Mplus Short Courses
Topic 2

Regression Analysis, Exploratory Factor Analysis, Confirmatory Factor Analysis, And Structural Equation Modeling For Categorical, Censored, And Count Outcomes

Linda K. Muthén
Bengt Muthén

www.statmodel.com

1

Table Of Contents

2

Table Of Contents (Continued)

3

Mplus Background

- Inefficient dissemination of statistical methods:
  - Many good methods contributions from biostatistics, psychometrics, etc are underutilized in practice
- Fragmented presentation of methods:
  - Technical descriptions in many different journals
  - Many different pieces of limited software
- Mplus: Integration of methods in one framework
  - Easy to use: Simple, non-technical language, graphics
  - Powerful: General modeling capabilities
- Mplus versions
  - V1: November 1998
  - V2: February 2001
  - V3: March 2004
  - V4: February 2006
  - V5: November 2007
- Mplus team: Linda & Bengt Muthén, Thuy Nguyen, Tihomir Asparouhov, Michelle Conn, Jean Maninger

4

Statistical Analysis With Latent Variables A General Modeling Framework

Statistical Concepts Captured By Latent Variables

Continuous Latent Variables

- Measurement errors
- Factors
- Random effects
- Frailties, liabilities
- Variance components
- Missing data

Categorical Latent Variables

- Latent classes
- Clusters
- Finite mixtures
- Missing data

5

Statistical Analysis With Latent Variables A General Modeling Framework (Continued)

Models That Use Latent Variables

Continuous Latent Variables

- Factor analysis models
- Structural equation models
- Growth curve models
- Multilevel models

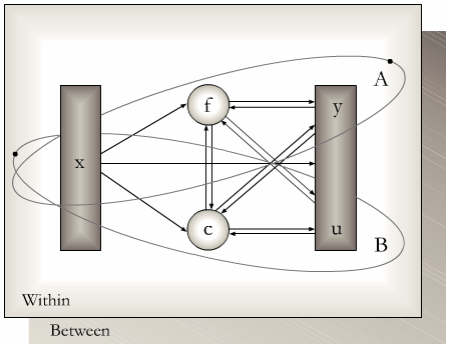
Categorical Latent Variables

- Latent class models
- Mixture models
- Discrete-time survival models
- Missing data models

Mplus integrates the statistical concepts captured by latent variables into a general modeling framework that includes not only all of the models listed above but also combinations and extensions of these models.

6

General Latent Variable Modeling Framework



- Observed variables
  - x background variables (no model structure)
  - y continuous and censored outcome variables
  - u categorical (dichotomous, ordinal, nominal) and count outcome variables
- Latent variables
  - f continuous variables
    - interactions among f's
  - c categorical variables
    - multiple c's

7

Mplus

Several programs in one

- Exploratory factor analysis
- Structural equation modeling
- Item response theory analysis
- Latent class analysis
- Latent transition analysis
- Survival analysis
- Growth modeling
- Multilevel analysis
- Complex survey data analysis
- Monte Carlo simulation

Fully integrated in the general latent variable framework

8

Overview Of Mplus Courses

- **Topic 1.** March 18, 2008, Johns Hopkins University: Introductory - advanced factor analysis and structural equation modeling with continuous outcomes
- **Topic 2.** March 19, 2008, Johns Hopkins University: Introductory - advanced regression analysis, IRT, factor analysis and structural equation modeling with categorical, censored, and count outcomes
- **Topic 3.** August 20, 2008, Johns Hopkins University: Introductory and intermediate growth modeling
- **Topic 4.** August 21, 2008, Johns Hopkins University: Advanced growth modeling, survival analysis, and missing data analysis

9

Overview Of Mplus Courses (Continued)

- **Topic 5.** November 10, 2008, University of Michigan, Ann Arbor: Categorical latent variable modeling with cross-sectional data
- **Topic 6.** November 11, 2008, University of Michigan, Ann Arbor: Categorical latent variable modeling with longitudinal data
- **Topic 7.** March 17, 2009, Johns Hopkins University: Multilevel modeling of cross-sectional data
- **Topic 8.** March 18, 2009, Johns Hopkins University: Multilevel modeling of longitudinal data

10

Analysis With Categorical Observed And Latent Variables

11

Categorical Variable Modeling

- Categorical observed variables
- Categorical observed variables, continuous latent variables
- Categorical observed variables, categorical latent variables

12

Categorical Observed Variables

13

Two Examples

Alcohol Dependence And Gender In The NLSY

	n	Not Dep	Dep	Prop	Odds (Prop/(1-Prop))
Female	4573	4317	256	0.056	0.059
Male	4603	3904	699	0.152	0.179
	9176	8221	955		

Odds Ratio = 0.179/0.059 = 3.019

Example wording: Males are three times more likely than females to be alcohol dependent.

Colds And Vitamin C

	n	No Cold	Cold	Prop	Odds
Placebo	140	109	31	0.221	0.284
Vitamin C	139	122	17	0.122	0.139

14

Categorical Outcomes: Probability Concepts

- Probabilities:
  - Joint: P (u, x)
  - Marginal: P (u)
  - Conditional: P ($u \mid x$)

Alcohol Example

Joint Conditional

	Not Dep	Dep	
Female	.47	. 03	.06
Male	.43	. 08	.15
Marginal	.90	.11	

- Distributions:
  - Bernoulli: $u = 0/1$; $E(u) = \pi$
  - Binomial: sum or prop. ($u = 1$), $E(prop.) = \pi$, $V(prop.) = \pi(1 - \pi)/n$, $\hat{\pi} = prop$
  - Multinomial (*#parameters* = *#cells* – 1)
  - Independent multinomial (product multinomial)
  - Poisson

15

Categorical Outcomes: Probability Concepts (Continued)

- Cross-product ratio (odds ratio):

	$u = 0$	$u = 1$
$x = 0$	π_{00}	π_{01}
$x = 1$	π_{10}	π_{11}

$$\pi_{00}\,\pi_{11} / (\pi_{01}\,\pi_{10}) = \frac{\pi_{11} / \pi_{10}}{\pi_{01} / \pi_{00}} =$$

$P(u = 1, x = 1) / P(u = 0, x = 1) \,/\, P(u = 1, x = 0) / P(u = 0, x = 0)$

- Tests:
  - Log odds ratio (approx. normal)
  - Test of proportions (approx. normal)
  - Pearson $\chi^2 = \Sigma(O - E)^2 / E$ (e.g. independence)
  - Likelihood Ratio $\chi^2 = 2\,\Sigma\, O\, log\, (O / E\,)$

16

Further Readings On Categorical Variable Analysis

Agresti, A. (2002). Categorical data analysis. Second edition. New York: John Wiley & Sons.

Agresti, A. (1996). An introduction to categorical data analysis. New York: Wiley.

Hosmer, D. W. & Lemeshow, S. (2000). Applied logistic regression. Second edition. New York: John Wiley & Sons.

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

17

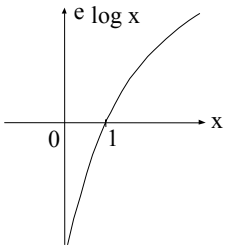
Logit And Probit Regression

- Dichotomous outcome
- Adjusted log odds
- Ordered, polytomous outcome
- Unordered, polytomous outcome
- Multivariate categorical outcomes

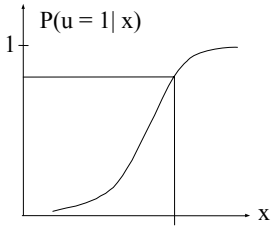
18

Logs

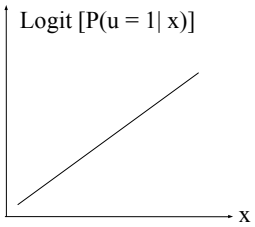
Logarithmic Function



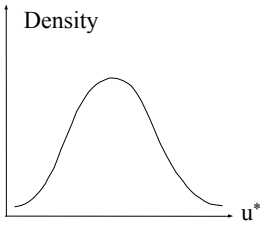
Logistic Distribution Function



Logit



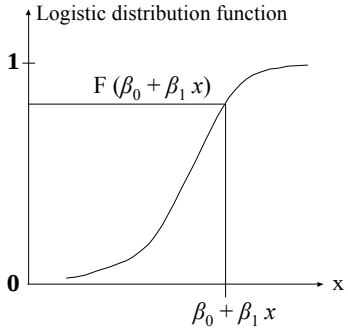
Logistic Density

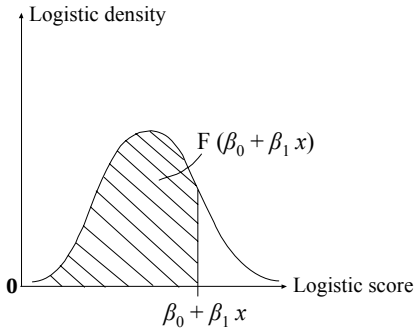


19

Binary Outcome: Logistic Regression

The logistic function $P(u = 1 \mid x) = \mathrm{F}(\beta_0 + \beta_1 x) = \dfrac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$.





Logistic density: $\delta \mathrm{F} / \delta z = \mathrm{F}(1 - \mathrm{F}) = \mathrm{f}(z; 0, \pi^2/3)$

20

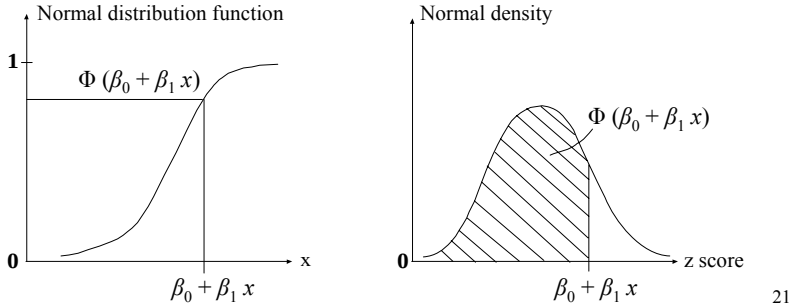
Binary Outcome: Probit Regression

Probit regression considers

$$P(u = 1 | x) = \Phi(\beta_0 + \beta_1 x), \qquad (60)$$

where Φ is the standard normal distribution function. Using the inverse normal function Φ^{-1}, gives a linear probit equation

$$\Phi^{-1}[P(u = 1 | x)] = \beta_0 + \beta_1 x. \qquad (61)$$



21

Interpreting Logit And Probit Coefficients

- Sign and significance
- Odds and odds ratios
- Probabilities

22

Logistic Regression And Log Odds

$Odds\ (u = 1 \mid x) = P(u = 1 \mid x) / P(u = 0 \mid x)$
$= P(u = 1 \mid x) / (1 - P(u = 1 \mid x))$.

The logistic function

$$P(u = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

gives a log odds linear in x,

$logit = log\ [odds\ (u = 1 \mid x)] = log\ [P(u = 1 \mid x) / (1 - P(u = 1 \mid x))]$

$$= \log \left[\frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \Big/ \left(1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}\right) \right]$$

$$= \log \left[\frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} * \frac{1 + e^{-(\beta_0 + \beta_1 x)}}{e^{-(\beta_0 + \beta_1 x)}} \right]$$

$$= \log \left[e^{(\beta_0 + \beta_1 x)} \right] = \beta_0 + \beta_1 x$$

23

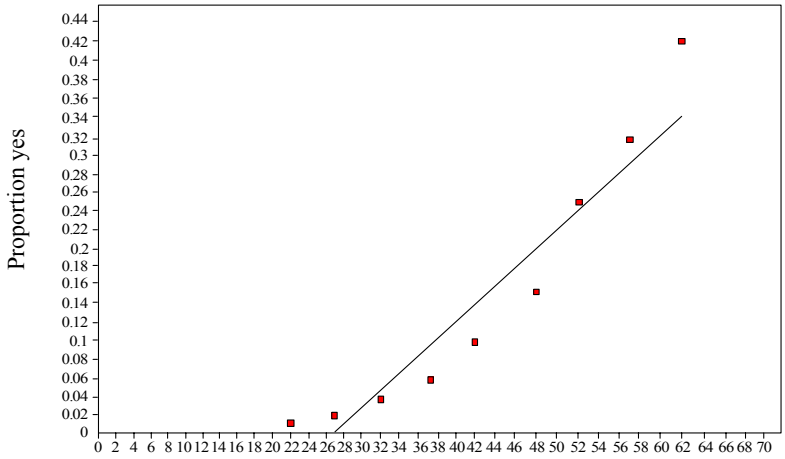
Logistic Regression And Log Odds (Continued)

- $logit = log\ odds = \beta_0 + \beta_1 x$
- When x changes one unit, the *logit* (*log odds*) changes β_1 units
- When x changes one unit, the *odds* changes e^{β_1} units

24

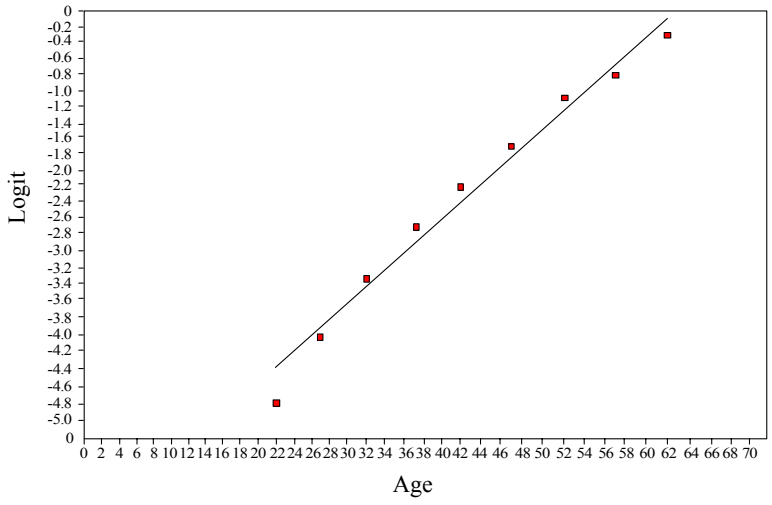
British Coal Miner Data

"Have you experienced breathlessness?"



25

Plot Of Sample Logits



Sample logit = log [proportion / (1 – proportion)]

26

British Coal Miner Data (Continued)

Age (x)	*N*	*N Yes*	*Proportion Yes*	*OLS Estimated Probability*	*Logit Estimated Probability*	*Probit Estimated Probability*
22	1,952	16	0.008	-0.053	0.013	0.009
27	1,791	32	0.018	-0.004	0.022	0.018
32	2,113	73	0.035	0.045	0.036	0.034
37	2,783	169	0.061	0.094	0.059	0.060
42	2.274	223	0.098	0.143	0.095	0.100
47	2,393	357	0.149	0.192	0.148	0.156
52	2,090	521	0.249	0.241	0.225	0.231
57	1,750	558	0.319	0.290	0.327	0.322
62	1.136	478	0.421	0.339	0.448	0.425
	18,282	2,427	0.130			

SOURCE: Ashford & Sowden (1970), Muthén (1993)

Logit model: $\chi^2_{LRT}(7) = 17.13$ ($p > 0.01$)

Probit model: $\chi^2_{LRT}(7) = 5.19$

27

Coal Miner Data

```
x     u     w
22    0     1936
22    1     16
27    0     1759
27    1     32
32    0     2040
32    1     73
37    0     2614
37    1     169
42    0     2051
42    1     223
47    0     2036
47    1     357
52    0     1569
52    1     521
57    0     1192
57    1     558
62    0     658
62    1     478
```

28

Mplus Input For Categorical Outcomes

- Specifying dependent variables as categorical – use the CATEGORICAL option

  CATEGORICAL ARE u1 u2 u3;

- Thresholds used instead of intercepts – only different in sign

- Referring to thresholds in the model – use $ number added to a variable name – the number of thresholds is equal to the number of categories minus 1

  u1$1 refers to threshold 1 of u1
  u1$2 refers to threshold 2 of u1

29

Mplus Input For Categorical Outcomes (Continued)

u2$1 refers to threshold 1 of u2
u2$2 refers to threshold 2 of u2
u2$3 refers to threshold 3 of u2

u3$1 refers to threshold 1 of u3

- Referring to scale factors – use { } to refer to scale factors

  {u1@1 u2 u3};

30

Input For Logistic Regression Of Coal Miner Data

```
TITLE:     Logistic regression of coal miner data
DATA:      FILE = coalminer.dat;
VARIABLE:  NAMES = x u w;
           CATEGORICAL = u;
           FREQWEIGHT = w;
DEFINE:    x = x/10;
ANALYSIS:  ESTIMATOR = ML;
MODEL:     u ON x;
OUTPUT:    TECH1 SAMPSTAT STANDARDIZED;
```

31

Input For Probit Regression Of Coal Miner Data

```
TITLE:     Probit regression of coal miner data
DATA:      FILE = coalminer.dat;
VARIABLE:  NAMES = x u w;
           CATEGORICAL = u;
           FREQWEIGHT = w;
DEFINE:    x = x/10;
MODEL:     u ON x;
OUTPUT:    TECH1 SAMPSTAT STANDARDIZED;
```

32

Output Excerpts Logistic Regression Of Coal Miner Data

Model Results

		Estimates	S.E.	Est./S.E.	Std	StdYX
U	ON					
	X	1.025	0.025	41.758	1.025	0.556
Thresholds						
	U$1	6.564	0.124	52.873		

Odds: $e^{1.025} = 2.79$

As x increases 1 unit (10 years), the odds of breathlessness increases 2.79

33

Estimated Logistic Regression Probabilities For Coal Miner Data

$$P(u=1|x) = \frac{1}{1+e^{-L}},$$

where $L = -6.564 + 1.025 \times x$

For $x = 6.2$ (age 62)

$L = -6.564 + 1.025 \times 6.2 = -0.209$

$$\mathrm{P}(\mathrm{u}=1|\,\text{age } 62) = \frac{1}{1+e^{0.209}} = 0.448$$

34

Output Excerpts Probit Regression Of Coal Miner Data

Model Results

		Estimates	S.E.	Est./S.E.	Std	StdYX
U	ON					
	X	0.548	0.013	43.075	0.548	0.545
Thresholds						
	U$1	3.581	0.062	57.866	3.581	3.581

R-Square

Observed Variable	Residual Variance	R-Square
U	1.000	0.297

35

Estimated Probit Regression Probabilities For Coal Miner Data

$$P(u = 1 \mid x = 62) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x)$$

$$= 1 - \Phi(\hat{\tau} - \hat{\beta}_1 x)$$

$$= \Phi(-\hat{\tau} + \hat{\beta}_1 x).$$

$$\Phi(-3.581 + 0.548 * 6.2) = \Phi(-0.1834) \approx 0.427$$

Note: $\text{logit}\ \hat{\beta} \approx \text{probit}\ \hat{\beta} * c$

where $c = \sqrt{\pi^2 / 3} = 1.81$

36

Categorical Outcomes: Logit And Probit Regression With One Binary And One Continuous X

$$P(u = 1 \mid x_1, x_2) = F[\beta_0 + \beta_1 x_1 + \beta_2 x_2], \qquad (22)$$

$P(u = 0 \mid x_1 , x_2) = 1 - P[u = 1 \mid x_1 , x_2]$, where $F[z]$ is either the standard normal ($\Phi[z]$) or logistic ($1/[1 + e^{-z}]$) distribution function.

Example: Lung cancer and smoking among coal miners

u lung cancer ($u = 1$) or not ($u = 0$)

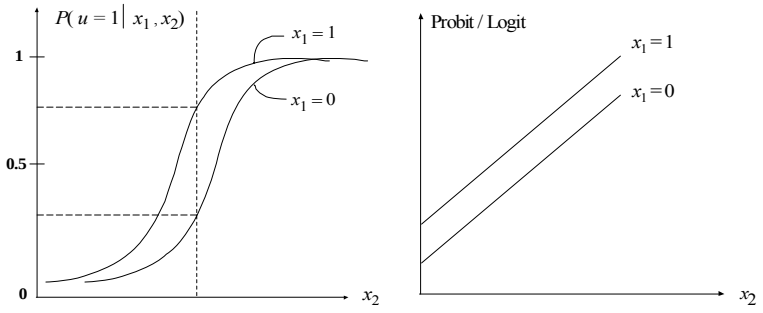
x_1 smoker ($x_1 = 1$), non-smoker ($x_1 = 0$)

x_2 years spent in coal mine

37

Categorical Outcomes: Logit And Probit Regression With One Binary And One Continuous X

$$P(u = 1 \mid x_1, x_2) = F [\beta_0 + \beta_1 x_1 + \beta_2 x_2], \qquad (22)$$



38

Logistic Regression And Adjusted Odds Ratios

Binary u variable regression on a binary x_1 variable and a continuous x_2 variable:

$$P(u = 1 | x_1, x_2) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}, \qquad (62)$$

which implies

$$log\ odds = logit\ [P(u = 1 | x_1, x_2)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 . \qquad (63)$$

This gives

$$log\ odds_{\{x_1 = 0\}} = logit\ [P(u = 1 | x_1 = 0, x_2)] = \beta_0 + \beta_2 x_2 , \qquad (64)$$

and

$$log\ odds_{\{x_1 = 1\}} = logit\ [P(u = 1 | x_1 = 1, x_2)] = \beta_0 + \beta_1 + \beta_2 x_2 . \qquad (65)$$

39

Logistic Regression And Adjusted Odds Ratios (Continued)

The log odds ratio for u and x_1 adjusted for x_2 is

$$log\ OR = log\ [\frac{odds_1}{odds_0}] = log\ odds_1 - log\ odds_0 = \beta_1 \qquad (66)$$

so that $OR = exp\ (\beta_1)$, constant for all values of x_2. If an interaction term for x_1 and x_2 is introduced, the constancy of the OR no longer holds.

Example wording:

"The odds of lung cancer adjusted for years is OR times higher for smokers than for nonsmokers"
"The odds ratio adjusted for years is OR"

40

Analysis Of NLSY Data: Odds Ratios For Alcohol Dependence And Gender

Adjusting for Age First Started Drinking (n=9176)

	Observed Frequencies, Proportions, and Odds Ratios				
	Frequency		Proportion Dependent		
Age 1st	Female	Male	Female	Male	OR
12 or <	85	223	.071	.233	3.98
13	105	180	.133	.256	2.24
14	198	308	.086	.253	3.60
15	331	534	.106	.185	1.91
16	800	990	.079	.152	2.09
17	725	777	.070	.170	2.72
18 or >	2329	1591	.030	.089	3.16

41

Analysis Of NLSY Data: Odds Ratios For Alcohol Dependence And Gender (Continued)

	Estimated Probabilities and Odds Ratios					
	Logit			Probit		
Age 1st	Female	Male	OR	Female	Male	OR
12 or <	.141	.304	2.66	.152	.298	2.37
13	.117	.260	2.66	.125	.257	2.42
14	.096	.220	2.66	.102	.220	2.48
15	.078	.185	2.66	.082	.186	2.55
16	.064	.154	2.66	.065	.155	2.63
17	.052	.127	2.66	.051	.128	2.72
18 or >	.042	.105	2.66	.040	.104	2.82

Logit model: $\chi^2_p(12) = 54.2$

Probit model: $\chi^2_p(12) = 46.8$

42

Analysis Of NLSY Data: Odds Ratios For Alcohol Dependence And Gender (Continued)

Dependence on Gender and Age First Started Drinking

	Logit Regression				Probit Regression				Unstd. Coeff Rescaled To Logit
	Unstd. Coeff.	s.e.	t	Std.	Unstd. Coeff.	s.e.	t	Std.	
Intercept	0.84	.32	2.6		-0.42	.18	-2.4		
Male	0.98	.08	12.7	0.51	0.50	.04	13.1	0.48	0.91
Age 1st	-0.22	.02	-11.6	-0.19	-0.12	.01	-11.0	-0.19	-0.22
R^2	0.12				0.08				

$OR = e^{0.98} = 2.66$

$\text{logit } \beta \approx \text{probit } \beta * c$

where $c = \sqrt{\pi^2 / 3} = 1.81$

43

NELS 88

Table 2.2 – 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	0.81*	0.73**	0.92
Race — ethnicity			
Asian vs. white	0.82	1.42**	0.59
Hispanic vs. white	2.09**	2.29**	2.01**
Black vs. white	2.23**	2.64**	2.23**
Native American vs. white	2.43**	3.50**	2.50**
Socioeconomic status			
Low vs. middle	1.90**	1.91**	3.95**
High vs. middle	0.46**	0.41**	0.39*

SOURCE: U.S. Department of Education, National Center for Education Statistics, National Education Longitudinal Study of 1988 (NELS:88), "Base Year and First Follow-Up surveys.

44

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:

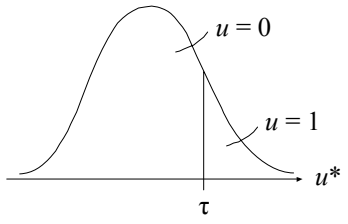
$$P(u = 1 \mid x) = F(\beta_0 + \beta_1 x), \tag{67}$$

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