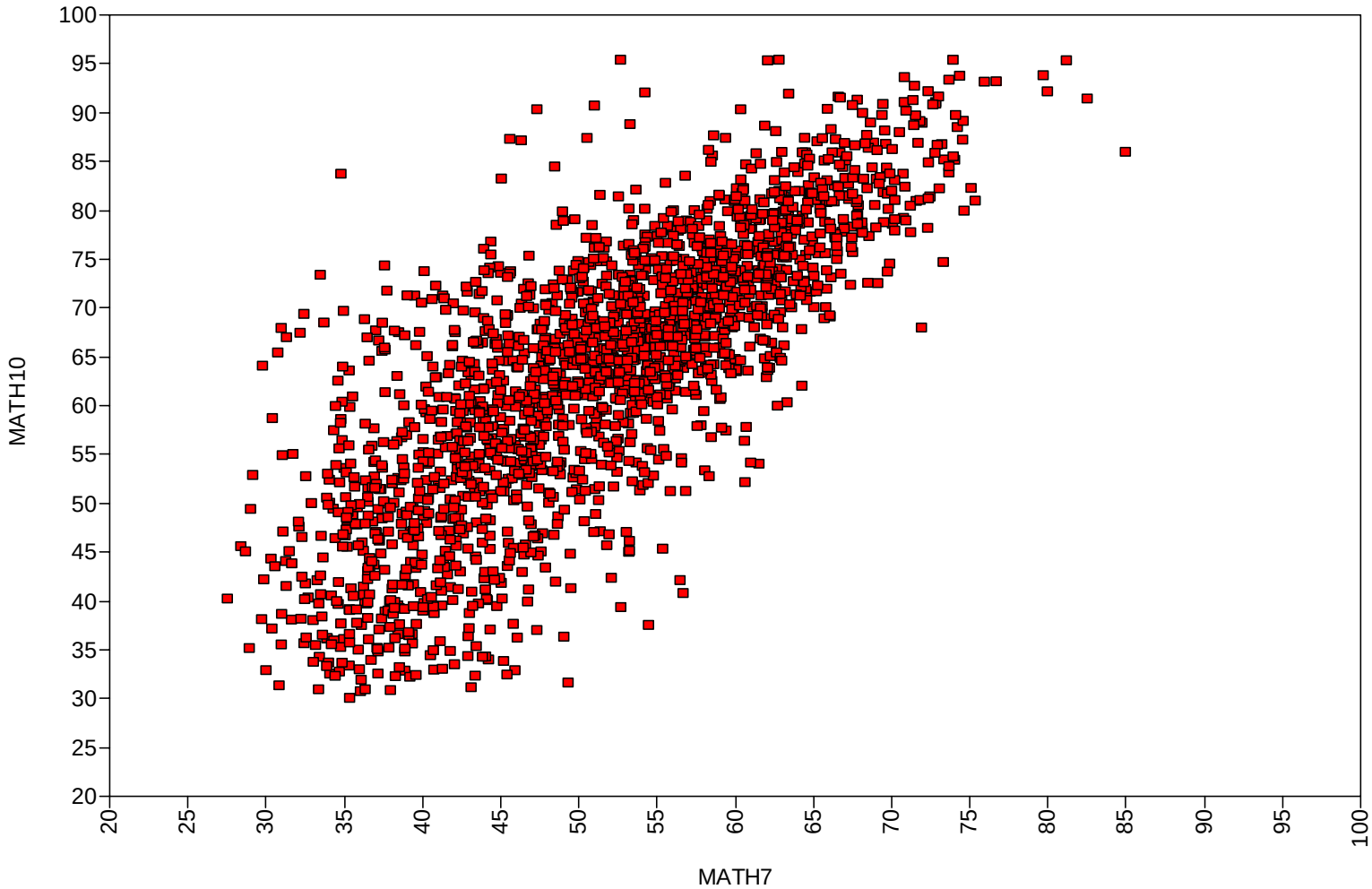
Regression Analysis

LSAY Math Regression



Regression Analysis

Regression model:

$$y_i = \alpha + \beta\, x_i + \varepsilon_i\,, \qquad (1)$$

$$E(\varepsilon_i \mid x_i) = E(\varepsilon_i) = E(\varepsilon) = 0 \text{ (x and ε uncorrelated)}, \qquad (2)$$

$$V(\varepsilon_i \mid x_i) = V(\varepsilon_i) = V(\varepsilon) \text{ (constant variance)}. \qquad (3)$$

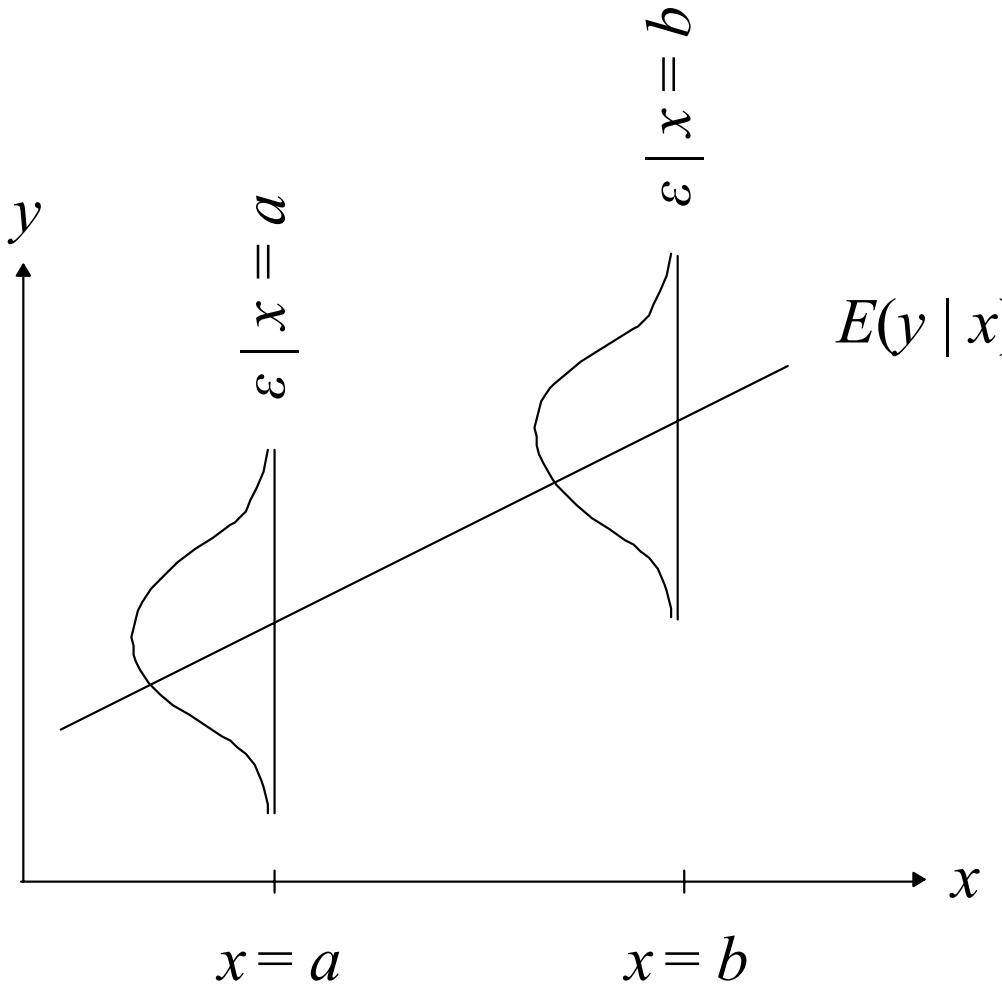
For inference and ML estimation, we also assume ε normal.

The model implies

$$E(y \mid x) = \alpha + \beta\, x \qquad \text{(conditional expectation function)}$$

$$V(y \mid x) = V(\varepsilon) \qquad \text{(homoscedasticity)}$$

Regression Analysis (Continued)



$$E(y \mid x) = \alpha + \beta x$$

Regression Analysis (Continued)

Population formulas:

$$y_i = \alpha + \beta\, x_i + \varepsilon_i\,, \tag{1}$$

$$E(y) = E(\alpha) + E(\beta\, x) + E(\varepsilon)$$

$$= \alpha + \beta\, E(x) \tag{2}$$

$$V(y) = V(\alpha) + V(\beta\, x) + V(\varepsilon)$$

$$= \beta^2\, V(x) + V(\varepsilon) \tag{3}$$

$$Cov(y, x) = E[y - E(y)]\,[x - E(x)] = \beta\, V(x) \tag{4}$$

$$R^2 = \beta^2\, V(x) \,/\, (\beta^2\, V(x) + V(\varepsilon)) \tag{5}$$

$$Stdyx\ \beta = \beta\, \frac{SD(x)}{SD(y)} \tag{6}$$

Regression Analysis (Continued)

The model has 3 parameters: α, β, and $V(\varepsilon)$
Note: $E(x)$ and $V(x)$ are not model parameters

Formulas for ML and OLS parameter estimates based on a random sample

$$\hat{\beta} = s_{yx} / s_{xx}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

$$\hat{V}(\varepsilon) = s_{yy} - \hat{\beta}^2 s_{xx}$$

Prediction

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}\,x_i$$

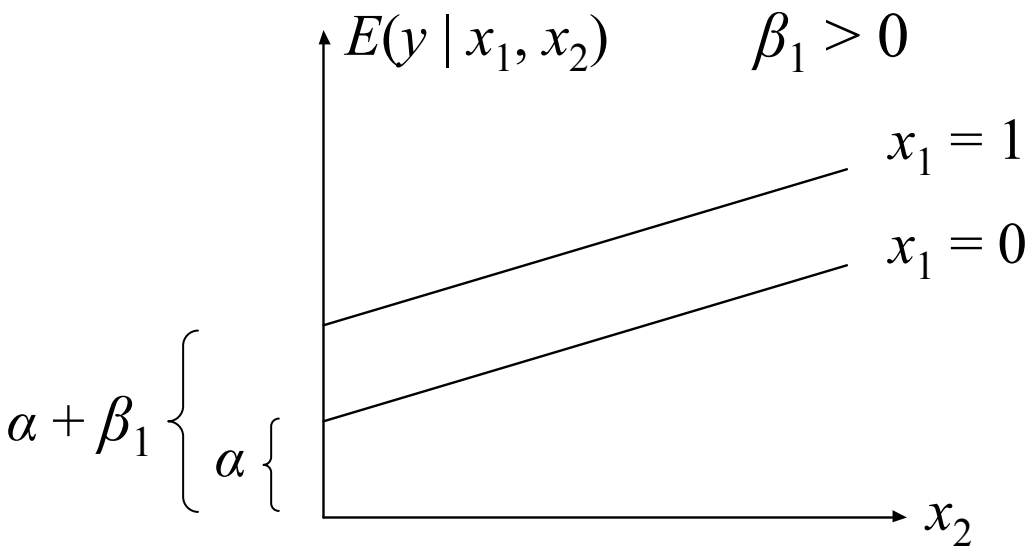
Regression Analysis (Continued)

x_1 0/1 dummy variable (e.g. gender), x_2 continuous variable

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

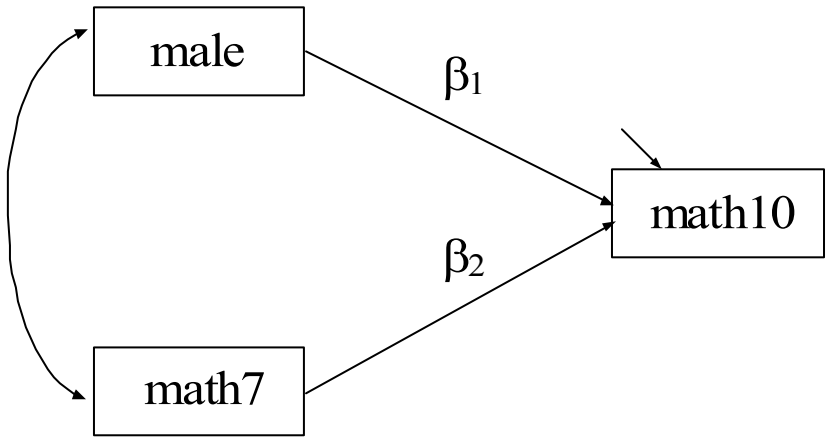
$$E(y \mid x_1 = 0, x_2) = \alpha + \beta_2 x_2$$

$$E(y \mid x_1 = 1, x_2) = \underbrace{\alpha + \beta_1}_{\text{intercept}} + \beta_2 x_2$$



Analogous to ANCOVA

Regression Of LSAY Math10 On Gender And Math7



Parameter estimates are produced for the intercept, the two slopes, and the residual variance.

Note: Variances and covariance for male and math7 are not part of the model

Input For Regression Of Math10 On Gender And Math7

```
TITLE:      Regressing math10 on math7 and gender
DATA:       FILE = dropout.dat;
            FORMAT = 11f8 6f8.2 1f8 2f8.2 10f2;
VARIABLE:   NAMES ARE id school gender mothed fathed fathsei ethnic
            expect pacpush pmpush homeres math7 math8 math9 math10
            math11 math12 problem esteem mathatt clocatn dlocatn
            elocatn flocatn glocatn hlocatn ilocatn jlocatn
            klocatn llocatn;
            MISSING = mothed (8) fathed (8)  fathsei (996 998)
                 ethnic (8) homeres (98) math7-math12 (996 998);
            USEVAR = math7 math10 male;
DEFINE:     male = gender - 1; ! male is a 0/1 variable created from
                             ! gender = 1/2 where 2 is male
MODEL:      math10 ON male math7;
ANALYSIS:   TYPE = MISSING;
OUTPUT:     TECH1 SAMPSTAT STANDARDIZED;
PLOT:       TYPE = PLOT1;
```

Output Excerpts For Regression Of Math10 On Gender And Math7

Estimated Sample Statistics

Means

	MATH10	MATH7	MALE
1	62.423	50.378	0.522

Covariances

	MATH10	MATH7	MALE
MATH10	186.926		
MATH7	109.826	103.950	
MALE	-0.163	-0.334	0.250

Correlations

	MATH10	MATH7	MALE
MATH10	1.000		
MATH7	0.788	1.000	
MALE	-0.024	-0.066	1.000

Output Excerpts For Regression Of Math10 On Gender And Math7 (Continued)

Model Results

	Estimates	S.E.	Est./S.E.	Std	StdYX
MATH10 ON					
MALE	0.763	0.374	2.037	0.763	0.028
MATH7	1.059	0.018	57.524	1.059	0.790
Intercepts					
MATH10	8.675	0.994	8.726	8.675	0.635
Residual Variances					
MATH10	70.747	2.225	31.801	70.747	0.378

R-SQUARE

Observed Variable	R-Square
MATH10	0.622

Further Readings On Regression Analysis

Agresti, A. & Finlay, B. (1997). Statistical methods for the social sciences. Third edition. New Jersey: Prentice Hall.

Amemiya, T. (1985). Advanced econometrics. Cambridge, Mass.: Harvard University Press.

Hamilton, L.C. (1992). Regression with graphics. Belmont, CA: Wadsworth.

Johnston, J. (1984). Econometric methods. Third edition. New York: McGraw-Hill.

Lewis-Beck, M. S. (1980). Applied regression: An introduction. Newbury Park, CA: Sage Publications.

Moore, D.S. & McCabe, G.P. (1999). Introduction to the practice of statistics. Third edition. New York: W.H. Freeman and Company.

Pedhazur, E.J. (1997). Multiple regression in behavioral research. Third Edition. New York: Harcourt Brace College Publishers.

Path Analysis

Path Analysis

Used to study relationships among a set of observed variables

- Estimate and test direct and indirect effects in a system of regression equations
- Estimate and test theories about the absence of relationships

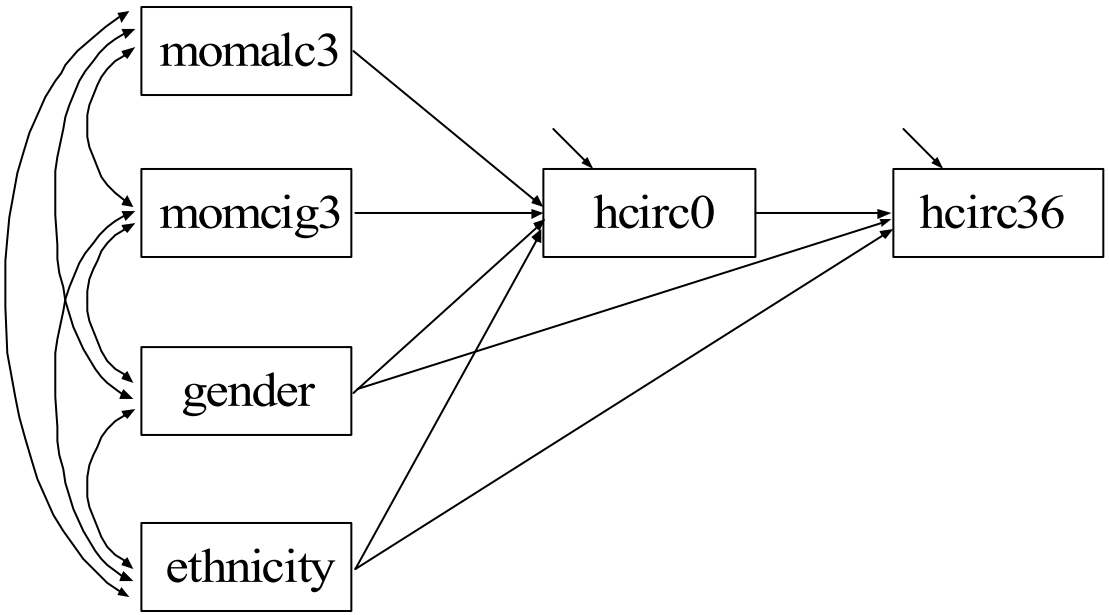
Maternal Health Project (MHP) Data

The data are taken from the Maternal Health Project (MHP). The subjects were a sample of mothers who drank at least three drinks a week during their first trimester plus a random sample of mothers who used alcohol less often.

Mothers were measured at the fourth and seventh month of pregnancy, at delivery, and at 8, 18, and 36 months postpartum. Offspring were measured at 0, 8, 18 and 36 months.

Variables for the mothers included: demographic, lifestyle, current environment, medical history, maternal psychological status, alcohol use, tobacco use, marijuana use, and other illicit drug use. Variables for the offspring included: head circumference, height, weight, gestational age, gender, and ethnicity.

Data for the analysis include mother's alcohol and cigarette use in the third trimester and the child's gender, ethnicity, and head circumference both at birth and at 36 months.


momalc3
momcig3
gender
ethnicity
hcirc0
hcirc36

Input For Maternal Health Project Path Analysis

```
TITLE:      Maternal health project path analysis

DATA:       FILE IS headalln.dat;
            FORMAT IS 1f8.2 47f7.2;

VARIABLE:   NAMES ARE id weight0 weight8 weight18 weigh36
            height0 height8 height18 height36 hcirc0 hcirc8
            hcirc18 hcirc36 momalc1 momalc2 momalc3 momalc8
            momalc18 momalc36 momcig1 momcig2 momcig3 momcig8
            momcig18 momcig36 gender eth momht gestage age8
            age18 age36 esteem8 esteem18 esteem36 faminc0
            faminc8 faminc18 faminc36 momdrg36 gravid sick8
            sick18 sick36 advp advm1 advm2 advm3;

            MISSING = ALL (999);

            USEV = momalc3 momcig3 hcirc0 hcirc36 gender eth;

            USEOBS = id NE 1121 AND NOT (momalc1 EQ 999 AND
                     momalc2 EQ 999 AND momalc3 EQ 999);
```

Input For Maternal Health Project
Path Analysis (Continued)

```
DEFINE:     hcirc0  = hcirc1/10;
            hcirc36 = hcirc36/10;
            momalc3 = log(momalc3 +1);

ANALYSIS:   TYPE = MISSING H1;

MODEL:      hcirc36 ON hcirc0 gender eth;
            hcirc0 ON momalc3 momcig3 gender eth;

OUTPUT:     SAMPSTAT STANDARDIZED;
```

Output Excerpts Maternal Health Project Path Analysis

Tests Of Model Fit

Chi-Square Test of Model Fit

Value	1.781	
Degrees of Freedom	2	
P-Value	.4068	

RMSEA (Root Mean Square Error Of Approximation)

Estimate	.000	
90 Percent C.I.	.000	0.079
Probability RMSEA <= .05	.774	

Output Excerpts Maternal Health Project Path Analysis (Continued)

Model Results

	Estimates	S.E.	Est./S.E.	Std	StdYX
HCIRC36 ON					
HCIRC0	.415	.036	11.382	.415	.439
GENDER	.762	.107	7.146	.762	.270
ETH	-.094	.107	-.879	-.094	-.033
HCIRC0 ON					
MOMALC3	-.500	.239	-2.090	-.500	-.084
MOMCIG3	-.013	.005	-2.604	-.013	-.108
GENDER	.495	.118	4.185	.495	.166
ETH	.578	.125	4.625	.578	.194

Output Excerpts Maternal Health Project Path Analysis (Continued)

Residual Variances					
HCIRC0	2.043	.119	17.107	2.043	.920
HCIRC36	1.385	.087	15.844	1.385	.697
Intercepts					
HCIRC0	33.729	.112	301.357	33.729	22.629
HCIRC36	35.338	1.227	28.791	35.338	25.069

R-Square

Observed Variable	R-Square
HCIRC0	.080
HCIRC36	.303

The MODEL INDIRECT Command

MODEL INDIRECT is used to request indirect effects and their standard errors. Delta method standard errors are computed as the default.

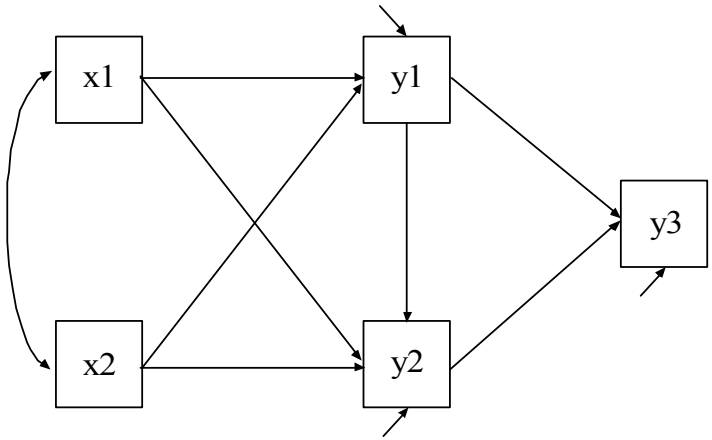
The BOOTSTRAP option of the ANALYSIS command can be used to obtain bootstrap standard errors for the indirect effects.

The STANDARDIZED option of the OUTPUT command can be used to obtain standardized indirect effects.

The MODEL INDIRECT Command (Continued)

The CINTERVAL option of the OUTPUT command can be used to obtain confidence intervals for the indirect effects and the standardized indirect effects. Three types of 95% and 99% confidence intervals can be obtained: symmetric, bootstrap, or bias-corrected bootstrap confidence intervals. The bootstrapped distribution of each parameter estimate is used to determine the bootstrap and bias-corrected bootstrap confidence intervals. These intervals take non-normality of the parameter estimate distribution into account. As a result, they are not necessarily symmetric around the parameter estimate.

The MODEL INDIRECT Command (Continued)



MODEL INDIRECT has two options:

- IND – used to request a specific indirect effect or a set of indirect effects
- VIA – used to request a set of indirect effects that includes specific mediators

MODEL INDIRECT

y3 IND y1 x1;	!x1 -> y1 -> y3
y3 IND y2 x2;	!x2 -> y2 -> y3
y3 IND x1;	!x1 -> y1 -> y3
	!x1 -> y2 -> y3
	!x1 -> y1 -> y2 -> y3
y3 VIA y2 x1;	!x1 -> y2 -> y3
	!x1 -> y1 -> y2 -> y3

Further Readings On Path Analysis

MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G. & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. Psychological Methods, 7, 83-104.

MacKinnon, D.P., Lockwood, C.M. & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. Multivariate Behavioral Research, 39, 99-128.

Shrout, P.E. & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. Psychological Methods, 7, 422-445.