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

Richard Williams, University of Notre Dame, https://academicweb.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 R^2 and on standardized (path) rather than metric (structural) coefficients. Do your best to explain why these problems occur.

Here is a copy of *evilstnd.do*:

```
version 12.1
* Evils of standardization:
* This program illustrates problems with the use of standardized
* variables and standardized coefficients.

* Hypothetical data: a sample of 1200 blacks and 1200 whites.  Assume
* that equal-sized samples of blacks and whites have been drawn, but
* that in this population, whites outnumber blacks by 8 to 1.  Hence,
* blacks have been oversampled by a factor of 8, and when the groups
* are analyzed together, each black should therefore be weighted only
* 1/8 as heavily as each white.  In this hypothetical example,
* all individuals conveniently have incomes of $3000, $3500, $4000,
* $4500, $5000, or $5500, hence cases are grouped together.

* Or, as an alternative, assume that 2 populations have been sampled
* from.  In one population, whites outnumber blacks by 8 to 1, whereas
* in the other population it is a 50-50 split.

* The vars are:
* ncases - the number of cases having a particular combination of values
* inc - income
* inc2 - income with random error.  The change in income conveniently
*     works out to be $1100 either way.  The variability in income
*     increases as a result.
* white - 0 = Black, 1 = white

* NOTE: esttab (available from SSC) need to be installed.

clear all
input ncases inc white inc2
200   3000  0  1900
200   3000  0  4100
200   3500  1  2400
200   3500  1  4600
200   4000  0  2900
200   4000  0  5100
200   4500  1  3400
200   4500  1  5600
200   5000  0  3900
200   5000  0  6100
200   5500  1  4400
200   5500  1  6600
```

```
end
* Expand back to the original 2400 cases.
expand ncases
drop ncases
order inc inc2 white

* Compute the correct relative weight. In a real sample, the weights might
* be more like 1,000 for blacks and 8,000 for whites, which would mean
* that a white had 1 chance in 8,000 of being selected while a black had
* 1 chance in a 1000. Or, equivalently, that each black in the sample
* represented 1000 blacks while each white represented 8000 whites.
gen wgtright = 1 if !white
replace wgtright = 8 if white

* The incorrect weight ignores the oversampling of blacks, and weights
* all cases equally.
gen wgtwrong = 1

*               E X A M P L E   1
* What effect does improper weighting of cases have on
* coefficients in a perfectly specified model?

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

* Now, use the correct and incorrect weights in regressions.
* What differences does it make?
reg inc white [pw = wgtright], beta
est store m1right
reg inc white [pw = wgtwrong], beta
est store m1wrong
* Show the metric coefficients side by side
esttab m1right m1wrong, mtitles r2 scalar(F)
* Show the standardized coefficients side by side
esttab m1right m1wrong, mtitles r2 scalar(F) beta

*               E X A M P L E   2
* Suppose variability in the dependent variable increases on a random
* basis. What effect does an increase in residual
* variance have on metric and standardized coefficients?
* To examine this, we compare inc and inc2. inc2 adds random variation
* to inc. We will use the correct weights.

reg inc white [pw = wgtright], beta
est store m2inc
reg inc2 white [pw = wgtright], beta
est store m2inc2
* Show the metric coefficients side by side
esttab m2inc m2inc2, r2 scalar(F)
* Show the standardized coefficients side by side
esttab m2inc m2inc2, r2 scalar(F) beta
```

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.