Mini Review

iMedPub Journals
http://www.imedpub.com

Journal of Drug Abuse
2471-853X

2016
Vol. 2 No. 2: 12

DOI: 10.21767/2471-853X.100021

On Correctly Adjusting the Squared Multiple Correlation Coefficient in Linear Regression: Effect Size Estimation and Significance Testing with Application to Substance Abuse Research

James B Hittner

Department of Psychology, College of Charleston, USA

Corresponding author: James B Hittner

hittnerj@cofc.edu

Department of Psychology, College of Charleston, 66 George Street Charleston, SC 29424, USA.

Tel: 8439536734
Fax: 8439537151

Citation: Hittner JB. On Correctly Adjusting the Squared Multiple Correlation Coefficient in Linear Regression: Effect Size Estimation and Significance Testing with Application to Substance Abuse Research. J Drug Abuse. 2016, 2:2.

Abstract

Linear regression analysis is ubiquitous in many areas of scholarly inquiry, including substance abuse research. In linear regression it is common practice to test whether the squared multiple correlation coefficient, R^2, differs significantly from zero. However, this test is misleading because the expected value of R^2 is not zero under the null hypothesis. In this brief methodological note I discuss the implications of this realization for calculating and interpreting the squared multiple correlation coefficient, R^2. In addition, I discuss and offer freely available software that calculates the expected value of R^2 under the null hypothesis that ρ-the population value of the multiple correlation coefficient-equals zero, an adjusted R^2 value and effect size measure that both take into account the expected value of R^2, and an F statistic that tests the significance of difference between the obtained R^2 and the expected value of R^2 under the null hypothesis that ρ=0.

Keywords: Multiple correlation; Regression; Effect size; Hypothesis testing; Computer program; SAS

Received: March 15, 2016; **Accepted:** March 31, 2016; **Published:** April 07, 2016

Suppose that a researcher is interested in predicting cocaine use frequency in adults from a number of putative risk factors such as sensation seeking tendencies, level of psychosocial stress, frequency of other illicit drug use, and number of negative life events. A commonly used statistical approach to modeling such data is linear multiple regression analysis, and when using regression it is standard practice to examine whether the squared multiple correlation coefficient, R^2 (i.e., the proportion of variance in the outcome variable accounted for by the predictors) is statistically significant. The intent of such a test is to determine whether R^2 differs significantly from zero, and the null hypothesis may be stated as H_o: $\rho^2=0$, where ρ^2 represents the population value (parameter) for the squared multiple correlation coefficient. Although this test is widely used, it is misleading because the expected value of R^2 is not zero when ρ=0 (where ρ represents the population value for the multiple correlation coefficient). Rather, as Morrison [1] pointed out, the expected value, or expected long-run mean, of R^2 is equal to p/n-1, where p is the number of predictor variables and n is the sample size. The implication of this equation is that R^2 should be examined in relation to the expected value of R^2, E(R^2), because the latter quantity is the value of R^2 that can be expected simply "by chance." In light of this realization, it seems more appropriate for researchers to test the null hypothesis, H_o: $\rho^2=\rho_o^2$, where $\rho_o^2=E(R^2)$. Unfortunately, commonly used statistical software programs such as IBM SPSS, Minitab, and SAS do not implement this adjusted hypothesis test whereby R^2 is tested against the expected value of R^2.

Huberty [2] recognized the importance of testing R^2 against its expected value and he proposed an adjusted R^2 index that takes into account the value of E(R^2). This adjusted R^2 index "corrects" the obtained squared multiple correlation coefficient by explicitly incorporating the expected value of R^2. Huberty [2] also presented an effect size measure for linear multiple regression studies that is calculated by subtracting E(R^2) from Huberty's adjusted R^2 index. Darlington [3] gave an F statistic for testing the null hypothesis that R^2 equals the expected value of R^2 (i.e., H_o: $\rho^2=\rho_o^2$). To calculate the aforementioned statistical quantities, I wrote a user-friendly computer program called DARHUB (short for Darlington and Huberty statistics). DARHUB calculates the expected value of R^2, Darlington's F statistic for testing the null hypothesis that $\rho^2=\rho_o^2$, the observed probability value for

Darlington's *F*, Huberty's adjusted R^2 index, and Huberty's effect size measure. The user specifies only the squared multiple correlation coefficient, R^2, from a linear regression analysis (i.e., the R^2 obtained by performing a regression analysis using SPSS, Minitab, SAS, etc.), the number of predictor variables, *p*, and the sample size, *n*. DARHUB is written in SAS data step language [4], and SAS is compatible with PC, Mac and Linux workstations. User documentation is included in the program listing. **Table 1** contains sample input, the full program listing (the SAS data step code), and output from the analysis **(Table 1)**. The DARHUB program and accompanying ReadMe file can be downloaded from the following academic website: http://psychology.cofc.edu/personal-pages/james-hittner,-ph.d.---software-page.php.

Table 1 The DARHUB SAS program with sample input and output.

***Input and program listing*:**

```
data R squared;

***  Below, specify the sample size "n", the number of predictor variables "p", and the
squared multiple correlation coefficient "Rsq"  ***;
n=60;
p=5;
Rsq=0.31;

*************************************************************************;
ERsq=((p / (n-1)));
a=((ERsq / (1-ERsq)));

******************Darlington's F below**********************************;
Dar_F=((Rsq / (1-Rsq)) * ((n-p-1) / p)) / (1 + (a * ((n-1) / p)));

*************************************************************************;
v=(((ERsq * (n-1)) + p))**2 / (((ERsq * (n-1) * (ERsq + 2)) + p));
Dar_p=((1 - probf(Dar_F, v, n-p-1)));

*************Huberty's adjusted Rsq value below***************************;
HubadRsq=((Rsq - ERsq) / (1 - ERsq));

*************Huberty's effect size index below****************************;
Hubefsiz=(HubadRsq - ERsq);

*************************************************************************;
proc print noobs;
   var Dar_F  Dar_p  ERsq  HubadRsq  Hubefsiz;

options nodate;
options nocenter;
title1
 'NOTE: Dar_F=F test reported in Darlington (1990) for testing the';
title2
 'null hypothesis that R-squared equals the expected value of R-squared.';
title3
 'Dar_p=Probability value for Dar_F.';
title4
 'ERsq=The expected value, or long-run mean, of R-squared under the null';
title5
 'hypothesis that rho=0.';
title6
 'HubadRsq=Adjusted R-squared value presented in Huberty (1994). Expressed';
title7
 'as a proportional-reduction-in-error or improvement-over-chance statistic.';
title8
 'Hubefsiz=Effect size index presented in Huberty (1994). Calculated by';
title9
 "subtracting the expected value of R-squared from Huberty's adjusted";
title10
 'R-squared value.';
run;
```

***Output*:**

NOTE: Dar_F=F test reported in Darlington (1990) for testing the
null hypothesis that R-squared equals the expected value of R-squared.
Dar_p=Probability value for Dar_F.
ERsq=The expected value, or long-run mean, of R-squared under the null
hypothesis that rho=0.
HubadRsq=Adjusted R-squared value presented in Huberty (1994). Expressed
as a proportional-reduction-in-error or improvement-over-chance statistic.
Hubefsiz=Effect size index presented in Huberty (1994). Calculated by
subtracting the expected value of R-squared from Huberty's adjusted
R-squared value.

Dar_F	Dar_p	ERsq	HubadRsq	Hubefsiz
2.31874	0.042009	0.084746	0.24611	0.16137

References

1 Morrison DF (1990) Multivariate statistical methods. McGraw-Hill, New York.

2 Huberty CJ (1994) A note on interpreting an R^2 value. Journal of Educational and Behavioral Statistics 19: 351-356.

3 Darlington RB (1990) Regression and linear models. McGraw-Hill, New York.

4 SAS Institute (1990) SAS/STAT User's Guide (Release 6.04). SAS, Cary, North Carolina.