What does it mean to have a low R-squared ? A warning about misleading interpretation xstatistics ximportant

Sunday, July 15, 2018 8:18 AM

Clipped from: https://humanvarieties.org/2014/03/31/what-does-it-mean-to-have-a-low-r-squared-a-warning-about-misleading-interpretation/

A common argument we read everytime, everywhere. All with the same common mistake. It consists in squaring the correlation. For example : "Your brain-IQ correlation is r=0.40, so if you square it, that only amounts to a tiny 16% (r^2=0.40* 0.40=0.16) of variance explained which is not impressive". Or something in this vein. R^2 use and abuse caused enough damage. It is more than time to put an end to this utter fallacy.

The problem is that a correlation is an effect size expressed in standard deviation so that a r=0.40 is equivalent to an increase in variable X (say, IQ) by 1 SD that is associated with an increase by 0.4 SD in variable Y (say, brain size) depending on which one is the independent and the dependent variable. This effect is not trivial, of course. The tentative interpretation from the R^2 is thus called into question regarding its meaning.

It is easy to understand the nonsensical concept of the R^2 if we know what a correlation is. So, in a regression model, we examine the effect of an independent variable income, measured in 10000 dollars/month per unit, with the dependent variable being hours of sleeping, with 10 hours/month per unit. Say, the unstandardized coefficient of income is -0.543, which means one unit increase in the income variable, or precisely 5430 dollars gain per month, is associated with 1 hour/month less of sleeping (given that linearity assumption holds). That's the real-world effect. No more, no less. But the R-squared will tell us something different. Now this time, we also want to examine the standardized coefficient of income. Suppose it was 0.20, the R^2 will be 0.04, considering that this variable was the only one entered in the regression model. So, the real-world effect size has been divided by 5. In other words, the R^2 tells us that much less than 5430 dollars/month is associated with 10 hours/month less of sleeping. This can't be serious.

Phil Birnbaum (On correlation, r, and r-squared, 2006) puts it comically :

The ballpark is ten miles away, but a friend gives you a ride for the first five miles. You're halfway there, right? Nope, you're actually only one quarter of the way there.

He rightly pointed out that r^2 expresses the effect size in a statistical sense, not in the real life sense. If only one is interested in the sums of the squares of the differences (i.e., deviations) the r^2 can make sense. But again, it's meaningless from the real life perspective. From the real life perspective, brain size would explain 40% of the variance in IQ, not 16%.

Given the serious flaws of R^2, some alternative metrics have been proposed. Here's how Sackett et al. (2008, p. 216) summarize it :

Prototypically, admissions tests correlate about .35 with first-year grade point average (GPA), and employment tests correlate about .35 with job training performance and about .25 with performance on the job. One reaction to these findings is to square these correlations to obtain the variance accounted for by the test (.25 accounts for 6.25%; .35 accounts for 12.25%) and to question the appropriateness of giving tests substantial weight in selection or admissions decisions given these small values (e.g., Sternberg, Wagner, Williams, & Horvath, 1995; Vasquez & Jones, 2006).

One response to this reaction is to note that even if the values above were accurate (and we make the case below that they are, in fact, substantial underestimates), correlations of such magnitude are of more value than critics recognize. As long ago as 1928, Hull criticized the small percentage of variance accounted for by commonly used tests. In response, a number of scholars developed alternate metrics designed to be more readily interpretable than "percentage of variance accounted for" (Lawshe, Bolda, & Auclair, 1958; Taylor & Russell, 1939). Lawshe et al. (1958) tabled the percentage of test takers in each test score quintile (e.g., top 20%, next 20%, etc.) who met a set standard of success (e.g., being an above-average performer on the job or in school). A test correlating .30 with performance can be expected to result in 67% of those in the top test quintile being above-average performers (i.e., 2 to 1 odds of success) and 33% of those in the bottom quintile being above-average performers (i.e., 1 to 2 odds of success). Converting correlations to differences in odds of success results both in a readily interpretable metric and in a positive picture of the value of a test that "only" accounts for 9% of the variance in performance. Subsequent researchers have developed more elaborate models of test utility (e.g., Boudreau & Rynes, 1985; Brogden, 1946, 1949; Cronbach & Gleser, 1965; Murphy, 1986) that make similar points about the substantial value of tests with validities of the magnitude commonly observed. In short, there is a long history of expressing the value of a test in a metric more readily interpretable than percentage of variance accounted for.

Taylor & Russell (1939) devised a set of tables on odds of success resulting from the correlations of cognitive test scores with selection in employment. Jensen (1980, pp. 306-308) already discussed it. I select some of the tables from Taylor and Russell. The ones below present the situation when 40% and 50% of the candidates present satisfactory characteristics. The next step is to select the best candidates among them, based on their cognitive/achievement test scores.

Proportion of Employees Considered Satisfactory = .40
Selection Ratio

Jacobian
DECEMBER 18, 2015 AT 10:55 PM

Some of the worst forms of R^2 abuse come when aggregating effects. If you have ten independent factors each having a 0.1 correlation with some outcome, improving all of them by 1 SD will improve the outcome by 1 SD on average. If you add the R-squares, which I've actually seen some shameless/ignorant people do, you conclude that even all factors taken together have no significant effect because 10*0.01 = 0.1.

From <https://humanvarieties.org/2014/03/31/what-does-it-mean-to-have-a-low-r-squared-a-warning-about-misleading-interpretation/#comments>

1. **pnard**
APRIL 2, 2014 AT 8:03 AM
I'm wondering if you could post some more examples of where it's been misused?
REPLY

Chuck
APRIL 7, 2014 AT 12:06 PM
Economists do this all of the time. Here was, for example, Greg Clark:
The estimated persistence rate for income in India of 0.58, however, is not much higher than those for the United Kingdom (0.5) or the United States (0.47). **The share of income variance in the next generation attributable to inheritance from parents in India is still only (0.58)squared, or 0.34.** This suggests that even in India, an individual's position in the income ranks is not primarily derived from inheritance.
REPLY

Meng Hu (Post author)
APRIL 2, 2014 AT 9:05 AM
Oh, easy. There are millions.
Beleaguered Pygmalion: A history of the controversy over claims that teacher expectancy raises intelligence (Spitz, 1999).
He [Rosenthal] also chided Jensen for writing that 6.4% of the variance had little practical importance, whereas it is "equivalent to increasing the success rate of a new treatment procedure from 37% to 63%, a change that can hardly be considered trivial" (Rosenthal, 1985, p. 49).
Intelligence: Knowns and Unknowns (Neisser 1996)
Intelligence *tests were originally devised by Alfred Binet to measure children's ability to succeed in school. They do in fact predict school performance fairly well: the correlation between IQ scores and grades is about .50. They also predict scores on school achievement tests, designed to measure knowledge of the curriculum. Note, however, that correlations of this magnitude account for only about 25% of the overall variance.
Intelligence (Nathan Brody 1992, page 66).
If a representative correlation of .5 is corrected for attenuation assuming test-retest correlations of .75 for their experimental measures (the value for test-retest measures of inspection time) and .9 for their measure of intelligence, the corrected correlation is .6, indicating that the experimental measures may account for approximately 36% of the variance in intelligence test scores. Jensen and Kranzler's meta-analysis of all of the inspection-time data suggests that inspection-time measures account for approximately 25% of the variance in scores on intelligence tests.
Educability & Group Differences (Jensen 1973).
The point-biserial correlation of 0.493 between race and IQ with SES partialled out corresponds to a mean IQ difference between the races of about 1σ. (Figure 8.1 shows the relationship between the point-biserial correlation, rpbs, and mean group difference, d, in sigma units, when the two groups have equal Ns and equal σs.) The correlation of SES and IQ with race partialled out is significantly smaller than the correlation between race and IQ with SES partialled out. All this can mean is that the environmental factors summarized in the SES index at most account for $(0.691)^2 - (0.493)_2 = 0.23$ of the total IQ variance which is associated with SES differences between races.
Notice that the correlation between SES and IQ (with race partialled out) is 0.312, so that SES accounts for about 0.10 (i.e., r_2) of the variance in IQ within racial groups – a value slightly greater than estimates of between-families environmental variance (e.g., Jensen, 1967).
In Jensen (1980) Bias in Mental Testing, you have a lot of studies that make use of ANOVA for studying DIF. Let aside the fact that it is proven ANOVA is not adequate to detect DIF (which is the topic of my forthcoming post) it is based on this old-fashioned "variance-based" index. Same thing for logistic regression. Even when you have a difference as large as 20 or 30% difference in probability of correct response in an item of a (cognitive) test, your R^2 amounts to no more than an absurd 1% of "variance accounted for". It's not only the usual habit of squaring correlations that is bad practice. Even

them, based on their cognitive/achievement test scores.

Proportion of Employees Considered Satisfactory = .40
Selection Ratio

r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.40	.40	.40	.40	.40	.40	.40	.40	.40	.40	.40
.05	.44	.43	.43	.42	.42	.42	.41	.41	.41	.40	.40
.10	.48	.47	.46	.45	.44	.43	.42	.42	.41	.41	.40
.15	.52	.50	.48	.47	.46	.45	.44	.43	.42	.41	.41
.20	.57	.54	.51	.49	.48	.46	.45	.44	.43	.41	.41
.25	.61	.58	.54	.51	.49	.48	.46	.45	.43	.42	.41
.30	.65	.61	.57	.54	.51	.49	.47	.46	.44	.42	.41
.35	.69	.65	.60	.56	.53	.51	.49	.47	.45	.42	.41
.40	.73	.69	.63	.59	.56	.53	.50	.48	.45	.43	.41
.45	.77	.72	.66	.61	.58	.54	.51	.49	.46	.43	.42
.50	.81	.76	.69	.64	.60	.56	.53	.49	.46	.43	.42
.55	.85	.79	.72	.67	.62	.58	.54	.50	.47	.44	.42
.60	.89	.83	.75	.69	.64	.60	.55	.51	.48	.44	.42
.65	.92	.87	.79	.72	.67	.62	.57	.52	.48	.44	.42
.70	.95	.90	.82	.76	.69	.64	.58	.53	.49	.44	.42
.75	.97	.93	.86	.79	.72	.66	.60	.54	.49	.44	.42
.80	.99	.96	.89	.82	.75	.68	.61	.55	.49	.44	.42
.85	1.00	.98	.93	.86	.79	.71	.63	.56	.50	.44	.42
.90	1.00	1.00	.97	.91	.82	.74	.65	.57	.50	.44	.42
.95	1.00	1.00	.99	.96	.87	.77	.66	.57	.50	.44	.42
1.00	1.00	1.00	1.00	1.00	1.00	.80	.67	.57	.50	.44	.42

Proportion of Employees Considered Satisfactory = .50
Selection Ratio

r	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.05	.54	.54	.53	.52	.52	.52	.51	.51	.51	.50	.50
.10	.58	.57	.56	.55	.54	.53	.53	.52	.51	.51	.50
.15	.63	.61	.58	.57	.56	.55	.54	.53	.52	.51	.51
.20	.67	.64	.61	.59	.58	.56	.55	.54	.53	.52	.51
.25	.70	.67	.64	62	.60	.58	.56	.55	.54	.52	.51
.30	.74	.71	.67	.64	.62	.60	.58	.56	.54	.52	.51
.35	.78	.74	.70	.66	.64	.61	.59	.57	.55	.53	.51
.40	.82	.78	.73	.69	.66	.63	.61	.58	.56	.53	.52
.45	.85	.81	.75	.71	.68	.65	.62	.59	.56	.53	.52
.50	.88	.84	.78	.74	.70	.67	.63	.60	.57	.54	.52
.55	.91	.87	.81	.76	.72	.69	65	.61	.58	.54	.52
.60	.94	.90	.84	.79	.75	.70	.66	.62	.59	.54	.52
.65	.96	.92	.87	.82	.77	.73	.68	.64	.59	.55	.52
.70	.98	.95	.90	.85	.80	.75	.70	.65	.60	.55	.53
.75	.99	.97	.92	.87	.82	.77	.72	.66	.61	.55	.53
.80	1.00	.99	.95	.90	.85	.80	.73	.67	.61	.55	.53
.85	1.00	.99	.97	.94	.88	.82	.76	.69	.62	.55	.53
.90	1.00	1.00	.99	.97	.92	.86	.78	.70	.62	.56	.53
.95	1.00	1.00	1.00	.99	.96	.90	.81	.71	.63	.56	.53
1.00	1.00	1.00	1.00	1.00	1.00	1.00	.83	.71	.63	.56	.53

Under all conditions, we see that when a test has r=0.00, the % of selection is always 40% or 50%, just the same % as for the situation where no cognitive test has been used for selection criteria. But when the correlation becomes positive and goes stronger and stronger, the probability of being selected is higher, especially more so when the selection ratio is more stringent (e.g., 0.05 instead of 0.95). Obviously, when the selection is not stringent, higher cognitive/achievement test scores don't benefit much for the candidates. But when the selection is accrued, the more intelligent candidates are the ones having the largest probability of being selected. This reminds the notion of cognitive elites and cognitive stratification mentioned by the authors of The Bell Curve (1994).

We should bear in mind the general picture that such correlations are usually underestimated by the presence of measurement errors and range restriction in ability. And sometimes sampling errors. Another is the deviation from perfect construct validity, as discussed by Hunter & Schmidt (2004, pp. 115-116).

Hunter & Schmidt (2004, pp. 289-291) give another illustration of why I dislike R^2.

r Versus r²: Which Should Be Used?

Chapter 3 focuses on the correlation coefficient as the statistic to be cumulated across studies. Some have argued, however, that it is the squared correlation — r^2 — that is of interest, not r itself. They argue that r^2 is the proportion of variance in one variable that is accounted for by the other variable, and this is the figure that provides the true description of the size of the relationship. Further, the advocates of r^2 typically hold that relationships found in the behavioral and social sciences are very small. For example, they maintain that r = .30 is small because r^2 = .09, indicating that only 9% of the variance in the dependent variable is accounted for. Even r = .50 is considered small: Only 25% of the variance is explained.

The "percentage of variance accounted for" is statistically correct but substantively erroneous. It leads to severe underestimates of the practical and theoretical significance of relationships between variables. This is because r^2 (and all other indexes of percentage of variance accounted for) are related only in a very nonlinear way to the magnitudes of effect sizes that determine their impact in the real world.

The correlation is the standardized slope of the regression of the dependent variable on the independent variable. If x and y are in standard score form, then $\hat{y} = rx$. Thus, r is the slope of the line relating y to x. As such, it indexes the predictability of y from x. For example, if r = .50, then, for each increase of 1 SD in x, there is an increase of .50 SD in y. The statistic r^2 plays no role in the regression equation. The

based on this old-fashioned "variance-based" index. Same thing for logistic regression. Even when you have a difference as large as 20 or 30% difference in probability of correct response in an item of a (cognitive) test, your R^2 amounts to no more than an absurd 1% of "variance accounted for". It's not only the usual habit of squaring correlations that is bad practice. Even the "model R^2" in usual regressions or other SEM analyses has the exact same problem. R^2 really should be banned. It means nothing and is, worse, misleading.
Everytime someone uses this kind of argument "A explains x% of variance in B" by referring to R^2 you can be confident at 100% that it is misused. I never saw, anyway, a reported R^2 that is not misused. I don't remember that.
REPLY

From <https://humanvarieties.org/2014/03/31/what-does-it-mean-to-have-a-low-r-squared-a-warning-about-misleading-interpretation/#comments>

same principle applies in raw score regression; here the slope again is based on r, not r^2. The slope is B = r (SDy/SDx). The raw score regression equation is

$$\hat{Y} = \left\{ r \frac{SD_y}{SD_x} \right\} X + C$$

where C is the raw score intercept.

The problem with all percentage variance accounted for indexes of effect size is that variables that account for small percentages of the variance often have very important effects on the dependent variable. Variance-based indexes of effect size make these important effects appear much less important than they actually are, misleading both researchers and consumers of research. Consider an example. According to Jensen (1980) and others, the heritability of IQ true scores is about .80. This means that 80% of the (true) variance is due to heredity and only 20% is due to environmental differences, yielding a ratio of "importance" of .80/.20 or 4 to 1. That is, based on percentage of variance accounted for indexes, heredity is 4 times more important than environment in determining intelligence. However, this picture is very deceptive. (For purposes of this example, we assume heredity and environment are uncorrelated; that is close to true, and in any event the principle illustrated here is not dependent on this assumption.) The functional relationships between these two variables and intelligence are expressed by their respective standard score regressions, not by the figures of .80 and .20. The correlation between IQ and heredity is $\sqrt{.80} = .894$, and the correlation between environment and intelligence is $\sqrt{.20} = .447$. Thus, the functional equation for predicting IQ from each (when all variables are in standard score form) is

$\hat{}$ YIQ = .894(H) + .447(E)

Thus, for each 1 SD increase in heredity (H), there is a .894 SD increase in IQ, and for each 1 SD increase in environment (E), there is a .447 SD increase in IQ. This is the accurate statement of the power of H and E to produce changes in IQ; that is, it is the true statement of their effects on IQ. The relative size of these effects is .894/.447 = 2. That is, the true impact of heredity on intelligence is only twice as great as that of environment, not 4 times as great, as implied by the percentage of variance accounted for indexes. The variance-based indexes underestimate the causal impact of environment relative to heredity by a factor of 2. Further, the absolute causal importance of environment is underestimated. The correct interpretation shows that if environment could be improved by 2 SDs, the expected increase in IQ (where SD IQ = 15) would be .447(2.00)(15) = 13.4. This would correspond to an increase from 86.6 to 100, which would have very important social implications. This correct analysis shows the true potential impact of environment, while the variance-based statement that environment accounts for only 20% of IQ variance leaves the false impression that environment is not of much importance. (Note: The fact that no one seems to know how to increase environment by 2 SDs is beside the point here.)

This is not an unusual case. For example, the Coleman Report (1966) concluded that, when other variables were controlled for, money spent per student by school districts accounted for only a small percentage of the variance of student achievement. The report concluded that financial resources and facilities, such as libraries and labs, were not very important because they provide little "leverage" over student achievement. Later analyses, however, showed that this small percentage of variance corresponded to a standardized regression coefficient for this variable that was much larger, and demonstrated that improvements in facilities could yield increases in student achievement that were significant socially and practically (Mosteller & Moynihan, 1972).

Variance-based interpretations have led to the same sort of errors in personnel selection. There it was said that validity coefficients of, for example, .40 were not of much value because only 16% of the variance of job performance was accounted for. A validity coefficient of .40, however, means that, for every 1 SD increase in mean score on the selection procedure, we can expect a .40 SD increase in job performance — a substantial increase with considerable economic value. In fact, a validity coefficient of .40 has 40% of the practical value to an employer of a validity coefficient of 1.00 — perfect validity (Schmidt & Hunter, 1998; Schmidt, Hunter, McKenzie, & Muldrow, 1979).

Variance-based indexes of effect size are virtually always deceptive and misleading and should be avoided, whether in meta-analysis or in primary research. In meta-analysis, such indexes have an additional disadvantage: They obscure the direction of the effect. Being nondirectional, they do not discriminate between an r of .50 and an r of −.50; both would enter the meta-analysis as $r^2 = .25$.

To illustrate that r, and not r^2, is the appropriate index of effect size and to show that "small" rs (e.g., .20–.30) indicate substantial relationships, Rosenthal and Rubin (1979b, 1982c) presented the binomial effect size display (BESD). Although this technique requires that both variables be dichotomous (e.g., treatment vs. control or "survived" vs. "died") and requires 50% on each side of each dichotomy, it does forcefully illustrate the practical importance of "small" correlations. For example, a correlation of .32 ($r^2 = .10$) between treatment with a particular drug and patient survival corresponds to a reduction in the death rate from 66% to 34% (Rosenthal, 1984, p. 130). Thus, a relationship that accounts for only 10% of the variance means a reduction in the death rate of almost 50%. Small correlations can indicate large impacts. The BESD uses a special case — that of truly dichotomous variables — to illustrate the same principle we have presented using the more general regression analysis method.

That Hunter & Schmidt admitted we don't know how to swing the environments by 2 SD reminds me of Jensen's cogent argument advanced in Educability & Group

Differences (1973, pp. 162-169) that has also been adopted by Herrnstein & Murray (1994) even though they arrive at a much less pessimistic view with regard to race differences. We prefer, this time, to assume a 70% heritability of IQ. But let's begin with Hunter & Schmidt operation. In that case, we are left with SQRT(0.30)=0.548 that needs to be multiplied by 15, which gives 0.548*15= 8.21 IQ points. Given Herrnstein & Murray operation, we take the standard deviation of IQ, being 15, its variance should be 225 (15^2), and because cognitive environment explains 30% of the variance, or more precisely 67.5 (225*30/100), with the SD of the distribution of the environmental component of IQ being the square root of 67.5, which is 8.21 (same as with Hunter & Schmidt), a difference in environments between groups should be 15/8.21, or 1.83 SD, if the B-W gap of 15 IQ points is entirely due to differences in cognitive environment. An environmental difference of this magnitude is difficult to imagine (Herrnstein & Murray, 1994, pp. 298-299, see footnote). Because this 1.83 SD gap implies that blacks would be at the 3.3th percentile (given this calculator, one-sided) of the distribution of environments among whites. And given the asian-black IQ gap of 21 points (106-85=21), thus implying a cognitive environmental gap of 2.56 SD (21/8.21), blacks should be at the 0.5th percentile of the distribution of environments among asians. If we take this time Jensen's (1970) estimate of 4.74 IQ produced by 1 SD of environmental effects, dividing 15/4.74= 3.16 and 21/4.74=4.43, we are left with blacks being at the 0.08th percentile of white environments and 0.0005th percentile of asian environments. To say that scenario is unrealistic is an euphemism.

It's interesting to see that what is perhaps the best hereditarian argument ever made by Jensen (1973, pp. 162-169) is left intact whereas the claim that environments explain "only" 30% of IQ variance is false because it explains roughly twice this effect if r is used instead of R^2. This illustrates the full extent of the damage caused by the R^2 fallacy. The day when scientists would come to understand what R^2 is and what it is not will be surely a great day.

Phil Birnbaum talked about "r-squared abuse" (October 30, 2007) in the scientific literature. Some authors make erroneous reports in the real world consequences of some given factors, claiming they produce "x% gain" instead of "variance of x% gain".

This, in itself, is important to note because the common readers don't know a slight thing about stats. They will just report what they read and what they heard from journalists, without taking a glance in the study itself. Because if the reported effect size is based on R^2, there is a problem. That must be more widely known, informed.

Please, stop the r-squared abuse.

References.

Jensen (1980). Bias in Mental Testing.

Hunter & Schmidt (2004). Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.

Lawshe, Bolda, Brune, Auclair (1958). Expectancy Charts II. Their Theoretical Development.

Sackett et al. (2008). High-Stakes Testing in Higher Education and Employment: Appraising the Evidence for Validity and Fairness.

Taylor & Russell (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection – discussion and tables.