

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

Richard Williams, University of Notre Dame, <https://academicweb.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.

First off, here is *nonlinhw.do*:

```
version 12.1
use https://academicweb.nd.edu/~rwilliam/statafiles/nonlinhw.dta, clear

***** Part 1.
* Plot of x1 with y1
estimates clear
scatter y1 x1, scheme(sj)
curvefit y1 x1, f(1 4)
* HW: Show another way to graph this model
* HW: Show how to estimate this model using regress and/or glm

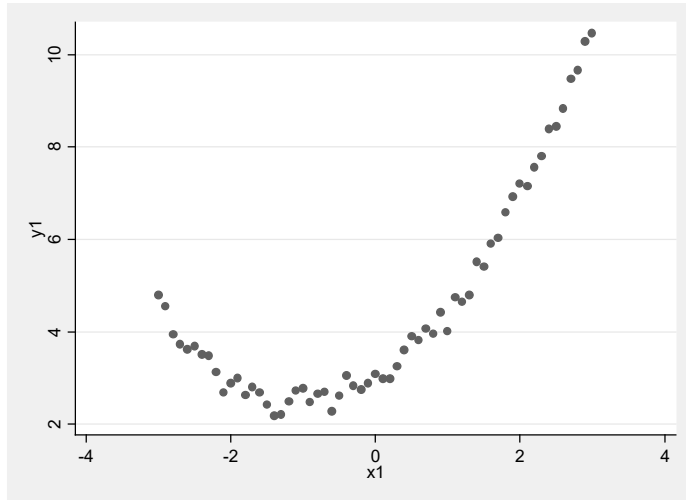
***** Part 2.
* Plot of x1 with y2
estimates clear
scatter y2 x1
curvefit y2 x1, f(1 0)
* HW: Show how to estimate this model using regress and/or glm

***** Part 3.
* Plot of x1 with y3.
estimates clear
scatter y3 x1
mkspline xle0 0 xgt0 = x1, marginal
reg y3 x1
predict linear
reg y3 xle0 xgt0
predict spline
scatter y3 x1 || line linear x1 || line spline x1, sort scheme(sj)

***** Part 4.
* As this shows, a polynomial model would also be plausible for y3.
* In practice, it is often hard to tell just from the scatterplot what
* transformation is best, so theory is important.
estimates clear
curvefit y3 x1, f(1 4)
```

The scatterplot of X1 with Y1 is

```
. use https://academicweb.nd.edu/~rwilliam/statafiles/nonlinhw.dta, clear
. ***** Part 1.
. * Plot of x1 with y1
. estimates clear
. scatter y1 x1, scheme(sj)
```



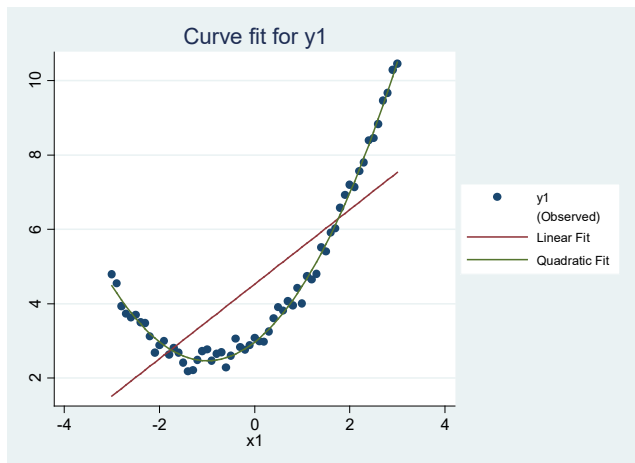
The U-shaped, curvilinear form suggests that a polynomial model is called for. There appears to be one “bend” so the model should include terms for $X1$ and $X1^2$. The `curvefit` command estimates the linear and quadratic models and generates the graph.

```
. curvefit y1 x1, f(1 4)
```

Curve Estimation between y1 and x1

Variable		Linear	Quadratic
b0	_cons	4.5251426	2.966442
		24.71	78.56
		0.0000	0.0000
b1	_cons	1.00287	1.00287
		9.64	70.16
		0.0000	0.0000
b2	_cons		.50280663
			55.37
			0.0000
Statistics			
N		61	61
r2_a		.60516076	.99254275

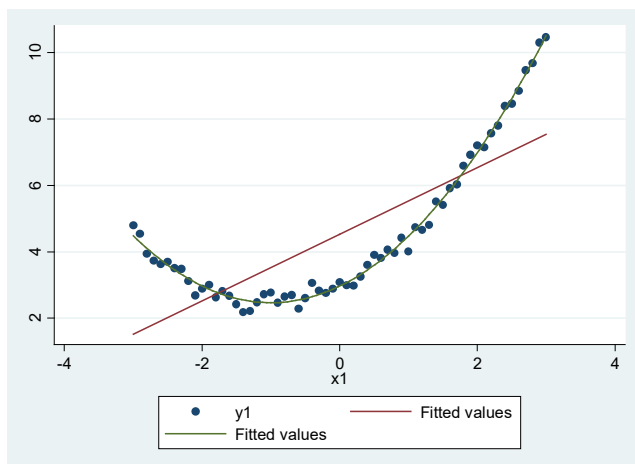
legend: b/t/p



As we can see, the linear model (where Y1 is regressed only on X1) at first underestimates several values of y1, then overestimates, then goes back to underestimating them. By way of contrast, the quadratic model (where Y1 is regressed on X1 and $X1^2$) matches the observed data almost perfectly.

We can also graph the relationship using these Stata commands:

```
. scatter y1 x1 || lfit y1 x1 || qfit y1 x1
```



To estimate the quadratic model using the `regress` command (which gives the same estimates that `curvefit` did),

```
. reg y1 x1 c.x1#c.x1
```

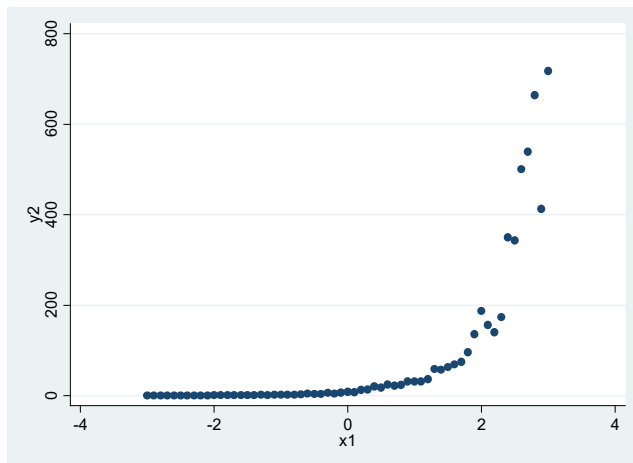
Source	SS	df	MS	Number of obs = 61		
Model	308.65331	2	154.326655	F(2, 58) = 3993.93		
Residual	2.24113791	58	.038640309	Prob > F = 0.0000		
Total	310.894448	60	5.18157413	R-squared = 0.9928		
				Adj R-squared = 0.9925		
				Root MSE = .19657		

y1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x1	1.00287	.0142947	70.16	0.000	.9742561	1.031484
c.x1#c.x1	.5028066	.0090808	55.37	0.000	.4846294	.5209838
_cons	2.966442	.037761	78.56	0.000	2.890855	3.042029

Such a model might explain the link between political ideology and political activism: the more extreme one is in his or her political views (in either direction), the more likely he or she is to be politically active. Those in the middle of the road are least active. The tendency is more pronounced for right-wing ideologues than for left-wing.

For Y2, the plot with X1 is

```
. ***** Part 2.
. * Plot of x1 with y2
. estimates clear
. scatter y2 x1
```



This suggests exponential growth. The points increase slowly at first, and then grow by larger and larger amounts. Plotting a linear and exponential model with `curvefit`,

```
. curvefit y2 x1, f(1 0)
```

Curve Estimation between y2 and x1

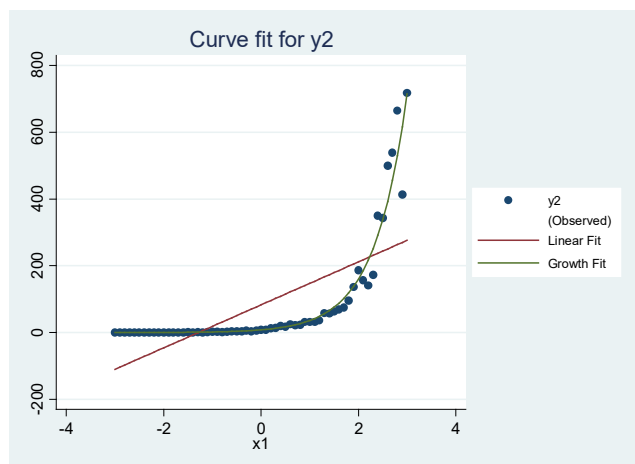
Variable		Linear	Growth

b0			
	_cons	83.005824	2.083302
		5.34	8.83
		0.0000	0.0000

b1			
	_cons	64.366035	1.496315
		7.29	17.34
		0.0000	0.0000

Statistics			
	N	61	61
	r2_a	.46518882	.9519786

legend: b/t/p			



As `curvefit` shows, if we just use a linear model with Y2 dependent, we will first underestimate the values of y2, then over-estimate them, then go back to underestimating them again.

We could estimate a model where we computed the log of y2 and regressed it on x. The potential problem with this approach is that the log of 0 is undefined; ergo, any cases with 0 (or for that matter negative) values will get dropped from the analysis. Further, most of us don't think in terms of logs of variables; we would rather see how X is related to the unlogged Y. It is therefore often better to estimate this model:

$$E(Y) = e^{(\alpha + \beta X)}$$

When you do this, Y itself can equal 0; all that is required is that its expected value be greater than zero. In Stata, we can estimate this as a *generalized linear model* with link log. The command and results are

```
. glm y2 x1, link(log)
```

```
Generalized linear models              No. of obs   =          61
Optimization      : ML                Residual df   =          59
                                      Scale parameter = 1631.748
Deviance          = 96273.1221         (1/df) Deviance = 1631.748
Pearson           = 96273.1221         (1/df) Pearson  = 1631.748

Variance function: V(u) = 1           [Gaussian]
Link function     : g(u) = ln(u)      [Log]

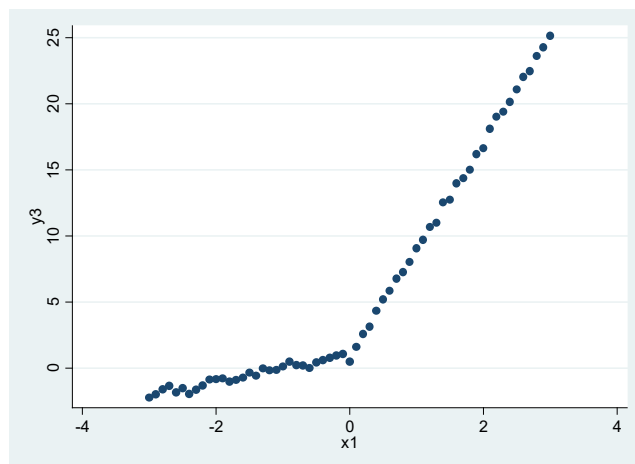
Log likelihood    = -311.1594035       AIC            = 10.26752
                                      BIC            = 96030.58
```

```
-----+-----
              |               OIM
              |               Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
              |
      x1      |   1.496316   .0824864   18.14   0.000   1.334645   1.657986
      _cons   |   2.0833    .2256518    9.23    0.000   1.641031   2.52557
-----+-----
```

Exponential/Growth models are good for variables such as income or sales that we expect to increase by percentage rather than absolute amounts. For example, we might predict that each dollar increase in cost would reduce sales by 5%.

For Y3, the plot is

```
. ***** Part 3.
. * Plot of x1 with y3.
. estimates clear
. scatter y3 x1
```



Here, there seem to be two pieces to the data. For $X < 0$, there is a small slope. Then, the slope becomes much greater. Ergo, a piecewise regression seems called for, and the `mkspline` command will make that possible:

```
. mkspline xle0 0 xgt0 = x1, marginal
. reg y3 x1
```

Source	SS	df	MS	Number of obs = 61		
Model	3852.18882	1	3852.18882	F(1, 59)	=	386.41
Residual	588.18379	59	9.96921678	Prob > F	=	0.0000
Total	4440.37261	60	74.0062102	R-squared	=	0.8675
				Adj R-squared	=	0.8653
				Root MSE	=	3.1574

y3	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x1	4.513444	.2296068	19.66	0.000	4.054001	4.972886
_cons	6.329963	.4042645	15.66	0.000	5.521032	7.138895

```
. predict linear
(option xb assumed; fitted values)
```

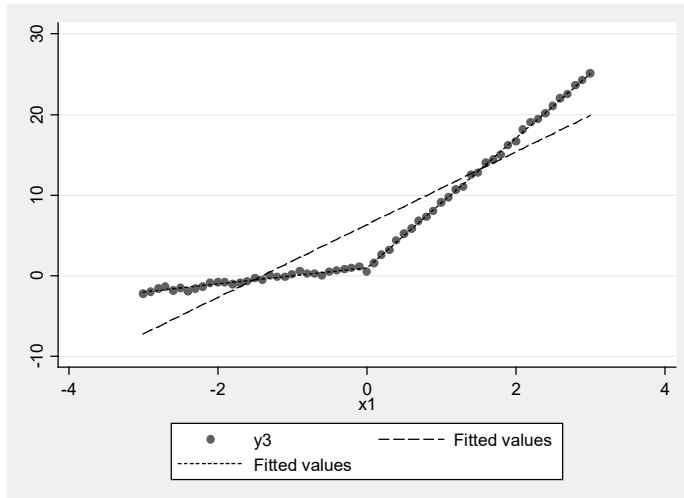
The linear model is a pretty good fit, but we can do better.

```
. reg y3 xle0 xgt0
```

Source	SS	df	MS	Number of obs = 61		
Model	4437.30505	2	2218.65252	F(2, 58)	=	41949.15
Residual	3.06756763	58	.052889097	Prob > F	=	0.0000
Total	4440.37261	60	74.0062102	R-squared	=	0.9993
				Adj R-squared	=	0.9993
				Root MSE	=	.22998

y3	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
xle0	.9967843	.0373837	26.66	0.000	.9219526	1.071616
xgt0	7.033319	.0668686	105.18	0.000	6.899467	7.167171
_cons	.968499	.0588672	16.45	0.000	.8506636	1.086334

```
. predict spline
(option xb assumed; fitted values)
. scatter y3 x1 || line linear x1 || line spline x1, sort scheme(sj)
```



As the graph shows, if we just estimate a linear model, we will first underestimate y_3 , then overestimate, then underestimate again. The fit of the piecewise model is near perfect in this case. Substantive example: Years of college education may have much more impact on earnings than years of elementary education.

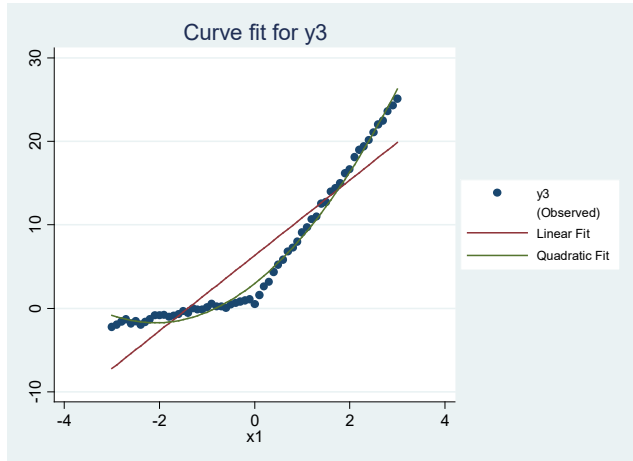
Finally, we present an alternative curvefit for Y_3 . Instead of doing piecewise regression, we fit a quadratic model:

```
. estimates clear
. curvefit y3 x1, f(1 4)
```

Curve Estimation between y_3 and x_1

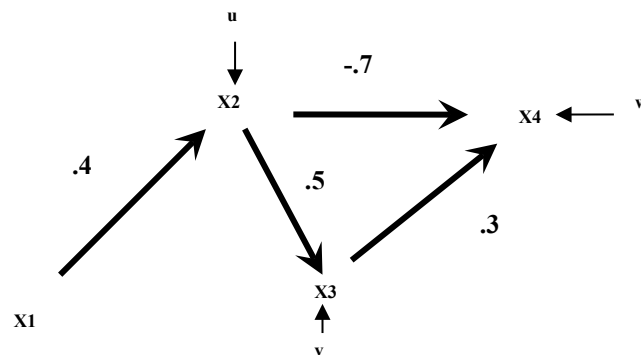
Variable		Linear	Quadratic
b0	_cons	6.3299633	2.9758679
		15.66	18.74
		0.0000	0.0000
b1	_cons	4.5134436	4.5134436
		19.66	75.09
		0.0000	0.0000
b2	_cons		1.0819662
			28.33
			0.0000
Statistics			
N		61	61
r2_a		.86529216	.99076762

legend: b/t/p



As we see, the quadratic model also seems to provide a very good fit to the observed data. Unless relationships are extremely strong, plots of the data may reveal that nonlinearity is present, but won't necessarily make it obvious what the best solution is. Theory should guide you as you attempt to determine what solution is most appropriate.

Problem 2. A sociologist believes that the following model describes the relationships between X_1 , X_2 , X_3 and X_4 . 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.

$$X_2 = \beta_{21}X_1 + u = .4X_1 + u$$

$$X_3 = \beta_{32}X_2 + v = .5X_2 + v$$

$$X_4 = \beta_{42}X_2 + \beta_{43}X_3 + w = -.7X_2 + .3X_3 + w$$

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

Here is the complete correlation matrix:

	x1	x2	x3	x4
x1	1.0000			
x2	0.4000	1.0000		
x3	0.2000	0.5000	1.0000	
x4	-0.2200	-0.5500	-0.0500	1.0000

To confirm, using normal equations (in this case though, it may be easier just to look at the diagram and use Sewell Wright):

$$\rho_{31} = \beta_{31} + \beta_{21}\beta_{32} = 0 + .4 \cdot .5 = .20$$

$$\rho_{32} = \beta_{32} + \beta_{31}\beta_{21} = .5 + 0 \cdot .4 = .5$$

$$\rho_{42} = \beta_{42} + \beta_{32}\beta_{43} + \beta_{41}\beta_{21} + \beta_{43}\beta_{31}\beta_{21} = -.7 + .5 \cdot .3 + 0 \cdot .4 + .3 \cdot 0 \cdot .4 = -.55$$

$$\rho_{43} = \beta_{43} + \beta_{41}\beta_{31} + \beta_{42}\beta_{32} + \beta_{41}\beta_{21}\beta_{32} + \beta_{42}\beta_{21}\beta_{31} = .3 + 0 \cdot 0 + -.7 \cdot .5 + 0 \cdot .4 \cdot .5 + -.7 \cdot .4 \cdot 0 = -.05$$

c. (5 pts) Decompose the correlation between X3 and X4 into

- Correlation due to direct effects

.3

- Correlation due to indirect effects

0

- Correlation due to common causes

-.35

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.

In the mis-specified model, the standardized coefficient will be the same as the bivariate correlation, -.05. The researcher will therefore conclude that increases in X3 lead to decreases in X4. However, in the correctly specified model, we see that the direct effect of X3 on X4 is .3, i.e. increases in X3 lead to increases in X4. By failing to take into account the common cause, X2, the research will not only mis-estimate the magnitude but also the direction of the effect of X3

on X4. If policy issues were involved, policy makers might do exactly the opposite of what they should do.

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.

We can use `pathreg` or `sem` (`pathreg` from UCLA must be installed),

```
. matrix input corr = (1,.4,.2,-.22\ .4,1,.5,-.55\ .2,.5,1,-.05\ -.22,-.55,-.05,1)
. corr2data x1 x2 x3 x4, n(100) corr(corr)
(obs 100)
. pathreg (x2 x1) (x3 x1 x2) (x4 x1 x2 x3)
```

x2	Coef.	Std. Err.	t	P> t	Beta
x1	.4	.092582	4.32	0.000	.4
_cons	3.07e-09	.0921179	0.00	1.000	.

n = 100 R2 = 0.1600 sqrt(1 - R2) = 0.9165

x3	Coef.	Std. Err.	t	P> t	Beta
x1	4.27e-09	.0959412	0.00	1.000	4.27e-09
x2	.5	.0959412	5.21	0.000	.5
_cons	1.36e-09	.0874908	0.00	1.000	.

n = 100 R2 = 0.2500 sqrt(1 - R2) = 0.8660

x4	Coef.	Std. Err.	t	P> t	Beta
x1	-8.76e-09	.0883883	-0.00	1.000	-8.76e-09
x2	-.7	.1	-7.00	0.000	-.7
x3	.3	.0935414	3.21	0.002	.3
_cons	-8.66e-09	.0806032	-0.00	1.000	.

n = 100 R2 = 0.3700 sqrt(1 - R2) = 0.7937

```
. sem (x2 <- x1) (x3 <- x1 x2) (x4 <- x1 x2 x3)
```

Endogenous variables

Observed: x2 x3 x4

Exogenous variables

Observed: x1

Fitting target model:

```
Structural equation model          Number of obs      =          100
Estimation method = ml
Log likelihood      = -519.3618
```

		OIM				
		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Structural						
x2 <-						
	x1	.4	.0916515	4.36	0.000	.2203663 .5796337
	_cons	3.07e-09	.0911921	0.00	1.000	-.1787332 .1787332
x3 <-						
	x2	.5	.0944911	5.29	0.000	.3148008 .6851992
	x1	4.27e-09	.0944911	0.00	1.000	-.1851992 .1851992
	_cons	1.36e-09	.0861684	0.00	1.000	-.168887 .168887
x4 <-						
	x2	-.7	.0979796	-7.14	0.000	-.8920364 -.5079635
	x3	.3	.0916515	3.27	0.001	.1203663 .4796337
	x1	-8.76e-09	.0866025	-0.00	1.000	-.1697379 .1697379
	_cons	-8.66e-09	.0789747	-0.00	1.000	-.1547875 .1547875
Variance						
	e.x2	.8316	.117606			.6302842 1.097217
	e.x3	.7425	.1050054			.5627537 .9796581
	e.x4	.6237	.0882045			.4727131 .8229128
LR test of model vs. saturated: chi2(0) = 0.00, Prob > chi2 = .						

The following post-estimation command after `sem` can confirm our estimates of direct, indirect and total effects (but not correlation due to common cause, alas):

Direct effects

		OIM				
		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Structural						
x2 <-						
	x1	.4	.0916515	4.36	0.000	.2203663 .5796337
x3 <-						
	x2	.5	.0944911	5.29	0.000	.3148008 .6851992
	x1	4.27e-09	.0944911	0.00	1.000	-.1851992 .1851992
x4 <-						
	x2	-.7	.0979796	-7.14	0.000	-.8920364 -.5079635
	x3	.3	.0916515	3.27	0.001	.1203663 .4796337
	x1	-8.76e-09	.0866025	-0.00	1.000	-.1697379 .1697379

Indirect effects

		Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
Structural							
x2 <-	x1	0	(no path)				
x3 <-	x2	0	(no path)				
	x1	.2	.0594018	3.37	0.001	.0835746	.3164253
x4 <-	x2	.15	.0283473	5.29	0.000	.0944402	.2055598
	x3	0	(no path)				
	x1	-.22	.066453	-3.31	0.001	-.3502455	-.0897545

Total effects

		Coef.	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
Structural							
x2 <-	x1	.4	.0916515	4.36	0.000	.2203663	.5796337
x3 <-	x2	.5	.0944911	5.29	0.000	.3148008	.6851992
	x1	.2	.0979796	2.04	0.041	.0079635	.3920365
x4 <-	x2	-.55	.1019979	-5.39	0.000	-.7499122	-.3500878
	x3	.3	.0916515	3.27	0.001	.1203663	.4796337
	x1	-.22	.09755	-2.26	0.024	-.4111945	-.0288055

The complete Stata code for problem 2 is

```

version 12.1
* Homework 7, Path analysis problem
clear all
matrix input corr = (1,.4,.2,-.22\ .4,1,.5,-.55\ .2,.5,1,-.05\ -.22,-.55,-.05,1)
corr2data x1 x2 x3 x4, n(100) corr(corr)
corr
pathreg (x2 x1) (x3 x1 x2) (x4 x1 x2 x3)
sem (x2 <- x1) (x3 <- x1 x2) (x4 <- x1 x2 x3)
estat teffects

```