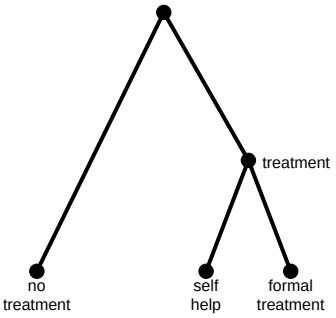
FITTING NESTED LOGIT MODELS TO COMPLEX SURVEY DATA

Moshe Feder[1], Alexander J. Cowell[1], Sarah Q. Duffy[2] and Weihua Shi[1]

[1]RTI International, P.O.Box 12194, Research Triangle Park, NC, 27709-2194
[2]Substance Abuse and Mental Health Services Administration, Rm 12-105 Parklawn Building, 5600 Fishers Ln, Rockville, MD 20857

KEY WORDS: Random Choice, Modeling, Unequal Weighting, Clustering

Evidence suggests that the number of individuals who enter treatment for substance abuse (SA) is far fewer than the number who could benefit from it. Among the many factors that may be preventing people from obtaining treatment is that insurance coverage, if available at all, is typically less generous for SA disorders than for other health problems. Thus, there is interest in examining the demand for SA treatment, including self-help treatment, among substance-abusing individuals, to see how perceived insurance coverage and other factors affect that demand. One possible individual's decision tree is shown in the diagram below.



Two natural modeling choices are 1. sequential logit, according to which an individual first chooses whether or not to enter treatment and then chooses the type of treatment, and 2. multinomial logit, where all ultimate choices are equally substitutable. A more general and flexible modeling approach is offered by nested logit (NL) models. These models include the sequential logit and the multinomial logit as special cases. Furthermore, research has shown that often NL models fit the data better than either the sequential or the multinomial logit models. (See, e.g., Fortney, Rost & Zhang 1998 in the context of choice of treatment for depression.)

We are interested in fitting NL models to data from the 2000 National Household Survey on Drug Abuse (NHSDA)[1]. The NHSDA sample selection employs a stratified multi-stage sampling design with oversampling younger age groups. Thus, the complex sample design of the NHSDA is characterized by unequal inclusion probabilities and by clustering. These characteristics — if ignored — may lead to biased point estimates (due to unequal inclusion probabilities), and biased variance estimates (due to clustering). We have used a pseudo-likelihood approach to account for the unequal weighting, and Taylor linearization approach to account for the clustering and obtain correct standard errors.

The Utility Model

NL models are commonly derived from utility models. In our case, the utility model has the form below:

$$\begin{aligned} U_{\text{no tx}} &= V_{\text{no tx}} + \varepsilon_{\text{no tx}} \\ U_{\text{self-help}} &= V_{\text{tx}} + V_{\text{self-help}} + \varepsilon_{\text{tx}} + \varepsilon_{\text{self-help}} \\ U_{\text{formal tx}} &= V_{\text{tx}} + V_{\text{formal tx}} + \varepsilon_{\text{tx}} + \varepsilon_{\text{formal tx}} \end{aligned} \quad (1)$$

where tx means 'treatment.' $V_{\text{no tx}}$, V_{tx}, $V_{\text{self-help}}$, $V_{\text{formal tx}}$ are the deterministic components of the utilities $U_{\text{no tx}}$, $U_{\text{self-help}}$ and $U_{\text{formal tx}}$ and are treated in our case as functions of the individual's characteristics.

The stochastic parts of the utilities, ε_{tx}, $\varepsilon_{\text{no tx}}$ and

[1]Now called the National Survey on Drug Use and Health (NSDUH)

$\varepsilon_{\text{formal tx}}$ are assumed to have Gumbel distributions.

The Probability of Choice Model
Under certain assumptions[2] regarding the joint distribution of stochastic components of the utility model we get our probability of choice (PC) model below:

$$\begin{aligned} Pr(\text{self-help}|\text{tx}) &= \frac{\exp(V_2)}{\exp(V_2)+1} \\ Pr(\text{no tx}) &= \frac{\exp(V_1)}{\exp(V_1)+[\exp(V_2)+1]^{\theta}} \end{aligned} \tag{2}$$

where θ is referred to as the inclusive value parameter. When the PC model (2) follows from the utility model (1), $\theta \in [0,1]$. However, when fitting models such as (2) to real data, the estimate θ might fall outside the $[0,1]$ range, providing evidence against the utility model (1). In the above, $V_1 = V_{\text{no tx}} - V_{\text{tx}} - V_{\text{formal tx}}$ and $V_2 = V_{\text{self-help}} - V_{\text{formal tx}}$. These are modeled as linear combinations of the individual-specific covariates $\mathbf{x}_1$ and $\mathbf{x}_2$:

$$V_1 = \boldsymbol{\beta}_1'\mathbf{x} \quad \text{and} \quad V_2 = \boldsymbol{\beta}_2'\mathbf{x}. \tag{3}$$

The covariates may be the same in both nodes—i.e., $\mathbf{x}_1 = \mathbf{x}_2$. The PC model (2), combined with the representation (3) of V_1 and V_2 is our NL model.

REMARK: The values $\theta = 1$ and $\theta = 0$ correspond to the multinomial and to the sequential logit models, respectively.

ESTIMATION

There are two possible ways to estimate the model parameters: two-step estimation and full information maximum pseudo likelihood estimation.

Two-Step Estimation
In what follows, we will denote the vector of the unknown model parameters by

$$\boldsymbol{\psi} = (\theta, \boldsymbol{\psi}_1', \boldsymbol{\psi}_2')'. \tag{4}$$

There is an easy way to obtain unbiased point estimates of $\boldsymbol{\psi}$ using existing software, by the following two-step approach. First, note that according to the PC model (2), conditional on choosing treatment, the choice between self-help and formal treatment follows a dichotomous logit model. Next, note that if we define a new variable $x_3 = \log[\exp(\boldsymbol{\beta}_2'\mathbf{x}_2 + 1]$ then the probability of no treatment is given by

$$Pr(\text{no tx}) = \frac{\exp(V_1)}{\exp(V_1) + \exp(\theta x_3)} \tag{5}$$

—a dichotomous logit model as well. In the first step, we subset the sample to those who chose treatment and fit the conditional model (first part of (2)) to obtain an estimate of $\boldsymbol{\beta}_2$. These estimates are then used to calculate estimates of x_3 for each individual. In the second step, the model (5) is fitted, providing estimates of θ and $\boldsymbol{\beta}_1$. Variance estimation using this approach is not straightforward because x_3 is not known without error (as it is calculated using an *estimate* $\hat{\boldsymbol{\beta}}_2$, rather than an exact value). We have tested this approach by a small simulation which confirmed its validity. However, because of its limitations, we have decided on the pseudo-likelihood approach (discussed in the next section), together with Taylor linearization variance estimation (described in the section after next). However, the two-step approach provides a good starting value for maximizing the pseudo-likelihood. Furthermore, estimates from the two methods may be compared for quality control purposes.

The Pseudo Log-likelihood and Point Estimates
Denote outcome variables y, y_1, y_2 by

$$y = \begin{cases} 0 & \text{No Treatment} \\ 1 & \text{Self-help Treatment} \\ 2 & \text{Formal Treatment} \end{cases} \tag{6}$$

and $y_1 = I(y \neq 0)$, $y_2 = I(y = 2)$.

The contribution of individual j to the log-likelihood, under a simple random sample is

$$\begin{aligned} \ell_j = \sum_{i \in s} \{ &(1-y_1)V_1 + y_1(1-y_2)V_2 \\ &+ y_1(\theta - 1)\log\left(\exp(V_2)+1\right) \\ &- \log\left[\exp(V_1) + (\exp(V_2)+1)^{\theta}\right]\}. \end{aligned}$$

Had the entire finite population U been observed, the (census) log-likelihood $\mathcal{L}_U = \sum_{j \in U} \ell_j$ could have been used to obtain maximum likelihood estimates (MLE) of the model parameters. Since only

[2]Details omitted for the sake of brevity.

a sample s is available, we use the weighted log-likelihood

$$\mathcal{L}_w = \sum_{j \in s} w_j \ell_j,$$

where w_j is the survey weight of individual j. We'll refer to $\mathcal{L}_w$ as the pseudo log-likelihood. We have used scoring to maximize the pseudo log-likelihood. To improve the estimates, we have followed the scoring by a numerical maximization of the pseudo log-likelihood.

REMARK: We have empirically found that the scoring performed better when the off-diagonal elements in the row and column in the Hessian which correspond to θ were replaced by zeros. We do not have an explanation yet for this phenomenon.

Variance Estimation

We have used a Taylor linearization (TL) approach to derive sandwich variance estimators (Binder, 1983).

Denoting the model parameters collectively by $\boldsymbol{\psi}$, we have

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\psi}}) = H(\hat{\boldsymbol{\psi}})^{-1} \hat{\Sigma}(\boldsymbol{\psi}_0) H(\hat{\boldsymbol{\psi}})^{-1}$$

where H is the Hessian matrix, and $\hat{\Sigma}(\boldsymbol{\psi}_0)$ is the estimated covariance matrix of the score function $\partial \mathcal{L}_w / \partial \boldsymbol{\psi}$ evaluated at $\hat{\boldsymbol{\psi}}$, treated as a total of the individual contributions to the score:

$$\frac{\partial \mathcal{L}_w}{\partial \boldsymbol{\psi}} = \sum_{j \in s} w_j \frac{\partial \ell_j}{\partial \boldsymbol{\psi}}.$$

(We calculated the covariance matrix assuming a WR first stage design.)

SIMULATION RESULTS

We have simulated 50 samples from our NL model, each of size 1000. An intercept (β_{10} in the upper node of the decision tree and β_{20} at the lower node) and two regression parameters (β_{11} and β_{12} in the upper node of the decision tree and β_{21} and β_{22} at the lower node) were assumed at each of the two nodes of the decision tree. The results are given below.

POINT ESTIMATES:

Parameter	TRUE	AVG	MEDIAN
θ	0.700	0.707	0.708
β_{10}	0.500	0.501	0.500
β_{11}	0.600	0.603	0.609
β_{12}	0.000	0.004	0.002
β_{20}	-0.300	-0.294	-0.288
β_{21}	1.000	1.011	0.995
β_{22}	0.300	0.297	0.293

Here, 'TRUE' means the simulation parameter, 'AVG' is the average of the 50 estimates, and 'MEDIAN' is their median.

STANDARD ERRORS:

Parameter	EMP	AVG	MEDIAN
θ	0.123	1.030	0.110
β_{10}	0.042	0.104	0.048
β_{11}	0.060	0.210	0.059
β_{12}	0.020	0.065	0.020
β_{20}	0.046	0.325	0.044
β_{21}	0.146	1.045	0.134
β_{22}	0.044	0.309	0.044

Here, 'EMP' means the empirical standard error, 'AVG' is the average of the 50 estimates, and 'MEDIAN' is their median.

Note that the average standard error is too high. We have traced this to a numerical overflow that affected the calculation of the Hessian in about 3–5% of the simulations, which yielded extremely high variance estimates. Indeed, the rest of the estimates were well-behaved, as can be seen in the median estimates of standard errors. We are currently working on numerical improvements to solve these rare cases.

ACKNOWLEDGEMENTS: We would like to thank Carol Council and Gordon Brown, both from RTI International, for their help on this project.

REFERENCES

Ben Akiva, M. and Lerman, S. (1985), Discrete Choice Analysis. Cambridge, MA: The MIT Press.

Binder, D.A. (1983), On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

Fortney J., Rost K. and Zhang M. (1998), A Joint Choice Model of the Decision to Seek Depression Treatment and Choice of Provider Sector. *Medical Care*, **36**(3):307–320, 1998.

Hensher, A.D. (1986), Sequential and full information maximum likelihood estimation of a nested logit model. *The Review of Economics and Statistics*, **86**, 657–667.

Hensher, A.D. and Johnson, L.W. (1981), Applied Discrete Choice Modelling. New York: Halsted.

Hensher, D.A. and Greene, W.H. (2002), Specification and estimation of the nested logit model: alternative normalisations. *Transportation Research Part B* **36**, 1–17.

Hunt, G.L. (2000), Alternative nested logit model structures and the special case of partial degeneracy. *Journal of Regional Science*, **40**, 89–113.

Koppelman, F.S. and Wen, C.-H. (1998), Alternative nested logit models: structure, properties and estimation. *Transportation Research Part B* **32**, 289–298.

Louviere, J.J., Hensher, D.A. and Swait, J.D. (2000), "Stated Choice Models," Cambridge: Cambridge University Press.

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