Minitab Software for Quality Improvement.

Contact Us

Blog Home | Data Analysis | Quality Improvement | Project Tools | Minitab.com

Search blog...

The Minitab Blog

2 0 19 48

February 17, 2014 by Patrick Runkel in Regression Analysis

R-Squared: Sometimes, a Square is just a Square

If you regularly perform regression analysis, you know that R^2 is a statistic used to evaluate the fit of your model. You may even know the standard definition of R^2: *the percentage of variation in the response that is explained by the model.*

Fair enough. With **Minitab Statistical Software** doing all the heavy lifting to calculate your R^2 values, that may be all you ever need to know.

But if you're like me, you like to crack things open to see what's inside. Understanding the essential nature of a statistic helps you demystify it and interpret it more accurately.

R-squared: Where Geometry Meets Statistics

So where *does* this mysterious R-squared value come from? To find the formula in Minitab, choose **Help > Methods and Formulas**. Click **General statistics > Regression > Regression > R-sq**.

$$r^2 = 1 - \frac{\text{SS Error}}{\text{SS Total}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Some spooky, wacky-looking symbols in there. Statisticians use those to make your knees knock together.

But all the formula really says is: "R-squared is a bunch of squares added together, divided by another bunch of squares added together, subtracted from 1."

$$r^2 = 1 - \frac{\text{SS Error}}{\text{SS Total}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2 \leftarrow \text{One bunch of squares}}{\sum(y_i - \bar{y})^2 \leftarrow \text{Another bunch of squares}}$$

What bunch of squares, you ask?

Recent | Authors | Categories

Revisiting the Relationship between Rushing and NFL Wins with Binary Fitted Line Plots

Gauging Gage Part 3: How to Sample Parts

Gauging Gage Part 2: Are 3 Operators or 2 Replicates Enough?

Gauging Gage Part 1: Is 10 Parts Enough?

Is Your Statistical Software FDA Validated for Medical Devices or Pharmaceuticals?

Unleash the Power of Linear Models with Minitab 17

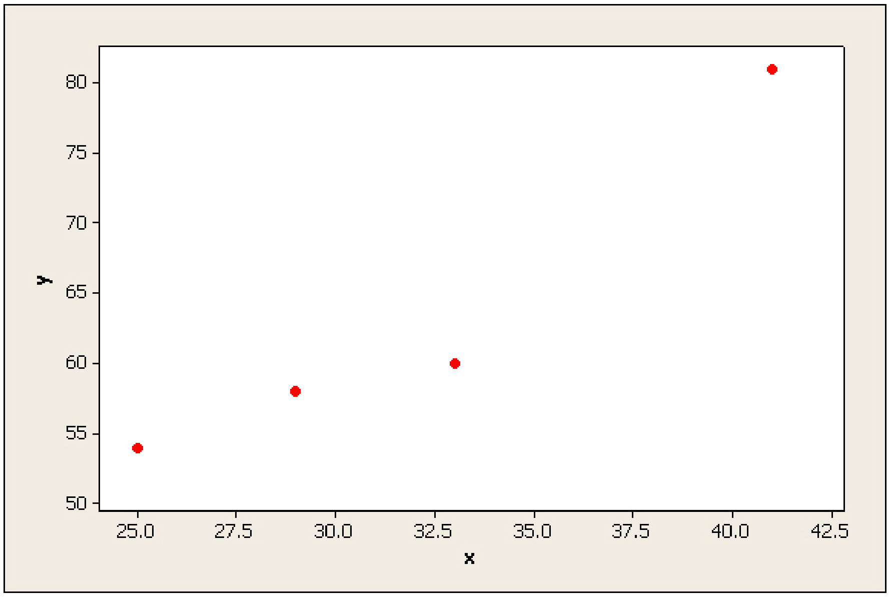
Histograms are Even Easier to Compare in Minitab 17

(We Just Got Rid of) Three Reasons to Fear Data Analysis

No, not them.

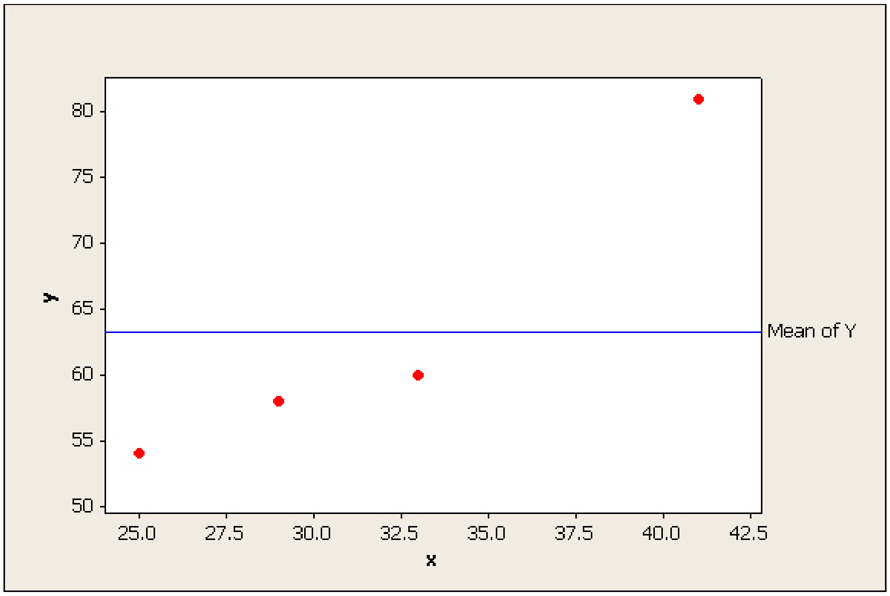
SS Total: Total Sum of Squares

First consider the "bunch of squares" on the bottom of the fraction. Suppose your data is shown on the scatterplot below:



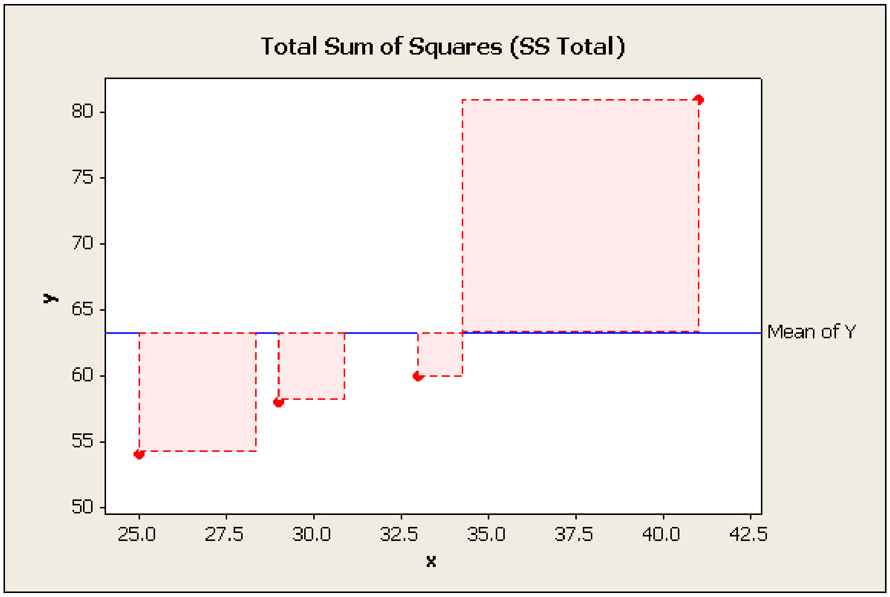
(Only 4 data values are shown to keep the example simple. Hopefully you have more data than this for your actual regression analysis!)

Now suppose you add a line to show the mean (average) of all your data points:



The line y = mean of Y is sometimes referred to the "trivial model" because it doesn't contain any predictor (X) variables, just a constant. How well would this line model your data points?

One way to quantify this is to measure the vertical distance from the line to each data point. That tells you how much the line "misses" each data point. This distance can be used to construct the sides of a square on each data point.



If you add up the pink areas of all those squares for all your data points you get the total sum of squares (SS Total), the bottom of the fraction.

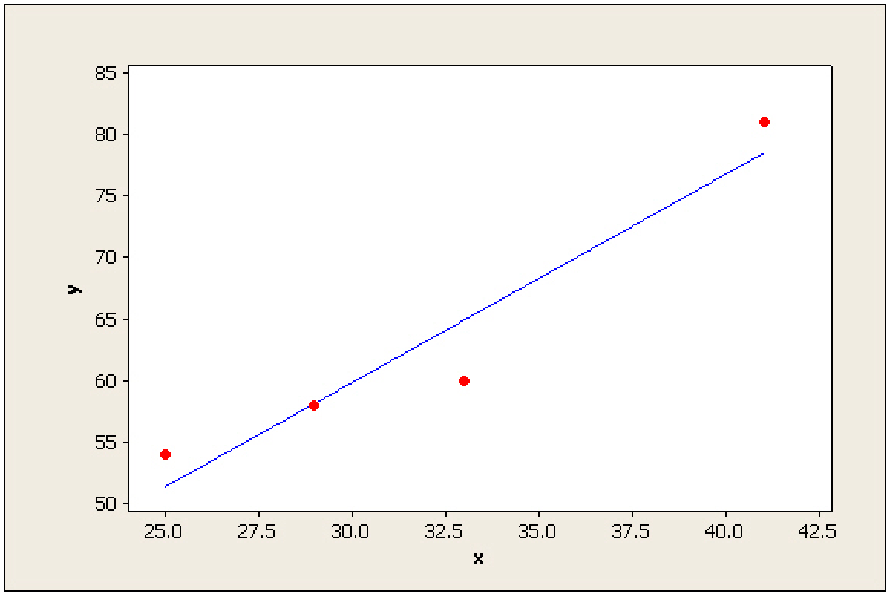
$$r^2 = 1 - \frac{SS\ Error}{SS\ Total} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\boxed{\sum(y_i - \bar{y})^2}}$$

One bunch of squares

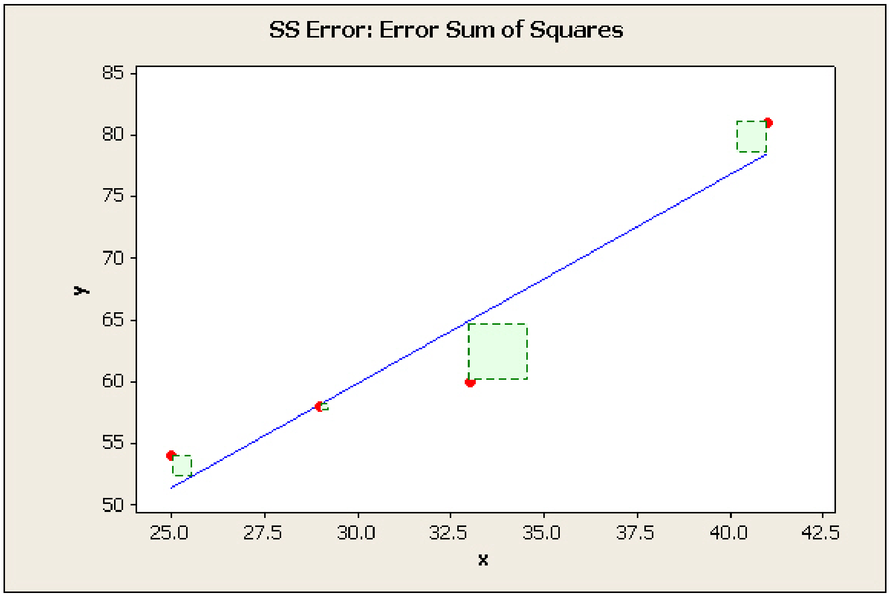
Another bunch of squares

SS Error: Error Sum of Squares

Now consider the model you obtain using regression analysis.



Again, quantify the "errors" of this model by measuring the vertical distance of each data value from the regression line and squaring it.



If you add the green areas of theses squares you get the SS Error, the top of the fraction.

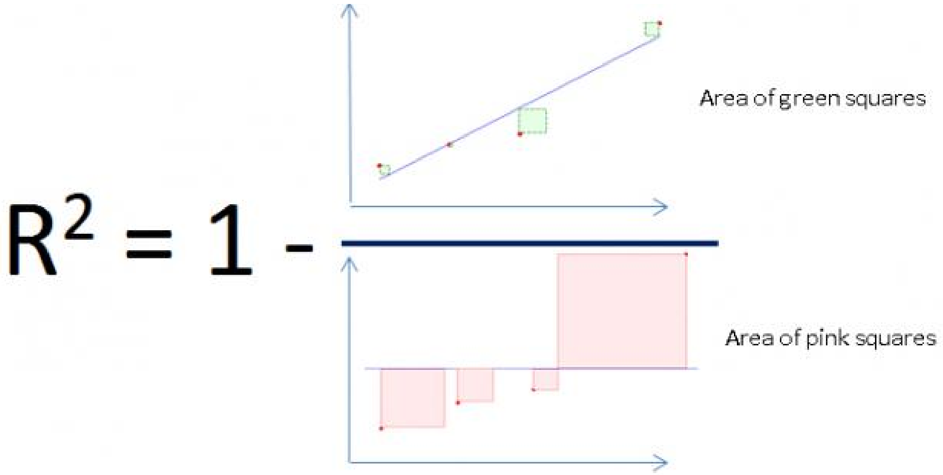
$$r^2 = 1 - \frac{\text{SS Error}}{\text{SS Total}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

One bunch of squares

Another bunch of squares

So R^2 basically just compares the errors of your regression model to the errors you’d have if you just used the mean of Y to model your data.

R-Squared for Visual Thinkers



The smaller the errors in your regression model (the green squares) in relation to the errors in the model based on only the mean (pink squares), the closer the fraction is to 0, and the closer R^2 is to 1 (100%).

That’s the case shown here. The green squares are much smaller than the pink squares. So the R^2 for the regression line is 91.4%.

But if the errors in your reqression model are about the same size as the errors in the trivial model that uses only the mean, the areas of the pink squares and the green squares will be similar, making the fraction close to 1, and the R^2 close to 0.

That means that your model, isn't producing a "tight fit" for your data, generally speaking. You’re getting about the same size errors you’d get if you simply used the mean to describe all your data points!

R-squared in Practice

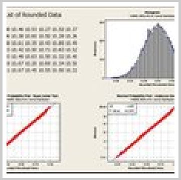
Now you know exactly what R^2 is. People have different opinions about **how critical the R-squared value is in regression analysis**. My view? No single statistic ever tells the whole story about your data. But that doesn't invalidate the statistic. It's always a good idea to evaluate your data using a variety of statistics. Then interpret the composite results based on the context and objectives of your specific application. If you understand how a statistic is actually calculated, you'll better understand its strengths and limitations.

Related link

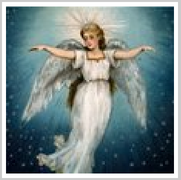
Want to see how another commonly used analysis, the t-test, really works? Read **this post** to learn how the t-test measures the "signal" to the "noise" in your data.

You might like:

Gauging Gage Part 1: Is 10 Parts Enough? - Fun with Statistics | Minitab


Normality Tests and Rounding - The Statistical Mentor | Minitab

Winning a Super Bowl Grid Pool: Frequency of Score Combinations in the NFL - Fun with Statistics | Minitab


Quantum Estimates: Where Angels Fear to Tread - Statistics and Quality Data Analysis | Minitab

Recommended by

<< Prev

Comments for *R-Squared: Sometimes, a Square is just a Square*

Name: Dr John M. Thompson
Time: Tuesday, February 18, 2014

The problem with use of R-squared is that most people seem unaware that there is an underlying assumption that the fit which is being assessed is a linear fit. If this assumption is invalid, the R-squared function has no meaning and is often, in such circumstances, very misleading. As someone who has often peer-reviewed scientific papers, I have found it is frequently misused and abused to claim that data do fit a linear model when clearly, if it relies on assuming a linear fit, it is incapable of such use!

Name: Patrick
Time: Wednesday, February 19, 2014

It’s interesting (and prescient) that you raised the issue about the assumption of linearity. Originally I drafted this post in response to a reader who asked why Minitab’s nonlinear regression analysis (Stat > Regression > Nonlinear Regression) did not include R-squared values in the output. To keep the post brief, I decided not to include the last part on why the calculation for R-squared only holds in linear regression, but not in nonlinear regression.

But for readers who are interested, I’ll give it a shot here.

The calculation of R-squared is based on the identify SS Total = SS Regression + SS Error. You can verify this by using algebraic substitution in the formula shown in this post. (SS Regression is the square formed using the vertical distance between the mean line and the regression line at each data point). Because of this basic identity of the sum of squares, R-squared can also be expressed as SSR/SST, and is always between 0 and 1 for linear models.

However, the equation for the sum of squares does not hold for nonlinear models. As a result, the value of R-squared is not always between 0 and 1. R-squared can take on negative values, making its interpretation problematic. Another issue is that R-squared is calculated by comparing the regression model to the constant (trivial) model—as seen in this post. However, there is no consistent, general definition of a constant model in nonlinear regression, observes Dr. HenSiong Tan, a senior researcher in Minitab’s Software Research Design department.

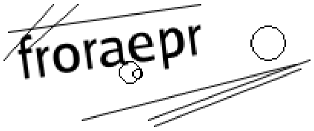
For these reasons, Minitab does not report an R-squared value for nonlinear regression. Instead, you can use the fitted line plot to visually examine how the nonlinear model fits the data across the range of values. You can also compare the value of S, the standard deviation of the residuals, for different nonlinear models. Smaller values of S (closer to 0) indicate a better fit.

Thanks for pointing out this important assumption.

Leave a comment

Name
Email
Your comment (No HTML)

Type the text from the image below



Submit

Blog Map Newsletter Contact Us Legal Privacy Policy

Quality. Analysis. Results.®