Overview of Modeling with Categorical Variables

- Examples:
  - Alcohol dependence
  - Randomized trials: colds and vitamins
  - NELS dropout
  - Breathing problems in coal miners
  - Alcohol consumption
  - Occupational destination
- Topics
  - Frequencies, probabilities, multinomial distribution
  - Odds, odds ratios
  - Logistic (logit) regression, adjusted odds ratios
  - Probit regression
  - Polytomous outcome: ordered and unordered
  - Multivariate outcomes
  - Polychoric, tetrachoric, polyserial correlations
  - Probit-based analysis
  - Path analysis with categorical outcomes
  - Factor analysis with covariates: categorical outcomes
  - Item Response Theory (IRT)

NLSY: Alcohol Dependence and Gender

	n	Not Dep	Dep	Prop	Odds
Female	4573	4317	256	0.056	0.059
Male	4603	3904	699	0.152	0.179
			955		
			(0.10)		

OR = 0.179/0.059 = 3.019

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

OR = 0.284/0.139 = 2.043

Categorical Outcomes: Probability Concepts

- Probabilities: joint, marginal, conditional
- Distributions:
  - Bernoulli: $y = 0/1$; $E(y) = \pi$
  - Binomial: sum or prop.$(y = 1)$, $E(prop.) = \pi$, $V(prop.) = \pi(1 - \pi)/n$, $\hat{\pi} = prop$
  - Multinomial $(\#parameters = cells - 1)$
  - Independent multinomial (product multinomial)
  - Poisson
- Cross-product ratio (odds ratio):

$$\pi_{00}\,\pi_{11}/(\pi_{01}\,\pi_{10}) \tag{55}$$

$$(\pi_{y=1,x=1}/\pi_{y=0,x=1})/(\pi_{y=1,x=0}/\pi_{y=0,x=0}) \tag{56}$$

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

Binary Outcome: Logit and Probit

Logistic (logit) regression versus log linear modeling

Hosmer and Lemeshow (1989)

$$Odds(y = 1|x) = \pi_{y=1|x}/\pi_{y=0|x} = \pi_{y=1|x}/(1 - \pi_{y=1|x}). \quad (57)$$

Logit regression. The logistic function gives log odds linear in x:

$$Prob(y = 1|x) = \pi_{y=1|x} = \pi|x = \frac{1}{1 + e^{-(\beta_0+\beta_1\, x)}}, \quad (58)$$

$$log[Odds(y = 1|x)] = log[\pi|x/(1 - \pi|x)] = logit = \beta_0 + \beta_1\, x. \quad (59)$$

Example: Logit = 0 gives Prob = 0.5, logit = −1 gives Prob = 0.27, logit = 1 gives Prob = 0.73.

Probit regression considers

$$P(y = 1|x) = \pi|x = \Phi(\beta_0 + \beta_1\, x), \quad (60)$$

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

$$\Phi^{-1}(\pi|x) = \beta_0 + \beta_1\, x. \quad (61)$$

Logistic Regression and Adjusted Odds Ratios

Binary y variable regressed on a binary x_1 variable and a continuous x_2 variable:

$$P(y=1|x) = \pi|x = \frac{1}{1+e^{-(\beta_0+\beta_1\, x_1+\beta_2\, x_2)}}, \tag{62}$$

or

$$logit(\pi|x) = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2. \tag{63}$$

Let π_0 denote the probability of $y = 1$ for $x_1 = 0$ and let π_1 denote the probability of $y = 1$ for $x_1 = 1$:

$$logit(\pi_0|x) = \beta_0 + \beta_2\, x_2, \tag{64}$$

and

$$logit(\pi_1|x) = \beta_0 + \beta_1 + \beta_2\, x_2. \tag{65}$$

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

$$log\ OR = log[\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}|x] = logit(\pi_1|x) - logit(\pi_0|x) = \beta_1 \tag{66}$$

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.

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

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

TABLE 3

Analysis of NLSY Data
Odds Ratios for 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

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

TABLE 4

Analysis of NLSY Data
Logit and Probit Regression of
Dependence on Gender and Age First Started Drinking
(n=9176)

	Logit Regression				Probit Regression				
	Unstd. Coeff.	s.e.	t	Std.	Unstd. Coeff.	s.e.	t	Std.	Rescaled to Logit
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				

NELS 88

It is important for the reader to keep in mind that the odds ratios presented in this report are not equivalent to the ratio of percentages. For example, the percentage of Hispanic students dropping out was 9.1 percent, while the percentage of white students dropping out was 4.8 percent. The ratio of the percentage of Hispanic students to white students dropping out was 9.1/4.8 or 1.90, while the odds ratio comparing Hispanics to whites was 2.01. In terms of the percentages, therefore, Hispanics were 90 percent more likely than whites to drop out, while in terms of odds they were 101 percent more likely to drop out. In this report we use the terms "more likely" and "less likely" to refer to the change in the odds and not the change in percentages.

In terms of odds ratios, females were slightly less likely than males to have low mathematics and reading skills, but were equally likely to have dropped out of school (table 2.2). Native American, Hispanic and black students were about twice as likely as white students to have performed below basic skill levels in mathematics and reading in the 8th grade and to have dropped out of school by the beginning of the 10th grade. Students from low-socioeconomic backgrounds were about twice as likely as middle class students to perform below basic skill levels and were almost four times as likely to have dropped out.

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*

† Not shown separately are persons whose race–ethnicity is unknown (approximately 2 percent of the unweighted sample).

NOTE: * indicates that the odds compared with the reference group are statistically significant at .05 level; ** at .01 level.

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.

NELS 88

However, it is well known that race–ethnicity and socioeconomic status are highly related and that students from minority backgrounds are also more likely to have low SES. Therefore, the increased likelihood of minority students being at risk may be due in part to their low-SES status and not their race–ethnicity per se. Table 2.3 presents odds ratios adjusted for socioeconomic status, race–ethnicity, and sex.[12] For example, when looking at dropout status, the adjusted odds ratio for the comparison of Hispanic versus white students is 1.12 and is no longer statistically significant. This adjusted figure indicates that when socioeconomic status and sex were held constant, in terms of odds, the likelihood of Hispanics dropping out was no greater than that of whites dropping out. That is, within levels of socioeconomic status and sex, Hispanics and whites dropped out at similar rates.

Table 2.3—Adjusted odds ratios[1] 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[2]			
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*

[1] Odds ratios after controlling for the student's socioeconomic status, race–ethnicity, and sex.

[2] Not shown separately are persons whose race–ethnicity is unknown (approximately 2 percent of the unweighted sample).

NOTE: * indicates that the odds compared with the reference group are statistically significant at .05 level; ** at .01 level.

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.

[12] Logistic regression equations were used to adjust for SES, race–ethnicity, and sex. See appendix A for a more detailed explanation of the adjustment methodology.

Latent Response Variable Formulation Versus Probability Curve Formulation

Probability curve formulation:

$$Prob(y = 1|x) = F(\alpha + \beta x), \qquad (67)$$

where F is the standard normal or logistic distribution function.

Latent response variable formulation (continuous y^*):

Define a threshold τ on y^* so that $y = 1$ is observed when y^* exceeds τ while otherwise $y = 0$ is observed,

$$y^* = \pi\ x + \delta. \qquad (68)$$

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

$$Prob(y = 1|x) = Prob(y^* > \tau|x) = 1 - Prob(y^* \leq \tau|x) = \qquad (69)$$

$$= 1 - \Phi[(\tau - \pi x)V(\delta)^{-1/2}] = \Phi[(-\tau + \pi x)V(\delta)^{-1/2}] \qquad (70)$$

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

Alternatively, a logistic density may be assumed for δ,

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

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

Polytomous Outcome: Ordered Case

A categorical variable y with C ordered categories,

$$y = c,\ if \quad \tau_{j,c} < y^* \leq \tau_{j,c+1} \tag{72}$$

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

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

Probit:

$$y^* = \pi\, x + \delta, \tag{73}$$

$$P(y=0|x) = \Phi(\tau_1 - \pi\, x), \tag{74}$$

$$P(y=1|x) = \Phi(\tau_2 - \pi\, x) - \Phi(\tau_1 - \pi\, x), \tag{75}$$

$$P(y=2|x) = 1 - \Phi(\tau_2 - \pi\, x) = \Phi(-\tau_2 + \pi\, x). \tag{76}$$

$$\begin{aligned} P(y = 1\ or\ 2|x) &= P(y=1|x) + P(y=2|x) && (77)\\ &= 1 - \Phi(\tau_1 - \pi\, x) && (78)\\ &= \Phi(-\tau_1 + \pi\, x) && (79)\\ &= 1 - P(y=0|x), && (80) \end{aligned}$$

with a linear probit for,

$$P(y=2|x) = \Phi(-\tau_2 + \pi\, x), \tag{81}$$

$$P(y = 1\ or\ 2|x) = \Phi(-\tau_1 + \pi\, x). \tag{82}$$

Note: same slope π, so parallel probability curves.

11

Logit for Ordered Categorical Outcome

$$P(y=2|x) = \frac{1}{1+e^{-(\beta_2+\beta\, x)}}, \qquad (83)$$

$$P(y=1 \; or \; 2|x) = \frac{1}{1+e^{-(\beta_1+\beta\, x)}}. \qquad (84)$$

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

$$logit[P(y=2|x)] = \qquad (85)$$

$$= log[P(y=2|x)/(1-P(y=2|x)] = \beta_2 + \beta\, x, \qquad (86)$$

$$logit[P(y=1 \; or \; 2|x)] = \qquad (87)$$

$$= log[P(y=1 \; or \; 2|x)/(1-P(y=1 \; or \; 2|x)] = \beta_1 + \beta\, x. \qquad (88)$$

Note: same slope β, so parallel probability curves.

When x is a 0/1 variable,

$$logit[P(y=2|x=1)] - logit[P(y=2|x=0)] = \beta, (89)$$

$$logit[P(y=1 \; or \; 2|x=1)] - logit[P(y=1 \; or \; 2|x=0)] = \beta, (90)$$

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

Regression With a Binary Dependent Variable

The first two examples discuss regression analysis with categorical dependent variables. As a starting point, consider the simple case of regression of a binary dependent variable y on an x variable. In terms of the general model of the previous section, probit regression is obtained with $\boldsymbol{\Lambda} = \boldsymbol{I}$, $\boldsymbol{\Theta}_\epsilon = \boldsymbol{0}$, $\boldsymbol{B} = \boldsymbol{I}$, so that

$$y^* = \gamma x + \zeta\,, \qquad [8]$$

obtaining the ML-estimated probit slope of $\boldsymbol{\Pi}$ in Equation 6 as γ. The variance of the residual ζ is standardized to one so that $\boldsymbol{\Omega}$ of Equation 7 is the scalar 1. The model expresses the conditional probability of y given x as the probability that y^* exceeds the threshold τ_c in Equation 1,

$$P(y = 1 \mid x) = \int_{\tau - \gamma x}^{\infty} \varphi(t)dt\,, \qquad [9]$$

where φ denotes the univariate standard normal density. Equivalently, conventional probit regression parameterization expresses the negative of τ as an intercept, while γ is the conventional slope.

$$P(y = 1 \mid x) = \Phi(-\tau + \gamma x) \qquad [10]$$

$$= \Phi(\alpha + \beta x)\,,$$

where Φ is the standard normal distribution function.

Example 1: Probit Regression

Table 9.1 Example 1: British Coal Miner Data

Age (x)	*N*	*N Yes*	*Proportion Yes*	*Probit Estimated Probability*	*Logit Estimated Probability*	*OLS Estimated Probability*
22	1,952	16	0.008	0.009	0.013	−0.053
27	1,791	32	0.018	0.018	0.022	−0.004
32	2,113	73	0.035	0.034	0.036	0.045
37	2,783	169	0.061	0.060	0.059	0.094
42	2,274	223	0.098	0.100	0.095	0.143
47	2,393	357	0.149	0.156	0.148	0.192
52	2,090	521	0.249	0.231	0.225	0.241
57	1,750	558	0.319	0.322	0.327	0.290
62	1,136	478	0.421	0.425	0.448	0.339
	18,282	2,427	0.130			

SOURCE: Data from Ashford and Sowden (1970).

Table 9.1 gives British coal miner data taken from Ashford and Sowden (1970). The x variable is age and the binary y variable is breathlessness. This is a case of grouped data in the sense that each distinct x value in the sample has more than a single observation. The sampling scheme may be considered as product-binomial so that the conditional probabilities of y given x are modeled. There are 9 different x values and for each x value a binomial variable is observed. Hence the unrestricted H_1 model for the data has one parameter, a probability, for each of 9 x values, giving a total of 9 parameters. In contrast, the linear probit model has two parameters, τ and γ. The latter model is nested within the former. This can be seen by considering a transformation of the 9 probability parameters π_j, $j = 1, 2, \ldots, 9$, into 9 (probit) parameters, z_j, where $\pi_j = \Phi(z_j)$: The probit model restricts the 9 z_js to be a linear function of x. In this way, a Pearson or likelihood ratio chi-square test of fit has 7 degrees of freedom. If a multinomial sampling scheme is instead considered, the result is the same. The unrestricted model then has 17 parameters because there are 9 × 2 cells of probabilities and these have to add to one. As is the case in log-linear modeling, 8 of these parameters correspond to the marginal distribution of x and should be added to the H_0 model. However, in line with ordinary regression, the marginal distribution of x is not restricted here. The likelihood ratio chi-square value is 5.19 with 7 degrees of freedom and the model is not rejected despite the huge sample size. Note that this may be considered a test against the data of the probit model family for the relationship between y and x. The corresponding test of the logit family results in a likelihood ratio chi-square of 17.13, which is not significant on the 1% level. Table 9.1 also gives the fitted, or predicted, probabilities of breathlessness for each x value. It can be seen that the probit family captures the observed proportions better than the logit family at low and high x values. In Example 1 there is no further structure imposed on the linear probit model parameters, but this could be envisioned as a case of testing $\gamma = 0$, or equality of γ slopes with more than one x variable. Such a test can be performed using as the alternative hypothesis the probit/logit model with unrestricted slopes. In this way, the test is done on the second level without involving the unrestricted multinomial model.

If data are not grouped as in Example 1, model testing against the data is more difficult, because the chi-square approximation may be poor, with many cells having zero or very low expected frequencies. This is the more common case and illustrates the difficulty of testing the categorical variable model against the data directly. Standard computer packages offer a "model test" also in this case, but it refers to the H_0 hypothesis of γs all being zero tested against the H_1 hypothesis of the γs not being zero. Such a test is what is here termed a *second-level test*. Although it is interesting to know that your predictors have significant influence on y, this procedure does not offer the desired test of the probit/logit family against the data. For ungrouped data, Agresti (1990) discusses more suitable goodness-of-fit tests related to residuals.

Regression With an Ordered Polytomous Dependent Variable

Example 2: Ordered Polytomous Regression

Muthén (1987) considered an example of alcohol consumption where y corresponds to the number of drinks a person has per day on average and the xs are age and income. The y categories are 0 (nondrinker), 1 (1-2 drinks per day), 2 (3-4 drinks per day), and 3 (5 or more drinks per day). A U.S. general population sample of 713 males with regular physical activity levels was considered. In this example there are four ordered response categories, where

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

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

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

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

The arguments of Φ are called (population) probits and are linear in the xs. For example, $(-\tau_3 + \gamma' x)$ is the probit for $P(y = 3 \mid x)$. The conditional

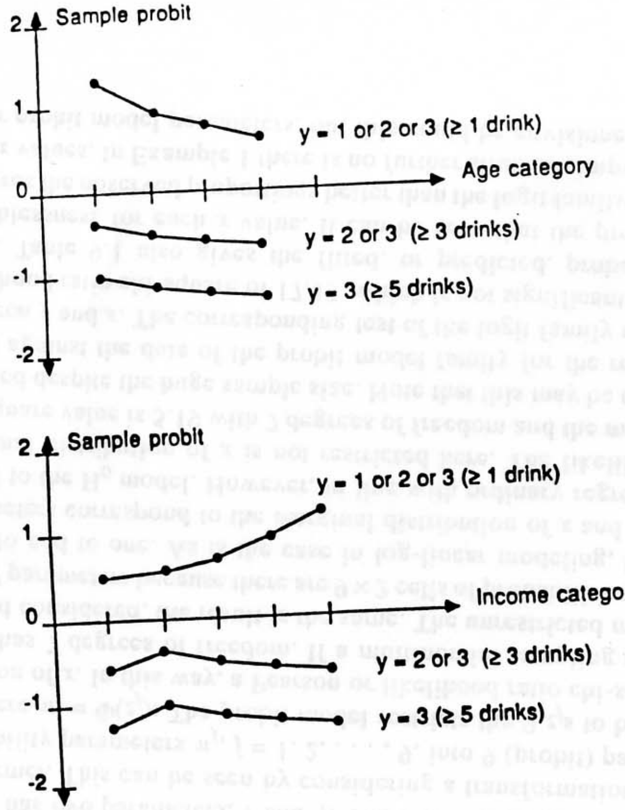


Figure 9.1. Example 2: Sample probit plots for alcohol data.

probabilities of the response categories imply that the probits are linear in the xs when the probabilities for the following three events are considered: $y = 3$, $y = 2$ or 3, $y = 1$ or 2 or 3. As an example, consider the event $y = 2$ or 3. Noting that $1 - \Phi(z) = \Phi(-z)$, the probit for this event is $-\tau_2 + \gamma' x$. The probits for these three events also have the same x slopes γ. These facts can be used to test the goodness of fit for the family of ordered four-category probit models against the data. Using grouped data, the corresponding sample probits based on observed proportions can be plotted against the x variables. Figure 9.1 shows the plots where age has been categorized into four categories and income has been categorized into five categories.

Polytomous Outcome: Unordered Case

Multinomial logistic regression:

$$P(y_i = c|x_i) = \frac{e^{\beta_{0c}+\beta_{1c}\ x_i}}{\sum_{c=1}^{K} e^{\beta_{0c}+\beta_{1c}\ x}}, \tag{91}$$

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

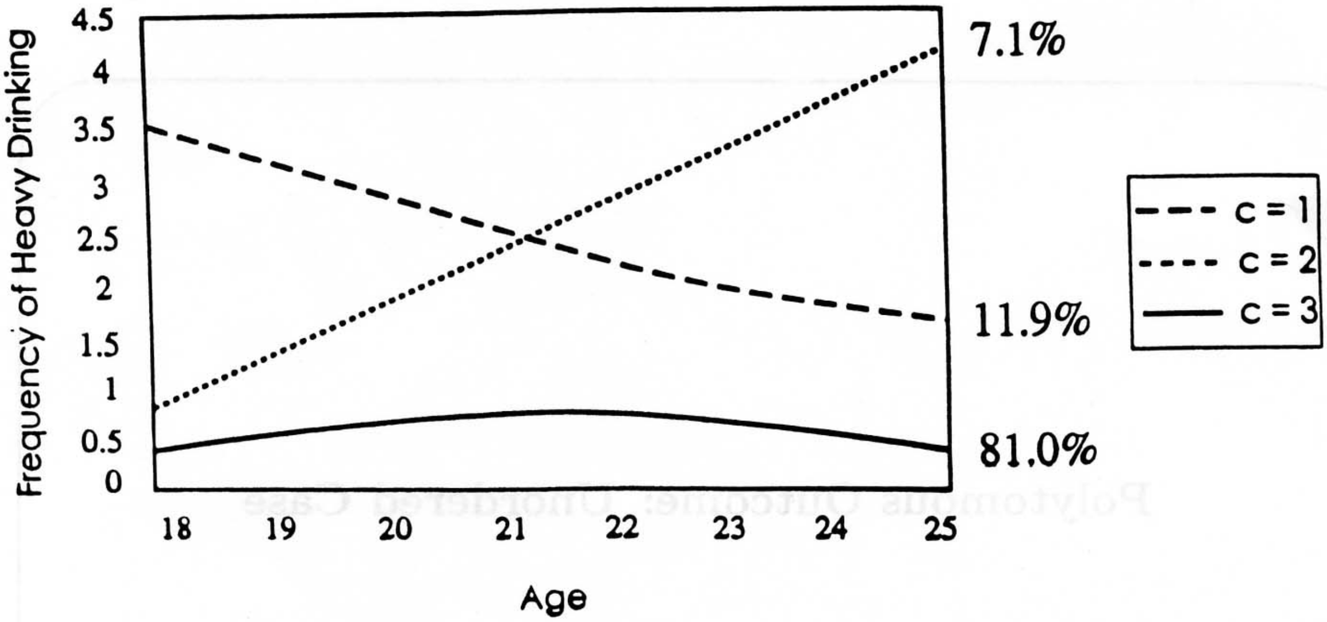
$$\beta_{0K} = 0, \tag{92}$$

$$\beta_{1K} = 0, \tag{93}$$

which gives the log odds

$$log[P(y_i = c|x_i)/P(y_i = K|x_i)] = \beta_{0c} + \beta_{1c}\ x_i, \tag{94}$$

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



Predicting Trajectory Class Membership

Estimated Logit Coefficients:

Covariate (x)	High vs Norm	Increase vs Norm
Male	**1.25**	**1.48**
Black	**-1.60**	-.67
Hispanic	-.22	.74
Early Onset	**1.07**	.62
FH123	.62	.68
Dropout	.22	**.80**
College	**-.61**	-.04