May 30, 2013 by Jim Frost in Regression Analysis

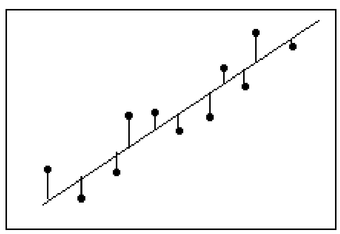
Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?

After you have fit a linear model using regression analysis, ANOVA, or design of experiments (DOE), you need to determine how well the model fits the data. To help you out, **Minitab statistical software** presents a variety of goodness-of-fit statistics. In this post, we'll explore the R-squared (R^2) statistic, some of its limitations, and uncover some surprises along the way. For instance, low R-squared values are not always bad and high R-squared values are not always good!

What Is Goodness-of-Fit for a Linear Model?

Linear regression calculates an equation that minimizes the distance between the fitted line and all of the data points. Technically, ordinary least squares (OLS) regression minimizes the sum of the squared residuals.

In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.



Definition: Residual = Observed value - Fitted value

Before you look at the statistical measures for goodness-of-fit, you should **check the residual plots**. Residual plots can reveal unwanted residual patterns that indicate biased results more effectively than numbers. When your residual plots pass muster, you can trust your numerical results and check the goodness-of-fit statistics.

What Is R-squared?

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

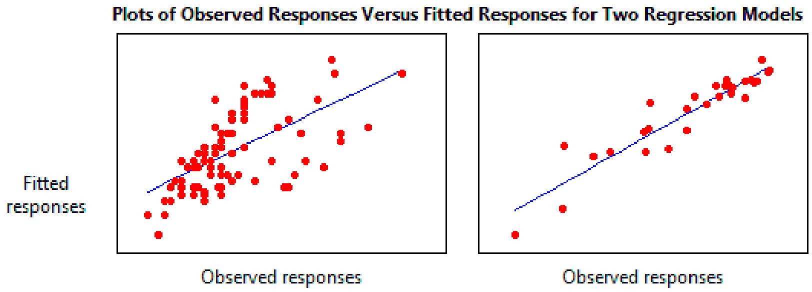
- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

In general, the higher the R-squared, the better the model fits your data. However, there are important conditions for this guideline that I'll talk about both in this post and my next post.

Recent | Authors | Categories

Graphical Representation of R-squared

Plotting fitted values by observed values graphically illustrates different R-squared values for regression models.



The regression model on the left accounts for 38.0% of the variance while the one on the right accounts for 87.4%. The more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line. Theoretically, if a model could explain 100% of the variance, the fitted values would always equal the observed values and, therefore, all the data points would fall on the fitted regression line.

A Key Limitation of R-squared

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

R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!

Are Low R-squared Values Inherently Bad?

No! There are two major reasons why it can be just fine to have low R-squared values.

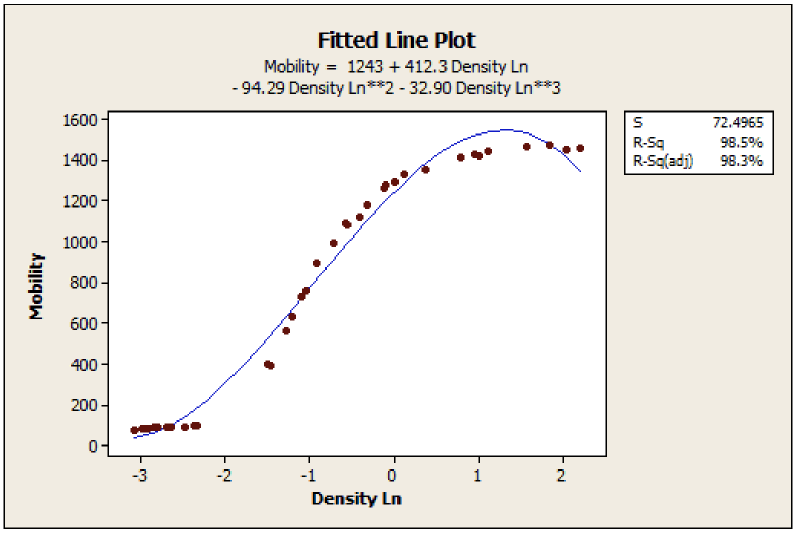
In some fields, it is entirely expected that your R-squared values will be low. For example, any field that attempts to predict human behavior, such as psychology, typically has R-squared values lower than 50%. Humans are simply harder to predict than, say, physical processes.

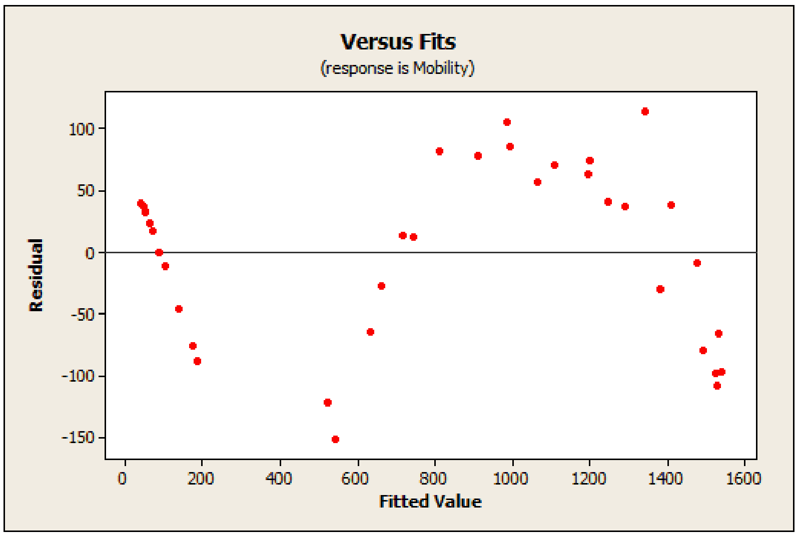
Furthermore, if your R-squared value is low but you have statistically significant predictors, you can still draw important conclusions about how changes in the predictor values are associated with changes in the response value. Regardless of the R-squared, the significant coefficients still represent the mean change in the response for one unit of change in the predictor while holding other predictors in the model constant. Obviously, this type of information can be extremely valuable.

A low R-squared is most problematic when you want to produce predictions that are reasonably precise (have a small enough **prediction interval**). How high should the R-squared be for prediction? Well, that depends on your requirements for the width of a prediction interval and how much variability is present in your data. While a high R-squared is required for precise predictions, it's not sufficient by itself, as we shall see.

Are High R-squared Values Inherently Good?

No! A high R-squared does not necessarily indicate that the model has a good fit. That might be a surprise, but look at the fitted line plot and residual plot below. The fitted line plot displays the relationship between semiconductor electron mobility and the natural log of the density for real experimental data.





The fitted line plot shows that these data follow a nice tight function and the R-squared is 98.5%, which sounds great. However, look closer to see how the regression line systematically over and under-predicts the data (bias) at different points along the curve. You can also see patterns in the Residuals versus Fits plot, rather than the randomness that you want to see. This indicates a bad fit, and serves as a reminder as to why you should always check the residual plots.

This example comes from my post about choosing between **linear and nonlinear regression**. In this case, the answer is to use nonlinear regression because linear models are unable to fit the specific curve that these data follow.

However, similar biases can occur when your linear model is missing important predictors, polynomial terms, and interaction terms. Statisticians call this specification bias, and it is caused by an underspecified model. For this type of bias, you can fix the residuals by adding the proper terms to the model.

Closing Thoughts on R-squared

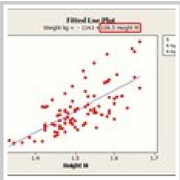
R-squared is a handy, seemingly intuitive measure of how well your linear model fits a set of observations. However, as we saw, R-squared doesn't tell us the entire story. You should evaluate R-squared values in conjunction with residual plots, other model statistics, and subject area knowledge in order to round out the picture (pardon the pun).

In my next blog, we'll continue with the theme that R-squared by itself is incomplete and look at two other types of R-squared: **adjusted R-squared and predicted R-squared**. These two measures overcome specific problems in order to provide additional information by which you can evaluate your regression model's explanatory power.

For more about R-squared, learn the answer to this eternal question: **How high should R-squared be?**

If you're learning about regression, read my **regression tutorial**!

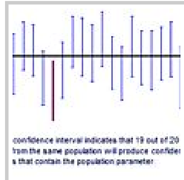
You might like:



How to Interpret Regression Analysis Results: P-values and Coefficients - Adventures in Statistics | Minitab

R-Squared: Sometimes, a Square is just a Square - Statistics and Quality Data Analysis | Minitab

How High Should R-squared Be in Regression Analysis? - Adventures in Statistics | Minitab



When Should I Use Confidence Intervals, Prediction Intervals, and Tolerance Intervals - Adventures in Statistics | Minitab

<< **Prev** **Next** >>

Comments for *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?*

Name: Alex
Time: Thursday, May 30, 2013

Hi, Jim! Nice post. What about lack-of-fit and pure error? There's one more usefull topic on the spot:
http://www.minitab.com/uploadedFiles/Shared_Resources/Documents/Articles/r2_misconceptions.pdf

---

Name: Jim Frost
Time: Friday, May 31, 2013

Hi Alex, thanks for the nice comment! I'm glad you enjoyed it.

Perhaps I'll cover lack-of-fit and pure error in a future post. And, I agree that is helpful topic written by some fellow Minitabers!

Jim

Name: Fawaz
Time: Thursday, July 25, 2013

Could you guide me to a statistics textbook or reference where I can find more explanation on how R-squared have different acceptable values for different study fields (e.g. social vs. engineering).

Thanks,
Fawaz

Name: Edgar de Paz
Time: Tuesday, October 1, 2013

THANK YOU!!!! I have had this question (Are Low R-squared Values Inherently Bad?) in my mind for a while...Working on a manufacturing project where human behavior have significant contribution; I see these typical low R-squared values, however I have significant contributions from some of my predictors (decent residuals). Aiming creating guidelines for standard work based on insight. Great article. Any bibliography that you can mention on this topic (low R-sq)?

Name: Jim Frost
Time: Wednesday, October 2, 2013

Hi Edgar, thanks for reading and I'm glad you found it helpful.

Unfortunately, I don't have a bibliography handy. However, the importantance of R-squared really depends on your field and what you want to do with your model.

If you just want to know what predictors are significant and how they relate to the response, then the coefficients and p-values are more important. A one unit increase in X is related to an average change in the response regardless of the R-squared value.

However, if you plan to use the model to make predictions for decision-making purposes, a higher R-squared is important (but not sufficient by itself). The biggest practical drawback of a lower R-squared value are less precise predictions (wider prediction intervals).

Keep in mind that a prediction is the mean response value given the inputs. You need to keep the variability around that mean in mind when using the model to make decisions.

This topic happens to be the subject of my next blog! That'll be out on October 3, 2013.

Jim

Name: Rafael
Time: Monday, December 16, 2013

Great Post, thank you for it. I'm trying to modeling a credit flow from a government bank that have political influence! Right now I'm trying to find texts like yours to show that R-square are not always above 80% in good models!

By the way, if you can sugest other texts that talks about that, I'd appreciate.

Thank you again for the info!

Name: Ruth
Time: Thursday, December 19, 2013

Thank you so much! I'm busy interpreting my results of my MA Psychology thesis and panicked when my R squared value was only 9.1%, despite all my predictors making significant contributions. Need an academic reference though (my university isn't keen on website references) so if you have any, that would be great!
Thanks again!

Name: tingting
Time: Monday, January 13, 2014

nice tutorial, really good for starters like me:P

Thank you so much, please carry on your great job.

Leave a comment

Name
Email
Your comment (No HTML)

Type the text from the image below

Submit