Paper 341-2011

Logit Models in Practice: B, C, E, G, M, N, O…

Joseph C. Gardiner, Zhehui Luo
Division of Biostatistics, Department of Epidemiology, Michigan State University, East Lansing, MI

ABSTRACT

Logit models appear in a variety of forms in applications in biostatistics, epidemiology, economics, marketing research and sociology. They are used to model the relationship between covariates and various types of discrete outcomes from the ubiquitous binary logit model for a two-level response to the conditional logit and multinomial (generalized) logit models concerning polytomous responses. Covariates may vary by characteristics of both the individual and response. For example, when assessing a consumer's choice of health insurance plan or health care provider, or selection of a treatment regime (surgery, medical management, or no treatment), the probability of choice depends on the consumer's own circumstances, utilities and preferences. Nested logit models allow for modeling the sequence of the decision process faced by the consumer by grouping alternatives at each stage into nests. Ordered logit models exploit the underlying ordinal structure of the response, whereas the exploded logit can be applied to rank ordered responses. We survey some enhancements in SAS/STAT® and SAS/ETS® software that can be used to fit various logit models.

INTRODUCTION

In many applications one encounters qualitative response data. The simplest binary outcome has two levels, for example a patient's response to treatment is success or failure; a voter supports, or does not support a piece of legislation. Polytomous outcomes with several levels may be ordinal such as the severity of pain recorded as none, mild, moderate or severe, or nominal (unordered) such as the choice of travel mode—car, bus, train or plane, for traveling between two cities. Rank-ordered response data arise when a consumer is provided a menu of alternatives such as several breakfast cereals, and asked to order their choice from best (most preferred) to worst (least preferred). There may be several nuances in the respondent data. The set of alternatives could vary across individuals; some choices may receive the same rank; only a subset of the offered alternatives may be ranked leaving the remaining choices unranked. Discrete choice models (DCMs) in which individuals make choices based on own tastes for attributes of the alternatives have applications in marketing research, health services research and behavioral and social sciences. See references.

Statistical models for analysis of qualitative observations should exploit their discrete nature while focusing on the inferential questions being addressed. Methods typically used to analyze quantitative, continuous responses are likely to be inadequate and inappropriate. For the models to be discussed in this paper the observations $\{(Y_i, \mathbf{x}_i) : 1 \le i \le n\}$ constitute a random sample from the target population, with Y_i denoting the response or the chosen alternative in DCMs and $\mathbf{x}_i$ a p×1 vector of explanatory variables (covariates) for the i-th individual or unit in the sample. Especially with DCMs the covariates will vary by characteristics of the alternatives. In this case $\mathbf{x}_i = \{\mathbf{x}_{ij} : j \in C_i\}$ where $\mathbf{x}_{ij}$ are the covariates for the j-th alternative in the choice set C_i for the i-th individual. Typically researchers wish to quantify the influence of the covariates on some feature of the distribution of Y_i, for example the probabilities of choosing alternative j. This quantification is through a regression model for an underlying unobserved continuous latent variable whose range of values is manifest in the observation Y_i. Although reference to a latent variable regression is not strictly necessary, it nevertheless provides a convenient primitive to frame the derivation of various models by changing the distribution assumption on the latent variable. If the latent variable has a meaning in a particular field of application, it has the advantage of providing a context that could help with interpretation of the model.

Binary Logit

The binary logit model is the mainstay for modeling a dichotomous response with applications in perhaps every research endeavor. The response Y_i is realized as a binary indicator $Y_i = [Y_i^* > 0]$ from the latent linear regression model $Y_i^* = \mathbf{x}_i'\beta + \varepsilon_i$ where the error ε_i has a logistic distribution $F(u) = (1 + e^{-u})^{-1}$, $u \in (-\infty, \infty)$ independent of $\mathbf{x}_i$. The response probability $\pi(\mathbf{x}_i) = P[Y_i = 1 \,|\, \mathbf{x}_i] = F(\mathbf{x}_i'\beta)$ when transformed by $\log\big(\pi(\mathbf{x}_i) / (1 - \pi(\mathbf{x}_i))\big) = \mathbf{x}_i'\beta$ provides an interpretation of the regression coefficients β as log odds ratios. The maximum likelihood estimator (MLE) of β is obtained by maximizing the log-likelihood $\ell(\beta) = \sum_{i=1}^{n}\big(Y_i \log F(\mathbf{x}_i'\beta) + (1 - Y_i)\log(1 - F(\mathbf{x}_i'\beta))\big) = \sum_{i=1}^{n} \log F(q_i \mathbf{x}_i'\beta)$ where $q_i = 2Y_i - 1$. The gradient (score) vector, $\mathbf{g} = \dfrac{\partial \ell(\beta)}{\partial \beta}$, outer product (OP) matrix $\mathbf{B} = \dfrac{\partial \ell(\beta)}{\partial \beta}\dfrac{\partial \ell(\beta)}{\partial \beta'}$ and Hessian matrix $\mathbf{H} = -\dfrac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta'}$ simplify to $\mathbf{g} = \sum_{i=1}^{n}\big(Y_i - F(\mathbf{x}_i'\beta)\big)\mathbf{x}_i$, $\mathbf{B} = \sum_{i=1}^{n}\big(Y_i - F(\mathbf{x}_i'\beta)\big)^2 \mathbf{x}_i\mathbf{x}_i'$, $\mathbf{H} = \sum_{i=1}^{n}\big(F(\mathbf{x}_i'\beta)(1 - F(\mathbf{x}_i'\beta))\big)\mathbf{x}_i\mathbf{x}_i'$. The MLE $\hat{\beta}$ of β is the solution to the normal equation $\mathbf{g}(\beta) = 0$. It is consistent and asymptotically normal with (estimated) asymptotic variance matrix $\hat{\mathbf{H}}^{-1}$ where $\hat{\mathbf{H}} = \mathbf{H}(\hat{\beta})$. Two other estimates of the asymptotic variance matrix are the OP estimate $\hat{\mathbf{B}}^{-1}$ and the robust-sandwich estimate $\hat{\mathbf{H}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{H}}^{-1}$, also referred to as the quasi (Q)-MLE variance matrix.

All three variances are computed by proc QLIM; only the Hessian variance is computed in proc LOGISTIC. Robust-sandwich (empirical) and Hessian variances are computed in proc GLIMMIX and proc GENMOD under the assumed set-up of the generalized linear model (GLM). The solution to the estimating equation (EE) for β, $\sum_{i=1}^{n} \dfrac{\partial \pi_i}{\partial \beta} \upsilon_i^{-1}(Y_i - \pi_i) = 0$, where $Var(Y_i \,|\, \mathbf{x}_i) = \upsilon_i = \pi_i(1 - \pi_i)$ is the same as the solution to the MLE normal equation. The GLM model-based and robust-sandwich estimators of the variance coincide with $\hat{\mathbf{H}}^{-1}$ and $\hat{\mathbf{H}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{H}}^{-1}$, respectively. In contrast the probit model with $F = \Phi$ (standard normal distribution) will yield slightly different variance estimators under the MLE and GLM theory although the MLE and EE estimators for β are the same.

Consistent estimation of β requires correct specification of $\pi(\mathbf{x}_i)$. Any of the following will make the MLE $\hat{\beta}$ inconsistent: (i) heteroscedasticity, i.e., $Var(\varepsilon_i \,|\, \mathbf{x}_i)$ being non-constant. (ii) endogeneity of covariates $\mathbf{x}_i$, i.e., one or more covariates are correlated with the error ε_i, (iii) incorrect distribution assumption on the error ε_i, and (iv) omitted covariates in $\mathbf{x}_i$ (even if they are orthogonal to those included). An example of (i) is $Var(\varepsilon_i \,|\, \mathbf{x}_i) = \sigma^2 \exp(\mathbf{z}_i'\gamma)$ where σ^2 is 1 for the probit model or $\pi^2/3$ for the logit model which lead to specifying $\pi(\mathbf{x}_i) = F(\mathbf{x}_i'\beta \exp(-\tfrac{1}{2}\mathbf{z}_i'\gamma))$. The covariates $\mathbf{z}_i$, typically a subset of $\mathbf{x}_i$, should be selected with guidance from subject-matter rather than statistical convenience. For (ii) we need additional models for the endogenous covariates. Both (i) and (ii) can be fitted in QLIM although for (ii) the model errors are assumed jointly normal. The logistic and normal distributional assumption on ε_i generally yield similar results for $\pi(\mathbf{x}_i)$. Moon (1988) and Mcdonald (2000) discuss other flexible forms for F concerning (iii). Wooldridge (2002) gives some insightful comments on the issue of neglected heterogeneity (iv) in the context of the probit model. Since all moments of the response Y_i are functions of $\pi(\mathbf{x}_i)$, in the single response context one might question the need for robust standard errors to guard against heteroscedasticity or misspecification.

Illustrative Example 1

The data set comprises 4483 respondents in year 1988 to the German Socioeconomic Panel Survey 1984-1995 on healthcare utilization (Riphahn *et al*, 2003). Self-reported assessment of health (HSAT) is recorded on a 0 to 10 scale with higher values indicative of better health. The covariates we will use in this analysis are the respondent's age (AGE), a measure of household income (HHNINC), education (EDUC) –all continuous, and the binary indicators for gender (FEMALE, 48%), having children in household (HHKIDS, 38%) and marital status (MARRIED, 75%). For purposes of illustration of various binary logit models we use the dichotomization Y=[$HSAT$≥7]. Approximately 60% have the event Y=1 which we will call "good health". The following formats might prove useful:

```
proc format;
value hsat low-<7='<7' 7-high='>=7';
value female 0='male' 1='female';
value affirm 0='no' 1=' yes';
run;
```

LOGISTIC and QLIM will produce identical results:

```
proc logistic data=c.healthcare(where=(year=1988));
class married(ref='no') hhkids(ref='no') /param=ref;
model hsat(event='>=7')=age educ hhninc married hhkids female/link=logit;
format female female. married hhkids affirm. hsat hsat.;
run;

proc qlim data=c.healthcare(where=(year=1988)); *covest=qml;
class hhkids married female;
endogenous hsat~discrete(dist=logistic order=formatted);
model hsat=age educ hhninc married hhkids female;
format female female. married hhkids affirm. hsat hsat.;
run;
```

Table 1 summarizes the estimation results. Although its need is questionable, the robust estimates of standard errors (column 4) are produced by the option `covest=qml` in the QLIM statement. The p-values (column 5) computed using either standard errors are practically the same. The heteroscedastic logit model (columns 6-8) is fitted by adding the HETERO statement to the QLIM syntax:

```
 hetero hsat~female HHNINC /link=exp noconst;
```

Model fit statistics at the bottom of Table 1 show that the heteroscedastic model is not significantly different from the homoscedastic model. The formal likelihood ratio (LR) test of $H_0 : \gamma = 0$ has $\chi^2 = 0.35$, 2 DF.

Table 1: Binary Logit Models

	Homoscedastic case				Heteroscedastic Case		
Parameter	**Estimate**	**Standard Error (Hessian)**	**Standard Error (QMLE)**	**P-value (Hessian)**	**Estimate**	**Standard Error (Hessian)**	**P-value (Hessian)**
Intercept	0.8091	0.24155	0.24287	0.0008	0.8146	0.23332	0.0005
AGE	–0.0328	0.00321	0.00321	<.0001	–0.0320	0.00589	<.0001
EDUC	0.0837	0.01503	0.01536	<.0001	0.0805	0.02032	<.0001
HHNINC	0.3487	0.20833	0.21234	0.0942	0.2224	0.37888	0.5572
MARRIED	–0.0518	0.08288	0.08339	0.5318	–0.0401	0.08622	0.6422
HHKIDS	0.1289	0.07557	0.07523	0.0881	0.1285	0.07690	0.0947
FEMALE	–0.0568	0.06388	0.06387	0.3738	–0.0304	0.08008	0.7040
_H.FEMALE					0.1212	0.27374	0.6579
_H.HHNINC					–0.3642	0.95830	0.7039
–2 Log L	5780.0				5779.6		
–2 Log L (null)	6020.8				6020.8		
–2 Log LR	240.8				241.2		

The results show that older age is associated with poor health, and more education with good heath. The sign on MARRIED suggests that the health status of married respondents was worse than their single counterparts. Fortunately the effect is not significant. Using the OUTPUT statement we can obtain predicted probabilities of response. This is useful in the heteroscedastic model because the standard interpretation of the β-coefficients as log odds ratios is not valid.

Cumulative Logit and Ordered Logit Models

There are many applications in which the categories of the outcome have a natural ordering. For example, the severity of pain recorded as none, mild, moderate or severe. Any categorical variable assessed on a Likert scale would also fit this type of response.

Suppose there are *J*- levels of the outcome Y_i with labels 1, 2, …,*J*. The response variable can be modeled in various ways. The cumulative probabilities of Y_i, $\gamma_j(\mathbf{x}_i) = P[Y_i \le j \mid \mathbf{x}_i]$, reflect the ordering, with $\gamma_1(\mathbf{x}_i) \le \gamma_2(\mathbf{x}_i) \le \cdots \le \gamma_J(\mathbf{x}_i) = 1$. Procedures LOGISTIC, GENMOD and GLIMMIX with the option link=cumlogit in the model statement will fit the model $\log\left(\gamma_j(\mathbf{x}_i) / (1 - \gamma_j(\mathbf{x}_i)\right) = \alpha_j + \mathbf{x}_i'\delta$, which is called the *cumulative logit* model (Agresti, 2002). Changing the link to cumprobit will fit the *cumulative probit* model. The α_j, *j*=1, 2, …,*J* are intercepts; a constant is not included in $\mathbf{x}_i$. The parameters δ describe the effect of a covariate on the log odds of response in the category *j* or below. When the corresponding δ>0, as the value of the associated covariate increases, the response is more likely to fall at the low end of the ordinal scale.

As in the aforementioned pain scale the response variable sometimes reflects an underling measure that is not observed in its entirety. Let $\mu_j, j = 0,\ldots,J$ be threshold-points that provide a partition of the entire real line, that is, $-\infty = \mu_0 < \mu_1 < \ldots < \mu_J = \infty$. The observed outcome is a categorization of a latent variable $Y_i^* = \mathbf{x}_i'\beta + \varepsilon_i$ such that $Y_i = j$ if and only if $\mu_{j-1} < Y_i^* \le \mu_j$. The probability of response is $\pi_j(\mathbf{x}_i) = P[Y_i = j \mid \mathbf{x}_i] = F(\mu_j - \mathbf{x}_i'\beta) - F(\mu_{j-1} - \mathbf{x}_i'\beta), \;\; j = 1,\ldots,J$ where F is the distribution of ε_i. The cumulative response probability is $\gamma_j(\mathbf{x}_i) = P[Y_i \le j \mid \mathbf{x}_i] = F(\mu_j - \mathbf{x}_i'\beta)$. By specifying F we get the two commonly used models: when F is the logistic distribution function we get the *ordered logit model*; when F is the standard normal distribution function Φ we get the *ordered probit model*.

In the ordered logit model $\log\left(\gamma_j(\mathbf{x}_i)/(1-\gamma_j(\mathbf{x}_i))\right) = \mu_j - \mathbf{x}_i'\beta = -\log\left((1-\gamma_j(\mathbf{x}_i))/\gamma_j(\mathbf{x}_i)\right), j = 1,\ldots,J-1.$ The parameters β describe the effect of a covariate on the log odds of response in the category above *j*, or equivalently the marginal effect of the covariate on $E[Y_i^* \mid \mathbf{x}_i]$. When β >0, as the value of the covariate increases, the response is more likely to fall at the high end of the ordinal scale, because $\dfrac{\partial E[Y_i^* \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \beta$.

Both the *cumulative logit* model and the *ordered logit* model have the *proportional odds property* because the odds ratio does not depend on the category to which the response variable belongs. Both models assume the effect of a covariate is identical for all *J*–1 cumulative logits. When this property holds, the model requires a single parameter rather than *J*–1 parameters to describe the effect of a covariate.

Proc QLIM fits the ordered logit and ordered probit models. It uses the latent variable formulation. By default an intercept is included in β and the first threshold parameter μ_1 is set to zero. The model option limit1=varying overrides the default.

Estimation of the parameters (μ_j, β) or (α_j, δ) in the ordered and cumulative models is via maximum likelihood. The log-likelihood is the same for two models except for the difference in the parameterization. For the ordered model the log-likelihood is $\ell(\mu,\beta) = \sum_{i=1}^{n}\sum_j [Y_i = j]\log\left(F(\mu_j - \mathbf{x}_i'\beta) - F(\mu_{j-1} - \mathbf{x}_i'\beta)\right)$. Standard errors can be obtained from the Hessian, OP or their combination as QMLE. The default Hessian is preferred. A heteroscedastic model can be also fitted using, for example the variance model $Var(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2 \exp(\mathbf{z}_i'\gamma)$ as we did in the binary logit case. Note that σ^2 is a constant.

Illustrative Example 2

In example 1 the self-reported health status (HSAT) has a range 0 to 10. Suppose we create an ordinal response using the categories reflected in the format:

```
value ohsat 0-<3='0' 3-<6='1' 6-<9='2' 9='3' 10='4';

proc logistic data=c.healthcare(where=(year=1988));
class married(ref='no') hhkids(ref='no') female(ref='male')/param=ref;
model HSAT=age educ hhninc married hhkids female /link=cumlogit;
format female female. married hhkids affirm. hsat ohsat.;
run;
```

The responses are cumulated over the lower formatted values. Table 2 shows the estimation results for the homoscedastic cumulative logit model (columns 3-5) fitted in proc LOGISTIC. The LR test (5 DF) is for the model's δ-parameters. The estimate for AGE, for example, is 0.0322, which indicates that as people grow older, they are more likely to be in the lower end of the observed ordinal scale, i.e., having worse health. The proportional odds assumption maintains the same slope parameter across the 4 response levels. Response-specific slope parameters increase the number of parameters by 18. Unfortunately, overall, the proportional odds assumption is violated (score test χ^2 = 66.6, 18 DF, p<.0001). Proc QLIM may be used to fit the equivalent homoscedastic ordered logit model. The results (not shown) are the same for the threshold parameters, but the signs for the covariates are reversed because the β-parameters here are $-\delta$.

A heteroscedastic ordered logit model with $Var(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2 \exp(\mathbf{z}_i'\gamma)$ is fitted in QLIM (columns 6-8). The LR test for no heteroscedasticity (χ^2 =20.42, 3DF) is significant, p<.0001. Comparison of coefficients between the two models is meaningless. Instead, predicted probabilities and marginal effects could be compared.

```
proc qlim data=c.healthcare(where=(year=1988));
endogenous HSAT~discrete(dist=logistic order=formatted);
model HSAT=age educ hhninc married hhkids female/limit1=varying;
format female female. married hhkids affirm. hsat ohsat.;
hetero HSAT~HHNINC female age/link=exp noconst;
test 'NOHETERO' _H.HHNINC, _H.female, _H.age/all;
run;
```

Table 2: Ordinal Logit Models

		Homoscedastic Cumulative Logit			**Heteroscedastic Ordered Logit**		
Parameter		**Estimate**	**Standard Error**	**p–value**	**Estimate**	**Standard Error**	**p–value**
Intercept 1	α_1, μ_1	–3.5070	0.2197	<.0001	–3.8870	0.3535	<.0001
Intercept 2	α_2, μ_2	–1.3858	0.2105	<.0001	–1.5145	0.2465	<.0001
Intercept 3	α_3, μ_3	0.9275	0.2099	<.0001	0.9868	0.2294	<.0001
Intercept 4	α_4, μ_4	1.8707	0.2129	<.0001	1.9965	0.2525	<.0001
AGE		0.0322	0.0029	<.0001	–0.0352	0.0040	<.0001
EDUC		–0.0650	0.0127	<.0001	0.0711	0.0142	<.0001
HHNINC		–0.4254	0.1820	0.0194	0.4166	0.1945	0.0322
MARRIED	yes	0.0636	0.0738	0.3884	–0.0661	0.0814	0.4173
HHKIDS	yes	–0.1144	0.0671	0.0884	0.1295	0.0724	0.0735
FEMALE	**F**	–0.0130	0.0570	0.8199	0.0152	0.0620	0.8063
_H.HHNINC					–0.5438	0.1611	0.0007
_H.FEMALE	**F**				0.0391	0.0571	0.4931
_H.AGE					0.0078	0.0026	0.0027
–2 Log L		11489.26			11477.84		
–2 Log L (null)		11750.19			11750.19		
–2 Log LR		251.93			272.35		

Since the proportional odds assumption is violated in this example one might consider fitting a model with level-specific coefficients for the covariates. But $\pi_j(\mathbf{x}_i) = F(\mu_j - \mathbf{x}_i'\beta_j) - F(\mu_{j-1} - \mathbf{x}_i'\beta_{j-1})$ must be between 0 and 1, and the only way to assure this for all covariate values is to have $\mu_j > \mu_{j-1}$ and $\beta_j = \beta_{j-1}$. This is tantamount to assuming the proportional odds model. SAS Usage Note 22954 uses NLMIXED to fit a fully non-proportional odds model wherein each of the covariates is crossed with the response levels. The likelihood for optimization is constructed from the cumulative response probabilities $\gamma_j(\mathbf{x}_i)$. Whenever '$\pi_j(\mathbf{x}_i) \le 0$' for an observation its contribution to the likelihood is set to near zero, whilst if '$\pi_j(\mathbf{x}_i) > 1$' the contribution is set to 1. In this way we can assure that estimates of the response probabilities are properly constrained. Stokes *et al* (2000) provide another approach based on generalized estimating equations (GEE) for the vector of binary responses $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ-1})$ where $Y_{ij} = [Y_i \le j], j = 1,...,J-1$. This makes the marginal responses highly correlated. The GEE model for the marginal response is $\text{logit}\left(P[Y_{ij} = 1 \mid \mathbf{x}_i\right) = \mathbf{x}_{ij}'\beta$. Although this method does not guarantee appropriately constrained response probability estimates, it is easy to implement and generally, with data sets of moderate size, will yield proper probability estimates $\pi_j(\mathbf{x})$ unless $\mathbf{x}$ lies in the fringes of the covariate space (McCullagh and Nelder, 1989). Another alternative with non-proportionality of odds is to abandon the ordinal model altogether and regard the response as nominal. The multinomial model is described next.

Multinomial logit (generalized logit)

The **multinomial logit model** (MLM) makes the parameters specific to the nominal outcome. With subject–specific covariates only, the probability of response $j \in \{0,1,\ldots,J-1\}$ is $\pi_{ij} = \pi_j(\mathbf{x}_i) = \dfrac{\exp(\mathbf{x}_i'\beta_j)}{\sum_{j=0}^{J-1}\exp(\mathbf{x}_i'\beta_j)}$ with $\beta_0 = 0$ for identification. An intercept is included in each β_j . The MLM has the property of *independence from irrelevant alternatives* (IIA) because $\pi_{ij} / \pi_{ik} = \exp(\mathbf{x}_i'(\beta_j - \beta_k))$ depends only on the two outcomes (*j*, *k*). Having too many parameters is a serious drawback of the MLM. Since one outcome (*j*=0) is used as a reference we will have *J*–1 intercepts and (*J*–1)*p* regression coefficients, a total (*J*–1)(*p*+1) parameters. In the previous example on health status at 5 nominal levels and 6 covariates we have a MLM with 28 parameters. The proporatinal odds model on the other hand has 10 parameters.

The MLM can be estimated in LOGISTIC or GLIMMIX using the link=glogit option. A single record file per subject is used with only the observed nominal response Y_i and covariates $\mathbf{x}_i$. Proc MDC in SAS/ETS could also be used but requires a multiple-record input file — one record for each of the *J* alternatives. The dependent variable is numeric with value 1 for the observed response and zero for all other alternatives. All subject covariates need to be made response level–specific (crossed effects). Essentially MDC is fitting a conditional logit model (see next). See MDC documentation example 'Binary Data Modeling' for a description of the binary logit model as a choice model.

Conditional logit

A series of logit models was popularized by McFadden (1984) in the context of discrete choice. A person (indexed by i) is presented with a set of discrete choices C_i—for example, choice of health insurance plan or health care provider; or different treatment regimes (surgery, medical management, no treatment). The observed option $Y_i = j$ that the individual chooses can be thought of as the person's attempt to optimize his or her utility function $\{U_{ij} : j \in C_i\}$. The selected choice $Y_i = j$ is made because the person believes $U_{ij} \geq \max\{U_{il} : l \in C_i, l \neq j\}$. Different classes of choice models are obtained from an underling latent random utility model (RUM) $U_{ij} = \mathbf{x}'_{ij}\beta + \varepsilon_{ij}$ where $\{\varepsilon_{ij} : j \in C_i\}$ are random variables with a specified distribution.

The **conditional logit model** (CLM) assumes $\{\varepsilon_{ij} : j \in C_i\}$ are independent identically distributed (iid) extreme–value random variables, with distribution function $F(u) = \exp(-\exp(-u))$, $-\infty<u<\infty$. The computation of $\pi_{ij} = P[Y_i = j \mid \mathbf{x}_i] = P[\max\{U_{il} : l \in C_i, l \neq j\} < U_{ij} \mid \mathbf{x}_i]$ leads to the expression $\pi_{ij} = \dfrac{\exp(\mathbf{x}'_{ij}\beta)}{\sum_{l \in C_i} \exp(\mathbf{x}'_{il}\beta)}$. The CLM has the IIA property, that is, for any two alternatives (*j*, *k*) we have $\pi_{ij} / \pi_{ik} = \exp((\mathbf{x}_{ij} - \mathbf{x}_{ik})'\beta)$ which depends only on the characteristics of the two alternatives (*j*, *k*). A covariate that does not vary across alternatives does not enter the model because it is a constant multiplier to both the numerator and denominator of π_{ij}.

Estimation in the CLM is via maximization of the log–likelihood $\ell(\beta) = \sum_{i=1}^{n} \sum_{j} [Y_i = j] \log \pi_{ij}$. This is exactly the same objective function that one obtains in *conditional logistic regression* for matched case–control studies. The analogy is that the revealed choice from the set C_i is a 'case' whilst all remaining alternatives in the choice set are 'controls'. Therefore the CLM can be analyzed in proc LOGISTIC using the strata statement to identify the matched sets, whereas in proc MDC the id statement serves the same functionality. Both procedures require a multiple-record input file—one record for each alternative in C_i.

Illustrative Example 3

Allison (1999) describes a study of 147 murder cases. Each of 50 trial judges were asked to read 14 or 15 murder cases and rank them from the most serious (rank=1) to the least (up to 15). All cases were ranked and ties were allowed with ties given the average rank. For example, ties in the three most serious cases received average rank=2; ties in 5–th and 6–th cases got average rank=5.5. Each case was ranked by 4 to 6 judges. The data set JUDGERNK is arrayed as one record per case with the following characteristics of each case:

BLACKD= indicator for defendant being black; WHITVIC= indicator for victim being white; DEATH= indicator for death penalty; JUDGID identifies judges. Allison adds CULP an ordinal variable for culpability on a scale 1 to 5 derived from prediction of the death penalty. CULP is used here as another covariate although it is a generated regressor (Wooldridge, 2002).

When ranking the cases for seriousness the judges did not receive information on race or penalty. We first consider the 35 judges who gave a unique top rank (=1). Other cases ranked may have ties. The objective is to assess the relative importance of characteristics of the case that was ranked as most serious. In the data set JUDGRNK2 the variables CHOSEN and CHOSEN2 are created for convenience:

```
chosen=(rank=1);
chosen2=(rank=1)+2*(rank>1);
```

In the parlance of the choice model, π_{ij} is the probability that judge i ranks case j as the most serious amongst his or her portfolio of cases C_i. The case characteristics $\mathbf{x}_i$ = (BLACKD, WHITVIC, DEATH, CULP) vary across cases in C_i and across judges. Proc MDC is dedicated to fitting discrete choice models. The CLM is invoked via the type=clogit option in the model statement. An equivalent model statement is also shown but it uses only the first ranked choice, all other ranks are ignored. Future enhancements will provide flexibility of analyzing rank–ordered responses.

```
proc mdc data=judgernk2 covest=hess;
id judgid;
model chosen=blackd whitvic death CULP/type=clogit choice=(rank);
*model rank=blackd whitvic death CULP/type=clogit choice=(rank) rank;
output out=stats_q pred=phat_q xbeta=xbeta_q;
run;
```

Exactly the same model is fitted by LOGISITC via

```
proc logistic data=judgernk2;
strata judgid;
model chosen(event='1')=blackd whitvic death CULP;
run;
```

Table 3: Choice models

	First ranked choice			**Ranked choices**		
Parameter	**Estimate**	**Standard Error**	**p–value**	**Estimate**	**Standard Error**	**p–value**
BLACKD	0.2043	0.4124	0.6204	0.1195	0.0971	0.2185
WHITVIC	0.3631	0.4266	0.3947	0.2370	0.1046	0.0235
DEATH	–0.4339	0.4821	0.3681	–0.1818	0.1377	0.1866
CULP	0.5311	0.1415	0.0002	0.2586	0.0423	<.0001

In Table 3 (columns 2–4) the only significant coefficient is CULP indicating that an increase in this variable is associated with an increase in the probability of a case being ranked as most serious. In fact the partial effects for continuous covariates are $\frac{\partial \pi_{ij}}{\partial \mathbf{x}_{ik}} = \beta\pi_{ij}([k=j]-\pi_{ik})$. For a discrete covariate the partial effect should be derived as differences in probabilities. The OUTPUT statement will compute the probability that each case in the input file is ranked first. For each judge these probabilities for the portfolio C_i must sum to 1. Note that only case characteristics are used in π_{ij}. This does not mean that cases with same values for BLACKD, WHITVIC, DEATH and CULP rated by different judges will have the same probability of receiving the most serious rank. The reason is that the choice set could be different for different judges.

Neither MDC nor LOGISTIC will compute a confidence interval for the choice probabilities. However, using a survival model that has the same likelihood as the choice model, PHREG would allow computation of confidence intervals. The variable CHOSEN2 is regarded as an event time with the first ranked case having value 1 and all other cases having value 2, which is treated as censored. Contribution to the partial log–likelihood by the potential times for cases $j \in C_i$ for judge i is the aforementioned $\ell(\beta)$. The absence of ties amongst the event times makes the likelihoods— Breslow, Efron, discrete all the same. In the parlance of survival analysis, the estimated cumulative hazard at time t is $\hat{H}_i(t\,|\,\mathbf{z}_0) = H_{i0}(t,\hat{\beta})\exp(\mathbf{z}_0'\hat{\beta})$ where $\mathbf{z}_0$ is a profile of a case, $H_{i0}(t,\hat{\beta}) = \int_0^t \{S_i^{(0)}(u,\hat{\beta})\}^{-1} dN_i(u)$, $S_i^{(0)}(t,\hat{\beta}) = \sum_{j\in C_i} Y_{ij}(t)\exp(\mathbf{x}_{ij}\hat{\beta})$, $Y_{ij}(t)$ is the indicator for cases to be ranked at time t, and $N_i(t)$ is the counting process for ranked cases up to time t. We have just one event time (=1) which yields the desired choice probability for case profile $\mathbf{z}_0$. Note that the profile $\mathbf{z}_0$ need not be one of the cases in the portfolio. This fact has important implications in application of discrete choice models in marketing research where the available constellation of choice characteristics could be extremely large.

PHREG computes a confidence interval for $S_i(1\,|\,\mathbf{z}_0) = \exp(-H_i(1\,|\,\mathbf{z}_0))$ from a confidence interval for $\log\left(-\log(S_i(1\,|\,\mathbf{z}_0)\right) = \log H_{i0}(1,\beta) + \mathbf{z}_0'\beta$. This can be salvaged to get the desired confidence interval via the approximation $1 - S_i(1\,|\,\mathbf{z}_0) = 1 - \exp(-H_i(1\,|\,\mathbf{z}_0)) \approx H_i(1\,|\,\mathbf{z}_0)$. As an alternative, one could use directly the variance of $\hat{H}_i(1\,|\,\mathbf{z}_0)$ to do the calculations, $Var(\hat{H}_i(1\,|\,\mathbf{z}_0)) \approx Var(\hat{S}_i(1\,|\,\mathbf{z}_0)) / \hat{S}_i^2(1\,|\,\mathbf{z}_0)$.

Exploded logit (rank–ordered logit) model

The exploded logit model uses the rankings of the utilities in the RUM $U_{ij} = \mathbf{x}_{ij}'\beta + \varepsilon_{ij}$ where we maintain the assumption that the errors $\{\varepsilon_{ij} : j \in C_i\}$ are distributed iid extreme–value. The observed responses are the rank order of the utilities of the choices, instead of the single choice that corresponds to the maximum utility. For example, without loss of generality suppose there are J alternatives and individual i ranks them as $Y_{i1} = \max\{U_{ij} : j \in C_i\} = U_{i1}$, $Y_{i2} = \max\{U_{ij} : j \in C_i, j > 1\} = U_{i2}, \ldots,$ $Y_{iJ} = U_{iJ}$. The observed response is only the rank order $U_{i1} > U_{i2} > \ldots > U_{iJ}$. The response probability is computed as $P[U_{i1} > U_{i2} > \ldots > U_{iJ}]$.We may allow for incomplete rankings with the first J_1 alternatives being ranked keeping the remaining $J-J_1$ unranked. The probability of response is then $P[U_{i1} > \ldots > U_{iJ_1} > \tilde{U}_{iJ_1}]$ where $\tilde{U}_{iJ_1} = \max\{U_{ij} : j > J_1\}$. Ties among ranks are theoretically not possible under the continuous utility specification. However, see below.

Use the fact that $X_{ij} = \exp(-U_{ij})$ has the exponential distribution with mean $(\lambda_{ij})^{-1}$ where $\lambda_{ij} = \exp(\mathbf{x}_{ij}'\beta)$. For a subset $A \subseteq C_i$, $\min\{X_{ij} : j \in A\}$ is exponentially distributed with inverse scale $\sum_{j\in A}\lambda_{ij}$. A pedestrian calculation yields $P[U_{i1} > \ldots > U_{iJ_1} > \tilde{U}_{iJ_1}] = \left(\frac{\lambda_{i1}}{\sum_{j\geq 1}\lambda_{ij}}\right)\left(\frac{\lambda_{i2}}{\sum_{j\geq 2}\lambda_{ij}}\right)\ldots\left(\frac{\lambda_{iJ_1}}{\sum_{j\geq J_1}\lambda_{ij}}\right)$. The structure makes the term exploded logit to describe this model quite appropriate. The overall likelihood is the product of such terms across the sample.

The form of this likelihood is exactly the same as the Breslow likelihood for observed survival times $T_{i1} < T_{i2} < \ldots < T_{iJ_1}$ in a sample of J potential events time of which the last $J-J_1$ are censored. Therefore to analyze these data on the preference ranks we can use PHREG with the survival times $1<2<\ldots<J_1$ for the first J_1 ranked alternatives and a censored value (=J_1+1) for the last $J-J_1$ unranked alternatives. Of course the actual "times" are immaterial as long as the order is preserved.

If there are ties amongst the preference ranks an acceptable approach is to modify the above likelihood terms as follows. Suppose alternatives $j_1,\ldots,j_p$ have the same rank r and R denotes all subsets of p alternatives amongst those that might receive a rank r or worse. Let $\mathbf{q}=(q_1,\ldots,q_p)$ denote subscripts for the p alternatives in a subset $\mathbf{q}\in R$. The corresponding term(s) in the likelihood is replaced by

$\left(\dfrac{\exp\left((\sum_{k=1}^{p}\mathbf{x}_{ij_k})'\beta\right)}{\sum_{\mathbf{q}\in R}\exp\left((\sum_{l=1}^{p}\mathbf{x}_{iq_l})'\beta\right)}\right)$. Allison suggests that this discrete logistic likelihood should be used with tied ranked data. Estimation is readily carried out in PHREG with the TIES=DISCRETE option to invoke use of this likelihood. The response times are the observed ranks 1, 2,…, allowing for ties.

Illustrative Example 4

Use the data set JUDGERNK with the syntax

```
proc phreg data=judgernk;
strata judgid;
model rank=blackd whitvic death CULP/ties=discrete;
output out=stats xbeta=xbeta logsurv=logsurv survival=survival/method=ch;
run;
```

The parameter estimates are shown in Table 3 (columns 5–7). Strictly speaking the results are not comparable with the analysis of the first ranked choice because inter alia the data sets used and the underlying models are different. Output statistics generated for the rank–ordered model must be interpreted with some caution because PHREG is operating in the context of a survival model.

Let us carry out a few calculations (see Table 4). JUDGID=11 provided unique ranks to his portfolio of 15 cases. The probability of case=4163 (obs=1) being ranked first is $P[U_{i1}>\max\{U_{ij}:j>1\}]=\dfrac{\exp(\mathbf{x}_{i1}'\beta)}{\sum_{j\geq 1}\exp(\mathbf{x}_{ij}'\beta)}$.

For obs=1 we get 3.1026/31.8052 =0.0976 which is the cumulative hazard $H(1)$= –logsurv. Survival is computed as $S(1)=P[RANK>1]=\exp(-H(1))=0.9071$.

Dividing each exp_xb=exp(xbeta) by the sum across all cases gives the probability of first rank for each case. Although the observed ranks differ in obs=11, 14 and 15, they have the same case characteristics and will therefore have the same probability of having the first rank. Obs=2 is a different case.

$P[U_{i1}>U_{i2}>\max\{U_{ij}:j>2\}]=\left(\dfrac{\exp(\mathbf{x}_{i1}'\beta)}{\sum_{j\geq 1}\exp(\mathbf{x}_{ij}'\beta)}\right)\left(\dfrac{\exp(\mathbf{x}_{i2}'\beta)}{\sum_{j\geq 2}\exp(\mathbf{x}_{ij}'\beta)}\right)$ =0.0976×2.2954/28.7026=0.0078.

For obs=2 we have $H(2)=\left((31.8052)^{-1}+(28.7026)^{-1}\right)\times 2.2954=0.15215$. Survival for this record is $S(2)=P[RANK>2]=\exp(-H(2))=0.8587$. We notice that we cannot easily use these results to compute the choice probabilities, other than for the first ranked choice. Moreover, if there were tied ranks at the first choice then $\Delta N_i(1)>1$ and we cannot use $H(1)$ directly to obtain choice probabilities.

Future enhancements to proc MDC for ranked choice response data are likely to address these issues. Currently, in SAS/ETS 9.2, MDC utilizes only the first ranked value (=1) in estimation, ignoring the rest, basically fitting a conditional logit model.

Table 4: Output statistics for ranked choice (exploded logit) model

Obs	judgid	rank	blackd	whitvic	death	culp	case	xbeta	survival	logsurv	exp_xb
1	11	1	1	1	0	3	4163	1.13224	0.90706	–0.09755	3.1026
2	11	2	0	1	1	3	2172	0.83093	0.85886	–0.15215	2.2954
3	11	3	1	1	0	1	1880	0.61508	0.82477	–0.19266	1.8498
4	11	4	1	0	1	5	2015	1.23062	0.60900	–0.49594	3.4234
5	11	5	0	0	0	1	1060	0.25858	0.77966	–0.24890	1.2951
6	11	6	1	1	0	1	2375	0.61508	0.63843	–0.44875	1.8498
7	11	7	1	0	1	2	1720	0.45487	0.62505	–0.46993	1.5760
8	11	8	1	0	1	5	1598	1.23062	0.29248	–1.22936	3.4234
9	11	9	1	0	1	2	197	0.45487	0.50295	–0.68727	1.5760
10	11	10	0	1	1	5	119	1.34809	0.13315	–2.01630	3.8501
11	11	11	1	0	0	1	3035	0.37809	0.38393	–0.95731	1.4595
12	11	12	1	0	0	2	4142	0.63668	0.21236	–1.54945	1.8902
13	11	13	0	0	0	1	4128	0.25858	0.25437	–1.36896	1.2951
14	11	14	1	0	0	1	1791	0.37809	0.12967	–2.04274	1.4595
15	11	15	1	0	0	1	463	0.37809	0.04770	–3.04274	1.4595
										Sum	**31.8052**

Nested logit

As an extension of the conditional logit model suppose the choice alternatives are partitioned into *K* non-overlapping nests, $B_1,\ldots,B_K$ (McFadden, 1984)**.** The observed response for the *i*-th subject is the revealed choice (*j*) within the nest (*k*), that is, $Y_i = j, j \in B_k$. Suppressing the subject index, the underlying latent utility model is $U_{kj} = V_{kj} + \varepsilon_{kj}$ where V_{kj} will be specified later. Within the nest B_k the errors $\boldsymbol{\varepsilon}_k = \{\varepsilon_{kj} : j \in B_k\}$ have a joint cumulative distribution $F(\mathbf{u}_k) = \exp\left(-\left(\sum_{j\in B_k} \exp(-u_{kj} / \theta_k)\right)^{\theta_k}\right)$, called the *generalized extreme-value distribution* (GEV, Train, 2003). To ensure that $F(\mathbf{u}_k)$ is a proper distribution we require $\theta_k \in (0,1]$ for *k*=1, 2,…, *K*. Across nests the errors are independent. The reasoning behind this specification originates from the marginal distribution of each $\varepsilon_{kj}, j \in B_k$, which is assumed to be the standard extreme-value distribution $\Lambda(u) = \exp(-\exp(-u)), -\infty < u < \infty$. We then generate the GEV distribution $F(\mathbf{u}_k) = C\left(\Lambda(u_{k1}), \Lambda(u_{k2}), \ldots\right)$ for $\boldsymbol{\varepsilon}_k$ via the Gumbel-Hougaard (GH) copula $C(\mathbf{v}) = \exp\left(-\left(\sum_{j\in B_k} (-\log v_j)^{1/\theta_k}\right)^{\theta_k}\right)$, $v_j \in [0,1]$ (Nelson, 1999).

The GH copula belongs to the Archimedean Family of Copulas which is generated by a continuous convex strictly decreasing function $\varphi : [0,1] \to [0,\infty]$. Then $\varphi(C(\mathbf{v})) = \sum_{j\in B_k} \varphi(v_j)$. For the GH copula $\varphi(v) = (-\log v)^{1/\theta_k}$. Kendall's tau (τ) assesses the association between two marginals $(\varepsilon_{k1}, \varepsilon_{k2})$. It is the difference of the probability of concordance $P[(\varepsilon_{k1}-\varepsilon'_{k1})(\varepsilon_{k2}-\varepsilon'_{k2})>0]$ and the probability of discordance $P[(\varepsilon_{k1}-\varepsilon'_{k1})(\varepsilon_{k2}-\varepsilon'_{k2})<0]$ where $(\varepsilon'_{k1},\varepsilon'_{k2})$ is an independent copy of $(\varepsilon_{k1},\varepsilon_{k2})$. For the GH copula $\tau=1+4\int_0^1\{\varphi(v)/\varphi'(v)\}dv=1-\theta_k$. Also $Corr(\varepsilon_{kj},\varepsilon_{kl})=1-\theta_k^2$ (Kotz and Nadarajah, 2000). Generally, we require $\theta_k\in(0,1]$ which makes the RUM consistent with utility maximization. If $\theta_k=1$ for all *k*, then $\{\varepsilon_{kj}: j\in B_k\}$ are iid extreme-value random variables.

With the specification of the GEV for $\boldsymbol{\varepsilon}_k$ the choice probability π_{kj} for a subject with choice (*j*) within the nest (*k*) is computed as $\pi_{kj}=P[Y_i=j\mid B_k]P[Y_i\in B_k]$, with the maximum utility across nests being in B_k. This is the *nested logit model* with level 1 for alternatives (*j*) and a level 2 layer for nests (*k*). For multiple layers the formulation becomes more complex in its notation (Hensher *et al*, 2005). Analogous to a tree structure a 4-layer nested model has alternatives (level 1) nested in branches (level 2) that are in limbs (level 3) of trunks (level 4) of the root.

The expression for a 2-level choice probability is $\pi_{kj}=\left(\dfrac{\exp(V_{kj}/\theta_k)}{\sum_{l\in B_k}\exp(V_{kl}/\theta_k)}\right)\dfrac{\left(\sum_{l\in B_k}\exp(V_{kl}/\theta_k)\right)^{\theta_k}}{\sum_{k=1}^{K}\left(\sum_{l\in B_k}\exp(V_{kl}/\theta_k)\right)^{\theta_k}}$.

For alternatives j,m in B_k we have $\pi_{kj}/\pi_{km}=\exp\left((V_{kj}-V_{km})/\theta_k\right)$ which maintains the IIA property within a nest. However, if $j\in B_k, m\in B_l, m\neq l$ then

$\pi_{kj}/\pi_{lm}=\exp\left((V_{kj}/\theta_k)-(V_{lm}/\theta_l)\right)\dfrac{\left(\sum_{h\in B_k}\exp(V_{kh}/\theta_k)\right)^{\theta_k-1}}{\left(\sum_{h\in B_l}\exp(V_{lh}/\theta_l)\right)^{\theta_l-1}}$ depends on alternatives in the nests B_k, B_l.

When $\theta_k=1$ for all *k*, the 2-level nested logit model reduces the conditional logit model.

To incorporate covariates let $V_{kj}=\mathbf{z}'_k\alpha+\mathbf{x}'_{kj}\beta$ with variables that depend only on the nest and variables that depend on alternatives (within the nest). Note that the additional subject index (*i*) is suppressed. Then the *utility-maximized nested logit model* (UMNL) is

$$\begin{aligned}\pi_{kj}&=\left(\frac{\exp(\mathbf{x}'_{kj}\beta/\theta_k)}{\sum_{l\in B_k}\exp(\mathbf{x}'_{kl}\beta/\theta_k)}\right)\frac{\exp(\mathbf{z}'_k\alpha)\left(\sum_{l\in B_k}\exp(\mathbf{x}'_{kl}\beta/\theta_k)\right)^{\theta_k}}{\sum_{k=1}^{K}\exp(\mathbf{z}'_k\alpha)\left(\sum_{l\in B_k}\exp(\mathbf{x}'_{kl}\beta/\theta_k)\right)^{\theta_k}}\\&=\left(\frac{\exp(\mathbf{x}'_{kj}\beta/\theta_k)}{\sum_{l\in B_k}\exp(\mathbf{x}'_{kl}\beta/\theta_k)}\right)\left(\frac{\exp(\mathbf{z}'_k\alpha+\theta_kI_k)}{\sum_{k=1}^{K}\exp(\mathbf{z}'_k\alpha+\theta_kI_k)}\right)\qquad\qquad UMNL\end{aligned}$$

where $I_k=\log\left(\sum_{l\in B_k}\exp(\mathbf{x}'_{kl}\beta/\theta_k)\right)$ $k=1,\ldots,K$ are called the *inclusive values*. The scale parameters θ_k in the GEV could be referred to as the inclusive value parameters. From the GEV distribution we can compute the expected maximum utility from alternatives in B_k: $E\left(\max(U_{kj}: j\in B_k)\right)=\theta_kI_k+\mathbf{z}'_k\alpha+\gamma$ where γ is a (Euler) constant. The first term in π_{kj} is the conditional probability of selected choice being j in the nest B_k, given that the maximum utility across nests is the nest B_k. The second term is the probability of $M_k=\max(U_{kj}: j\in B_k)$ being the maximum utility across nests, i.e., $P[M_k>\max(M_l: l\neq k, 1\le l\le K)]$.

Proc MDC does not fit the UMNL (Silberhorn *et al*, 2008). Instead it fits the *non-normalized nested logit model* (NNNL) where β / θ_k is replaced by β yielding the formula:

$$\pi_{kj} = \left(\frac{\exp(\mathbf{x}'_{kj}\beta)}{\sum_{l \in B_k} \exp(\mathbf{x}'_{kl}\beta)} \right) \left(\frac{\exp(\mathbf{z}'_k \alpha + \theta_k I_k)}{\sum_{k=1}^{K} \exp(\mathbf{z}'_k \alpha + \theta_k I_k)} \right) \qquad \textit{NNNL}$$

The NNNL places no restrictions on the parameters θ_k. If all θ_k are constrained to be equal, the UMNL model reduces to the NNNL model. In MDC we could estimate nest-specific coefficients by replacing β / θ_k with β_k. The corresponding UMNL beta coefficients are $\beta_k \times \theta_k$. This means that we are estimating alternative-specific effects that differ by nest. Unless the intended application can support a complex structure, fitting such a model would be unwieldy and its interpretation a challenge.

For each individual i in a random sample, the observation is the revealed choice and the nest to which it belongs, i.e., $Y_i = j$ and the nest B_k with $j \in B_k$. With individual covariate values $(\mathbf{x}_{kji}, \mathbf{z}_{ki})$ the log-likelihood is $\ell(\alpha, \beta) = \sum_{i=1}^{n} \sum_{k,j} [Y_i = j, j \in B_k] \log \pi_{kj}(\mathbf{x}_{kji}, \mathbf{z}_{ki})$.

Illustrative Example 4

Brownstone and Small (1989) describe a study of 527 automobile commuters from home to their work place. The choice of arrival time at work consists of 12 alternatives based on their preference for arriving at work early, on-time or late relative to the official work-start time. Early arrivals (ALT 1-8) have a schedule delay (SD) of between –40 min to –5 min in 5 min increments; on-time arrivals (ALT 9) of course have SD=0; and late arrivals (ALT 10-12) have SD 5, 10 or 15 min. The binary DECISION (0 or 1) revealed the arrival time choice. About 35% (n=187) of commuters chose on-time arrival, and only 5% (n=22) favored late arrival. Data on travel time (TTIME) in minutes were obtained from actual work-arrival time, official start-time at work supplemented by calculations for each commuter for each alternative. For the chosen alternative (DECISION=1), as expected TTIME was on average slightly longer for carpoolers (CP=1, n=156) than non-carpoolers (CP=0, n=371). Some individuals had the flexibility of late arrival at work without any consequence. The binary variable D2L=[SD≥FLEX] indicates schedule delay in excess of flex time (FLEX in mins). So D2L=0 for ALT 1-8. The variable SDLX=(SD–FLEX)/10, if SD>FLEX, and 0 otherwise, measures schedule delay in excess of allowed FLEX. We define an indicator FL=[FLEX>0] for commuters who had flexibility of late arrival (FL=1, n=193), and those who did not (FL=0, n=334).

Four variables describe characteristics of alternatives only: SDE=(–SD/10)×[SD<0] for schedule delay for early arrival; SDL =(SD/10)×[SD>0] for schedule delay for late arrival. Note that (SDE, SDL)=(0,0) only for ALT=9, on-time arrival. Binary variables R15=[SD∈{–30, –15, 0, 15}] and R10 =[SD∈{–40, –30, –20, –10, 0, 10}] capture the tendency of respondents to round off answers to their schedule delay time to 15 minutes and 10 minutes, respectively.

The figure depicts the tree structure for schedule delay. This is a 2-level model. Level 1 at the bottom shows the alternatives which are nested at level 2 in three nests. The nests are joined at the top of the tree.

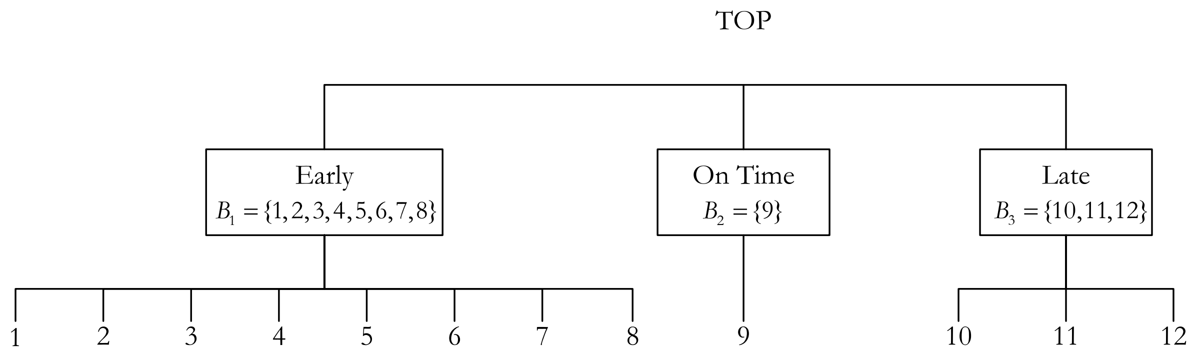
The NEST statement in proc MDC for the nesting of level 1 alternatives in level 2 nests is

```
nest level(1) = (1 2 3 4 5 6 7 8 @ 1, 9 @ 2, 10 11 12 @ 3),
     level(2) = (1 2 3 @ 1);
```

Covariates for the model are specified through the UTILITY statement. The general specification is

```
utility u(level,alternatives@nest)=;
```

Figure: Tree Structure for Schedule Delay



Although proc MDC permits some flexibility in covariate specifications, having too many alternative-specific covariates builds an unwieldy model that is likely at best to be un-interpretable, let alone being able to fit properly (convergence problems). In most applications one would use a set of covariates that are common to all alternatives. All covariates in the model must appear in the MODEL statement.

The data set SMALL may be accessed from the SAS Sample Library for the MDC procedure. Individual commuters are identified by ID, there are 12 records per individual corresponding to ALT= 1-12, for a total of 527×12= 6324 records. For additional description and analysis of this data set see Brownstone and Small (1989), Small (1982) and the documentation example 'Nested Logit Analysis' in MDC. For illustrative purposes and demonstration of different nested logit models we will use the following covariates:

Level 1: R10, R15, TTIME, SDE, SDL, SDLX, D2L
Level 2: CP_2, FL_2, CP_3, FL_3.

The level 2 variables are indicators for CP and FL specific to nest=2 (ALT 9), and nest=3 (ALT=10-12). For nest=1 (ALT 1-8) all four variables are zero. Note that these level 2 variables are subject-specific. They are constant across the alternatives within each nest.

Table 5 summarizes the output from fitting different NNNL models.

Model A: Covariates at level 1 only.

```
proc mdc data=small maxit=200 covest=hess;
model decision = r15 r10 ttime sde sdl sdlx d2l/
            type=nlogit
            choice=(alt);
id id;
utility u(1, )= r15 r10 ttime sde sdl sdlx d2l;
nest level(1) = (1 2 3 4 5 6 7 8 @ 1, 9 @ 2, 10 11 12 @ 3),
     level(2) = (1 2 3 @ 1);
run;
```

The labeling of the parameters β in the output is self-explanatory. The inclusive value parameters $\theta_1, \theta_2, \theta_3$ are named INC_L2G1C1, INC_L2G1C2, INC_L2G1C3. The three nests at level 2 (L2) form a single group (G1) at the top of the tree (see Figure).

Model B: Covariates at level 1 only with the restriction $\theta_1 = \theta_2 = \theta_3$.

As previously noted, model B will be consistent with utility maximization. The restriction is accomplished by adding the option SAMESCALE to the model statement. The LR test for model B versus model A has –2 log LR=8.03. The 2 DF chi-square test is significant (p=.018). The test can be carried out within the invocation for fitting model A by adding the TEST statement:

```
test "SAME SCALE"  INC_L2G1C1=INC_L2G1C2=INC_L2G1C3/LR;
```

Model C: Covariates at level 1 only with the restriction $\theta_2 = 1$.

Nest 2 is degenerate because it has a single alternative associated with it. The inclusive value parameter θ_2 is not defined for the UMNL but identifiable in the NNNL. To impose the restriction, add the RESTRICT statement to model A:

```
restrict "THETA2=1" INC_L2G1C2=1;
```

The LR test for model C versus model A is not significant.

Model D: Covariates at level 1 and level 2.

The syntax modifies the UTILITY statement, imposes bounds on θ_1 and θ_3 via a BOUNDS statement, and restricts $\theta_2 = 1$ as before.

```
proc mdc data=small maxit=250 covest=hess;
bounds 0<INC_L2G1C1<=1, 0<INC_L2G1C3<=1;
model decision = r15 r10 ttime sde sdl sdlx d2l cp_2 FL_2 cp_3 FL_3/
            type=nlogit
            choice=(alt);
id id;
utility u(1, ) = r15 r10 ttime sde sdl sdlx d2l,
        u(2, 1 2 3@1)=cp_2 fl_2 cp_3 fl_3;

   nest level(1) = (1 2 3 4 5 6 7 8 @ 1, 9 @ 2, 10 11 12 @ 3),
        level(2) = (1 2 3 @ 1);
restrict "THETA2=1" INC_L2G1C2=1;
run;
```

The utility specification for Model D is $U_{kj} = \mathbf{z}_k'\alpha + \mathbf{x}_{kj}'\beta + \varepsilon_{kj}$ where

$$\mathbf{x}_{kj}'\beta = \beta_1 \mathrm{R10} + \beta_2 \mathrm{R15} + \beta_3 TTIME + \beta_4 SDE + \beta_5 SDL + \beta_6 SDLX + \beta_7 D2L \,,$$

$$\mathbf{z}_k'\alpha = \alpha_{11} CP\_2 + \alpha_{12} FL\_2 + \alpha_{21} CP\_3 + \alpha_{22} FL\_3 \,.$$

An increase of 1 min in travel time is associated with an expected disutility β_3 =–0.0933, whereas arriving at work a minute earlier has a disutility of β_4 =–0.6490/10 (recall that SDE is scaled by 10). The marginal rate of substitution is $\Delta TTIME_{kj} / (-\Delta SLE_{kj}) \approx (\frac{\partial EU_{kj}}{\partial SLE_{kj}}) / (\frac{\partial EU_{kj}}{\partial TTIME_{kj}}) = 10^{-1} \beta_4 / \beta_3 = 0.70$. So a commuter will incur 0.70 minutes of extra travel time to avoid arriving an extra minute early. In all models the negative sign on the β-coefficients for variables associated with time signify their disutility. The α-coefficients in model D are subject-specific. For example, with all other variables held constant, α_{21} is the difference in utility between a commuter who carpools and arrives late, and a commuter who does not carpool and arrives late. The negative sign on the estimate seems plausible, reflecting perhaps the perceived inconvenience of having to travel with others. A Wald test is not significant (p=.4052). By default MDC produces Wald tests for all parameters in the model. The TEST statement carries out hypotheses tests for linear combinations of model parameters through the LR, Wald, or Lagrange multiplier (score) chi-square tests.

Estimates of choice probabilities are computed for each record in the data set from an OUTPUT statement:

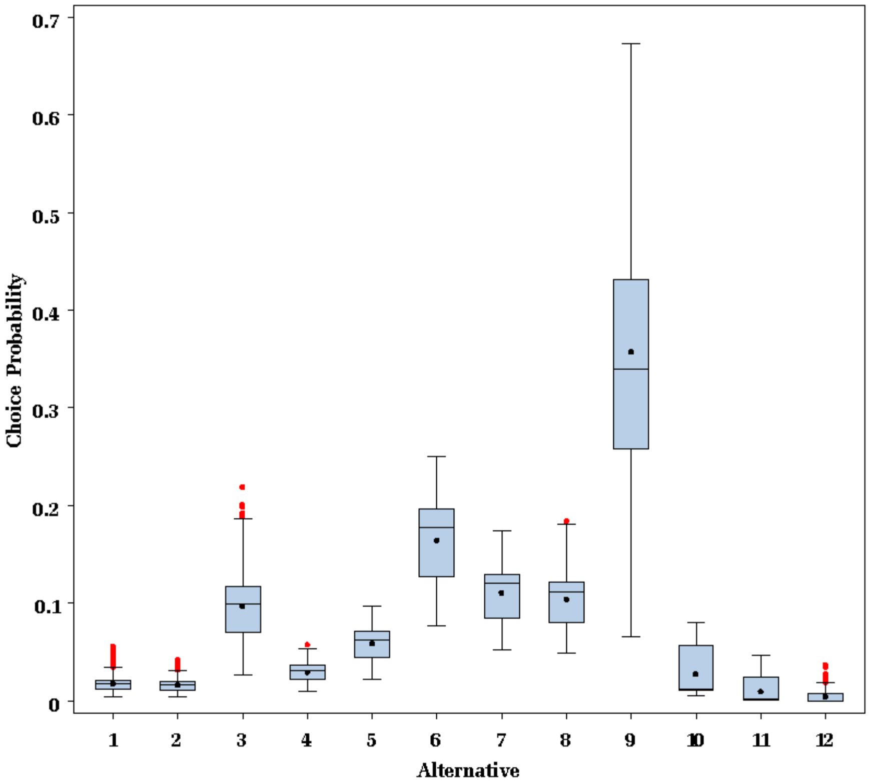
```
output out=stats_mdc predicted=phat;
```

The distribution of values of phat by alternative are shown in the boxplot. Each box represents 527 estimates.

Table 5: Non-normalized Nested Logit Models

	Model A		Model B		Model C		Model D	
Parameter	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
r15_L1	1.1455	0.1234	1.1404	0.1104	1.1300	0.1118	1.0996	0.1260
r10_L1	0.4344	0.1202	0.4260	0.1096	0.4203	0.1105	0.3862	0.1250
ttime_L1	–0.0803	0.0361	–0.1072	0.0441	–0.0752	0.0290	–0.0933	0.0367
sde_L1	–0.6711	0.0760	–0.6765	0.0572	–0.6623	0.0693	–0.6490	0.0710
sdl_L1	–2.1683	0.5036	–2.1960	0.4994	–2.1146	0.4649	–2.3154	0.6865
sdlx_L1	–3.4391	1.5077	–3.1042	1.3509	–3.3737	1.4740	–2.5152	1.7652
d2L_L1	–1.2057	0.3665	–1.3962	0.3640	–1.1183	0.1897	–0.7994	0.2630
INC_L2G1C1	0.5992	0.2547	0.7471	0.1521	0.6574	0.1735	0.7641	0.1304
INC_L2G1C2	0.9133	0.2782	0.7471	0.1521	1.0000		1.0000	
INC_L2G1C3	0.7436	0.1543	0.7471	0.1521	0.7694	0.1387	0.8730	0.1803
CP_2_L2G1							–0.7075	0.2268
FL_2_L2G1							0.4282	0.2862
CP_3_L2G1							–0.4096	0.4921
FL_3_L2G1							0.5630	0.7200
–Log L	993.53		997.54		993.58		988.19	

Boxplot: Distribution of estimates of choice probabilities (Model D)



PROC NLP is harnessed to carry out the maximum likelihood estimation for two UMNL models E and F (Table 6) considered as counterparts to NNNL models C and D. In model F having covariates at level 2 appears to be detrimental as most alternative-specific variables are not significant. All alternative-specific coefficients are scaled by the corresponding inclusive value parameters θ_1 or θ_3, but θ_2 is not defined and thus fixed at value 1. Initial parameters for the NLP procedure (inest= option) were from the NNNL models, and initial results from NLP were used in subsequent iterations of NLP with the hope of improving convergence and precision (i.e., small gradients). Although the results in Table 6 are satisfactory, we are unsure if additional improvements are possible using the myriad of options available in NLP.

SUMMARY

SAS Usage Note 22871 summarizes the types of logit models that can be fitted with SAS software. In this paper we described some of the capabilities of SAS procedures LOGISTIC, GENMOD, PHREG, QLIM and MDC in fitting a variety of logit models. We covered the binary logit for a dichotomous response, the ordinal and cumulative logit for ordered responses, the multinomial (or generalized) logit for nominal responses, and the exploded logit model for ranked responses. The latter used PHREG for analysis by exploiting the analogy between the ranked outcomes and a discrete time survival model. For discrete choice models, the conditional logit and nested logit models were discussed. The conditional logit model (CLM) is structurally similar to conditional logistic regression (CLR) for matched case-control data. However, important differences exist in interpretation of results from CLR and CLM because of differences in study design. For all models discussed in this paper estimation of model parameters is via maximization of an appropriate objective function, which is generally a log-likelihood function.

Although we focused on a single categorical response, there are natural extensions to longitudinal and clustered data. In specific contexts GLIMMIX and GENMOD could be used to account for correlation in repeated measures. CATMOD performs categorical data analyses for data structures that are presented as multidimensional contingency tables, using weighted least-squares for estimation. Some logit models not discussed in this paper are the continuation-ratio, adjacent-category models for ordinal responses, the stereotype models for ordered and multinomial responses, and mixed-logit model in the context of discrete choice. Finally, we note that using the term *logit* broadly to describe structurally very different models might seem overly simplistic.

Table 6: Utility Maximized Nested Logit Models

	Model E			**Model F**		
Parameter	**Estimate**	**Standard Error**	**p-value**	**Estimate**	**Standard Error**	**p-value**
r15_L1	0.7868	0.2951	0.0079	0.6852	0.5418	0.2066
r10_L1	0.2879	0.1369	0.0359	0.2328	0.2117	0.2720
ttime_L1	–0.0765	0.0365	0.0369	–0.0696	0.0512	0.1745
sde_L1	–0.4698	0.1804	0.0095	–0.4069	0.3216	0.2063
sdl_L1	–1.8759	0.9548	0.0500	–1.8602	1.1545	0.1077
sdlx_L1	–2.3989	0.8865	0.0070	–2.7819	1.7467	0.1118
d2L_L1	–1.0429	0.1592	<.0001	–0.7970	0.1812	<.0001
INC_L2G1C1	0.6866	0.2693	0.0111	0.6177	0.4988	0.2161
INC_L2G1C2	1.0000			1.0000		
INC_L2G1C3	0.8872	0.5545	0.1102	0.9357	0.6083	0.1246
CP_2_L2G1				–0.7849	0.2241	0.0005
FL_2_L2G1				0.4140	0.2169	0.0568
CP_3_L2G1				–0.4661	0.4896	0.3416
FL_3_L2G1				0.0860	1.0276	0.9333
– Log L	997.50			990.89		

DATA SOURCES

The German Socioeconomic Panel Survey 1984-1995 on healthcare utilization used in examples 1 and 2 is discussed extensively in Greene and Hensher (2010). The judge rank data set used in example 3 is from Allison (1999). The travel time data set of commuters used in example 4 can be obtained from the SAS Sample Program Library for the MDC procedure.

REFERENCES

Agresti A. *Categorical Data Analysis, Second edition*. New York: John Wiley & Sons; 2002.

Allison PD. *Logistic Regression Using the SAS System*. Cary, NC: SAS Institute Inc; 1999.

Greene WG, Hensher DA. *Modeling Ordered Choices: A Primer*. New York, NY: Cambridge University Press; 2010.

Hensher DA, Rose JM, Greene WH. *Applied Choice Analysis: A Primer*. New York, NY: Cambridge University Press; 2005.

Kotz S, Nadarajah S. *Extreme Value Distributions: Theory and Applications*. London, UK: Imperial College Press; 2000.

Kuhfeld WF. *Marketing Research Methods in SAS: Experimental Design, Choice, Conjoint, and Graphical Techniques.* Cary, NC: SAS Institute Inc; 2009.

McFadden D. Econometric analysis of qualitative response models. In: Griliches Z, Intriligator MD, eds. *Handbook of Econometrics, Volume 2*. Amsterdam: North-Holland; 1984:1395-1457.

Moon CG. Simultaneous specification test in a binary logit model - Skewness and Heteroscedasticity. *Communications in Statistics-Theory and Methods.* 1988;17(10):3361-3387.

McDonald JB, Hansen JV. An application and comparison of some flexible parametric and semiparametric qualitative-response models with heteroskedasticity. *International Journal of Systems Science.* 2000;31(1):27-33.

Nelson R. *An Introduction to Copulas*. New York, NY: Springer-Verlag; 1999.

Riphahn RT, Wambach A, Million A. Incentive effects in the demand for health care: A bivariate panel count data estimation. *Journal of Applied Econometrics.* 2003;18(4):387-405.

SAS Institute Inc. What kinds of logistic (or logit) models can be fit using SAS? Usage Note 22871. Available at: http://support.sas.com/kb/22/871.html. Accessed 01/18/2011.

SAS Institute Inc. The PROC LOGISTIC proportional odds test and how to fit a partial proportional odds model. Usage Note 22954. Available at: http://support.sas.com/kb/22/954.html Accessed 01/18/2011.

Silberhorn N, Boztug Y, Hildebrandt L. Estimation with the nested logit model: specifications and software particularities. *OR Spectrum.* 2008;30(4):635-653.

Stokes ME, Davis CS, Koch GG. *Categorical Data Analysis Using the SAS System. Second edition.* Cary, NC: SAS Institute Inc; 2000.

Train K. *Discrete Choice Methods with Simulation*. New York, NY: Cambridge University Press; 2003.

Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press; 2002.

ACKNOWLEDGMENTS

This study was supported by the Agency for Healthcare Research & Quality under grant 1R01 HS14206.

CONTACT INFORMATION

We welcome your comments and questions. Please contact

Joseph C. Gardiner
Division of Biostatistics
Department of Epidemiology
B629 West Fee Hall
Michigan State University
East Lansing, MI 48824
jgardiner@epi.msu.edu