Soc 63993, Homework #7:
Nonlinear effects/ Intro to path analysis

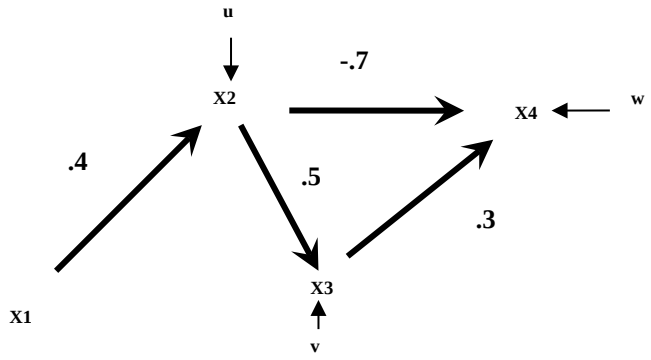
Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised February 20, 2015

Problem 1. The files *nonlinhw.do* and *nonlinhw.dta* will generate the computer runs you need for this problem. Copy them from the course web page. You will also need to install `curvefit`, available from SSC. (You will need to refer to curvefit's help file so you know what the functions are.) Run the program a few lines at a time; otherwise you will always be erasing your graphs.

There are 4 variables in nonlinhw.dta: X1 (the IV), Y1, Y2, and Y3 (the DVs). The Stata program does scatterplots of X1 versus each DV and then generates other graphs that model the nonlinear relationship. For each DV in turn, you are to do the following:

- Examine each scatterplot. Explain why the relationship is nonlinear and what type of nonlinearity appears to be present. Put another way, explain the rationale for the followup graph of the nonlinear relationship.
  - For Part I only, show a different set of Stata commands that could graph the nonlinear relationship.
  - For Parts I and II only, show how the same models could be estimated using the `regress` and/or `glm` commands.
  - For Y3 only, two different Curvefits are presented (Parts III and IV). Explain why, based on the graphics only, it would be difficult to decide which nonlinear specification was most appropriate, and how theory might help you to choose.
- Discuss what problems result from a linear (mis)specification. The graphs will help you here.
- For Parts I, II, III, present a substantive example, real or hypothetical, that the model you have estimated might be appropriate for. Explain why it is appropriate. Do not use any of the examples already given in class.

Problem 2. A sociologist believes that the following model describes the relationships between X1, X2, X3 and X4. All variables are in standardized form. The hypothesized value of each path is included in the diagram.



a. Write out the structural equation for each endogenous variable, using both the names for the paths (e.g. β_{42}) and the estimated value of the path coefficient.

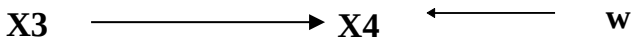
b. Part of the correlation matrix is shown below. Determine the complete correlation matrix. (Remember, variables are standardized. You can use either normal equations or Sewell Wright, but you might want to use both as a double-check.)

	x1	x2	x3	x4
x1	1.0000			
x2	0.4000	1.0000		
x3	?	?	1.0000	
x4	-0.2200	?	?	1.0000

c. Decompose the correlation between X3 and X4 into

- Correlation due to direct effects
- Correlation due to indirect effects
- Correlation due to common causes

d. Suppose the above model is correct, but instead the researcher believed in and estimated the following model:



What conclusions would the researcher likely draw? In particular, what would the researcher conclude about the effect of changes in X3 on X4? Why would he make these mistakes? Discuss the consequences of this mis-specification.

e. [Optional] Confirm your answer to 2b using Stata, i.e. create a pseudo-replication of the data using corr2data and then use one of the methods described in the notes for making sure that you can reproduce the estimates of the path coefficients given in the diagram.