NELS 88

Table 2.3 – Adjusted odds ratios of eighth-grade students in 1988 performing below basic levels of reading and mathematics in 1988 and dropping out of school, 1988 to 1990, by basic demographics

Variable	Below basic mathematics	Below basic reading	Dropped out
Sex			
Female vs. male	0.77**	0.70**	0.86
Race — ethnicity			
Asian vs. white	0.84	1.46**	0.60
Hispanic vs. white	1.60**	1.74**	1.12
Black vs. white	1.77**	2.09**	1.45
Native American vs. white	2.02**	2.87**	1.64
Socioeconomic status			
Low vs. middle	1.68**	1.66**	3.74**
High vs. middle	0.49**	0.44**	0.41*

45

Latent Response Variable Formulation Versus Probability Curve Formulation

Probability curve formulation in the binary u case:

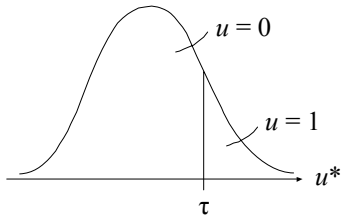
$$P(u = 1 \mid x) = F(\beta_0 + \beta_1 x), \tag{67}$$

where F is the standard normal or logistic distribution function.

Latent response variable formulation defines a threshold τ on a continuous u^* variable so that $u = 1$ is observed when u^* exceeds τ while otherwise $u = 0$ is observed,

$$u^* = \gamma x + \delta, \tag{68}$$

where $\delta \sim N(0, V(\delta))$.



46

Latent Response Variable Formulation Versus Probability Curve Formulation (Continued)

$$P(u = 1 \mid x) = P(u^* > \tau \mid x) = 1 - P(u^* \leq \tau \mid x) = \quad (69)$$

$$= 1 - \Phi[(\tau - \gamma x)\, V(\delta)^{-1/2}] = \Phi[(-\tau + \gamma x)\, V(\delta)^{-1/2}]. \quad (70)$$

Standardizing to $V(\delta) = 1$ this defines a probit model with intercept $(\beta_0) = -\tau$ and slope $(\beta_1) = \gamma$.

Alternatively, a logistic density may be assumed for δ,

$$f[\delta\,;\, 0, \pi^2/3] = dF/d\delta = F(1 - F), \quad (71)$$

where in this case F is the logistic distribution function $1/(1 + e^{-\delta})$.

47

Latent Response Variable Formulation: R^2, Standardization, And Effects On Probabilities

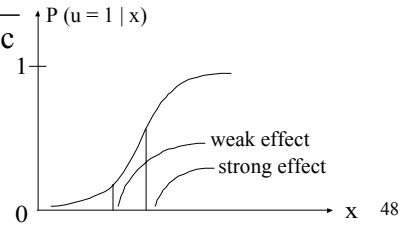
$$u^* = \gamma x + \delta$$

- $R^2(u^*) = \gamma^2 V(x) / (\gamma^2 V(x) + c)$, where $c = 1$ for probit and $\pi^2 / 3$ for logit (McKelvey & Zavoina, 1975)
- Standardized γ refers to the effect of x on u^*,

$$\hat{\gamma}_s = \hat{\gamma}\, SD(x) / SD(u^*),$$

$$SD(u^*) = \sqrt{\hat{\gamma}^2\, V(x) + c}$$

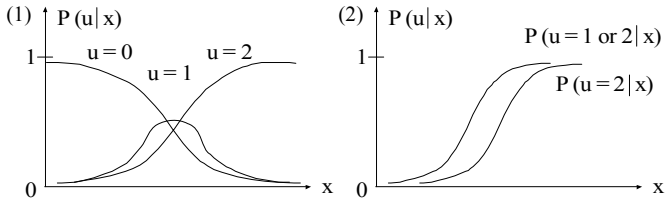
- Effect of x on $P(u = 1)$ depends on x value

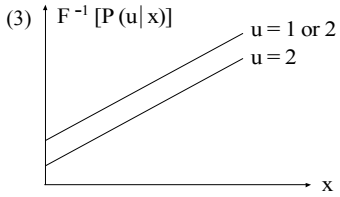


48

Modeling With An Ordered Polytomous u Outcome

u polytomous with 3 categories

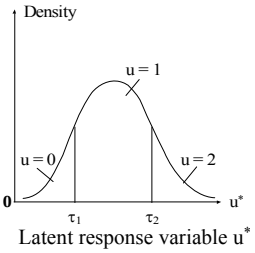




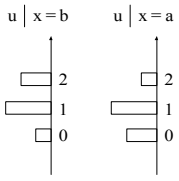
Proportional odds model (Agresti, 2002)

49

Ordered Polytomous Outcome Using A Latent Response Variable Formulation

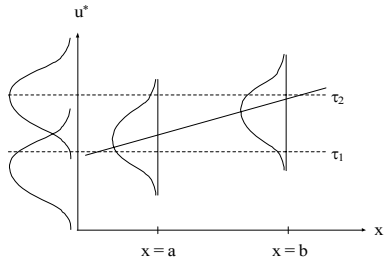


Latent response variable u*



Latent response variable regression:

$$u_i^* = \gamma x_i + \delta_i$$



50

Ordered Polytomous Outcome Using A Latent Response Variable Formulation (Continued)

A categorical variable u with C ordered categories,

$$u = c, \text{ if } \tau_{j,c} < u^* \leq \tau_{j,c+1} \quad (72)$$

for categories $c = 0, 1, 2, \ldots, C-1$ and $\tau_0 = -\infty$, $\tau_C = \infty$.

Example: a single x variable and a u variable with three categories. Two threshold parameters, τ_1 and τ_2.

Probit:

$$u^* = \gamma x + \delta, \text{ with } \delta \text{ normal} \quad (73)$$

$$P(u = 0 \mid x) = \Phi(\tau_1 - \gamma x), \quad (74)$$

$$P(u = 1 \mid x) = \Phi(\tau_2 - \gamma x) - \Phi(\tau_1 - \gamma x), \quad (75)$$

$$P(u = 2 \mid x) = 1 - \Phi(\tau_2 - \gamma x) = \Phi(-\tau_2 + \gamma x). \quad (76)$$

51

Ordered Polytomous Outcome Using A Latent Response Variable Formulation (Continued)

$$P(u = 1 \text{ or } 2 \mid x) = P(u = 1 \mid x) + P(u = 2 \mid x) \quad (77)$$

$$= 1 - \Phi(\tau_1 - \gamma x) \quad (78)$$

$$= \Phi(-\tau_1 + \gamma x) \quad (79)$$

$$= 1 - P(u = 0 \mid x), \quad (80)$$

that is, a linear probit for,

$$P(u = 2 \mid x) = \Phi(-\tau_2 + \gamma x), \quad (81)$$

$$P(u = 1 \text{ or } 2 \mid x) = \Phi(-\tau_1 + \gamma x). \quad (82)$$

Note: same slope γ, so parallel probability curves

52

Logit For Ordered Categorical Outcome

$$P(u = 2 \mid x) = \frac{1}{1 + e^{-(\beta_2 + \beta x)}}, \quad (83)$$

$$P(u = 1 \text{ or } 2 \mid x) = \frac{1}{1 + e^{-(\beta_1 + \beta x)}}. \quad (84)$$

Log odds for each of these two events is a linear expression,

$$logit\,[P(u = 2 \mid x)] = \quad (85)$$

$$= log[P(u = 2 \mid x)/\,(1 - P(u = 2 \mid x))] = \beta_2 + \beta x, \quad (86)$$

$$logit\,[P(u = 1 \text{ or } 2 \mid x)] = \quad (87)$$

$$= log[P(u = 1 \text{ or } 2 \mid x)/\,(1 - P(u = 1 \text{ or } 2 \mid x))] = \beta_1 + \beta x. \quad (88)$$

Note: same slope β, so parallel probability curves

53

Logit For Ordered Categorical Outcome (Continued)

When x is a 0/1 variable,

$$logit\,[P(u = 2 \mid x = 1)] - logit\,[P(u = 2 \mid x = 0)] = \beta \quad (89)$$

$$logit\,[P(u = 1 \text{ or } 2 \mid x = 1)] - logit\,[P(u = 1 \text{ or } 2 \mid x = 0)] = \beta \quad (90)$$

showing that the ordered polytomous logistic regression model has constant odds ratios for these different outcomes.

54

Alcohol Consumption: Ordered Polytomous Regression

u: "On the days that you drink, how many drinks do you have per day, on the average?"

Ordinal u: ("Alameda Scoring")
0 non-drinker
1 1-2 drinks per day
2 3-4 drinks per day
3 5 or more drinks per day

x's: Age: whole years 20 – 64

Income: 1 ≤ $4,999
2 $5,000 – $9,999
3 $10,000 – $14,999
4 $15,000 – $24,999
5 ≥ $25,000

N = 713 Males with regular physical activity levels

Source: Golden (1982), Muthén (1993)

55

Alcohol Consumption: Ordered Polytomous Regression (Continued)

$$P(u = 0 \mid x) = \Phi\,(\tau_1 - \gamma' x) \qquad (11)$$

$$P(u = 1 \mid x) = \Phi\,(\tau_2 - \gamma' x) - \Phi\,(\tau_1 - \gamma' x)\,,$$

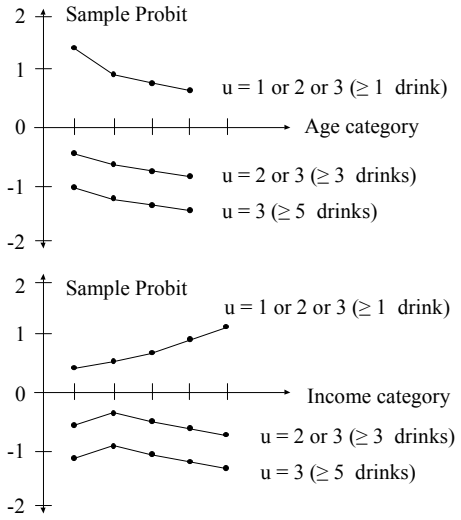
$$P(u = 2 \mid x) = \Phi\,(\tau_3 - \gamma' x) - \Phi\,(\tau_2 - \gamma' x)\,,$$

$$P(u = 3 \mid x) = \Phi\,(-\tau_3 + \gamma' x).$$

Ordered *u* gives a single slope

56

Alcohol Consumption: Ordered Polytomous Regression (Continued)



57

Polytomous Outcome: Unordered Case

Multinomial logistic regression:

$$P(u_i = c \mid x_i) = \frac{e^{\beta_{0c} + \beta_{1c} x_i}}{\sum_{k=1}^{K} e^{\beta_{0k} + \beta_{1k} x_i}}, \qquad (91)$$

for $c = 1, 2, \ldots, K$, where we standardize to

$$\beta_{0K} = 0, \qquad (92)$$
$$\beta_{1K} = 0, \qquad (93)$$

which gives the log odds

$$log[P(u_i = c \mid x_i) / P(u_i = K \mid x_i)] = \beta_{0c} + \beta_{1c} x_i, \qquad (94)$$

for $c = 1, 2, \ldots, K - 1$.

58

Multinomial Logistic Regression
Special Case Of K = 2

$$P(u_i = 1 \mid x_i) = \frac{e^{\beta_{01} + \beta_{11} x_i}}{e^{\beta_{01} + \beta_{11} x_i} + 1}$$

$$= \frac{e^{-(\beta_{01} + \beta_{11} x_i)}}{e^{-(\beta_{01} + \beta_{11} x_i)}} * \frac{e^{\beta_{01} + \beta_{11} x_i}}{e^{\beta_{01} + \beta_{11} x_i} + 1}$$

$$= \frac{1}{1 + e^{-(\beta_{01} + \beta_{11} x_i)}}$$

which is the standard logistic regression for a binary outcome.

59

Input For Multinomial Logistic Regression

```
TITLE:          multinomial logistic regression
DATA:           FILE = nlsy.dat;
VARIABLE:       NAMES = u x1-x3;
                NOMINAL = u;
MODEL:          u ON x1-x3;
```

60

Output Excerpts
Multinomial Logistic Regression:
4 Categories Of ASB In The NLSY

		Estimates	S.E.	Est./S.E.
U#1	ON			
AGE94		-.285	.028	-10.045
MALE		2.578	.151	17.086
BLACK		.158	.139	1.141
U#2	ON			
AGE94		.069	.022	3.182
MALE		.187	.110	1.702
BLACK		-.606	.139	-4.357
U#3	ON			
AGE94		-.317	.028	-11.311
MALE		1.459	.101	14.431
BLACK		.999	.117	8.513
Intercepts				
U#1		-1.822	.174	-10.485
U#2		-.748	.103	-7.258
U#3		-.324	.125	-2.600

61

Estimated Probabilities
For Multinomial Logistic Regression:
4 Categories Of ASB In The NLSY

Example 1: x's = 0

	exp	probability = exp/sum
log odds (u=1) = -1.822	0.162	0.069
log odds (u=2) = -0.748	0.473	0.201
log odds (u=3) = -0.324	0.723	0.307
log odds (u=4) = 0	1.0	0.424
sum	2.358	1.001

62

Estimated Probabilities For Multinomial Logistic Regression: 4 Categories Of ASB In The NLSY (Continued)

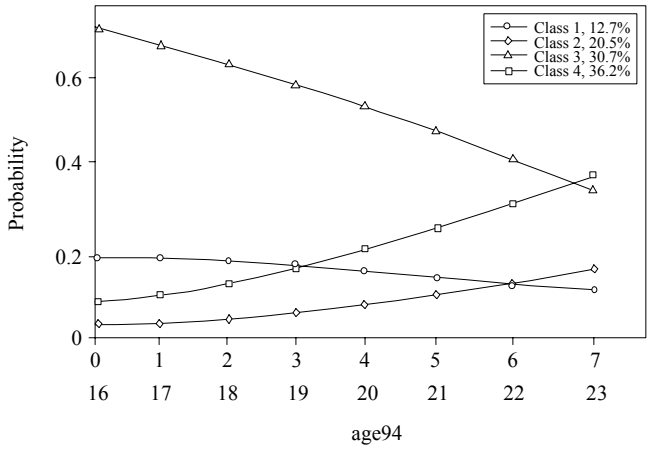
Example 2: x = 1, 1, 1

```
log odds (u=1) = -1.822 + (-0.285*1) + (2.578*1) + (0.158*1
               = 0.629
log odds (u=2) = -0.748 + 0.069*1 + 0.187*1 + (-0.606*1)
               = -1.098
log odds (u=3) = -0.324 + (-0.317*1) + 1.459*1 + 0.999*1
               = 1.817
```

	exp	probability = exp/sum
log odds (u=1) = 0.629	1.876	0.200
log odds (u=2) = -1.098	0.334	0.036
log odds (u=3) = 1.817	6.153	0.657
log odds (u=4) = 0	1.0	0.107
sum	9.363	1.000

63

Estimated Probabilities For Multinomial Logistic Regression: 4 Categories Of ASB In The NLSY (Continued)



64

Censored-Normal (Tobit) Regression

$y^* = \pi_0 + \pi x + \delta$ $\qquad$ $V(\delta)$ identifiable

Continuous – unlimited: $\quad y = y^*$

Continuous-censored: $\quad y = \begin{cases} c_L, \text{ if } y^* \le c_L \\ y^*, \text{ if } c_L < y^* < c_U \\ c_U, \text{ if } y^* \ge c_U \end{cases}$

Censoring from below, $c_L = 0$, $c_U = \infty$:

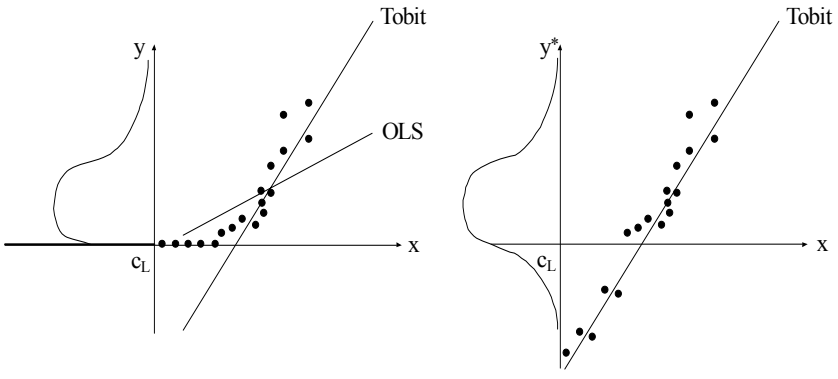
$$P(y > 0 \mid x) = F\left(\frac{\pi_0 + \pi x}{\sqrt{V(\delta)}}\right) \quad \text{(Probit Regression)}$$

$$E(y \mid y > 0, x) = \pi_0 + \pi x + f/F\sqrt{V(\delta)}$$

Classical Tobit

65

OLS v. Tobit Regression For Censored y But Normal y^*



66

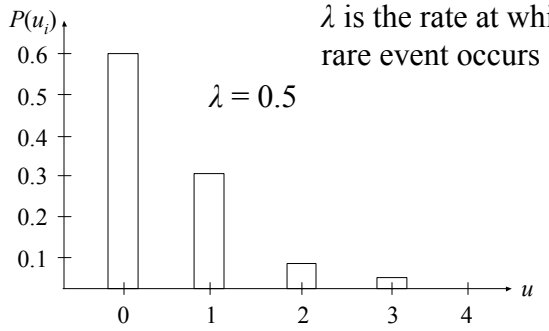
Regression With A Count Dependent Variable

67

Poisson Regression

A Poisson distribution for a count variable u_i has

$$P(u_i = r) = \frac{\lambda_i^r \, e^{-\lambda_i}}{r!}, \text{ where } u_i = 0, 1, 2, \ldots$$



λ is the rate at which a rare event occurs

Regression equation for the log rate:

$$^{e} \log \lambda_i = \ln \lambda_i = \beta_0 + \beta_1 x_i$$

68

Zero-Inflated Poisson (ZIP) Regression

A Poisson variable has mean = variance.

Data often have variance > mean due to preponderance of zeros.

$\pi = P$ (being in the zero class where only $u = 0$ is seen)

$1 - \pi = P$ (not being in the zero class with u following a Poisson distribution)

A mixture at zero:

$$P(u = 0) = \pi + (1 - \pi) \underbrace{e^{-\lambda}}_{\text{Poisson part}}$$

The ZIP model implies two regressions:

$$\text{logit}\,(\pi_i) = \gamma_0 + \gamma_1 x_i\,,$$

$$\ln \lambda_i = \beta_0 + \beta_1 x_i$$

69

Negative Binomial Regression

Unobserved heterogeneity ε_i is added to the Poisson model

$$ln\,\lambda_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,,\ \text{where}\ exp\,(\varepsilon) \sim \Gamma$$

Poisson assumes	Negative binomial assumes
$E(u_i \mid x_i) = \lambda_i$	$E(u_i \mid x_i) = \lambda_i$
$V(u_i \mid x_i) = \lambda_i$	$V(u_i \mid x_i) = \lambda_i(1 + \lambda_i\,\alpha)$

NB with $\alpha = 0$ gives Poisson. When the dispersion parameter $\alpha > 0$, the NB model gives substantially higher probability for low counts and somewhat higher probability for high counts than Poisson.

Further variations are zero-inflated NB and zero-truncated NB (hurdle model or two-part model).

70

Mplus Specifications

Variable command	Type of dependent variable	Variance/ residual variance
CATEGORICAL = u;	Binary, ordered polytomous	No
NOMINAL = u;	Unordered, polytomous (nominal)	No
CENSORED = y (b);	Censored normal (Tobit)	Yes
= y (a);	Censored from below or above	
COUNT = u;	Poisson	No
= u (i);	Zero-inflated Poisson	No

71

Further Readings On Censored and Count Regressions

Hilbe, J. M. (2007). Negative binomial regression. Cambridge, UK: Cambridge University Press.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34, 1-13.

Long, S. (1997). Regression models for categorical and limited dependent variables. Thousand Oaks: Sage.

Maddala, G.S. (1983). Limited-dependent and qualitative variables in econometrics. Cambridge: Cambridge University Press.

Tobin, J (1958). Estimation of relationships for limited dependent variables. Econometrica, 26, 24-36.

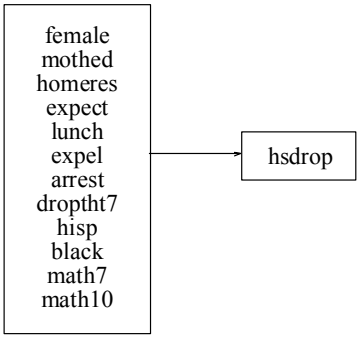
72

Path Analysis With Categorical Outcomes

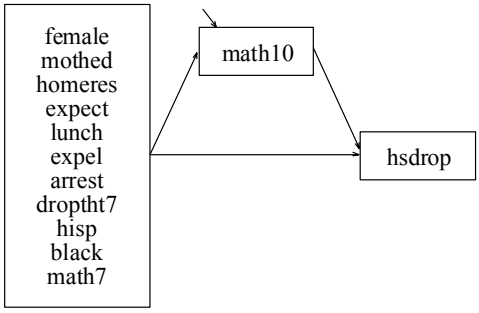
73

Path Analysis With A Binary Outcome And A Continuous Mediator With Missing Data

Logistic Regression



Path Model



74

Input For A Path Analysis With A Binary Outcome And A Continuous Mediator With Missing Data Using Monte Carlo Integration

```
TITLE:     Path analysis with a binary outcome and a continuous
           mediator with missing data using Monte Carlo integration
DATA:      FILE = lsaydropout.dat;
VARIABLE:  NAMES ARE female mothed homeres math7 math10 expel arrest
           hisp black hsdrop expect lunch droptht7;
           MISSING = ALL(9999);
           CATEGORICAL = hsdrop;
ANALYSIS:  ESTIMATOR = ML;
           INTEGRATION = MONTECARLO(500);
MODEL:     hsdrop ON female mothed homeres expect math7 math10 lunch
           expel arrest droptht7 hisp black;
           math10 ON female mothed homeres expect math7
           lunch expel arrest droptht7 hisp black;
OUTPUT:    PATTERNS STANDARDIZED TECH1 TECH8;
```

75

Output Excerpts Path Analysis With A Binary Outcome And A Continuous Mediator With Missing Data Using Monte Carlo Integration

```
           MISSING DATA PATTERNS FOR Y
                          1     2
MATH10                    x
FEMALE                    x     x
MOTHED                    x     x
HOMERES                   x     x
MATH7                     x     x
EXPEL                     x     x
ARREST                    x     x
HISP                      x     x
BLACK                     x     x
EXPECT                    x     x
LUNCH                     x     x
DROPTHT7                  x     x

     MISSING DATA PATTERN FREQUENCIES FOR Y
          Pattern     Frequency       Pattern     Frequency
                1          1639             2           574
```

76

Output Excerpts Path Analysis With A Binary Outcome And A Continuous Mediator With Missing Data Using Monte Carlo Integration (Continued)

Tests Of Model Fit

Loglikelihood

H0 Value	-6323.175

Information Criteria

Number of Free Parameters	26
Akaike (AIC)	12698.350
Bayesian (BIC)	12846.604
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	12763.999

77

Output Excerpts Path Analysis With A Binary Outcome And A Continuous Mediator With Missing Data Using Monte Carlo Integration (Continued)

Model Results

	Estimates	S.E.	Est./S.E.	Std	StdYX
HSDROP ON					
FEMALE	0.336	0.167	2.012	0.336	0.080
MOTHED	-0.244	0.101	-2.421	-0.244	-0.117
HOMERES	-0.091	0.054	-1.699	-0.091	-0.072
EXPECT	-0.225	0.063	-3.593	-0.225	-0.147
MATH7	-0.012	0.015	-0.831	-0.012	-0.058
MATH10	-0.031	0.011	-2.816	-0.031	-0.201
LUNCH	0.005	0.004	1.456	0.005	-0.053
EXPEL	1.010	0.216	4.669	1.010	0.129
ARREST	0.033	0.314	0.105	0.033	0.003
DROPTHT7	0.679	0.272	2.499	0.679	0.067
HISP	-0.145	0.265	-0.548	-0.145	-0.019
BLACK	0.038	0.234	0.163	0.038	0.006

78

Output Excerpts Path Analysis With A Binary Outcome And A Continuous Mediator With Missing Data Using Monte Carlo Integration (Continued)

	Estimates	S.E.	Est./S.E.	Std	StdYX
MATH10 ON					
FEMALE	-0.973	0.410	-2.372	-0.973	-0.036
MOTHED	0.343	0.219	1.570	0.343	0.026
HOMERES	0.486	0.140	3.485	0.486	0.059
EXPECT	1.014	0.166	6.111	1.014	0.103
MATH7	0.928	0.023	39.509	0.928	0.687
LUNCH	-0.039	0.011	-3.450	-0.039	-0.059
EXPEL	-1.404	0.851	-1.650	-1.404	-0.028
ARREST	-3.337	1.093	-3.052	-3.337	-0.052
DROPTHT7	-1.077	1.070	-1.007	-1.077	-0.016
HISP	-0.644	0.744	-0.866	-0.644	-0.013
BLACK	-0.809	0.694	-1.165	-0.809	-0.019

79

Output Excerpts Path Analysis With A Binary Outcome And A Continuous Mediator With Missing Data Using Monte Carlo Integration (Continued)

	Estimates	S.E.	Est./S.E.	Std	StdYX
Intercepts					
MATH10	10.941	1.269	8.621	10.941	0.809
Thresholds					
HSDROP$1	-1.207	0.521	-2.319		
Residual Variances					
MATH10	65.128	2.280	28.571	65.128	0.356

Observed Variable	R-Square
HSDROP	0.255
MATH10	0.644

80

Path Analysis Of Occupational Destination

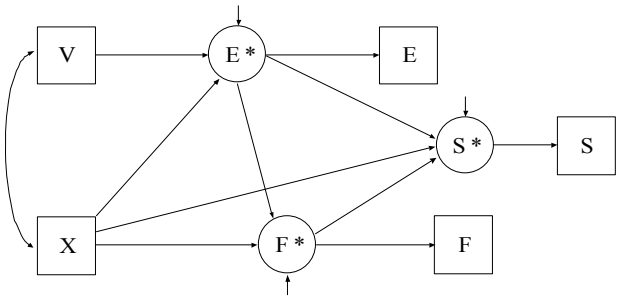


Figure 3: Structural Modeling of the Occupational Destination of Scientist or Engineer, Model 1

Reference: Xie (1989)
Data source: 1962 OCG Survey. The sample size is 14,401.
V: Father's Education. X: Father's Occupation (SEI)

81

Path Analysis Of Occupational Destination (Continued)

Table 2. Descriptive Statistics of Discrete Dependent Variables

Variable	Code	Meaning	Percent
S: Current Occupation	0	Non-scientific/engineering	96.4
	1	Scientific/engineering	3.6
F: First Job	0	Non-scientific/engineering	98.3
	1	Scientific/engineering	1.7
E: Education	0	0-7 years	13.4
	1	8-11 years	32.6
	2	12 years	29.0
	3	13 and more years	25.0

82

Differences Between Weighted Least Squares And Maximum Likelihood Model Estimation For Categorical Outcomes In Mplus

- Probit versus logistic regression
  - Weighted least squares estimates probit regressions
  - Maximum likelihood estimates logistic or probit regressions
- Modeling with underlying continuous variables versus observed categorical variables for categorical outcomes that are mediating variables
  - Weighted least squares uses underlying continuous variables
  - Maximum likelihood uses observed categorical outcomes

83

Differences Between Weighted Least Squares And Maximum Likelihood Model Estimation For Categorical Outcomes In Mplus (Continued)

- Delta versus Theta parameterization for weighted least squares
  - Equivalent in most cases
  - Theta parameterization needed for models where categorical outcomes are predicted by categorical dependent variables while predicting other dependent variables
- Missing data
  - Weighted least squares allows missingness predicted by covariates
  - Maximum likelihood allows MAR
- Testing of nested models
  - WLSMV uses DIFFTEST
  - Maximum likelihood (ML, MLR) uses regular or special approaches

84

Further Readings On Path Analysis With Categorical Outcomes

MacKinnon, D.P., Lockwood, C.M., Brown, C.H., Wang, W., & Hoffman, J.M. (2007). The intermediate endpoint effect in logistic and probit regression. Clinical Trials, 4, 499-513.

Xie, Y. (1989). Structural equation models for ordinal variables. Sociological Methods & Research, 17, 325-352.

85

Categorical Observed And Continuous Latent Variables

86