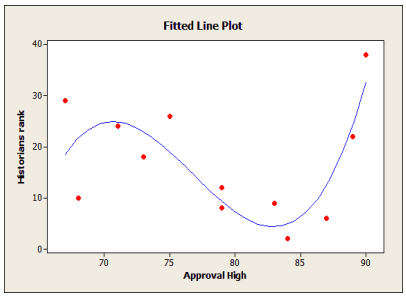
June 13, 2013 by Jim Frost in Regression Analysis

Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables



Multiple regression can be a beguiling, temptation-filled analysis. It's *so* easy to add more variables as you think of them, or just because the data are handy. Some of the predictors *will* be significant. Perhaps there is a relationship, or is it just by chance? You can add higher-order polynomials to bend and twist that fitted line as you like, but are you fitting real patterns or just connecting the dots? All the while, the R-squared (R^2) value increases, teasing you, and egging you on to add more variables!

Previously, I showed how **R-squared can be misleading** when you assess the goodness-of-fit for linear regression analysis. In this post, we'll look at why you should resist the urge to add too many predictors to a regression model, and how the adjusted R-squared and predicted R-squared can help!

Some Problems with R-squared

In my last post, I showed how R-squared *cannot* determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots. However, R-squared has additional problems that the adjusted R-squared and predicted R-squared are designed to address.

Problem 1: Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases. Consequently, a model with more terms may appear to have a better fit simply because it has more terms.

Problem 2: If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. This condition is known as overfitting the model and it produces misleadingly high R-squared values and a lessened ability to make predictions.

What Is the Adjusted R-squared?

The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors.

Suppose you compare a five-predictor model with a higher R-squared to a one-predictor model. Does the five predictor model have a higher R-squared because it's better? Or is the R-squared higher because it has more predictors? Simply compare the adjusted R-squared values to find out!

The adjusted R-squared is a modified version of R-squared that has been adjusted for the

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

number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared.

In the simplified Best Subsets Regression output below, you can see where the adjusted R-squared peaks, and then declines. Meanwhile, the R-squared continues to increase.

Vars	R-Sq	R-Sq(adj)
1	72.1	71.0
2	85.9	84.8
3	87.4	85.9
4	89.1	82.3
5	89.9	80.7

You might want to include only three predictors in this model. In my last blog, we saw how an under-specified model (one that was too simple) can produce biased estimates. However, an overspecified model (one that's too complex) is more likely to reduce the precision of coefficient estimates and predicted values. Consequently, you don't want to include more terms in the model than necessary. (Read an example of **using Minitab's Best Subsets Regression**.)

What Is the Predicted R-squared?

The predicted R-squared indicates how well a regression model predicts responses for new observations. This statistic helps you determine when the model fits the original data but is less capable of providing valid predictions for new observations. (Read an example of **using regression to make predictions.**)

Minitab calculates predicted R-squared by systematically removing each observation from the data set, estimating the regression equation, and determining how well the model predicts the removed observation. Like adjusted R-squared, predicted R-squared can be negative and it is always lower than R-squared.

Even if you don't plan to use the model for predictions, the predicted R-squared still provides crucial information.

A key benefit of predicted R-squared is that it can prevent you from overfitting a model. As mentioned earlier, an overfit model contains too many predictors and it starts to model the random noise.

Because it is impossible to predict random noise, the predicted R-squared must drop for an overfit model. If you see a predicted R-squared that is much lower than the regular R-squared, you almost certainly have too many terms in the model.

Examples of Overfit Models and Predicted R-squared

You can try these examples for yourself using this Minitab **project file** that contains two worksheets. If you want to play along and you don't already have it, please download the free 30-day trial of Minitab **Statistical Software**!

There's an easy way for you to see an overfit model in action. If you analyze a linear regression model that has one predictor for each degree of freedom, you'll always get an R-squared of 100%!

In the random data worksheet, I created 10 rows of random data for a response variable and nine predictors. Because there are nine predictors and nine degrees of freedom, we get an R-squared of 100%.

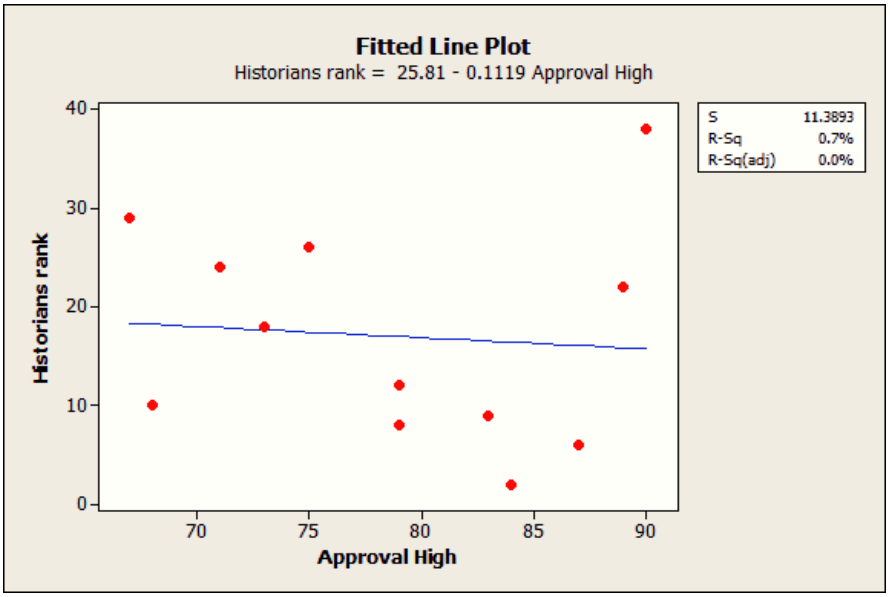
```
Summary of Model

S = *       R-Sq = 100.00%   R-Sq(adj) = *%
PRESS = *   R-Sq(pred) = *%
```

It appears that the model accounts for all of the variation. However, we know that the random

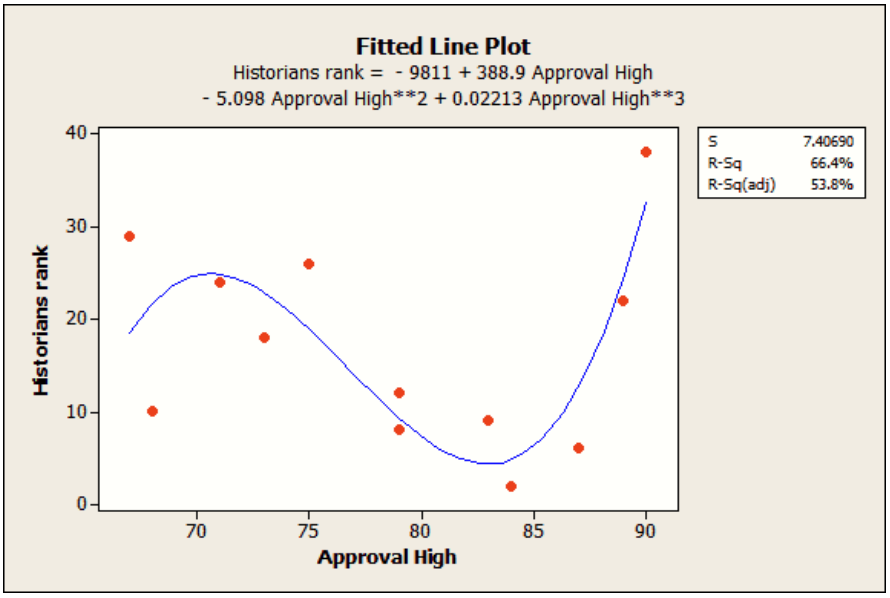
predictors do not have any relationship to the random response! We are just fitting the random variability.

That's an extreme case, but let's look at some real data in the President's ranking worksheet.

These data come from my post about **great Presidents**. I found no association between each President's highest approval rating and the historian's ranking. In fact, I described that fitted line plot (below) as an exemplar of no relationship, a flat line with an R-squared of 0.7%!



Let's say we didn't know better and we overfit the model by including the highest approval rating as a cubic polynomial.



Coefficients

Term	Coef	SE Coef	T	P
Constant	-9811.20	3568.01	-2.74977	0.025
Approval High	388.93	137.59	2.82665	0.022
Approval High*Approval High	-5.10	1.76	-2.89633	0.020
Approval High*Approval High*Approval High	0.02	0.01	2.96268	0.018

Summary of Model

S = 7.40690 R-Sq = 66.39% R-Sq(adj) = 53.79%
PRESS = 1588.39 R-Sq(pred) = -21.62%

Wow, both the R-squared and adjusted R-squared look pretty good! Also, the coefficient estimates are all significant because their p-values are less than 0.05. The residual plots (not shown) look good too. Great!

Not so fast...all that we're doing is excessively bending the fitted line to artificially connect the dots rather than finding a true relationship between the variables.

Our model is too complicated and the predicted R-squared gives this away. We actually have a negative predicted R-squared value. That may not seem intuitive, but if 0% is terrible, a negative percentage is even worse!

The predicted R-squared doesn't have to be negative to indicate an overfit model. If you see the predicted R-squared start to fall as you add predictors, even if they're significant, you should begin to worry about overfitting the model.

Closing Thoughts about Adjusted R-squared and Predicted R-squared

All data contain a natural amount of variability that is unexplainable. Unfortunately, R-squared doesn't respect this natural ceiling. Chasing a high R-squared value can push us to include too many predictors in an attempt to explain the unexplainable.

In these cases, you *can* achieve a higher R-squared value, but at the cost of misleading results, reduced precision, and a lessened ability to make predictions.

Both adjusted R-squared and predicted R-square provide information that helps you assess the number of predictors in your model:

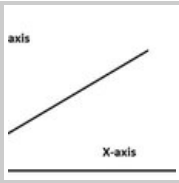
- Use the adjusted R-square to compare models with different numbers of predictors
- Use the predicted R-square to determine how well the model predicts new observations and whether the model is too complicated

Regression analysis is powerful, but you don't want to be seduced by that power and use it unwisely!

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

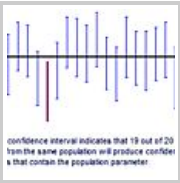
You might like:

Regression Smackdown: Stepwise versus Best Subsets! - Adventures in Statistics | Minitab



Regression Analysis: How to Interpret the Constant (Y Intercept) - Adventures in Statistics | Minitab

Applied Regression Analysis: How to Present and Use the Results to Avoid Costly Mistakes, part 2 - Adventures in Statistics | Minitab



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

<< Prev Next >>

Comments for *Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables*

Name: Enogwe Samuel Ugochukwu
Time: Tuesday, September 3, 2013

compliments. I suggest that you tell us the implication of R squared in the presence of multicollinearity .

Name: Jim Frost
Time: Wednesday, September 4, 2013

Hi, thanks for writing!

I write about this issue in an earlier blog. Multicollinearity doesn't affect how well the model fits, so R-squared is unaffected.

You can read more about this here:
http://blog.minitab.com/blog/adventures-in-statistics/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them

Jim

Name: prasshanth
Time: Wednesday, January 22, 2014

How can you add "predicted R-squared" result in the "Fitted line" plot above. Please help I am using minitab 16 and not finding this option.

Name: Jim Frost
Time: Wednesday, January 22, 2014

Hi Prasshanth,

I wish there was a simple option to choose, but you can add it to legend yourself.

First, you'll need to obtain the predicted R-squared by running your model in General Regression. Note the predicted R-squared, and then add it to the legend in the Fitted Line Plot.

To add it to the legend, single-click the legend box in the plot, and drag the bottom edge down to make the box large enough for one more row of text.

Click on the large T in the Graph Annotation Tools toolbar to add a text box. You'll see a crosshairs. Position that over the blank space in the legend and click. (If the Graph Annotation toolbar isn't visible, go to Tools > Toolbars and check Graph Annotation Tools.)

In the Add Text dialog that appears, type in R-Sq (pred) and the value. For best spacing results, put 5 spaces between R-sq (pred) and the value. Click OK.

Now, you need to edit the font to make it the right size. Double click the text you just added and click the Font tab in the dialog that appears. Change the font to Tahoma and size 8.

The text you added should match the text that appears in the legend by default.

Jim

Name: Sachin Wagh
Time: Monday, February 24, 2014

What should be the optimum level of values for assessment of best fitment of model.
1) All R-Sq, Adj R-Sq, Pred R Sq be Max
2) All minimum
3) All contradicting
Can you please put the inferences based on these different combinations.

Name: Jim Frost
Time: Monday, February 24, 2014

Hi Sachin,

I'll take a stab at some general guidelines, but you should always use subject area knowledge as well as all of the model statistics and diagnostic graphs to get a complete picture of how well your model fits the data and what you can learn from the model.

1) Usually if all of the R-squared values are high, it's a good thing. Your model explains a lot of the variance, and you're probably not overfitting the model, or including too many predictors. Predictions based on the model are likely to be precise. However, keep in mind that none of the R-squared values can tell you whether your estimates are biased. So, you still need to check your residual plots.

Read this post for more details for this point:
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

2) If all of the R-squared values are low, it indicates that your model doesn't explain much of the variance in the response variable. That doesn't sound good, but if you happen to have significant predictors and satisfy the assumptions, you are still learning useful information about how the predictors and response are related. Low R-squared values are not always bad, and are even expected in some fields. However, you won't be able to predict new observations that precisely.

I suggest you read my post "How high should R-squared be?" for more information about this point:
http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis

3) Contradicting R-squared values. The adjusted and predicted R-squared values must be lower than the R-squared value. So, if you have a high R-squared value but a low adjusted R-squared, you're probably including too many predictors. Try reducing the number of predictors.

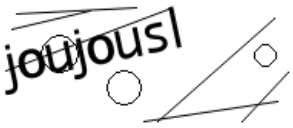
If the predicted R-squared is low but the regular R-squared is high, you know that the your model explains the existing data set well but you can't predict new observations that well. This often happens when you're overfitting the model, which is generally caused by having too many predictors for the number of data points.

I hope this helps! Thanks for reading!
Jim

Leave a comment

Name
Email
Your comment (No HTML)

Type the text from the image below



Submit