When is R squared negative?

My understanding is that R squared cannot be negative as it is the square of R. However I ran a simple linear regression in SPSS with a single independent variable and a dependent variable. My SPSS output give me a negative value for R-squared. If I was to calculate this by hand from R then R squared would be positive. What has SPSS done to calculate this as negative?

```
R=-.395
R squared =-.156
B (un-standardized)=-1261.611
```

Code I've used:

```
DATASET ACTIVATE DataSet1.
REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA
           /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN
           /DEPENDENT valueP /METHOD=ENTER ageP
```

I get a negative value. Can anyone explain what this means?

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	-.395	-.156	-.050	60237.05361

Thanks.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	137278.4	46816.646		2.932	.019
	agep	-1261.611	1038.015	-.395	-1.215	.259

a. Dependent Variable: valuep

spss r-squared

edited Jul 16 '11 at 5:22

asked Jul 11 '11 at 17:07

Anne
342 3 7 13

3 Does this answer your question? stats.stackexchange.com/questions/6181/… If not, then please provide more information: this is the "SPSS output" of what procedure? – whuber ♦ Jul 11 '11 at 17:14

2 Does your linear regression model have an intercept? – NPE Jul 11 '11 at 17:59

2 @Anne Again, **which SPSS procedure are you using?** – whuber ♦ Jul 11 '11 at 18:19

1 @Anne I suggest you disregard the time series reply, because your data are not time series and you're not using a time series procedure. Are you really sure the R squared is given as a negative value? Its magnitude is correct:

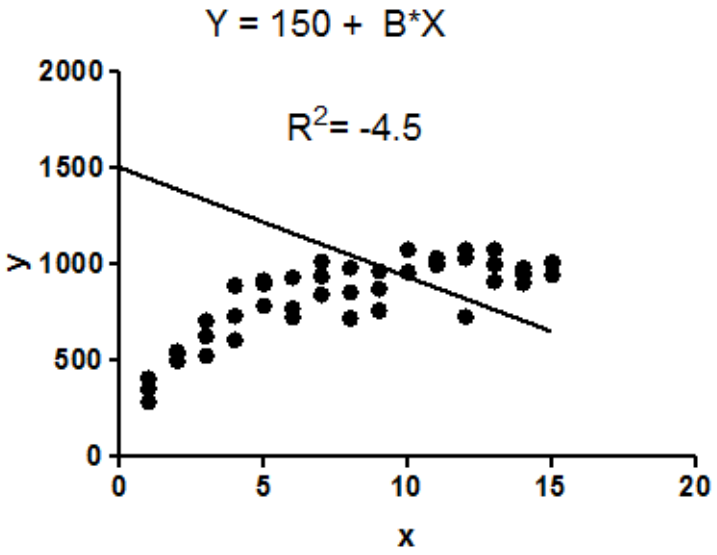
$(-0.395)^2 = 0.156$. I have looked through SPSS help to see whether perhaps as a convention the R-squared value for negative R's is negated, but I don't see any evidence that this is the case. Perhaps you could post a screen shot of the output where you are reading the R-squared? – whuber ♦ Jul 11 '11 at 20:26

1 The dependent variable is price of houses so it is feasible that the 95% CI maybe 120,000. Unfortunately I cannot post the data here as it would be contrary to data use conditions. – Anne Jul 16 '11 at 5:26

5 Answers

R^2 compares the fit of the chosen model with that of a horizontal straight line (the null hypothesis). If the chosen model fits worse than a horizontal line, then R^2 is negative. Note that R^2 is not always the square of anything, so it can have a negative value without violating any rules of math. R^2 is negative only when the chosen model does not follow the trend of the data, so fits worse than a horizontal line.

Example: fit data to a linear regression model constrained so that the Y intercept must equal 1500.



The model makes no sense at all given these data. It is clearly the wrong model, perhaps chosen by accident.

The fit of the model (a straight line constrained to go through the point (0,1500)) is worse than the fit of a horizontal line. Thus the sum-of-squares from the model (SS_{reg}) is larger than the sum-of-squares from the horizontal line (SS_{tot}). R^2 is computed as $1 - \frac{SS_{\text{reg}}}{SS_{\text{tot}}}$. When SS_{reg} is greater than SS_{tot}, that equation computes a negative value for R^2.

With linear regression with no constraints, R^2 must be positive (or zero) and equals the square of the correlation coefficient, r. A negative R^2 is only possible with linear regression when either the intercept or the slope are constrained so that the "best-fit" line (given the constraint) fits worse than a horizontal line. With nonlinear regression, the R^2 can be negative whenever the best-fit model (given the chosen equation, and its constraints, if any) fits the data worse than a horizontal line.

Bottom line: a negative R^2 is not a mathematical impossibility or the sign of a computer bug. It simply means that the chosen model (with its constraints) fits the data really poorly.

edited May 6 '14 at 2:26

Nick Stauner
8,550 5 24 53

answered Jul 13 '11 at 15:07

Harvey Motulsky
7,003 1 30 65

3 @JMS That's the opposite of what my Googling indicates: "/ORIGIN" fixes the intercept at 0; "/NOORIGIN" "tells SPSS not to suppress the constant" (An Introductory Guide to SPSS for Windows) – whuber ♦ Jul 13 '11 at 18:13

7 @whuber Correct. @harvey-motulsky A negative R^2 value **is** a mathematical impossibility (and suggests a computer bug) for regular OLS regression (with an intercept). This is what the 'REGRESSION' command does and what the original poster is asking about. Also, for OLS regression, R^2 **is** the squared correlation between the predicted and the observed values. Hence, it must be non-negative. For simple OLS regression with one predictor, this is equivalent to the squared correlation between the predictor and the dependent variable -- again, this must be non-negative. – Wolfgang Jul 14 '11 at 7:17

1 @whuber Indeed. My bad; obviously I don't use SPSS - or read, apparently :) – JMS Jul 14 '11 at 16:56

1 @whuber. I added a paragraph pointing out that with linear regression, R2 can be negative only when the intercept (or perhaps the slope) is constrained. With no constraints, the R2 must be positive and equals the square of r, the correlation coefficient. – Harvey Motulsky Jul 16 '11 at 15:55

1 @HarveyMotulsky, in this case the intercept or slope were not constrained. It seems that you are saying that Rsquared can only be negative if these are constrained. Can you elaborate on what might have occurred in this particular case? – Anne Jul 16 '11 at 21:56

Have you forgotten to include an intercept in your regression? I'm not familiar with SPSS code, but on page 21 of Hayashi's Econometrics:

> If the regressors do not include a constant but (as some regression software packages do) you nevertheless calculate R^2 by the formula
>
> $$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$
>
> then the R^2 can be negative. This is because, without the benefit of an intercept, the regression could do worse than the sample mean in terms of tracking the dependent variable (i.e., the numerator could be greater than the denominator).

I'd check and make sure that SPSS is including an intercept in your regression.

answered Jul 12 '11 at 7:04

jefflovejapan
303 1 11

4 NOORIGIN subcommand in her code tells that intercept was included in the model – ttnphns Jul 12 '11 at 10:12

1 that's weird. I would have guessed that `NOORIGIN` would mean that intercept was not included in the model, just going off the name. – Matt O'Brien Nov 8 '15 at 4:29

This can happen if you have a time series that is N.i.i.d. and you construct an inappropriate ARIMA model of the form(0,1,0) which is a first difference random walk model with no drift then the variance (sum of squares - SSE) of the residuals will be larger than the variance (sum of squares SSO) of the original series. Thus the equation 1-SSE/SSO will yield a negative number as SSE exceeedS SSO . We have seen this when users simply fit an assumed model or use inadequate procedures to identify/form an appropriate ARIMA structure. The larger message IS that a model can distort (much like a pair of bad glasses) your vision. Without having access to your data I would otherwise have a problem in explaining your faulty results. Have you brought this to the attention of IBM ?

The idea of an assumed model being counter-productive has been echoed by Harvey Motulsky. Great post Harvey !

edited Jul 13 '11 at 17:57

answered Jul 11 '11 at 18:11

IrishStat
12.9k 1 14 28

1 stat. Thanks. No I have not spoken to IBM. The data is not time series. It is from point in time data. – Anne Jul 11 '11 at 19:55

4 @Anne and others: Since your data are not time series and you're not using a time series procedure please disregard my answer. Others who have observed negative R Squares when involved with time series might find my post interesting and tangentially informative. Others unfortunately may not. – IrishStat Jul 11 '11 at 21:36

I encountered a similar problem when implementing a Least Squares solution in Python. The problem turned out to be a failure on my part to normalize the inputs to R^2 when I had also normalized the inputs to the Least Squares method. The resulting negative R^2 values where caused by the disparity between the larger real values of the original inputs versus the smaller normalized inputs.

answered Jan 4 '15 at 22:22

jjh
11 1

It seems to me that this doesn't answer the question, since this was a simple *miscalculation* of R^2 due to code that didn't implement R^2 correctly. It might be better suited to be a comment. – Silverfish Sep 30 '15 at 0:53

If adj R square negative that means sample size less than number of parameters if increase sample size the matter would solve.

answered Feb 10 at 6:50

Shabir ahmad
1 1

I'm not sure that's the case, can anybody else confirms? – Student T Feb 15 at 5:11