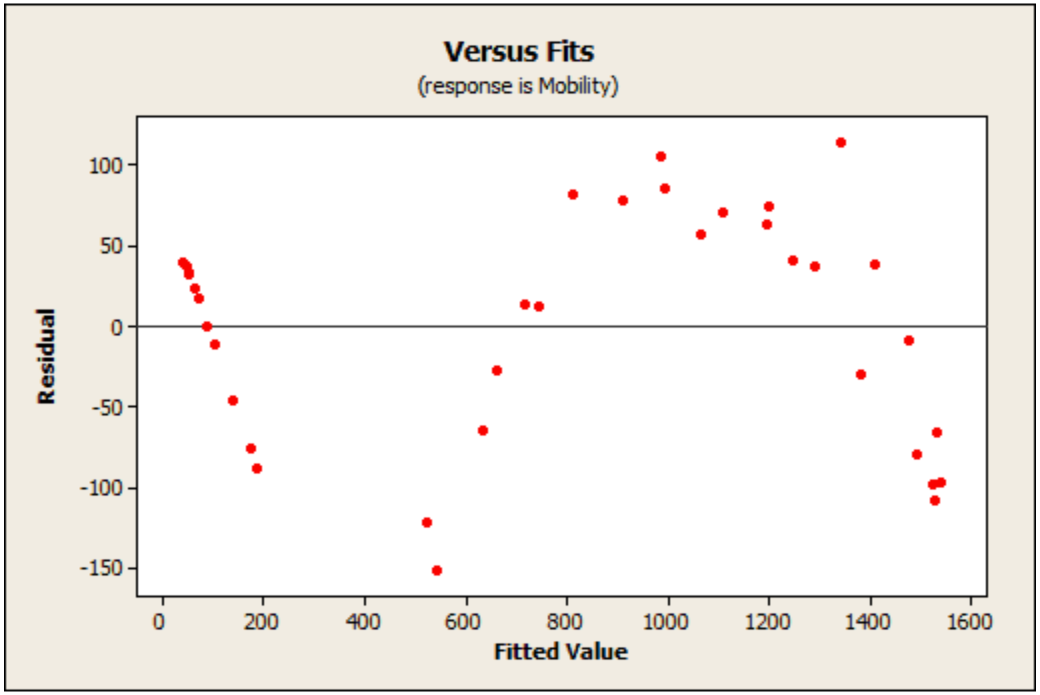
Related posts: Understand Precision in Applied Regression to Avoid Costly Mistakes and Mean Squared Error (MSE)

X

probably expect that a high R^2 indicates a good model but examine the graphs below. The fitted line plot models the association between electron mobility and density.



X



The data in the fitted line plot follow a very low noise relationship, and the R-squared is 98.5%, which seems fantastic. However, the regression line consistently under and over-predicts the data along the curve, which is bias. The Residuals versus Fits plot emphasizes this unwanted pattern. An unbiased model has residuals that are randomly scattered around zero. Non-random residual patterns indicate a bad fit despite a high R^2. Always check your residual plots!

This type of specification bias occurs when your linear model is underspecified. In other words, it is missing significant independent variables, polynomial terms, and interaction terms. To produce random residuals, try adding terms to the model or fitting a nonlinear model.

X

A variety of other circumstances can artificially inflate your R^2. These reasons include overfitting the model and data mining. Either of these can produce a model that looks like it provides an excellent fit to the data but in reality the results can be entirely deceptive.

An overfit model is one where the model fits the random quirks of the sample. Data mining can take advantage of chance correlations. In either case, you can obtain a model with a high R^2 even for entirely random data!

Related post: Five Reasons Why Your R-squared can be Too High

R-squared Is Not Always Straightforward

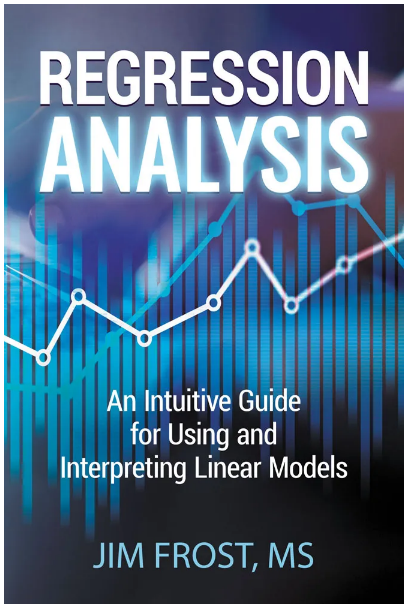
At first glance, R-squared seems like an easy to understand statistic that indicates how well a regression model fits a data set. However, it doesn't tell us the entire story. To get the full picture, you must consider R^2 values in combination with residual plots, other statistics, and in-depth knowledge of the subject area.

I'll continue to explore the limitations of R^2 in my next post and examine two other types of R^2: adjusted R-squared and predicted R-squared. These two statistics address particular problems with R-squared. They provide extra information by which you can assess your regression model's goodness-of-fit.

You can also read about the standard error of the regression and the root mean square error, which are different typed of goodness-of-fit measures.

X

If you're learning regression and like the approach I use in my blog, check out my Intuitive Guide to Regression Analysis book! You can find it on Amazon and other retailers.



Note: I wrote a different version of this post that appeared elsewhere. I've completely rewritten and updated it for my blog site.

Filed Under: Regression

Tagged With: conceptual, interpreting results

X

Odeke David Onyebuchi says

August 23, 2022 at 5:23 pm

Thanks a lot sir it really helps me

Reply

Rhea Elliv says

June 23, 2022 at 6:45 am

Good day, Mr. Jim
Here is my question after I read the entire article of yours, and its very informative.
The best test of the performance of two different regression equations is their respective values of R2?

Thank you.

Reply

Jim Frost says

X

Evaluating two (or more) different models can be complex and definitely involves more than any one statistic, such as R-squared. I've written an entire article about how to evaluate and choose models. Please click the link to read about it.

Reply

Mehnaz Arain says
June 21, 2022 at 5:11 pm

Hello experts
My regression r square is .446 what i can do next which model i can use please guide me

Reply

Jim Frost says
June 21, 2022 at 11:16 pm

Hi Mehnaz,

X

statistically significant.

Reply

Norainny Yunitasari says

May 19, 2022 at 10:32 am

hello sir, let me ask, is the value of R-square used to determine whether or not a linear equation can be used? and is there a categorization of R-square values?

Reply

X

Jim Frost says

May 25, 2022 at 3:24 pm

Hi,

No, the R-squared cannot be used to determine whether the linear equation is sufficient. I write about that in this article specifically. To learn how to evaluate a linear model, you particular need to pay attention to the residual plots. Also, read my post about choosing the best regression model for more details.

Reply

Hanna Claire Schwant says

May 18, 2022 at 8:16 pm

Dear Jim,
I'm a huge fan of this article. I've been thinking about it for days. I'm truly inspired by your work ethic and knowledge and hope to one day achieve the same. It was an honour to read such beautiful writing.

X

Jim Frost says

May 18, 2022 at 9:53 pm

Hi Hanna! Thanks so much, you’re making me blush! Your kind words made my day! 😳

Reply

Adam Levitt says

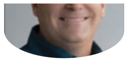
April 7, 2022 at 1:10 pm

Hello Jim,

Me and a few other engineering students at Purdue are working on a project which involves data modeling. The dataset is non-linear, but we need to find the initial slope. We are working on a program that cuts the data down until it reaches a certain r^2 value. The model resembles a ln(x) graph. From your vast experience, what would be a reasonable r^2 value to use to determine the slope? Thank you so much.

Reply

X



April 8, 2022 at 8:08 pm

Hi Adam,

I don’t fully understand what your project seeks to do but using R-squared to find a slope is probably not the best way.

There are several complications here. First, there’s no single R-squared value that is acceptable for all subject areas. You’d really need to see what R-squared values are normal for your specific subject area. Also, as I show in this post, high R-squared values do not necessarily indicate that your slope is correct nor do low R-squared values indicate that the slope is incorrect. So, R-squared isn’t a perfect measure for your task. There frequently is no relationship between R-squared and finding the slope.

Another complication is that it looks like you’re transforming the data. In that case, R-squared applies to the transformed response (outcome) variable rather than the raw data. That makes understanding what R-squared represents a bit murkier.

My suggestion is to look through many examples of the analysis that you’re performing. Perform it many times and see if a relationship between R-squared values and getting a good slope exists. If it does, you might be able to empirically find a good cutoff value for R-squared. Understand that if you can find a good R-squared for your task, that it is very specific to the subject area. In other words, if you’re limiting the procedure to work in a very specific subject area, there *might* be an R-squared value that tends to work for it. Maybe

X

Mo Hashem says

January 9, 2022 at 3:03 am

Hi Jim…my regression model has better results when I included an intercept dummy variable. However, the dummy variable is insignificant… why is that? How can I interpret that?
Thank you for your help

Reply

Jim Frost says

January 14, 2022 at 5:13 pm

Hi Mo,

Anytime you add a new variable, R-squared will increase. That's one of the shortcomings I mention about R-squared. I focus on this in my post about adjusted R-squared. This problem occurs because any chance correlation between the new DV and the IV causes R-squared to increase. Consequently, it's not a good idea to use R-squared by itself to determine whether to include a variable in your model.

For interpretation, you'd just say that the dummy variable is not significant. When

Reply

Janet says

January 8, 2022 at 5:57 pm

I have a VAR model ,can i use the R-square values to explain how good the model explains the dependent variable (explanatory power of the model) and if yes how will the values of the R-square be interpreted.
I have R-squared values for all variables both dependent and the 3 independents, how do i interprete them?

Reply

X

Jim Frost says

January 14, 2022 at 5:15 pm

Hi Janet,

I haven't used VAR models myself, so I'm not an expert in them. However, I believe that R-squared has the same interpretation in them as linear regression because it's a form of linear regression. But I'd double-check that!

Reply

Yonas says

December 27, 2021 at 10:49 am

Hi Jim,
I would like to ask:
What would be the value of R-squared, in the case of a regression model with a constant term and no explanatory variables . For instance y = β0 + u
Thank you in advance

X

Ben Smith says

November 19, 2021 at 12:06 pm

Hi Jim,
I have a variable of interest (lung function) that varies in the population and I would like to quantify important sources of this variation (e.g. age, sex, smoking, etc…). Measurement error and/or intrinsic ‘normal’ day-to-day fluctuations also contribute to some of the variation in lung function. My question: Is there a way to estimate the expected ceiling of R2 given that measurement error will be ‘unexplainable’? I’m wondering if calculating the intra-class correlation coefficient from repeated lung function measurements estimates this.
Thanks,
Ben

Reply

Jim Frost says

November 21, 2021 at 8:16 pm

Hi Ben,

The only way I can think of would be to look at similar studies if they exist and see

Reply

Andee says

September 3, 2021 at 3:45 pm

Thank you for asking this. I too am an appraiser and see that all the time and don’t understand why they use it, especially because most don’t understand it to be a relationship between data, which they can’t explain. It’s also not the figure used to explain adjustments. I think most are just at a point where they want to show a picture but don’t know what it means and figure “everyone else is doing it…”.

I have found that the mean and median are better for our industry and dropping a graph with a trend line helps sell the explanation of “Yes, it actually is going up or down”. I don’t actually see another use for it and find it to be more work added to our already busy day.

Reply

X

Thank you again for your insights!

The part I am struggling with is that I do not use random samples.

Being a real estate appraiser I delineated to my specific competitive market segment of competing sales for the property I am appraising, so my competitive market segment is not a random sample…so am I correct that inferential statistics, such as p-values, r squared, etc…are not relevant as they pertain to models built on a random sample?

Descriptive statistics…mean, median, etc…appear to be what are relevant for non-random samples.

I see so many people in my field taking non-random samples and then noting the relevance of their p-values and their r squared values within their analysis as proof and support for their conclusions, but to me this appears to be misleading and incorrect.

People in my industry appear to be running incorrect models and passing the outputs of the software (p-values, r squared, etc…) as meaningful but to me, just because excel or another software program spits out such statistics doesn't mean the outputs are relevant to their analysis.

Your thoughts?

Reply

X

It can be hard to collect a truly random sample. There's often some sort of approximation and it's important to understand how your sample differs from a representative sample. Sometimes you can compare your sample statistics to other, fuller datasets to get an idea. Sometimes it's an educated guessed based on knowledge about how you acquired your sample (i.e., what observations will be missed/excluded based on your methodology).

When it comes to estimating the relationships in the data, your coefficient estimates will reflect the range of data in your sample. Consequently, if the relationship changes throughout the full population space and your sample only contains a portion of the full range, the estimated relationships will be for that portion rather than the full population. I show an example of how this works in the section about interpreting the constant (y-intercept) where I explain how a relationship can be locally linear but curvilinear overall.

So, you need to understand how representative, or not, your sample is and how that could affect the estimates. You should use estimated relationships only within the range of data you collect. The relationships can change outside your sample. Conversely, you can safely trust the estimates within the range of your sample even if you're not randomly sampling the entire population, assuming your satisfying the usual regression assumptions of course. The coefficients and their p-values would apply for within your sample space and they can be wrong outside that space.

In terms of the goodness-of-fit, when you collect a random sample, the sample variability of the independent variable values should reflect the variability in the

X

adjusted R-squared tends to underestimate the population goodness-of-fit. Conversely, if the variability of the sample is greater than the population variability, adjusted R-squared tends to overestimate goodness-of-fit.

That’s the general gist. And, knowing how much to trust or not trust the statistics depends on a good understanding of the subject area and how the data were collected. And, you need to always remember to not use these statistics outside the sample space!

Reply

Ed Bedinotti says
August 10, 2021 at 10:37 pm

Jim,

I am well into your Regression Analysis book and enjoy your writing style of explaining things in plain English.

Goodness of fit….if I’m reading correctly, it too is an aspect of inferential statistics, when the model is for a random sample…?

X

Jim Frost says

August 10, 2021 at 11:01 pm

Hi Ed,

Thanks so much for buying my book and I'm so glad to hear that it's been helpful!

Yes, that's absolutely correct. At least, it can be a population property that you estimate using a sample. Like many statistics, it can simply describe your sample or, when you have a representative sample, it can estimate a characteristic of your population.

However, there is a key difference between using R-squared to estimate the goodness-of-fit in the population versus, say, the mean. The mean is a unbiased estimator, which means the population estimate won't be systematically too high or too low. However, R-squared is a biased estimator. It tends to be higher than the true population value. Adjusted R-square corrects this problem by shrinking the R-squared down to a value where it becomes an unbiased estimator. We usually think of adjusted R-squared as a way to compare the goodness-of-fit for models with differing numbers of IVs. However, it's also the unbiased estimator of GOF in the population. I write about this briefly in chapter 11 of the book and show how it works.

Reply

X

Amy says

August 5, 2021 at 1:36 pm

Hello! I would like to say thank you for this great insights on regression. I have used your explanation on low R-square for my thesis. However, I am having a bit trouble on how I should recite your blog as my reference (APA style). Do you have any other resources which I could use to recite for my reference?

Reply

X

Jim Frost says

August 6, 2021 at 11:58 pm

Hi Amy,

You're very welcome and I'm so glad it's been helpful!

For website references, I always recommend checking Purdue's Online Writing Lab (OWL) resources. Here's their page on the APA style for electronic resources, such as websites.

I also have books, ebooks and printed, which can be cited.

I hope that helps!

Reply

Ed Bedinotti says

July 25, 2021 at 10:15 am

I am a real restate appraiser and am able to delineate to my competitive market segment…I am able to export a csv file of all sales that were available and competed with my subject as of the effective date of my valuation, so I am dealing with the entire population of competitive sales. So, to me, descriptive statistics are all I really need.

I primarily use regression for market price indexing. I scatter plot a linear regression line and then a 3rd order polynomial line over the linear line so I can visually see if and when a change occurred. If a change occurred…say 6 months back from my effective date…if the 3rd order polynomial line curves up, illustrating an increasing sale price trend, I then use a spline date from the point in time when the sales prices began increasing, to my effective date, and plot a linear regression line so I can establish the value increase per day over that time period. Sales price is on the y axis and sale date is on the x axis.

I am seeing a lot of people plotting 4th, 5th, 6th, etc…polynomial lines, but it seems to me like that is overfitting as the line simply follows more dots and you don't really learn anything and the results seem bias as the increased polynomial lines appear to over and under illustrate the actual sales point directions.

Wanted to get your thought on overfitting.

I look forward to reading your book as well.

Ed

Reply

X

July 26, 2021 at 1:23 am

Ed, it sounds like you have your method down!

I agree that using 4th and higher order polynomials is overkill. I'd consider it overfitting in most any conceivable scenario. I've personally never even used third-order terms in practice. You might be fine with just a squared term. But, given that you're not stuffing the model with other variables, using a cubed term is probably fine–especially given that you're just using it to denote an increasing trend and to split your data for the specific date range. Cubed terms imply there are two bends/changes in direction in the curve over the range of the data. These bends should actually exist and have a strong theoretical basis supporting them. As you say, it's not a good idea to include unnecessarily high order terms just to follow the dots more closely. The problem with using unnecessarily high order terms (and overfitting in general) is that they tend to fit the noise in the data rather than the real relationship.

I write about polynomial terms and overfitting in my regression book. In fact, I specifically use a regression example example with real data where the real relationship has an R-squared of about zero but when you include a cubed term, it fits the noise in the data and produces a decent R-squared (60-70% IIRC).

Jim

Reply

X

Ed Bedinotti says

July 24, 2021 at 8:47 am

Jim, little r squared, big R squared, p-values, etc.. while relevant for inferential statistics, they aren't really relevant when you have the data for your entire population and aren't dealing with random samples.

Correlation isn't necessarily causation and I see people not understanding the difference.

Does your book, or do you have a book or articles that deal with non-parametric regression?

Ed Bedinotti

Reply

Jim Frost says

July 25, 2021 at 12:49 am

Hi Ed,

Yes, you're correct. R-squared (or more appropriately adjusted R-squared, which is data for the entire population!

I haven’t written about nonparametric regression yet. Although, nonparametric procedures are still inferential tests with p-values, population estimators, etc. I do write extensively about how correlation isn’t necessarily causation, when it might be, and how to tell in my introduction to statistics book.

Reply

Vincent Schantz says
July 21, 2021 at 3:36 pm

Hey Jim,

I’m trying to forecast future sales – if the R2 for advertising spend and net sales is 0.955 (the closest to 100% I’ve found), does that mean it is the most accurate independent/dependent variable combo to forecast sales? In summary, using the forecast.linear function to find out what future (larger) amount of ad spend will yield for net sales? The data set I’m using to feed the regression goes back to 11/2018. Thanks!

Reply

X

Hi Vincent,

Finding the model with the highest R-squared isn't the best approach. For an overview of identifying the best model, I'd read my post about choosing the correct regression model. Additionally, evaluating the model mainly by choosing the one with the highest R-squared is a form of data dredging. Click that link to understand the problems associated with focusing on that approach.

Additionally, my Regression Analysis book contains an entire chapter about using regression analysis to make predictions. You might check it out!

Reply

François says
April 7, 2021 at 5:03 am

Thank you so much for this very clear explanation of the concept and applications!

Reply

X

February 28, 2021 at 8:17 am

Hello Sir, This is one of the best useful articles on regression I found till now.Thank you so much posting

Reply

Terika Z Ray says
February 11, 2021 at 7:23 pm

Hi Jim

I am relatively new to statistics. I am working on a research project to analyze financial charts and determine a relationship between two variables. I set the criteria for the stocks to weed out anomalies, like bear markets. Anyway I collected 34 data points so far and I am trying to work my way up to 50. My questions are:
1. How do I know how many data points to collect to represent an accurate model?
2. I added 3 data points and my R squared adj. value increased from 74.5% to 75%, but my RSME decreased from .98% to .71%, and my MAPE increased from 18.67% to 22%. What does this mean?
3. I was going to look at all the outliers (via the IQR method) after I collect 50 data points, then group the outliers in a separate group to model separately. These are the

X

< 10% is ideal.

Reply

Sergio says
January 19, 2021 at 10:47 am

Hi Jim,

awesome article, thank you so much!
Only thing is, I still don't understand what you exactly mean with 'terms' in broader sense. Do you have an example of such a term?

Reply

Jim Frost says
January 19, 2021 at 3:22 pm

Hi Sergio,

Often people talk about adding independent variables (IVs) to a model. However, I'll see in the equation. For example, the model below has three terms. The first two are IVs and the third term is the interaction between them:

a + b + a*b

I hope that clarifies what I mean!

Reply

Joseph Lombardi says

December 26, 2020 at 2:48 pm

Jim, thanks for the reply, and that was perfectly clear. You have given me new stuff to consider. Now, I have a rational basis to do so.

As it turns out, the added IV didn't move the "goodness-of-fit" needle very much. The improvement to the Standard Error was negligible, and the other coefficients didn't change much. Moreover, the sign of the added IV was negative, which in the context of my system makes no sense. I realize analysis can produce counter-intuitive results, sometimes, but a negative sign here is almost silly. And guess what: When I regressed ONLY that predictor against the DV, the correlation was positive, so I am pretty sure I need to leave it out of my model!

Reply

Jim Frost says

December 27, 2020 at 1:36 am

Hi Joe,

Thanks for the follow-up message. It's always interesting to hear about these stories in progress!

It sure doesn't sound like it's worthwhile including. Best case scenario, it's just not improving the model. A simpler model that provides a very similar goodness-of-fit is usually a good thing.

Yeah, that negative sign is big, flashing warning sign! It's hard to say exactly what's going on but something isn't right.

Thanks for writing!

Reply

X

Jim, I could have posted this under any of your blog posts about R-Squared, Standard Error, Data Mining, etc., but here seems appropriate.

I have stopped chasing high R-Squared. (I used to do that mostly by using polynomials of varying degrees when there was no theoretical basis to do so!) Then I would add IVs willy-nilly, which ALWAYS increases R-Squared. Now, I concentrate mainly on the SE of the regression. I think the beauty of SE is that it's in the same units as the DV. My "customers" can get their brains around that. If they look at the SE and see a 95% Confidence Interval too large to be useful, they know it intuitively, which is good.

But here's something that makes me scratch my head. Maybe you can help me.

If I add an IV to my model, the R-Squared necessarily goes up, as you have stated, but the SE might not. But what if the SE improves even though the IV that I added to the model proves not to be statistically significant (P-value over 0.05)? Did my model really "improve"? Notwithstanding the improved SE, should I drop that newly added IV because it's not Stat Sig?

It's worth mentioning that when I look at the scatter plot of the residuals, I see a random distribution of dots. (Average is zero, and the slope of the line through the dots is zero, too.)

What else should I be considering?

Reply

X

Hi Joseph,

Thanks so much for writing with the great observations! And it's always great to hear when someone isn't chasing a high R-squared. That's so tempting but yet potentially causes problems. And it's always great to chat with another fan of the standard error of the regression! So much good stuff in your comment! 🙂

Here's what's going on. First, the standard error of the regression uses the adjusted mean square error in its calculations. This adjusted means square error is the same used for adjusted R-squared. So, both the adjusted R-squared and standard error of the regression use the same adjustment for the DF your model uses. And when you add a predictor to the model, it's not guaranteed that either measure (adj. R-sq or S) will improve. Of course, R-squared will go up by at least a little.

Here's what's going on about when these measures go up but the predictor is not significant. Both adjusted R-squared and S will improve when the absolute value of the t-value for the predictor is greater than or equal to 1. Depending on your DF, t needs to have an absolute value of approximately 1.96 to be significant. So, you have this t-value range from 1 – ~1.96 where the predictor is not significant but those two measures will improve.

So, that's what's going. Now, what to do about it! This is a case where theory should be your guide. If the variable itself, along with the sign and magnitude of its coefficient makes theoretical sense, you might leave it in. If you're uncertain, it's generally better to include an unnecessary variable than to remove a variable that overfitting the model or going hog wild adding variables. If you really want to get into the weeds, you can look at the CIs for the coefficient estimates of the other variables with and without the non-significant variable. If those CIs are widening, you might exclude the non-significant variable. On the other hand, if the CIs don't change or even improvement, it's ok to include.

Also, consider the magnitude of the improvement of the goodness-of-fit measures. If the improvement is small, and the other coefficients don't change much, and your residuals look good without the extra variable, you're probably fine leaving it out. On the other hand, if the other coefficients change notably, then you have to worry about the possibility of omitted variable bias.

That's probably not entirely clear! But, it's a bit of a grey area. But consider the size of the improvement, the change in the coefficients and CIs of the coefficients for the other variables, and theoretical issues. Theoretical issues can override the other statistical issues when you have solid theoretical reasons for including a variable or not. In my regression analysis book, which you have, the beginning portion of chapter 7 (Specify Your Model) has some tips for what to consider.

I hope that helps!

Reply

X

Hello dear in my python calculation of forecasting R-squared is 97.56% but my examiners asks me to justify it. How can I justify it please any one who have idea help me?

Reply

Jana Aksamitova says
November 26, 2020 at 6:16 pm

Hello Jim,

I came across your article and your comments regarding regressions.
I wanted to ask if you would be able to explain something for me – I am lost with this.

I have run a hierarchical multiple regression – and am comparing model one and model two. I understand your comment regarding comparing models solely based on adjusted r squared may not be a good practice.
I am trying to however get my head around the following:

If you have a adjusted R squared value in model one with 3 predictors included that sits at .340 and then you add 4th predictor into the model – which is not model two. After adding the 4th predictor your adjusted R squared value drops to .337 – meaning there is about 0.3% variance. What does it mean? How do I interpret this? I understand if the change much when you compare model one and model 2.Some of the predictors are statistically significant, two of them are not – so I would understand that the ones that are statistically significant will influence the dependent variable more greatly than the ones that are not.

But back to the beginning – my adjusted squared is lower in the second model and the change is not statistically significant – how do i interpret this?

Reply

Lamessa says

November 19, 2020 at 3:41 pm

Hi my dear! My r square value is 0.397,Adjusted R square is 0.366 and SEE is 0.738 respectively.please suggest me how to inter prate them

Reply

Jim Frost says

November 19, 2020 at 11:31 pm

X

the dependent variable around its mean. Typically, you only interpret adjusted R-squared when you're comparing models with different numbers of predictors.

The SEE is the typical distance that observations fall from the predicted value. It measures the precision of the model's predictions. How wrong is the model typically? I write a post about how to interpret the SEE. In that post, I refer to it as the standard error of the regression, which is the same as the standard error or the estimate (SEE). Please read that post to see how to interpret it.

Reply

Laurie says

July 23, 2020 at 6:10 pm

Thank you. All my models have the exact same predictors, they are standardized test scores, but each model's scores are normed with a different a demographic variable/combination of demographic variables. I guess it makes sense that they are not wildly different in that aspect. I appreciate your perspective and will read up on the resources you suggested.

Reply

X

July 23, 2020 at 9:23 pm

The changes in demographic variables are still changes in the model. But, it does sound like they're not that much different. A good rule of thumb is to go with the simplest model if everything else is equal. So, if the R-squares are similar, and the residual plots are good for all them, then pick the simplest model of those. Then, proceed on with the incremental validity test for your variables of focus.

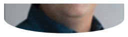
Reply

Laurie says
July 23, 2020 at 4:04 pm

Thank you. I also used Akaike information criterion to confirm the findings. My trouble is that I don't know how to make an argument that such a small difference would actually be of clinical or practical significance in the real world because I'm not a clinician and this is research on standardized testing. It was suggested by a colleague that I read up on Incremental validity. It is a step in the right direction. But if you have any other suggestions it would be beneficial.

Reply

X



July 23, 2020 at 6:03 pm

Hi Laurie,

I'd say that you can't make an argument that the differences between the models are meaningful based on R-squared values. Even if your R-squared values had a greater difference between them, it's not a good practice to evaluate models solely by goodness-of-fit measures, such as R-squared, Akaike, etc.

Also, if your models have different numbers of predictors, you should be looking at adjusted R-squared and not the regular R-squared. But, even then, it's just one factor to consider.

You do need to consider other factors, such as residual plots and theory. I talk about all those considerations in the post about model specification that I linked to previously. You won't be able to choose the best model from R-squared alone (even with Akaike IC).

Now, the question about whether your treatment is clinically significant is a different but related matter. And, that's where incremental validity comes in–sort of.

First, you need to choose the best model. That involves all the factors I mentioned. After picking your final model, you can test for incremental validity.

Incremental validity assesses how much R-squared increases when you add your treatment variable to the model last. There is an F-test to use that can determine if

I don’t write about the F-test for incremental validity specifically, however you can read about what you can learn from assessing the R-squared increased caused by adding the variable to the model last in my post about identifying the most important variable in a regression model. The only aspect I don’t touch on is that F-test. But, you’ll read about what you can learn from that approach.

Finally, the “sort of” part. This test for incremental validity determines whether the improvement caused by your treatment variable is statistically significant. However, there is a difference between statistical significance and practical significance. You can have something that is statistically significant but it won’t necessarily be practically/clinically significant in the real world. For practical significance, you need to evaluate the effect size. Read that article for more information about the process.

Reply

Laurie says
July 22, 2020 at 8:46 pm

I just ran my first regression analysis for my dissertation. All models were built from standardized test scores scored with 4 different variations of demographic correction

X

interpretations for differences that are so small? I can argue that theoretically one model is clearly superior, because it captures the most variance in the outcome variable, but do you have any resources that would help me interpret such small differences in effect size and how I can formulate clinical or practical significance when applied to testing the general population?

Reply

Jim Frost says

July 22, 2020 at 9:54 pm

Hi Laurie,

Evaluating the models by R-squared alone is not the recommended method. I've written an article about how to choose the best model, which looks at various approaches. Read that. If you have more questions afterwards, post them there!

Reply

BHABANI PRASAD MONDAL says

X

Reply

Jim Frost says

July 21, 2020 at 3:42 am

Hi, you need look at the overall F-test.

Reply

Katja says

July 14, 2020 at 12:44 am

Thank you for a very good summary. You might want to clarify that to get the percentage fit the R-squared needs to be multiplied by 100. It’s obvious in hindsight but I read the whole page and didn’t realise until I checked the wikipedia page and they specified that R-squared is always between 0 and 1. I was thinking my value was even lower than it actually is! (Very low)

Reply

X

Jung Han Lee says

July 1, 2020 at 10:04 am

Hello sir, thank you for the nice explanation.

Btw, what does ‘terms’ mean in ‘Are High R-squared Values Always Great?’ section? Such as polynomial terms, interactive terms..

Reply

Jim Frost says

July 1, 2020 at 10:56 pm

Hi Jung,

Yes, that’s exactly what I mean. “Term” is a broader word in this context. I didn’t want to say variable because that really limits it to just the independent variables. However, terms encompasses the independent variables as well as polynomial terms and interaction terms.

Reply

X

Renu says

June 6, 2020 at 12:32 pm

Does regression represent predictor relationship? If not what ?
How do we write discussion for linear regression?

Reply

Jim Frost says

June 8, 2020 at 3:39 pm

Hi Renu,

I have an article about that–when to use regression analysis. That'll explain what regression can tell you. If you have more specific questions after reading that article, please post them in the comments section there.

Reply

X

Hello, Sir. I just wanna ask when did you publish this. I will be citing some parts of this blog for our Predictive Statistics research paper. Thank you!

Reply

Jim Frost says

May 2, 2020 at 11:33 pm

I'm glad this article was helpful for you! When you cite web pages, you actually use the date you accessed the article. See this link at Purdue University for Electronic Citations. Look in the section for "A Page on a Website." Thanks!

Reply

Hamster says

April 6, 2020 at 10:43 am

Is it possible to fit a linear regression for which r2 is exactly equal to zero?

Reply

X

MAHESWAR SETHI says

April 3, 2020 at 11:41 am

If the R squared in a regression model is 80%. What does it mean? Does it mean that the 80% of total variation in the dependent variable is explained by the independent variables together? or the independent variables together are capable of explaining 80% of the variations caused by them only out of 100% variation caused by them?

Reply

Jim Frost says

April 5, 2020 at 7:08 pm

Your model collectively explains 80% of the variability of the dependent variable around it’s mean.

Reply

X

hallo thanks for your explanation. i want to make sure of my confused about R square, you write that
“For the same data set, higher R-squared values represent smaller differences between the observed data and the fitted values.”

i run my data thesis using STATA. i use XTGLS to estimate robust FE model (because when i use xtreg, fe ro, one of dummy variable is omitted). but when i use XTGLS instead of XTREG FE, there is no R2 appear (see https://www.stata.com/support/faqs/statistics/r-squared-after-xtgls/), but i need to add r2 to describe my research model.

and from this (https://www.researchgate.net/post/Does_anyone_know_about_goodness_of_fit_in_generalized_least_squares_estimation)
i find that we can use the correlation between observed value of DV and fitted value of DV.

i fine the correlation between my observed value of DV and fitted value of DV is 0.8027.

if u can give me an advise, can i tell that my model is good by looking the high correlation of observed value of DV and fitted value of DV? that show the smaller diferences between the observed data and the fitted values? mean the model have high R2?

thanks in advance

Reply

X

February 13, 2020 at 5:35 am

Hello Jim your post helped me very much thanks a lot! I'm currently working on my bachelors thesis where I examine the low beta anomaly on the cryptocurrency-market. I calculate the Betas via the dataanalysis tool on excel over one year and compare the quarter returns from high and low beta cryptos. As I continue from quarter to quarter with the beta calculation my R-squared gets smaller and smaller and even the significance (p-value) rises. How could I interpret my low R-squared? Could it be that behavioral biases of investors influence my results? I have 365 observations (daily) and my p-value went from 3.08E-23 for the first quarter to 1.4313E-08. My R-squared went from 0.23815 to 0.08463222. (For Bitshares)

Bitcoin for example started with an R-squared of 0.8059 and ended in the last quarter with 0.59338. P-value shrinks with lower R-squared aswell from 2.539E-131 to 4.042E-73. The Adjusted R-square only differs by 0.001 to maybe 0.01 in some cases (very small difference).

How could I interprete this results? Its obvious there are behavioral biases, could this lead to a shrinage of the R-squared? And could the heavy weight of 45% from the Bitcoin in comparison to other cryptos 1% to 10% to the big difference in R-sqauared, as the index CCi30 is highly dependent of the Bitcoin due to its weight?

It would help me a lot! 🙂 an thanks for the intuitive posts on this page!

Best regards
Kevin

Reply

X

Josh says

December 11, 2019 at 9:50 pm

Can I calculate predicted r-squared using excel?

Reply

Jim Frost says

December 13, 2019 at 10:27 am

Hi Josh,

Unfortunately, I don't believe that Excel calculates predicted R-squared out of the box.

Reply

Teryn Lee says

November 17, 2019 at 10:45 am

X

mean that I can assume that it is only natural to assume that profit will increase and with the outliers removed, the r^2 and adjusted r^2 then becomes smaller as I am now assuming the business doesn't make the sale hence not making the profit therefore the decrease in r^2 and adjusted r^2. Just like the reason in the article, "People are just harder to predict than things like physical processes." it is the same for businesses as sales are hard to predict as there may be external factors that might actually contribute to low sales?

Reply

Jim Frost says
November 17, 2019 at 10:42 pm

Hi Teryn,

If I understand correctly, you have two regression models where sales and profit are the dependent variables?

Typically, when you remove outliers, your model will fit the data better, which should increase your r-squared values. However, outliers are a bit more complicated in regression because you can have unusual X values and unusual Y values. They can a different effect on the model. I cover this in much more detail in my ebook about regression analysis. It's impossible to say exactly what impact your outliers are having with the limited information. You can fit the model with and

X

more from a hypothesis testing point of view, but the guidelines in general are still applicable. My regression ebook covers it in depth from a regression standpoint.

I haven't use regression to predict sales or profit, so I can't really say where it falls in terms of predictability. In fact, it might well vary from business to business. If there's literature you can review on the subject, that should provide some helpful information about what other businesses find.

I hope this helps!

Reply

G Love says

October 5, 2019 at 5:17 pm

Thanks for your reply. What I'm trying to intuit is whether variance explained as an estimator of risk explained (Korn & Simon, 1991) can be extended to: Risk Ratio ≈ R2total / R2base.

Reply

X

Hi, I’m not familiar with that article and will have to check it out. R-squared indicates the amount of variance around the mean of the dependent variable that the model explains. I suppose you can interpret unaccounted variance as a risk. It really affects the precision of the prediction. If imprecise predictions are a risk (which they can be), I suppose R-squared can represent that–although that’s not how it’s usually discussed.

And, if you need to understand the precision of your model’s predictions, the standard error the regression (S) and prediction intervals both assess that directly.

Reply

G Love says
October 5, 2019 at 1:30 pm

Suppose you calculate the R-squared for a linear model, regressing a response on a treatment and control variables (call this R2total). Then you calculate R-squared for an identical model with only the controls and not the treatment on the right-hand side (call this R2base).

I assume that you might reasonably say that the quantity R2treatment = R2total –

X

R2base) represent? What I'm looking for is a measure of (relative) risk.

Reply

Jim Frost says

October 5, 2019 at 4:04 pm

Hi, that difference between the R-squared for just the controls and the R-squared for the controls plus treatment is the percentage of variation for which the treatments uniquely account. R-squared is a relative measure of model precision and not directly linked to risk. You might be able to assess risk in binary logistic regression if you have a dependent variable that represents a condition you want to avoid and include the control and treatment variables. You can then assess the changes in probability associated with the condition associated with the treatment and non-treatment conditions.

Reply

ARUN SINGH says

September 9, 2019 at 1:24 am

linear regression models if the underlying data has a linear relationship.

Thanks

Reply

Gui says
September 2, 2019 at 5:01 pm

Gotcha! Thanks for clarifying, it was really helpful!
Best,
Gui

Reply

Gui says
August 31, 2019 at 2:38 pm

Hi Jim,

X

One thing about your answer to my second question wasn’t completely clear to me, though. You mentioned that “for the same dataset, as R-squared increases the other (MAPE/S) decreases”, and in your post “How High Does R-squared Need to Be?” you mentioned that “R2 is relevant in this context because it is a measure of the error. Lower R2 values correspond to models with more error”. I understand S’s value, specially in regards to the precision interval, but I also like MAPE because it offers a “dimension” of the error, meaning its proportion vs the observed value.

Having that in mind, I picture a model with a very high R-sqr as a model for which the predicted curve is a close fit to the observed values and, therefore, I expect MAPE to be small; a model with a very low R-sqr, on the other hand, would have predictions distant to the actual observations and therefore MAPE would be high. I guess in that sense that I would expect a negative correlation between R-sqr and MAPE.

Is that a fair assessment? Is it possible to have a case with say R-sqr = 90% and MAPE = 50%, or R-sqr = 15% and MAPE = 2% (these numbers don`t matter per se, they are merely to illustrate my point)?

Thanks once more for your help,
Gui

Reply

Jim Frost says

X

Yes, that's the general rule. High R-squared values tend to have lower MAPE and S values. They're all essentially measuring the error in different ways. A large relative amount of error will both decrease R-squared and increase MAPE and S.

I'd have to work through the math for you last question but my sense is that yes it's possible with the 90% R-squared example. It all depends on the magnitude of the total variation and how the unexplained portion of it relates to the magnitude of the residuals. I'm not sure if the 15% R-squared example is possible. Again, I'd have to check the math, which I don't have time at the moment. I don't use MAPE myself so I don't have those answers at hand. I do know that there are some unusual issues with it. Low predictions have an error that can't exceed 100% while high predictions don't have an upper boundary. Using MAPE to compare models will tend to choose the model that predicts too low. I suspect that the proportion of low to high predictions might feed into the unusual results you're asking about.

I hope this helps!

Reply

Guilherme says
August 30, 2019 at 1:45 pm

X

model. My question is not about when to use one or another but rather using them together and how to interpret the possible combinations of results.

True enough, there's no such thing as "the right number for R-squared/MAPE" or "R-squared should always be above X%" but, for the sake of simplicity, let's just consider a high R-squared (let's say maybe >80%) as a good thing, as well as a low MAPE (let's say maybe <10).

With these assumptions, I'd say that a model that has a high R-squared and low MAPE is a model that has a good fit, and a model that has a low R-squared and high MAPE is not a good one. Now, to my questions:

1. Is that a fair assessment?
2. How about the other two possible combinations, a low R-squared with a low MAPE and a high R-squared with a high MAPE, is it possible for them to occur? How to interpret those cases?

Thanks a mill, Gui

Reply

Jim Frost says
August 30, 2019 at 3:29 pm

X

I'm a big fan of the standard error of the regression (S), which is similar to MAPE. While R-squared is a relative measure of fit, S and MAPE are absolute measures. S and MAPE are calculated a bit differently (squared errors vs mean absolute errors) but get at the same idea of describing how wrong the model tends to be using the units of the dependent variable. Read my post about the standard error of the regression for more information about it.

I agree with the idea that higher R-squared values tend to (but not necessarily) represent a better fit and that lower MAPE/S values represent better fits. For the same dataset, as R-squared increases the other (MAPE/S) decreases. However, across different datasets, R-squared and MAPE/S won't necessarily follow in lockstep like that–but the tendency will be there.

It is possible to obtain what you define as a good R-squared but yet obtain a bad MAPE using your definition. Or vice versa. The issue is that there's not a direct mapping between these values that you can apply across different models generally. You might have a general concept of what is good for both measures, but the measures can disagree.

Analysts tend to use R-squared and MAPE/S in different contexts. R-squared tends to be used when you want to compare one study to another. It's easier to make the comparison across studies when they're looking at a similar research question but they might be using different outcome measures.

On the other hand, I see S (and I presume MAPE) used more often to determine values for the predictions to be useful. The researcher needs to define that acceptable margin of error using their subject area knowledge. I talk about this in my article about the standard error of the regression.

So, using that answer your second question, interpretation depends on which context you're using your model. If you want to compare your study to other similar studies and the R-squared is in a good range and you're not using the predictions to make decisions, it probably doesn't matter what MAPE/S are. Conversely, if R-squared is considered to low, it probably doesn't matter what MAPE/S are.

On the other hand, if you're using your model to make predictions and assessing the precision of those predictions, MAPE/S reign supreme. If the margin of error around the predictions are sufficiently small as measured by MAPE/S, your model is good regardless of the R-squared. Conversely, if the precision of the predictions (MAPE/S) are not sufficiently precise, your model is inadequate regardless of the R-squared.

It really depend what question you're trying to answer. Read my article about S for more information. Also, read my article How High Does R-squared Need to Be? This article addresses a similar question but focuses on the issue of defining your objectives so you can answer this question and choose between R-squared and S.

I hope this helps!

Reply

X

Omoleye Ojuri says

August 8, 2019 at 6:59 am

Thanks Jim for such a wonderful explanation. Please, am I to use the coefficients of the independent variables to plot the residual plots? Responses on this confusing issue will be appreciated.

Reply

Jim Frost says

August 8, 2019 at 3:54 pm

Hi Omoleye,

A residual is the difference between observed value and the fitted value for an observation. In other words, it's how wrong the model was for each observation. Residual plots will graph all the residuals for your dataset.

Statistical software should do this for you using a command. You should not have to calculate the fitted value for each observation and do the subtraction yourself. Your software should do this for you. Check the documentation for your software.

But, yes, the software plugs in the values of the independent variables for each obtain the residual. It repeats this process for all observations in your dataset and plots the residuals.

For more information, please see my post about residual plots.

Reply

Hitesh says
July 17, 2019 at 11:50 pm

Hello Jim… Thanks for such a wonderful explanation about R2..
I would like to know the references like book or journal which can give explain the limitations of R2 as you have explained.

Reply

Jim Frost says
July 18, 2019 at 12:49 am

Hi Hitesh,

Reply

Luca romen says

June 20, 2019 at 12:44 pm

Hello Jim thanks for your help can you talk pleas about discrete variables regression ? How does it work? In which is it different from normaL regression?

Reply

Jeff says

May 24, 2019 at 7:11 pm

Thank you for all Jim, you can explain difficult concept so easley

Reply

X

Thanks Jim for the article.

I am wondering if there is any way in stat to enhance R2s in the linear regression?

Reply

Jim Frost says
May 14, 2019 at 2:59 pm

Hi, I'm not sure what you mean by "enhance?" If you're asking how to increase R-squared, you can do that by adding independent variables to your model, properly modeling curvature, and considering interaction terms where appropriate.

However, be sure that if you take any of these actions you're doing so because they are appropriate variables to add given subject-area knowledge and theory. Don't simply make your model more complicate to chase a higher R-squared. For any given study area, there's an inherent amount of unexplainable uncertainty, which represent a ceiling for R-squared. As I mention in this post, you can push past the ceiling but at he risk of producing results that you can't trust!

And, remember that a lower R-squared isn't necessarily bad!

Reply

X

Thomas says

May 13, 2019 at 12:23 pm

Hello, thanks for the great article! There is a concept I can't wrap my head around for some reason and I'm hoping you can shed some light!
What do we mean when we say that a model "explains" a percentage of a dependent variable's variation? The variation of the variable y is the squared differences of the actual observations from their mean ! Now, I think that the word explain is used metaphorically but still Im not exactly sure what it actually means! Thanks in advance Jim and thanks again for the article

Reply

Jim Frost says

May 14, 2019 at 1:03 am

Hi Thomas,

That's a great question. I actually answer it in my ebook. You can get the free sample version which has the complete first two chapters. The reason I mention that is because I talk about this issue in the 2nd chapter, which is included in the free sample book. It's a free download and you don't need a credit card. If you get

X

I personally prefer saying the model "accounts" for variability. It might be a semantics issue, but to me "explains" implies more a casual relationship. You can actually have a model with proxy variables that are only indirectly correlated with the dependent variable and "explains" doesn't sound right for that situation either! Read that section of the ebook and if you still have questions, let me know!

I hope this helps!

Reply

Dana says

April 20, 2019 at 3:05 pm

In a hierarchical regression, would R2 change for, say, the third predictor, tell us the percentage of variance that that predictor is reponsible for? I seem to have things that way for some reason but I'm unsure where I got that from or if it was a mistake.

Reply

Jim Frost says

X

In general, I find that determining how much R-squared changes when you add a predictor to the model last can be very meaningful. It indicates the amount of variance that a variable accounts for uniquely. You can read more about it in my post about identifying the most important variables in a model.

Reply

Takunda says

April 7, 2019 at 1:54 pm

Hi Jim,

I used random effects model to perform my regression analysis. Would r squared and adjusted r squared serve as appropriate reliability tests. I have all the results from Stata

Reply

Jim Frost says

April 7, 2019 at 6:35 pm

X

discussion thread, which might be helpful.

Reply

X

Jim says

April 6, 2019 at 11:33 pm

Wow Jim, thank you so much for this article, I've been banging my head against the wall for a while now watching every youtube video I could find trying to understand this. I finally actually feel like I can relate a lot of what you've said to my own regression analysis, which is huge for me…… thank you so much.

Reply

Jim Frost says

April 6, 2019 at 11:46 pm

You're very welcome! Your comment really makes my day because I strive to make statistics more relatable. Because you're using regression analysis, you might consider my ebook about regression analysis.

Reply

X

Very interesting discussion. And I actually understood most of it! I was using Microsoft Excel to chart some daily variances we are experiencing in our fuel storage tanks. (we operate gas bars & convenience stores).
I see that we are experiencing day to day variances (both gains and losses), but I wanted to graph these variances, and run a trend line, to see if we were losing or gaining fuel – over time. Excel has a few options for trend lines (linear, logarthimetic & polynomial). Based on your discussion, I used the option with the highest R-squared value, thinking it would be the best predictor. However, all the trend line options had extremely low R-square values…ranging from .5% to 3%. I thought perhaps my data variances were too extreme to allow for a predictive trend line. I was curious as to what a high r-square trend line might look like, so I created a "mock" table of data, covering 30 days, and used numbers that were in a fairly tight range (95 to 105). I graphed this range of data & ran the trend line. I expected the R-square value to be close to 100% – but its only at 10%.
Long story short – I can't figure out why my R-square values are so low!

Reply

Jim Frost says

March 19, 2019 at 2:36 pm

Hi Greg,

X

It might be that your variances aren't related to time. Or, perhaps they are but your data don't cover enough time to capture it. In other words, your predictor (time it sounds like) just aren't explaining the variances.

Make sure that your trend line follows the data. You can get a lower R-squared when your model isn't fitting the data.

You also want to check for something called heteroscedasticity. If you're measuring variances using plus and minus values, and the absolute value of the variances increases over time, you could see a flat trend but increases in the spread of values around the fitted line over time. You'd see a cone shape in your data. You can also see that in a residuals plot. You can use the search box on my website to find my post about heteroscedasticity if you see that fan/cone shape in the graph of your data over time. That could happen if both plus and minus variances grew in magnitude over time.

Those are the types of things I'd look into. You could also try other variables instead/in addition to time such as temperature. Anything that might be related to the variances. Maybe it's related to other conditions rather than the passage of time. Or maybe you need a longer time frame for the time effect to reveal itself?

Reply

X

Indeed a clear and precise way to understand the concept of R square.I was bit worried as my R square value was coming .037.Thanks for your assistance over Multiple regression and its related parameters. I Would like to get benefited more from coming online study materials on statistics.

Thanks & Regards

Reply

Angie says

November 23, 2018 at 8:16 pm

Very easy to follow. Glad I found your site.

Reply

X

Charles says

November 13, 2018 at 7:29 am

Thanks Jim, i wonder whether you youtube videos. Second, i would to see an explanation of how to reshape data to have it, in a time to event nature, in STATA.

Reply

Luyando says

November 10, 2018 at 5:59 pm

thank you

Reply

Alexandros says

October 19, 2018 at 5:21 am

X

R-squared Values Always a Problem”.

I am writing a report concerning my research (field of asphalt properties) and I’m experiencing lower R square (from 0.21 to 0.469 for different models). In my case, I really believe that asphalt can be as complex as a human and therefore when you try to fit properties in a regression model the interpretation of the result can be similar to the case you give as an example concerning human behavior.
Again thank you for sharing this article and looking forward to your reply!
Cheers!

Reply

Jim Frost says
October 19, 2018 at 11:07 am

Hi Alexandros,

I don’t have a specific reference for that issue about low R-squared values not always being a problem other than it is based on the equations and accepted properties of R-squared that you’ll find in any regression/linear model text book.

A couple of thoughts for you. One, if you haven’t read it already, you should probably read my post about how to interpret regression models with low R-squared values and significant independent variables. It sounds like this situation

X

difficult to specify the correct model and/or obtain all the necessary data. On the other hand, I think you can probably argue that you can have a simple subject area that is hard to predict. For example, rolling a fair die, you can only predict the outcome accurately 1/6 of the time! Of course, you can be in a subject area that is both complex and unpredictable!

I mention this distinction because you'll need to determine whether your subject area is predictable rather than just the complexity. I don't know your field so I can't answer that but typically physical properties are more predictable than human behavior.

The practical aspect you need to determine is whether your R-squared is low because it's inherently unpredictable or because you're not including an important variable, modeling curvature, modeling an interaction, or possibly using imprecise measurements? If it's inherently unpredictable, then you've hit a brick wall that you can't legitimately get passed. However, if it's one of the other issues, you can legitimately improve your model. The trick is to determine which case you fall under!

I hope this helps at least a bit!

Reply

X

Hi Jim,

Your works are amazing. It’s very clear and useful. Please keep up the good work.

Reply

Jim Frost says

October 12, 2018 at 1:57 pm

Thank you so much for your kind words, Kesinee! It’s great motivation to keep going! 🙂

Reply

Narasimha murthy says

October 1, 2018 at 5:49 am

Hi Jim,

Excellent articulation and the language is simple..I enjoy reading your blog.

X

Jim Frost says

October 1, 2018 at 9:35 pm

Thank you very much, Narasimha!

Reply

Nic says

September 6, 2018 at 12:12 am

Hi Jim,

What a fantastic, clear article explaining the ideas behind R-squared;

Reply

Jim Frost says

September 6, 2018 at 12:37 am

X

the similarity.

Reply

Mahogany Hartley says
September 5, 2018 at 5:15 pm

What would be considered a low R2

Reply

Jim Frost says
September 6, 2018 at 12:43 am

Hi Mahogany,

That depends on the subject matter. If you are working in the physical sciences and has a low noise, predictable process, then an R-squared of 60% would be considered to be extremely low and represent some sort of problem with the study. However, if you’re predicting human behavior, the same R-squared would be very

I hope this helps!

Reply

Don says

August 26, 2018 at 4:18 am

Sir I would like to ask, if I acquire is R^2 = .1027 (10.27%) can I assumed that the dependent and indepent variable are inversely proportional to each other?

Reply

Jim Frost says

August 26, 2018 at 11:12 pm

Hi Don,

From the R-squared value, you can’t determine the direction of the relationship(s) between the independent variable(s) and dependent variable. The 10% value indicates that the relationship between your independent variable and dependent

X

post about understanding correlation.

Reply

Ajay verma says
May 12, 2018 at 10:25 am

Hello Sir,
In some situation adjusted R square may be negative then how we interpret them?

Reply

Jim Frost says
May 13, 2018 at 3:24 pm

Hi Ajay,

Yes, it’s entirely possible for adjusted R-squared (and predicted R-squared) to be negative. Some statistical software will report a 0% for these cases while other software returns the negative value a negative value is even worse, but that doesn't change what you'd do. If you have a zero value (or negative), you know that your model is unusable. The next step is to check the regular R-squared. Is it much higher? If so, your problem might be only that you're including too many independent variables and you need to use a simpler model. However, if the regular R-squared is similarly low, then you know that your model just isn't explaining much of the variance.

The only additional information that you might glean from a negative value, as opposed to a small or zero value, is that there is a higher probability that you're working with a small sample and including too many variables. If you have a large sample size, it's harder to get a negative value even when your model doesn't explain much of the variance. So, if you obtain a negative value, be aware that you are probably working with a particularly small sample, which severely limits the degree of complexity for your model that will yield valid results.

I hope this helps!

Reply

Kamala says

April 25, 2018 at 10:08 pm

X

I will go through your reference for the low R-squared values and get back to you.

Thank You,
Kamala.

Reply

Kamala says
April 19, 2018 at 2:45 am

:)….
am just seeing the relationship between variance and regression…is it so that for more variance does the data points are closer to the regression line??? I doubt on this statement….
Can you explain ….

Regards
Kamala

Reply

Jim Frost says

X

R-squared measures the amount of variance around the fitted values. If you have a simple regression model with one independent variable and create a fitted line plot, it measures the amount of variance around the fitted line. The lower the variance around the fitted values, the higher the R-squared. Another way to think about it is that it measures the strength of the relationship between the set of independent variables and the dependent variable. Either way, the closer the observed values are to the fitted values for a given dataset, the higher the R-squared.

I've written a couple of other posts that illustrate this concept in action. In this post about low R-squared values, I compare models with high and low R-squared values to show how they're different. And, in this post about correlation, I show how the variance around a line that indicates the strength of the relationship.

I hope this helps!

Reply

Kamala says

April 18, 2018 at 9:17 pm

X

Jim Frost says
April 19, 2018 at 2:12 am

Hi Kamala, thanks so much! Your kind comments made my day!

Reply

Qayoom Khachoo says
March 22, 2018 at 11:33 am

Hi sir thank you very much for the informative post. It would be more enriching if you could kindly highlight the causes of low R-square in panel data. I'm struggling to defend my model because of low R-square value. Thank you

Reply

Jim Frost says
March 22, 2018 at 11:43 am

think you'll find the concepts helpful. Best of luck with your model!

Reply

Akhilesh Gupta says
March 11, 2018 at 6:07 am

Sir how can i calculate R-square for time series models and how to interpret that R-square

Reply

Miteya says
March 3, 2018 at 2:17 pm

Hello Sir, Thank you for the data. Can you please suggest some methods like the R-square method to compare the results I get by using R square methods.

X

March 3, 2018 at 9:59 pm

Hi Miteya, R-squared is an example of a goodness-of-fit statistic. There are other goodness-of-fit statistics that you can use. I have written about some of them. In the last section of this post, look for and click the links for two posts that are about:
adjusted and predicted R-squared, and
standard error of the regression.

These three statistics all assess the goodness-of-fit, like R-squared, but they are different.

The Akaike information criterion (AIC) is another goodness-of-fit statistic. I haven’t written about that one yet, but you can search for it.

I hope this helps!

Reply

Dharmendra Dubey says
February 24, 2018 at 1:36 pm

X

Jim Frost says

February 24, 2018 at 3:46 pm

You’re very welcome! I’m glad you found it to be helpful!

Reply

Comments and Questions

X

Meet Jim

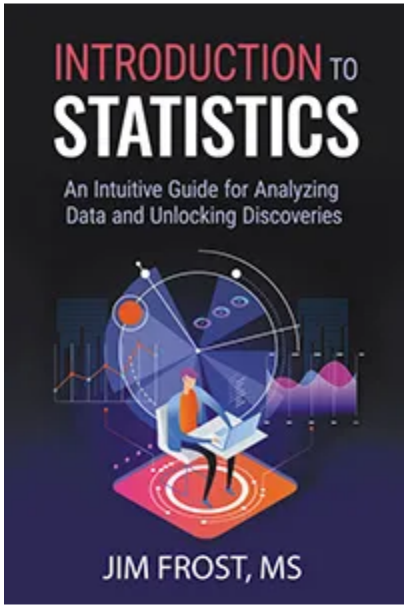
I'll help you intuitively understand statistics by focusing on concepts and using plain English so you can concentrate on

X

Search this website

X

Buy My Introduction to Statistics Book!

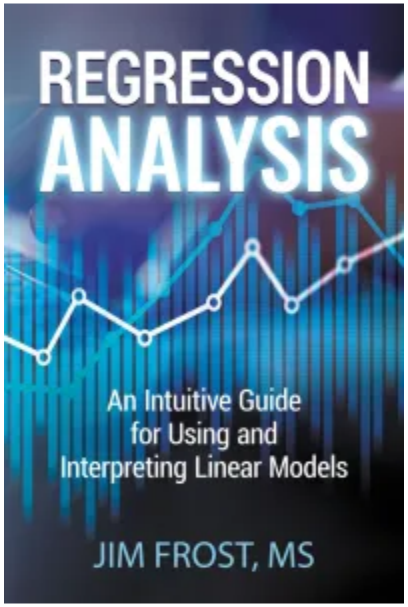

INTRODUCTION TO
STATISTICS
An Intuitive Guide for Analyzing Data and Unlocking Discoveries
JIM FROST, MS

Buy My Hypothesis Testing Book!

X

X

X



Subscribe by Email

Enter your email address to receive notifications of new posts by email.

X

First Name

Subscribe

I won't send you spam. Unsubscribe at any time.

X

X

How to Interpret P-values and Coefficients in Regression Analysis

How To Interpret R-squared in Regression Analysis

Weighted Average: Formula & Calculation Examples

Multicollinearity in Regression Analysis: Problems, Detection, and Solutions

How to do t-Tests in Excel

Cronbach’s Alpha: Definition, Calculations & Example

Choosing the Correct Type of Regression Analysis

Mean, Median, and Mode: Measures of Central Tendency

F-table

X

Wilcoxon Signed Rank Test Explained

What is P Hacking: Methods & Best Practices

Likert Scale: Survey Use & Examples

Correlation Coefficient Formula Walkthrough

Two-Way Table Explained

Kruskal Wallis Test Explained

Recent Comments

Matt on Qualitative Research: Goals, Methods & Benefits

Kerry on Populations, Parameters, and Samples in Inferential Statistics

Yu Gao on ANCOVA: Uses, Assumptions & Example

X

Jim Frost on Multivariate ANOVA (MANOVA) Benefits and When to Use It

X

X

Copyright © 2023 · Jim Frost · Privacy Policy

X