Soc 63993, Homework #8 Answer Key: Structural Coefficients/R^2

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised March 28, 2015

Problems 1 and 2. The file *evilstnd.do* will generate the computer runs you need for this problem. Copy it from my web page. You will also need to install `esttab` from SSC if you haven't already.

This program contains two examples (which you will treat as problem 1 and problem 2). In each example, two regressions are run. Indicate what each example tells you about the problems that can arise if you focus on R2 and on standardized (path) rather than metric (structural) coefficients. Do your best to explain why these problems occur.

First, note that incorrect weighting causes the means, and the standard errors of the means, to be mis-estimated. The estimate of the mean of inc is too low with incorrect weighting, because the lower income of blacks is being weighted too heavily. The presence of random variability causes inc2 to have a larger standard error than does inc (which also implies that the variance of inc2 is greater). Incorrect weighting also results in a larger standard error (and larger variance) for white.

```
. * First, compute the means using the correct and
. * incorrect weights. Observe the differences.
. mean  inc inc2 white [pw=wgtright]

Mean estimation                     Number of obs    =    2400

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         inc |   4444.444   21.21481      4402.843    4486.046
        inc2 |   4444.444   35.49031       4374.85    4514.039
       white |   .8888889   .0040329      .8809805    .8967973
--------------------------------------------------------------

. mean  inc inc2 white [pw=wgtwrong]

Mean estimation                     Number of obs    =    2400

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
         inc |       4250   17.43405      4215.813    4284.187
        inc2 |       4250   28.43102      4194.248    4305.752
       white |         .5   .0102083      .4799819    .5200181
--------------------------------------------------------------
```

In problem 1, two regressions are estimated. In both cases, the model is correctly specified. However, in one of the regressions, cases have been incorrectly weighted. This is not that atypical of a situation: complicated or flawed sampling schemes, nonresponse, and other factors can cause researchers to be unsure what weights are correct. Or, a similar situation might occur in which there were two different populations which had different racial compositions, but the effect of race on income was the same in both populations. Or, within the same population, there might be changes in racial composition across time. Following is a side by side comparison of the correctly weighted and incorrectly weighted results:

```
. * Show the metric coefficients side by side
. esttab m1right m1wrong, mtitles r2 scalar(F)

--------------------------------------------
                      (1)             (2)   
                  m1right         m1wrong   
--------------------------------------------
white               500.0***          500***
                  (14.99)         (14.99)   

_cons              4000.0***         4000***
                 (169.63)        (169.63)   
--------------------------------------------
N                    2400            2400   
R-sq                0.036           0.086   
F                   224.8           224.8   
--------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

. * Show the standardized coefficients side by side
. esttab m1right m1wrong, mtitles r2 scalar(F) beta

--------------------------------------------
                      (1)             (2)   
                  m1right         m1wrong   
--------------------------------------------
white               0.189***        0.293***
                  (14.99)         (14.99)   
--------------------------------------------
N                    2400            2400   
R-sq                0.036           0.086   
F                   224.8           224.8   
--------------------------------------------
Standardized beta coefficients; t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

The incorrect weighting scheme results in incorrect estimates of R^2 and standardized coefficients. However, the metric coefficients are not affected. In a more “real world” example, they probably would be, but still, if the model is indeed perfectly specified, the metric coefficients would remain unbiased even under a flawed weighting scheme.

In this case, the differential weighting causes the variance of the IV, white, to differ between the two regressions. The variance is greater with incorrect weighting, which in turn causes R^2 to be greater. As Duncan notes, such a seemingly simple change in one variance can affect all of the underlying variances and covariances, even though the underlying structural model is the same. Standardized coefficients are not invariant against such changes in the IV variance, but structural coefficients are.

Obviously, incorrect weighting is undesirable and will lead to many bad consequences. But, if the model is good, the metric coefficients will not be as affected as other parameters (e.g. standardized coefficients and R^2) will be.

In problem 2, two regressions are again run. In the second regression, there is more random variability in the DV than there is in the first. This might occur because (a) there is more random measurement error in the 2nd DV (b) one population is more affected by random influences than is another (c) across time, random influences on a DV may increase in the population. Following is a side by side comparison:

```
. * Show the metric coefficients side by side
. esttab m2inc m2inc2, r2 scalar(F)

--------------------------------------------
                      (1)             (2)   
                      inc            inc2   
--------------------------------------------
white               500.0***        500.0***
                  (14.99)          (8.94)   

_cons              4000.0***       4000.0***
                 (169.63)        (101.11)   
--------------------------------------------
N                    2400            2400   
R-sq                0.036           0.013   
F                   224.8           79.86   
--------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

. * Show the standardized coefficients side by side
. esttab m2inc m2inc2, r2 scalar(F) beta

--------------------------------------------
                      (1)             (2)   
                      inc            inc2   
--------------------------------------------
white               0.189***        0.114***
                  (14.99)          (8.94)   
--------------------------------------------
N                    2400            2400   
R-sq                0.036           0.013   
F                   224.8           79.86   
--------------------------------------------
Standardized beta coefficients; t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

We see that higher levels of random variability result in smaller R^2 values, smaller standardized coefficients, and smaller values for the significance test statistics. However, the metric coefficients remain unchanged. Again, in a more “real world” example, they probably wouldn’t be identical, but the metric coefficients would be unbiased either way. Looking at the standardized coefficients alone would distort the underlying similarities between the two regressions.

When random variability is added to the DV, its variance increases. Because there is more random variance, the proportion of explained variance, R^2, goes down. Similarly, the standardized effects go down (because their computation involves the variance of the DV). Estimates of the underlying structural relationships are not biased by random measurement error in the DV though. When we talked about random measurement error earlier, we also noted that random error in the DV did not bias the metric coefficients but did result in larger standard errors and smaller T values. We are seeing the same thing now, with the added insight that standardized coefficients and other statistics are affected by random measurement error in the DV.

Problem 3. Present a substantive example, real or hypothetical, in which the value of R^2 may be high but inaccurate or misleading, e.g. the model is mis-specified in some way. If possible, cite a real example from the published literature, but otherwise make one up. If hypothetical, try to make the example "reasonable," i.e. the model might at first glance seem to make sense but on closer examination it really isn't correct.

I already gave my examples in class. But, I'll be happy to comment on yours if you want to bring them to me.