gesis Leibniz-Institut für Sozialwissenschaften

GESIS *Papers* 2017|16

Detailed description of the implementation the multinomial logit model with fixed effects (femlogit)

Klaus Pforr

GESIS Papers 2017|16

Detailed description of the implementation the multinomial logit model with fixed effects (femlogit)

Klaus Pforr

GESIS – Leibniz-Institut für Sozialwissenschaften 2017

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Postfach 12 21 55
68072 Mannheim
Telefon: (0621) 1246 - 231
Telefax: (0621) 1246 - 100
E-Mail: klaus.pforr@gesis.org

ISSN: 2364-3781 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

1 Introduction[1]

Fixed effect models have become increasingly popular in the field of sociology. The possibility to control for unobserved heterogeneity makes these models a prime tool for causal analysis (Gangl 2010; Brüderl and Ludwig 2015). Fixed effects models for continuous, dichotomous, and count dependent variables are widely used and available in Stata as well as many other software packages. A fixed effects estimator for polytomous discrete dependent variables, however, is not yet available for any statistical software package (Allison 2009, 44). The available alternatives for such dependent variables are the pooled multinomial logistic or probit regression (Wooldridge 2010; Rabe-Hesketh and Skrondal 2012) and the multinomial logistic or probit regression with random-effects (Wooldridge 2010; Rabe-Hesketh and Skrondal 2012). For both models we have to assume that any unobserved heterogeneity is independent of the observed covariates.

In this paper, I present an implementation of the multinomial logistic regression with fixed effects (femlogit) in Stata. The `femlogit` command implements an estimator due to Chamberlain (1980). The implementation draws on the native Stata multinomial logit and conditional logit model implementations. The actual ml evaluator utilizes mata functions to implement the conditional likelihood function.

Possible applications of the fixed effects estimator include analyses of effects on employment status with special consideration of part-time or irregular employment, and analyses of the effects on voting behavior that implicitly control for stable individual differences in party preference rather than having to measure it directly.

After explaining the mathematical background and the implementation of the model, I will discuss the syntax of `femlogit`. Afterwards, I show the application of the ado and the interpretation of its results with a model of voting behavior with British election panel data and a model of the effect smoking on pre-term, full term, and post-term birth with multi-level data.

[1] This manuscript is the modified version of the first submission of Pforr (2014), which includes parts that were cut in the course of the review process. The code fragments shown here are from the most recent version available.

2 Statistical model

The statistical model was first proposed by Chamberlain (1980, 231). More extensive expositions are found in Lee (2002, 143ff.) and Pforr (2014). I assume a sample of individuals $i = 1, \ldots, N$ with observations across time $t = 1, \ldots, T_i$.[2] The outcome variable o_j with $j = 1, \ldots, J$ is a polytomous categorical variable with J identical levels for all individuals and observation times. The values of the outcome levels are unrestricted: $\forall j : o_j \in \mathbb{R}$. For each individual i and each observation time t, the chosen outcome y_{it} is measured as the dependent variable and a vector of M independent variables $\mathbf{x}_{it} = x_{it1}, \ldots, x_{itM}$. Next to the realized choices, I define y^*_{itj} to be the latent propensity for each individual i at time t to choose outcome j. With this notation at hand, I assume this relation between the propensities y^*_{itj} and the independent variables $\mathbf{x}_{it}$:

$$\forall j \in \{1, \ldots, J\} : y^*_{itj} = \alpha_{ij} + \mathbf{x}_{it}\boldsymbol{\beta}_j + \varepsilon_{itj}. \tag{1}$$

In this equation, $\boldsymbol{\beta}_j$ is the coefficient vector, which has to be estimated. On the other hand, α_{ij} is a random variable. The error term ε_{itj} is a Type I (Gumbel-type) extreme-value random variable, i.i.d. across all outcomes j. The link to the chosen outcome is defined by:

$$\forall j \in \{1, \ldots, J\} : \Pr\left(y_{it} = o_j \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta}, \mathbf{x}_{it}\right) = \Pr\left(\max_{k \in \{1, \ldots, J\}} y^*_{itk} = y^*_{itj} \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta}, \mathbf{x}_{it}\right). \tag{2}$$

With these assumptions, the probabilities of each outcome can be derived. To guarantee identifiability, an arbitrarily chosen outcome $B \in \{1, \ldots, J\}$ is defined as the base outcome, and the respective coefficients are restricted to zero: $\alpha_{iB} = 0$, $\boldsymbol{\beta}_B = \mathbf{0}$. From this follows:

$$\Pr\left(y_{it} = o_j \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta}, \mathbf{x}_{it}\right) = \begin{cases} \dfrac{\exp\left(\alpha_{ij} + \mathbf{x}_{it}\boldsymbol{\beta}_j\right)}{1 + \sum_{k \neq B} \exp\left(\alpha_{ik} + \mathbf{x}_{it}\boldsymbol{\beta}_k\right)} & j \neq B \\ \dfrac{1}{1 + \sum_{k \neq B} \exp\left(\alpha_{ik} + \mathbf{x}_{it}\boldsymbol{\beta}_k\right)} & j = B \end{cases}. \tag{3}$$

Up to this point, I have set up the assumptions for the pooled multinomial logistic regression, although I would have to rule out unobserved heterogeneity: $\forall j : \alpha_{ij} = \alpha_j$.

The advantage of the femlogit model is that it allows for individual unobserved heterogeneity with respect to the intercepts. The heterogeneity terms α_{ij} are random variables with no restrictions on the joint distribution with the independent variables $\mathbf{x}_{it}$. Direct estimation of the individual α_{ij} creates an incidental parameters problem, which leads to inconsistent estimators with asymptotics solely based on $N \to \infty$. However, with additional assumptions it is possible to consistently estimate the

[2] The subscript i at T_i means that in principle the model allows for analyzing unbalanced panel data. However, the attrition process has to be at least at random, i.e. attrition is completely at random, once conditioning for the independent variables (Wooldridge 2010).

coefficient vector $\boldsymbol{\beta}$. Firstly, we assume that the observed covariates are strictly exogenous conditional on the unobserved heterogeneity:

$$\forall t \in \{1,\ldots,T_i\}, j \in \{1,\ldots,J\} : f_{y_{it}|\alpha_{ij},\mathbf{x}_i} \equiv f_{y_{it}|\alpha_{ij},\mathbf{x}_1,\ldots,\mathbf{x}_{T_i}} = f_{y_{it}|\alpha_{ij},\mathbf{x}_t} \ . \tag{4}$$

Secondly, we assume that the error terms are independent across time. That is, autocorrelation is ruled out:

$$\forall s,t \in \{1,\ldots,T_i\}, \forall j \in \{1,\ldots,J\} : \varepsilon_{isj} \perp \varepsilon_{itj} \, . \tag{5}$$

Chamberlain (1980) states that under these additional assumptions the term $\theta_{ij} \equiv \sum_{t=1}^{T_i} \delta_{y_{it}o_j}$, where δ denotes the Kronecker delta function with respect to y_{it} and o_j, is a sufficient statistic for the unobserved heterogeneity α_{ij}. The intuitive interpretation of this relation is that the sum of occurrences of an outcome j for an individual i across time is a sufficient statistic for her inclination towards that outcome.

The existence of a sufficient statistic for the unobserved heterogeneity means that one can reformulate the likelihood function in such a way that the estimands α_{ij} disappear. The probability mass function for the sequence of chosen outcomes across time for individual i conditional on the sufficient statistic is

$$f_{\mathbf{y}_i|\boldsymbol{\alpha}_i,\boldsymbol{\beta},\mathbf{x}_i,\boldsymbol{\theta}_i} = \frac{\prod_{t=1}^{T_i}\prod_{j=1}^{J} \Pr\left(y_{it}=o_j \mid \boldsymbol{\alpha}_i,\boldsymbol{\beta},\mathbf{x}_i,\boldsymbol{\theta}_i\right)^{\delta_{y_{it}o_j}}}{\sum_{\mathbf{v}_i \in \Upsilon_i}\left(\prod_{t=1}^{T_i}\prod_{j=1}^{J} \Pr\left(v_{it}=o_j \mid \boldsymbol{\alpha}_i,\boldsymbol{\beta},\mathbf{x}_i,\boldsymbol{\theta}_i\right)^{\delta_{v_{it}o_j}}\right)} \ . \tag{6}$$

The summation in the denominator is taken over all "potential" sequences of chosen outcomes $\mathbf{v}_i \equiv \left(v_{i1},\ldots,v_{iT_i}\right)$ that fulfill the condition of the sufficient statistic $\boldsymbol{\theta}_i$. The set Υ_i contains all sequences $\mathbf{v}_i$ for which the sum of occurrences of each outcome j is the same as for the realized sequence y_i. Formally, this means:

$$\Upsilon_i \equiv \left\{ \left(v_{i1},\ldots,v_{iT_i}\right) \mid \forall j \in \{1,\ldots,J\} : \sum_{t=1}^{T_i} \delta_{v_{it}o_j} = \sum_{t=1}^{T_i} \delta_{y_{it}o_j} = \theta_{ij} \right\}. \tag{7}$$

Technically, the set Υ_i is the set of all permutations of the realized sequence of chosen outcomes $\mathbf{y}_i$. With some algebra taking into account the assumptions and definitions above, equation (6) can be written as:

$$f_{\mathbf{y}_i|\boldsymbol{\alpha}_i,\boldsymbol{\beta},\mathbf{x}_i,\boldsymbol{\theta}_i} = \frac{\exp\left(\sum_{t=1}^{T_i}\sum_{j=1,j\neq B}^{J} \delta_{y_{it}o_j}\mathbf{x}_{it}\boldsymbol{\beta}_j\right)}{\sum_{\mathbf{v}_i \in \Upsilon_i}\exp\left(\sum_{t=1}^{T_i}\sum_{j=1,j\neq B}^{J} \delta_{v_{it}o_j}\mathbf{x}_{it}\boldsymbol{\beta}_j\right)}. \tag{8}$$

Having derived the probability mass function, the simplified expression of the log-likelihood function of the femlogit model follows from its definition. The contribution to log-likelihood of individual i is:

$$\ln \ell_i\left(\boldsymbol{\beta} \mid \mathbf{y}_i, \mathbf{x}_i\right) = \ln f_{\mathbf{y}_i \mid \boldsymbol{\alpha}_i, \boldsymbol{\beta}, \mathbf{x}_i, \boldsymbol{\theta}_i} = \\ = \sum_{t=1}^{T_i} \sum_{j=1, j \neq B}^{J} \delta_{y_{it} o_j} \mathbf{x}_{it} \boldsymbol{\beta}_j - \ln \sum_{\upsilon_i \in \Upsilon_i} \exp\left(\sum_{t=1}^{J} \sum_{j=1, j \neq B}^{J} \delta_{\upsilon_{it} o_j} \mathbf{x}_{it} \boldsymbol{\beta}_j \right). \tag{9}$$

Therefore the overall log-likelihood function for the sample – given a simple random sample of panel groups – is:

$$\ln L\left(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{x}\right) = \sum_{i=1}^{N} \ln \ell_i\left(\boldsymbol{\beta} \mid \mathbf{y}_i, \mathbf{x}_i\right). \tag{10}$$

For the ML estimation, the gradient matrix of the log-likelihood function and the Hessian matrix is needed. Although these could be determined in the practical estimation process numerically, this would reduce numerical precision and considerably slow down the estimation process (cf. Gould, Pitblado, and Poi 2010, 20–24). I derived analytical expressions for the gradient matrix and the Hessian matrix.

The gradient matrix is the $N \times (J-1)M$ matrix of the partial derivatives of all individual contributions to the log-likelihood with respect to all coefficients β_{jm}. That is, the element in row a and column b in this matrix is the partial derivative of contribution individual a to the log-likelihood with respect to the b-th coefficient. Note, that there are only $(J-1)M$ coefficients, as the coefficients of the outcome B are constrained to zero.

$$\mathbf{g} = \left(\frac{\partial \ln \ell_i\left(\boldsymbol{\beta} \mid \mathbf{y}_i, \mathbf{x}_i\right)}{\partial \beta_{jm}} \right) \tag{11}$$

The element for individual i and the coefficient β_{jm} in the gradient matrix is:

$$\frac{\partial \ln \ell_i\left(\boldsymbol{\beta} \mid \mathbf{y}_i, \mathbf{x}_i\right)}{\partial \beta_{jm}} = \sum_{t=1}^{T_i} \delta_{y_{it} o_j} x_{itm} - \\ \frac{\sum_{\upsilon_i \in \Upsilon_i} \left(\left(\sum_{t=1}^{T_i} \delta_{\upsilon_{it} o_j} x_{itm} \right) \exp\left(\sum_{t=1}^{T_i} \sum_{j=1, j \neq B}^{J} \delta_{\upsilon_{it} o_j} \mathbf{x}_{it} \boldsymbol{\beta}_j \right) \right)}{\sum_{\upsilon_i \in \Upsilon_i} \exp\left(\sum_{t=1}^{T_i} \sum_{j=1, j \neq B}^{J} \delta_{\upsilon_{it} o_j} \mathbf{x}_{it} \boldsymbol{\beta}_j \right)}. \tag{12}$$

The Hessian matrix is the $(J-1)M \times (J-1)M$. matrix of the partial derivatives of the overall log-likelihood of the sample with respect to all pairs of coefficients β_{jm} and β_{kn}.

$$\mathbf{H} = \left(\frac{\partial^2 \ln L\left(\boldsymbol{\beta} \middle| \mathbf{y}, \mathbf{x}\right)}{\partial \beta_{jm} \partial \beta_{kn}} \right) \tag{13}$$

Here, the element of the Hessian matrix for the coefficients β_{jm} and β_{kn} is:

$$\frac{\partial^2 \ln L\left(\boldsymbol{\beta}|\mathbf{y},\mathbf{x}\right)}{\partial\beta_{jm}\partial\beta_{kn}} = \Bigg[\sum_{\upsilon_i\in\Upsilon_i} \left(\left(\sum_{t=1}^{T_i} \delta_{\upsilon_{it}o_j} x_{itm} \right) \exp\left(\sum_{t=1}^{T_i} \sum_{j=1,j\neq B}^{J} \delta_{\upsilon_{it}o_j} \mathbf{x}_{it}\boldsymbol{\beta}_j \right) \right) \cdot$$
$$\cdot \sum_{\upsilon_i\in\Upsilon_i} \left(\left(\sum_{t=1}^{T_i} \delta_{\upsilon_{it}o_k} x_{itn} \right) \exp\left(\sum_{t=1}^{T_i} \sum_{j=1,j\neq B}^{J} \delta_{\upsilon_{it}o_j} \mathbf{x}_{it}\boldsymbol{\beta}_j \right) \right) \Bigg/$$
$$\left(\sum_{\upsilon_i\in\Upsilon_i} \exp\left(\sum_{t=1}^{T_i} \sum_{j=1,j\neq B}^{J} \delta_{\upsilon_{it}o_j} \mathbf{x}_{it}\boldsymbol{\beta}_j \right) \right)^2 \Bigg] - \Bigg[\sum_{\upsilon_i\in\Upsilon_i} \left(\left(\sum_{t=1}^{T_i} \delta_{\upsilon_{it}o_j} x_{itm} \right) \cdot . \right. \quad (14)$$
$$\left. \cdot \left(\sum_{t=1}^{T_i} \delta_{\upsilon_{it}o_k} x_{itn} \right) \exp\left(\sum_{t=1}^{T_i} \sum_{j=1,j\neq B}^{J} \delta_{\upsilon_{it}o_j} \mathbf{x}_{it}\boldsymbol{\beta}_j \right) \right) \Bigg/$$
$$\sum_{\upsilon_i\in\Upsilon_i} \exp\left(\sum_{t=1}^{T_i} \sum_{j=1,j\neq B}^{J} \delta_{\upsilon_{it}o_j} \mathbf{x}_{it}\boldsymbol{\beta}_j \right) \Bigg]$$

From standard ML theory, it follows that the estimates of the coefficients of interest as described by (Wooldridge 2010) are:

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} \equiv \max_{\boldsymbol{\beta}} \left(N^{-1} \sum_{i=1}^{N} \ln \ell_i \left(\boldsymbol{\beta} \,|\, \mathbf{y}_i, \mathbf{x}_i \right) \right),$$
$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} \xrightarrow{\;p\;} \boldsymbol{\beta},$$
$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}} - \boldsymbol{\beta} \right) \xrightarrow{\;d\;} \mathrm{N}\left(0, -\mathrm{E}\left(\frac{\partial^2 \ln \ell_i \left(\boldsymbol{\beta} \,|\, \mathbf{y}_i, \mathbf{x}_i \right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} \right)^{-1} \right), \quad (15)$$
$$\widehat{\mathrm{Avar}\left(\hat{\boldsymbol{\beta}}_{\mathrm{ML}} \right)} = -\sum_{i=1}^{N} \left. \frac{\partial^2 \ln \ell_i \left(\boldsymbol{\beta} \,|\, \mathbf{y}_i, \mathbf{x}_i \right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} \right|_{\hat{\boldsymbol{\beta}}_{\mathrm{ML}}}.$$

This means that – provided that the aforementioned assumptions are valid – the ML estimation of the coefficient vector $\boldsymbol{\beta}$ of the structural model described in equation (3) is the maximum of the sample analogue of the overall log-likelihood function conditional on the dependent and independent variables with respect to $\boldsymbol{\beta}$. This estimator $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ converges in probability to the true coefficient vector $\boldsymbol{\beta}$ with $N \longrightarrow \infty$ and fixed T_i. It converges in distribution with $N \longrightarrow \infty$ and fixed T_i to a multivariate normal distribution with the true coefficient vector as the mean and a variance-covariance-matrix as described in equation (15).

3 Implementation

In this section, I explain the implementation of the femlogit model. Starting with a short introduction to how ML estimators are implemented in Stata in general, afterwards I show, how the statistical model derived in the first section corresponds to the notational framework of the `ml/moptimize( )` command suites. Next, I explicate the concrete implementation of the femlogit model in the femlogit command. Here, I start with a detailed description of the evaluator, and conclude with the general layout of the ado.

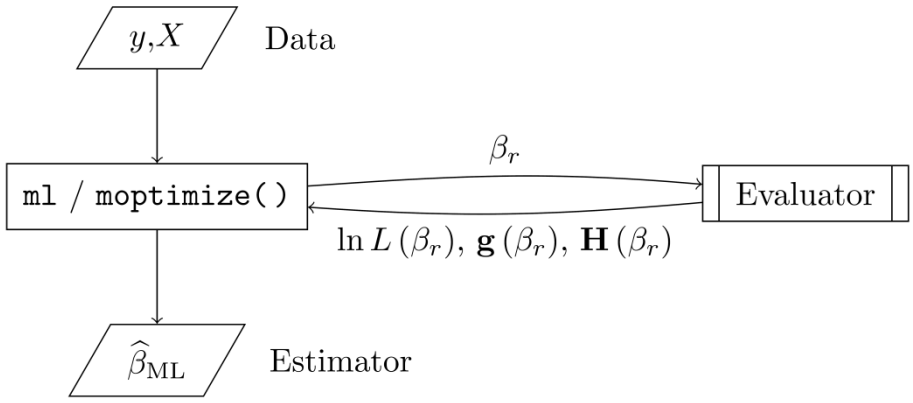

Figure 1: Schematic illustration of ML estimation in Stata

Maximum likelihood estimation in Stata – outside of existing estimation commands~– generally works along the structure depicted in figure 1 (Gould, Pitblado, and Poi 2010). At this point the log-likelihood function $\ln L\left(\boldsymbol{\beta}|\mathbf{y},\mathbf{x}\right)$ and optionally the gradient and Hessian matrices have already been derived. The first step is to give Stata the dependent variable $\mathbf{y}$ and the independent variables $\mathbf{x}$, from which the coefficients of interest can be estimated. This information is passed to Stata via the existing Stata command suite `ml` or via the Mata analogue `moptimize( )`. In the second step, the algebraic expression of the log-likelihood function is translated so that it can be interpreted by Stata. This is accomplished by programming the so-called "evaluator" either as a Stata-ado or a Mata-function. Generally, the evaluator expects to be given from `ml` or `moptimize( )` the dependent and independent variables and a coefficient vector. The coefficient vector is either the initial vector, from which the iterative ML algorithm starts, or the vector of the previous ML iteration step. With this input, the evaluator calculates the log-likelihood, and optionally the gradient and the Hessian matrices. This output is given back to `ml` or `moptimize( )`, from which the coefficient vector of the next ML iteration step is calculated.

The femlogit model described above can best be implemented in Stata with the `moptimize( )` Mata suite and the evaluator as a Mata function. With the mathematical derivations of the individual contributions to the overall log-likelihood function for the sample (eq. (9)), and the gradient and Hessian matrices (eqs. (11)and (13)) at hand, I organize them in the `moptimize( )` notation, following subsection "Mathematical statement of the `moptimize( )` problem" in StataCorp LP (2009a, 639ff.).

Find coefficients

$$b=((b_1), \ldots, (b_{B-1}), (b_{B+1}), \ldots, (b_J))$$

where

b_1: 1 x M,

...

b_{B-1}: 1 x M,

b_{B+1}: 1 x M,

...

b_J: 1 x M,

that maximize

$$\ln L(b|y,x)$$

as defined in equation (10) where

y: $\sum_{i=1..N} N_i$ x 1,

x: $\sum_{i=1..N} N_i$ x M.

This layout shows that the column vector $\mathbf{y}$ represents the single dependent variable and the matrix $\mathbf{x}$ represents the M independent variables. Overall, there are N panel groups, each with T_i observations. The coefficient column vectors are organized as $J-1$ equations in the `moptimize( )` terminology. Note that the coefficient vectors do not include terms for the constants.

This abstract description provides the basis for the implementation of the femlogit model into Stata. In the following, I present the concrete implementation of the evaluator, the ML-call, and the ado wrapper, which allows to estimate the model for general problems.

3.1 Evaluator

The programming of the evaluator depends on the chosen method or type, where the first term is used for Stata evaluators and the latter term is used for Mata evaluators (StataCorp Lp 2009a; Gould, Pitblado, and Poi 2010, 48–51). If the individual contributions to the overall log-likelihood function for the sample depend on individual rows of the data matrix and the overall log-likelihood is the sum of these contributions across the rows of the data matrix, it is generally preferable to use a method or type from the `lf`-family. If the overall log-likelihood cannot be constructed as a sum across the rows of the data matrix, a method or type from the `d`-family is recommended. If the individual contributions to the overall log-likelihood function are derived from groups of cases in the data matrix, and the overall log-likelihood function is the sum across the groups of cases, a method or type from the gf-family should be chosen. The latter case applies to most panel-data and multi-level models.

For the implementation of `femlogit`, I use a `gf2`-type evaluator. Besides the more straightforward consideration of the panel-data structure, this type allows an easier integration of the implemented ado into the `svy` command suite.[3] The `gf2`-type evaluator expects to use the coefficient row vector

[3] Note, that in the current version of the implementation, there is no support for weights and there is no connection to the \stcmd{svy} command suite.

$\boldsymbol{\beta}: 1\times(J-1)M$, the independent variable matrix $\mathbf{x}: \sum_{i=1}^{N} T_i \times M$ and the dependent variable column vector $\mathbf{y}: \sum_{i=1}^{N} T_i \times 1$ as input, and returns the column vector $\left(\ln \ell_i(\boldsymbol{\beta})\right): N\times 1$ of the individual contributions for all panel-groups, the gradient matrix $\mathbf{g}: N\times(J-1)M$, as defined in equation (11), and the Hessian matrix $\mathbf{H}: (J-1)M\times(J-1)M$, as defined in equation (13).

Before laying out the programming of the evaluator, I define auxiliary terms, from which the expressions in the equations (10), (12), and (14) can be constructed. For each panel group i and for each of its permutations $\boldsymbol{\upsilon}_i$, I define the scalar $Z_{1i}: 1\times 1$ and the row vector $Z_{2i}: 1\times(J-1)M$ as:

$$\begin{aligned} Z_{1i} &\equiv \exp\left(\sum_{t=1}^{T_i}\sum_{j=1, j\neq B}^{J} \delta_{\upsilon_{it} o_j} \mathbf{x}_{it}\boldsymbol{\beta}_j\right), \\ Z_{2i} &\equiv \left(Z_{2ijm}\right), \\ Z_{2ijm} &\equiv \sum_{t=1}^{T_i} \delta_{\upsilon_{it} o_j} x_{itm}. \end{aligned} \tag{16}$$

With these expressions as building blocks for each panel group i and for each of its permutations $\boldsymbol{\upsilon}_i$, I define the scalars $A_i: 1\times 1$ and $B_i: 1\times 1$, the row vectors $C_i: 1\times(J-1)M$ and $D_i: 1\times(J-1)M$, and the matrix $E_i: (J-1)M\times(J-1)M$ as:

$$\begin{aligned} A_i &\equiv \sum_{t=1}^{T_i}\sum_{j=1, j\neq B}^{J} \delta_{y_{it} o_j} \mathbf{x}_{it}\boldsymbol{\beta}_j, \\ B_i &\equiv \sum_{\upsilon_i \in \Upsilon_i} Z_{1i}, \\ C_i &\equiv \left(C_{ijm}\right), \\ C_{ijm} &\equiv \sum_{t=1}^{T_i} \delta_{y_{it} o_j} \mathbf{x}_{itm}, \\ D_i &\equiv \left(D_{ijm}\right), \\ D_{ijm} &\equiv \sum_{\upsilon_i \in \Upsilon_i} Z_{1i} Z_{2i}, \\ E_i &\equiv \left(E_{ijmkn}\right), \\ E_{ijmkn} &\equiv \sum_{\upsilon_i \in \Upsilon_i} Z_{1i} Z_{2i}' Z_{2i}. \end{aligned} \tag{17}$$

With these auxiliary terms, the column vector of the individual contributions to the overall log-likelihood function $\left(\ln \ell_i(\boldsymbol{\beta})\right)$ – in Stata terminology `lnfj` – and the gradient matrix $\mathbf{g}$ or `S` and Hessian matrix $\mathbf{H}$ or `H` can be constructed in the following way:

$$\begin{pmatrix} \ln \ell_i(\boldsymbol{\beta}) \\ \vdots \\ \ln \ell_N(\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} A_1 - \ln B_1 \\ \vdots \\ A_N - \ln B_N \end{pmatrix},$$
$$\mathbf{g} = \begin{pmatrix} C_1 - \dfrac{D_1}{B_1} \\ \vdots \\ C_N - \dfrac{D_N}{B_N} \end{pmatrix}, \qquad (18)$$
$$\mathbf{H} = \left(\sum_{i=1}^{N} \frac{D_i' D_i}{B_i^2} - \frac{E_i}{B_i} \right).$$

This completes the description of the correspondence between the statistical model and the `moptimize( )` command suite, so that I can map the model into a `gf2`-type evaluator.

Code 1: femlogit_eval_gf2( ) – part 1

```
*! version 1.0.0 16apr2014 13:35 mata_eval_moptgf2.mata
version 11.0
mata:
mata set matastrict on
void femlogit_eval_gf2(transmorphic scalar ML, real scalar todo, /*
  */ real rowvector b, real colvector lnfj, real matrix S, /*
  */ real matrix H) {

  // declare variables
  real colvector touse, id, yi, upsiloni
  real matrix panelinfo, Xi, out2eq, X, E, T, Hc, Sc
  real scalar N, M, J, i, A, B, j, m, Z1
  real rowvector C, D, permuteinfo, Z2
```

First I have to declare the objects that are used in the function. This step is necessary, as I use strict Mata syntax.

Code 2: femlogit_eval_gf2( ) – part 2

```
// get things from Stata
st_view(touse=.,.,st_local("touse"))
st_view(id=.,.,st_local("group"),st_local("touse"))
st_view(X=.,.,st_local("rhs"),st_local("touse"))

// auxilary matrix
out2eq=st_matrix(st_local("out2eq"))
```

The estimation sample indicator, the panel group indicator, and the independent variables are accessed by referencing, which are stored in the macros `touse`, `group`, and `rhs`. I copy the matrix `out2eq`, which maps the indexes j to the outcomes o_j, from Stata to Mata:

$$\text{out2eq} = \begin{pmatrix} o_1 & 1 \\ \vdots & \vdots \\ o_{B-1} & B-1 \\ o_B & 0 \\ o_{B+1} & B+1 \\ \vdots & \vdots \\ o_J & J-1 \end{pmatrix}. \tag{19}$$

Code 3: femlogit_eval_gf2( ) – part 3

```
// derived information
J=rows(out2eq)
M=cols(X)
panelinfo=panelsetup(id,1)
N=panelstats(panelinfo)[1]
```

From these objects, I derive the constants that make up the sample, i.e. the number of alternatives J, the number of independent variables M, and the number of panel groups N. For easier management of the panel groups, I use the function `panelsetup( )`, which gives us a matrix `panelinfo( )` that marks the cases, which belong to each panel group.

Code 4: femlogit_eval_gf2( ) – part 4

```
// init lnfj, S, H
lnfj=J(N,1,0)
if (todo>0) {
  S=J(N,(J-1)*M,0)
  if (todo==2) {
    H=J((J-1)*M,(J-1)*M,0)
  }
}
```

The central outputs of the evaluator are `lnfj`, `S`, and `H`. These objects are initialized to contain zeros as elements. Note that the scalar `todo` is used to streamline the computation, following Gould et al. (2010, 79, 111) and the subsection "Example using type d" in StataCorp LP (2009a). This streamlining is also applied in the rest of code.

Code 5: femlogit_eval_gf2( ) – part 5

```
// calculate lnfj, S, H
for(i=1;i<=N;i++) { // loop over panels
  // create panel-wise variables (only one call per panel)
  yi=moptimize_util_depvar(ML,1)[|panelinfo[i,1]\panelinfo[i,2]|]
  Xi=X[|panelinfo[i,1],.\panelinfo[i,2],.|]
```

To compute the column vector `lnfj`, I compute each of its N elements. Analogously for the matrix `gi`, each of its N row vectors are computed. For the matrix `Hi`, each of the N addends of the sum shown in equation (18) are computed. Therefore, I loop over all N panel groups. For each panel group $i = 1,\ldots,N$, I use the utility function `moptimize_util_depvar()` from the `moptimize()`-suite and the matrix `panelinfo` created above, to create the column vector y_i that contains the sequence of chosen outcomes across time for panel group i. Accordingly, I create the matrix $\mathbf{x}_i$ that contains the sequences of all independent variables across time for panel group i.

Code 6: femlogit_eval_gf2( ) – part 6

```
// init major auxiliary variables (A,B,C,D,E)
A=0
B=0
if (todo>0) {
  C=J(1,(J-1)*M,0)
  D=J(1,(J-1)*M,0)
  if (todo==2) {
    E=J((J-1)*M,(J-1)*M,0)
  }
}
```

To compute `lnfj`, `S`, and `H` from the defined auxiliary variable described in equations (16), (17), and (18), I first initialize the auxiliary terms $A_i = 0$, $B_i = 0$, $C_i = \mathbf{0}$, $D_i = \mathbf{0}$, and $E_i = \mathbf{0}$ to contain zeros as elements.

Code 7: femlogit_eval_gf2( ) – part 7

```
// calculate A,C
for(j=1;j<=J;j++) { // loop over outcomes
  if (out2eq[j,2]!=0) { // exclude base outcome
    A=A+quadcolsum((yi:==out2eq[j,1]):* /*
    */ ((Xi*(colshape(b,M)'))[.,out2eq[j,2]]))
    if (todo>0) {
      for(m=1;m<=M;m++) { // loop over indep. vars
        C[1,(out2eq[j,2]-1)*M+m]=quadcolsum((yi:==out2eq[j,1]) /*
        */ :*(Xi[.,m]))
      }
    }
  }
}
```

Here, the auxiliary scalar A_i and the auxiliary row vector C_i is computed, as defined in equation (17).

Code 8: femlogit_eval_gf2( ) – part 8

```
// calculate B,D,E
// generate Upsilon_i=Set of permutations of y_i
permuteinfo=cvpermutesetup(yi)
// loop over permutations of y_i (upsilon_i in Upsilon_i)
while((upsiloni=cvpermute(permuteinfo))!=J(0,1,.)) {
  // init minor auxiliary variables
  Z1=0
  if (todo>0) {
    Z2=J(1,(J-1)*M,0)
  }
```

To calculate the auxiliary term B_i, D_i, and E_i, I have to define the auxiliary terms Z_{1i} and Z_{2i}. These objects are constructed from the set of all permutations of outcome sequences for each panel group i. Therefore, I use the functions `cvpermutesetup()` and `cvpermute()` to loop over of all permutations υ_i for each panel group i. For each permutation υ_i within each panel group i, I first initialize the auxiliary terms $Z_{1i}=0$ and $Z_{2i}=\mathbf{0}$ to contain zeros.

Code 9: femlogit_eval_gf2( ) – part 9

```
  // calculate Z1,Z2
  for(j=1;j<=J;j++) { // loop over outcomes
    if (out2eq[j,2]!=0) { // exclude base outcome
      Z1=Z1+quadcolsum((upsiloni:==out2eq[j,1]):*((Xi* /*
      */ (colshape(b,M)'))[.,out2eq[j,2]]))
      if (todo>0) {
        for(m=1;m<=M;m++) {
          Z2[1,(out2eq[j,2]-1)*M+m]= /*
          */ quadcolsum((upsiloni:==out2eq[j,1]):*(Xi[.,m]))
        }
      }
    }
  }
  Z1=exp(Z1)
```

Here, I compute the auxiliary terms Z_{1i} and Z_{2i}, as defined in equation (16). Note that most of the computation is structurally analogous to the computation of A_i and C_i above, where $\mathbf{y}_i$ is exchanged with $\boldsymbol{\upsilon}_i$.

Code 10: femlogit_eval_gf2( ) – part 10

```
    // fill up B,D,E with minor aux. var's
    B=B+Z1
    if (todo>0) {
      D=D+Z2:*Z1
      if (todo==2) {
        E=E+(quadcross(Z2,Z2)):*Z1
      }
    }
  }
```

With the auxiliary terms Z_{1i} and Z_{2i} at hand, I compute the auxiliary terms B_i, D_i, and E_i, as defined in equation (17). To safe computation time, the terms are summed up in passing within the same loop over all permutations $\boldsymbol{\upsilon}_i$.

Code 11: femlogit_eval_gf2( ) – part 11

```
  // fill up lnfj,S,H with major aux. var's A,B,C,D,E
  lnfj[i]=A-ln(B)
  if (todo>0) {
    S[i,.]=C-D:/B
    if (todo==2) {
      // Sum up H
      H=H+((quadcross(D,D)):/(B^2))-(E:/B)
    }
  }
}
```

After completion of the loop over all permutations $\boldsymbol{\upsilon}_i$, the auxiliary terms A_i, B_i, C_i, D_i, and E_i are computed. With these, I can fill in the elements of the column vector `lnfj` and the rows of the matrix `S` and sum up the cell entries of the matrix `H` for each panel group i, as described in equation (18). After completion of the loop over all panel groups i, all elements of `lnfj`, all rows of `S`, and all addends of `H` are computed.

Code 12: femlogit_eval_gf2( ) – part 12

```
  // Push out scores and Hessian for robust variance matrix (preci-
sion issues!)
  if (st_local("robust")!="" & st_local("constraints")=="") {
    if (cols(S)==rows(H) & rank(H)==rows(H)) {
      st_matrix(st_local("rvm"),(N/(N-1)):* /*
      */ (invsym(-H)*quadcross(S,S)*invsym(-H)))
    }
  }
  if (st_local("robust")!="" & st_local("constraints")!="") {
    T=st_matrix(st_local("T"))
    if (cols(S)==rows(T)) {
      Sc=quadcross(S',T)
      if (cols(H)==rows(T) & rows(H)==rows(T)) {
        Hc=quadcross(T,H)*T
        if (cols(Sc)==rows(Hc) & rank(Hc)==rows(Hc)) {
          st_matrix(st_local("rvm"),T*((N/(N-1)):* /*
          */ (invsym(-Hc)*quadcross(Sc,Sc)* /*
          */ invsym(-Hc)))*T')
        }
      }
    }
  }
}
end
```

At the end of the evaluator, the robust variance-covariance matrix is computed and put into a Stata matrix, where it can be accessed, when the ML algorithm has reached convergence.

3.2 ML-call and ado wrapper

In this section I briefly describe, how the ML problem is initialized, how the evaluator is called, and how this part is embedded in an ado wrapper, which makes it possible to apply the estimator to general problems. The exposition of the ado wrapper proceeds from outside inwards, i.e. it starts with the command that a user posts to Stata.

At first, the command posted by the user is parsed with `syntax`. This creates a reference to the dependent variable, the independent variables, and the panel group indicator. Further, the optional base-

outcome and the optional set of linear constraint equations are passed. Finally, the optional `difficult` instruction on how to deal non-concave regions of the log-likehood function is passed. Afterwards, observations with missing values on the dependent, independent or on the panel group indicator variables are marked for list-wise deletion with `markout`. Next, collinear independent variables are marked for exclusion in the model with `_rmcoll`. Following the implementation of `mlogit`, with the option `mlogit` indicators about the base-outcome are created. Subsequently, following the implementation of `clogit` panel groups without variance across time in the dependent variables and independent variables without variance across time in all panel groups are marked for exclusion. Thereafter, a Stata matrix as defined in equation (19) is created, which maps the level indices of the dependent variables with the respective values. Afterwards, following the implementation of `clogit`, the overall log-likelihood function for the sample for a model without any independent variables is computed, which is as the baseline for a likelihood-ratio test. Next, initial values for the coefficient vector $\boldsymbol{\beta}$ are computed. Following the implementation of `clogit`, the initial values are the estimated coefficients of a pooled multinomial logit model. Subsequently, the submitted constraints are preprocessed and checked for correct specification. Thereafter, the ML problem is defined as described above, i.e. the dependent, the independent and the panel group indicator variables are passed the ML problem definition. Further, the equation structure, the estimation sample indicator and the linear constraints definition are passed to the ML problem definition. Afterwards, the actual femlogit model is estimated using the implemented evaluator described above. Finally, the output is computed, returned to Stata and displayed.

3.3 Data structure

The implementation expects that the data are organized in long format, i.e. from the panel data perspective each observation represents a time points of one person. An illustrative example is shown with a modified version of the example data used in StataCorp Lp (2009b, 325ff.):[4]

```
. use femlogitid.dta
. list in 1/11
     ┌──────────────────────────┐
     │   id   y   x1   x2 │
     ├──────────────────────────┤
  1. │ 1014   3    0    4 │
  2. │ 1014   0    1    4 │
  3. │ 1014   2    1    6 │
  4. │ 1014   1    1    8 │
  5. │ 1017   0    0    1 │
     ├──────────────────────────┤
  6. │ 1017   2    0    7 │
  7. │ 1017   1    1   10 │
  8. │ 1019   0    0    1 │
  9. │ 1019   2    1    7 │
 10. │ 1019   1    1    7 │
     ├──────────────────────────┤
 11. │ 1019   1    1    9 │
     └──────────────────────────┘
```

The first four observations belong to the person with the `id` $=1014$. The independent variables are `x1` and `x2` and `y1` is the categorical dependent variable with four levels $\{0,1,2,3\}$. Note, that the

[4] The data `femlogitid.dta` and syntax `femlogit_example1.do` can found in the online appendix.

different levels of the categorical dependent variable are stored in one variable and one case similarly to `mlogit`. In contrast, the implementation of `clogit` expects that the outcomes of the dependent variable for each time point are stored in long format.

4 Syntax

The command femlogit is called with the following syntax.

femlogit *depvar* [*indepvars*] [*if*] [*in*] [, group(*varlist*) baseoutcome(#)
constraints(*clist*) difficult or robust]

depvar and *indepvars* may not contain factor variables or time-series operators. No prefix commands are allowed. Weights and vce() are not allowed at this point.

4.1 Options

group(*varlist*) specifies one or more identifier variables (numeric or string) for the matched groups. If not specified, the identifier variables from xtset are used.

baseoutcome(#) specifies the value of depvar to be treated as the base outcome. The default is to choose the mode outcome.

constraints(*clist*) specifies the linear constraints to be applied during estimation. The default is to perform unconstrained estimation. *clist* has the form # [-#] [, # [-#] ...].

difficult specifies that in non-concave regions of the likelihood function the ``hybrid'' method is used instead of the default "modified marquart" method (Gould, Pitblado, and Poi 2010, 15–17).

or reports odds ratio effects.

robust gives back Huber-White-sandwich estimator of the variance-covariance matrix

4.2 Saved results

Femlogit saves the following in e():

Scalars

e(rank)	rank of e(V)	e(df_m)	model degrees of freedom
e(N)	number of observations	e(chi2)	χ^2
e(ic)	number of iterations	e(p)	significance
e(k)	number of parameters	e(N_drop)	Number of observations dropped because of invariant dependent variable
e(k_eq)	number of equations in e(b)	e(N_group_~p)	number of groups dropped because of invariant dependent variable
e(k_dv)	number of dependent variables	e(r2_p)	pseudo-R-squared
e(converged)	1 if converged, 0 otherwise	e(ibaseout)	index of the base outcome

`e(rc)`	return code	`e(baseout)`	the value of *depvar* to be treated as the base outcome
`e(ll)`	log likelihood	`e(k_out)`	number of outcomes
`e(k_ew_model)`	number of equations in overall model test		
Macros			
`e(cmdline)`	command as typed	`e(predict)`	`_predict`
`e(cmd)`	`femlogit`	`e(user)`	`femlogit eval gf2( )`
`e(eqnames)`	names of equations	`e(ml_method)`	`gf2`
`e(group)`	name of `group()` variable	`e(technique)`	`nr`
`e(chi2type)`	`Wald` or `LR`; type of model χ^2 test	`e(which)`	`max`
`e(vce)`	`oim`	`e(depvar)`	name of dependent variable
`e(title)`	title in estimation output	`e(properties)`	`b V`
`e(crittype)`	`log likelihood`	`e(marginsn~k)`	`stdp stddp`
`e(opt)`	`moptimize`	`e(marginsok)`	`xb`
Matrices			
`e(b)`	coefficient vector	`e(out)`	outcome values
`e(V)`	variance-covariance matrix of the estimator	`e(ilog)`	iteration log (up to 20 iterations)
`e(Cns)`	constraints matrix	`e(gradient)`	gradient vector
Functions			
`e(sample)`	marks estimation sample		

5 Applications

In this section, I show with two illustrative examples, how the `femlogit` command can be used and how the results are interpreted.

The first showcase example illustrates a typical application with panel data. Drawing on Skrondal and Rabe-Hesketh (2003), I analyze a model of the effect of the distance between the voter's and the parties' positions on the left–right political dimension on the voter's electoral choice, applied to the 1987–1992 panel of the British Election Study (Heath et al. 1993). The second example shows, how `femlogit` can be used with multi-level data. I build on Abrevaya (2006) and analyze a model of the effect of smoking during pregnancy on pre-term, full term, and post-term birth. The model is estimated with multi-level data of children nested in mothers. Note, that the described models are deliberately simple. The main intention is to show the specific advantage of the femlogit model.

Both examples use data with a general multi-level structure: in the voting example, panel waves are nested within voters, and in the smoking example, children are nested within mothers. In comparison to alternative models, such as the pooled multinomial logistic regression or multinomial logistic regression with random effects, the femlogit model has the specific advantage that it controls for possibly confounding, unobserved heterogeneity at the top level. More explicitly, in the first example the femlogit model controls for unobserved heterogeneity at the level of the voters, and in the second example it controls for heterogeneity at the level of the mothers.

In the rest of this section, I begin with the voting example and continue with the smoking-effect-example. For each example, I shortly describe the data and the estimated model, and afterwards discuss the results of the `femlogit` in comparison to the pooled multinomial logistic regression and multinomial logistic regression with random effects. Finally, I generally discuss possible interpretation strategies of the femlogit model.

5.1 Effect of ideological distance on voting behavior with British election panel data

The first example takes up the example that Skrondal and Rabe-Hesketh (2003) use to illustrate the application of multilevel random-effects models for polytomous and ordinal dependent variables. They analyze data from the 1987–1992 panel of British Election Study (Heath et al. 1993) to estimate a model of the recalled vote choice for the Conservative, Labour, or Liberal party and a model of the rank order of the parties. Here, I concentrate on the recalled vote choice, and use the `femlogit` command to estimate the effect of the distance on the left–right policy dimension between the voter and the party on the vote choice. I control for the time-varying rating of perceived inflation and implicitly for all time-variant factors at the voter-level. The analysis syntax for the first example is found in `femlogit_example2.do`, provided in the online appendix.

The raw data is taken from Rabe-Hesketh and Skrondal (2012, 2:680f.). Cleaning and preparation leads to these analysis data:

```
. describe
Contains data
  obs:         2,458
 vars:             9
 size:        46,702
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
serialno        int     \%8.0g                Respondent number
rldist2         float   \%9.0g                Dist(Labour)-Dist(Conservative)
rldist3         float   \%9.0g                Dist(Liberal)-Dist(Conservative)
male            byte    \%8.0g                Male
manual          byte    \%8.0g                Manual worker
inflation       byte    \%8.0g                Perceived inflation
age             float   \%9.0g                Age in 10 yr units
yr92            byte    \%8.0g                1992 election indicator
choice          byte    \%12.0g     choice    Recalled vote for party
-------------------------------------------------------------------------------
Sorted by:  serialno
     Note:  dataset has changed since last saved
```

The dependent variable `choice` is a discrete variable with three alternatives "Conservatives", "Labour", and "Liberal". In the femlogit model, four independent variables are used: the difference of the distance between the voter and the Labour party and the distance between the voter and the Conservative party (`rldist2`), the difference of the distance between the voter and the Liberal party and the distance between the voter and the Conservative party (`rldist3`), a rating of the perceived inflation (`inflation`), and a wave dummy (`yr92`).

The data is in long format. As the summary command for panel data `xtdes` shows, the data contain information of 1,344 persons across both elections. For 1,114 persons, the time series across both waves is complete. For the remaining 230 persons, information at least for one wave is missing.

```
. xtset serialno yr92
       panel variable:  serialno (unbalanced)
        time variable:  yr92, 0 to 1
                delta:  1 unit
. xtdes
serialno:  2, 11, ..., 5997                                 n =       1344
    yr92:  0, 1, ..., 1                                     T =          2
           Delta(yr92) = 1 unit
           Span(yr92)  = 2 periods
           (serialno*yr92 uniquely identifies each observation)
Distribution of T_i:   min      5\%     25\%       50\%       75\%     95\%     max
                         1       1       2          2          2       2        2
     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
     1114     82.89   82.89 |  11
      121      9.00   91.89 |  1.
      109      8.11  100.00 |  .1
 ---------------------------+---------
     1344    100.00         |  XX
```

The differences in the policy distances vary not only across voters and waves, but also across alternatives. This allows us to specify the model as a mixed-logit model (Cameron and Trivedi 2009). That is, I estimate one coefficient for the alternative-varying policy distances and alternative-specific coeffi-

cients for the alternative-invariant voters' rating of inflation and the wave dummy. In order to do this, I define constraints for the effects of the policy distances:

```
. constraint 1 [Labour]rldist3=0
. constraint 2 [Liberal]rldist2=0
. constraint 3 [Labour]rldist2=[Liberal]rldist3
```

With these constraints, the effect of the relative policy distance between the voter and the Liberal party plays no role for the propensity to vote for Labour in comparison to the Conservative party and vice versa, the relative policy distance between the voter and the Labour party is irrelevant for the propensity to vote for the Liberal party instead of the Conservative party. The third constraint guarantees that the relative policy distances have the same effect on both propensities.

The estimation output of `femlogit` for this model is this:

```
. femlogit choice rldist2 rldist3 inflation yr92, group(serialno) const(1/3) b(
> 1)
note: 1097 groups (1964 obs) dropped because of all positive or
      all negative outcomes.
Iteration 0:   log likelihood = -156.16844
Iteration 1:   log likelihood = -139.49392
Iteration 2:   log likelihood = -138.19403
Iteration 3:   log likelihood = -138.19006
Iteration 4:   log likelihood = -138.19006
Fixed-effects multinomial logistic regression     Number of obs    =        494
                                                  Wald chi2(5)     =      45.69
Log likelihood = -138.19006                       Prob > chi2      =     0.0000
 ( 1)  [Labour]rldist3 = 0
 ( 2)  [Liberal]rldist2 = 0
 ( 3)  [Labour]rldist2 - [Liberal]rldist3 = 0
```

choice	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conservative	(base outcome)					
Labour						
rldist2	-.0590691	.0145332	-4.06	0.000	-.0875536	-.0305846
rldist3	(omitted)					
inflation	.8354586	.3692285	2.26	0.024	.111784	1.559133
yr92	.6791261	.2734095	2.48	0.013	.1432534	1.214999
Liberal						
rldist2	(omitted)					
rldist3	-.0590691	.0145332	-4.06	0.000	-.0875536	-.0305846
inflation	.5786913	.305657	1.89	0.058	-.0203854	1.177768
yr92	-.2315669	.2188483	-1.06	0.290	-.6605018	.1973679

The output header shows that 1,097 voters and respectively 1,964 observations are dropped, as for these voters there is no variance in the dependent variable across waves. That is, the model is estimated with 247 voters and 494 observations. The iteration log shows that the ML-algorithm converged after four steps. The log likelihood for the first step is the derived from the initial coefficient vector, which is the result of pooled multinomial logit with same variable structure. The header shows also the Wald test statistic of 45.69. The 5 degrees of freedom reflect the reduced number of free number of

parameters. Note that the command returns a Wald test instead of a LR test, as constraints were specified.

The coefficient table shows the logarithm of the relative risk ratios for a one-unit change in the corresponding variables. That is, with an increase in the relative distance between a voter and the Labour party by one unit ceteris paribus, the logarithm of the probability to vote for Labour divided by the probability to vote for the Conservative party decreases by 0.059. Equivalently, if ceteris paribus this relative distance increases by one unit, the odds to vote for Labour vs. voting Conservative increase by a factor of $\exp(-0.059) = 0.943$, that is, they decrease by 6.7 percent. Similarly, with each unit increase in the inflation rating ceteris paribus, the odds to vote for Labour vs. voting Conservative $\exp(\beta)$ increase by 130.6 percent, and the odds to vote Liberal vs. voting Conservative increase by 78.4 percent. The odds effects for other contrasts are interpreted by looking at the respective coefficient or variable differences. For example, if the inflation rating increases by unit ceteris paribus, the odds to vote Labour vs. voting Liberal increase by a factor of $\exp(0.835 - 0.579) = 1.293$ or 29.3 percent.

As mentioned previously, the femlogit model allows for possibly confounding unobserved heterogeneity at the level of the voter with respect to the preferences for a specific party. Alternative models have to rule this out or have to measure the heterogeneity. In table 1, I show the respective effects for the pooled multinomial logistic regression and the multinomial logistic regression with random effects. For the first model, panel-robust standard errors are used to take into account possible correlation across waves. The latter model is estimated with `gsem`, as described in StataCorp Lp (2009c, 407ff.). In the alternative models, heterogeneity is captured in the time-invariant variables `male`, `age`, and `manual`.

Table 1: Pooled, random-, and fixed-effects models for voting example

	pomlogit exp(beta) / se	remlogit exp(beta) / se	femlogit exp(beta) / se
Labour			
Relat. policy dist.	0.896***	0.818***	0.943***
	(0.005)	(0.011)	(0.014)
Inflation	2.134***	3.812***	2.306*
	(0.236)	(0.815)	(0.851)
1992 election	1.153	1.564*	1.972*
	(0.112)	(0.346)	(0.539)
Male	0.452***	0.261***	
	(0.068)	(0.082)	
Age	0.702***	0.499***	
	(0.037)	(0.056)	
Manual worker	1.952***	5.188***	
	(0.302)	(1.767)	
Constant	0.059***	0.007***	
	(0.029)	(0.007)	
Liberal			
Relat. policy dist.	0.896***	0.818***	0.943***
	(0.005)	(0.011)	(0.014)
Inflation	1.735***	2.938***	1.784
	(0.185)	(0.584)	(0.545)
1992 election	0.808*	0.771	0.793
	(0.080)	(0.159)	(0.174)

	pomlogit exp(beta) / se	remlogit exp(beta) / se	femlogit exp(beta) / se
Male	0.493***	0.304***	
	(0.073)	(0.092)	
Age	0.810***	0.632***	
	(0.039)	(0.066)	
Manual worker	0.900	1.235	
	(0.132)	(0.393)	
Constant	0.102***	0.013***	
	(0.048)	(0.012)	
Var($alpha_{Lab.}$)		14.672***	
		(2.988)	
Var($alpha_{Lab.}$)		13.915***	
		(2.325)	
Cov($alpha_{Lab.}$, $alpha_{Lab.}$)		11.441***	
		(2.377)	
log. likelihood	-1946.269	-1764.331	-138.190
N obs.	2458	2458	494
N groups	1344	1344	247

Notes: * p < .05, ** p < .01, *** p < .001; base outcome: Conservative party; reference categories: 1987 election, female, not manual worker; pomlogit: pooled multinomial logisitic regression, remlogit: multinomial logistic regression with random-effects.

5.2 Effect of smoking on birthweight with multi-level data

The second example builds on Abrevaya (2006), whose data is used in condensed form as an example in Rabe-Hesketh and Skrondal (2012, 2:123ff.). Abrevaya uses a multi-level data of children nested in mothers. The data contain information about children's birthweight and gestation age at birth and the mothers' smoking behavior during pregnancy, prenatal care for the child, and other sociodemographic information. With this information, Abrevaya uses fixed-effects models for continuous dependent variables to estimate the effect of smoking on the children's birthweight.

In this example, I switch perspective and analyze the timing of birth (pre-term/full term/post-term). I use the `femlogit` command to estimate the effect of smoking on the odds of pre-term birth vs. full term birth vs. post-term birth. The fixed-effects model implicitly controls for all constant variables at the level of the mother. Also, I control for prenatal care and prenatal visits to doctors, which vary within mothers across children. Finally, I control for birth year cohort dummies.

The analysis syntax for the second example is found in `femlogit_example3.do` and provided in the online appendix. I use the shortened version of the data from Rabe-Hesketh and Skrondal (2012, 2:123ff.). After cleaning and preparation, the analysis data is this:

```
. des
Contains data from http://www.stata-press.com/data/mlmus3/smoking.dta
  obs:         8,604
 vars:            19                          21 Jul 2007 11:49
 size:       189,288
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
momid           float   \%9.0g                 Mother id
idx             byte    \%9.0g                 Child number
smoke           byte    \%9.0g      s          Smoke
married         byte    \%9.0g                 Married
hsgrad          byte    \%9.0g                 12 years of education
somecoll        byte    \%9.0g                 13-15 years of education
collgrad        byte    \%9.0g                 16+ years of education
black           byte    \%9.0g      b          African-American
novisit         byte    \%9.0g                 No prenatal visit to doctor
gestage         byte    \%9.0g      gestage    Categorial gestation age at birth
y1              byte    \%8.0g                 1990 birth cohort
y2              byte    \%8.0g                 1991 birth cohort
y3              byte    \%8.0g                 1992 birth cohort
y4              byte    \%8.0g                 1993 birth cohort
y5              byte    \%8.0g                 1994 birth cohort
y6              byte    \%8.0g                 1995 birth cohort
y7              byte    \%8.0g                 1996 birth cohort
y8              byte    \%8.0g                 1997 birth cohort
kessner1        byte    \%9.0g                 Adequate prenatal care (Kessner
                                                index)
-------------------------------------------------------------------------------

Sorted by:  momid
     Note:  dataset has changed since last saved
```

The data contain 8,604 observations. The dependent variable `gestage` is generated from gestation age in weeks, which is provided in the raw data, following the WHO-definition (World Health Organization 2011, 151). The main independent variable `smoke` is a dummy variable that indicates, if the mother smoked during pregnancy with the respective child. The first control variable for the femlogit model is the dummy variable `novisit`, which indicates if the mother visited the doctor during pregnancy. The second control variable is the dummy variables `kessner1`, which indicates if the overall prenatal care was adequate according to the Kessner index in contrast to inadequate or intermediate (Kessner et al. 1973). I also control for a set of birth cohort dummies `y1`, ..., `y8`. The birth cohort of 1998 is the reference category.

As in the first example, the data is in long format. `xtdes` shows that the data contain information of 3,978 mothers with up to three children. For 3,330 mothers, there is information for two children and for 648 mothers there is information for three children.

```
. xtset momid idx
       panel variable:  momid (unbalanced)
        time variable:  idx, 1 to 3
                delta:  1 unit
. xtdes
   momid:  14, 25, ..., 109039                             n =       3978
     idx:  1, 2, ..., 3                                    T =          3
           Delta(idx) = 1 unit
           Span(idx)  = 3 periods
           (momid*idx uniquely identifies each observation)
```

```
Distribution of T_i:   min      5\%     25\%       50\%       75\%     95\%     max
                         2        2        2          2          2        3       3
     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
     3330     83.71   83.71 |  11.
      648     16.29  100.00 |  111
 ---------------------------+---------
     3978    100.00          |  XXX
```

The estimation results for the described femlogit model are these:

```
. femlogit gestage smoke kessner1 novisit y1-y8, /*
> */ b(2) group(momid) difficult
note: 2919 groups (6275 obs) dropped because of all positive or
      all negative outcomes.
Iteration 0:   log likelihood = -817.11022
Iteration 1:   log likelihood = -813.24731
Iteration 2:   log likelihood = -813.23023
Iteration 3:   log likelihood = -813.23023
Fixed-effects multinomial logistic regression     Number of obs   =       2329
                                                  LR chi2(22)     =      32.14
                                                  Prob > chi2     =     0.0750
Log likelihood = -813.23023                       Pseudo R2       =     0.0194
------------------------------------------------------------------------------
     gestage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pre_term     |
       smoke |   .2835948   .3039032     0.93   0.351    -.3120446    .8792343
    kessner1 |    -.21959   .1683691    -1.30   0.192    -.5495875    .1104074
     novisit |   .2727011   .5076698     0.54   0.591    -.7223135    1.267716
          y1 |   1.066149   1.165668     0.91   0.360    -1.218518    3.350816
          y2 |   .9058572   1.162697     0.78   0.436    -1.372986    3.184701
          y3 |   .8379417     1.1621     0.72   0.471    -1.439731    3.115615
          y4 |    1.27221    1.16112     1.10   0.273    -1.003543    3.547962
          y5 |    1.21856   1.169092     1.04   0.297    -1.072818    3.509937
          y6 |   1.633868   1.231724     1.33   0.185    -.7802663    4.048003
          y7 |   1.422654   1.254915     1.13   0.257    -1.036934    3.882243
          y8 |   1.354895   1.257938     1.08   0.281    -1.110619    3.820409
-------------+----------------------------------------------------------------
full_term    |  (base outcome)
-------------+----------------------------------------------------------------
post_term    |
       smoke |   .3272824    .249783     1.31   0.190    -.1622832     .816848
    kessner1 |  -.2699509   .1473989    -1.83   0.067    -.5588475    .0189457
     novisit |  -.3393401   .5737103    -0.59   0.554    -1.463792    .7851115
          y1 |  -.0426524     .92764    -0.05   0.963    -1.860793    1.775489
          y2 |   -.024454   .9218789    -0.03   0.979    -1.831304    1.782396
          y3 |  -.1014507   .9235822    -0.11   0.913    -1.911639    1.708737
          y4 |  -.3240081   .9224428    -0.35   0.725    -2.131963    1.483947
          y5 |  -.3278039   .9279563    -0.35   0.724    -2.146565    1.490957
          y6 |  -.2718321   .9930388    -0.27   0.784    -2.218152    1.674488
          y7 |  -.5344754   .9997496    -0.53   0.593    -2.493949    1.424998
          y8 |  -.8619275   1.062914    -0.81   0.417    -2.945201    1.221346
------------------------------------------------------------------------------
```

The information in the header shows that 2,919 mothers and respectively 6,275 children are taken out of the analysis, as there is no variation in the dependent variable across children for these mothers. That is, 1,059 mothers with 2,329 children remain in the analysis. The iteration log, the log likelihood and the χ^2-statistic hold essentially the same information as in the first example. However, note that here a LR-test-statistic and McFadden-Pseudo-R^2 are reported.

The table of coefficients looks different than in the first example, as I chose the second alternative full term birth as base outcome. Otherwise the interpretation is the same as in the first example. Smoking during pregnancy ceteris paribus increases the odds of pre-term birth vs. full term birth by the factor $\exp(.286) = 1.328$ or 32.8 percent. Similarily, it increases the odds of post-term birth by the factor $\exp(.327) = 1.387$ or 38.7 percent.[5]

The femlogit model implicitly controls for all factors at the level of the mother that do no vary across children. Alternative models as the pooled or random-effects models have to rely on complete measurement of these factors. Table \ref{tab2} shows the results for the femlogit and the alternative models.

For the pooled and the random-effects model, heterogeneity at the level of the mother is controlled with five variables: the dummy variable `married` indicates that the mother is married, the set of dummies `hsgrad`, `somecoll`, and `colgrad` indicate several levels of education in contrast to less than 12 years of education. The dummy `black` indicates that the mother is African-American. For all models, the set of dummies `y1`, ..., `y8`, which indicate the birth cohorts from 1990 to 1997, are also controlled for. Further, for the pooled model, cluster-robust standard-errors are reported, to control for auto-correlation within mothers across children.

Table 2: Pooled, random-, and fixed-effects models for smoking example

	pomlogit exp(beta) / se	remlogit exp(beta) / se	femlogit exp(beta) / se
Pre-term birth			
Smoked during pregnancy	1.352*	1.399*	1.328
	(0.175)	(0.205)	(0.404)
Adequate care (Kessner index)	0.730**	0.703**	0.803
	(0.076)	(0.084)	(0.135)
No visit of doctor	2.481**	2.763**	1.314
	(0.699)	(1.005)	(0.667)
African-American	1.902***	2.168***	
	(0.290)	(0.406)	
Married	0.735*	0.687*	
	(0.109)	(0.122)	
12 years of educ.	0.849	0.816	
	(0.124)	(0.142)	
13–15 years of educ.	0.750	0.697	
	(0.130)	(0.136)	
> 15 years of educ.	0.585**	0.543**	
	(0.101)	(0.106)	
Constant	0.116**	0.052***	
	(0.088)	(0.045)	
Post-term birth			
Smoked during pregnancy	1.223	1.254	1.387
	(0.137)	(0.157)	(0.346)
Adequate care (Kessner index)	0.748**	0.725**	0.763
	(0.068)	(0.073)	(0.113)
No visit of doctor	1.185	1.142	0.712
	(0.407)	(0.469)	(0.409)

[5] The effects are significant for the same model with the complete data of Abrevaya (2006).

	pomlogit exp(beta) / se	remlogit exp(beta) / se	femlogit exp(beta) / se
African-American	1.094	1.089	
	(0.168)	(0.193)	
Married	0.887	0.869	
	(0.117)	(0.134)	
12 years of educ.	0.774	0.750	
	(0.101)	(0.111)	
13–15 years of educ.	0.751*	0.722*	
	(0.108)	(0.116)	
> 15 years of educ.	0.563***	0.527***	
	(0.082)	(0.086)	
Constant	0.107**	0.081**	
	(0.079)	(0.065)	
$Var(alpha_{Pre\text{-}term\ birth})$		1.815***	
		(0.326)	
$Var(alpha_{Post\text{-}term\ birth})$		1.155***	
		(0.229)	
$Cov(alpha_{Pre\text{-}term\ birth}, alpha_{Post\text{-}term\ birth})$		-0.089	
		(0.235)	
log. likelihood	-4600.120	-4548.395	-813.230
N obs.	8604	8604	2329
N groups	3978	3978	247

Notes: * p < .05, ** p < .01, *** p < .001; base outcome: Full term birth; birth cohort dummies y1, . . . , y8 included; reference categories: Not smoked during pregnancy, intermediate or inadequate prenatal care, at least one visit of doctor, not African-American, not married, less than 12 years of education, birth cohort 1998; pomlogit: pooled multinomial logisitic regression, remlogit: multinomial logistic regression with random-effects.

5.3 Interpretation

Odds ratio and logit effects are criticized as unintuitive. Moreover, with this interpretation approach due to the arbitrary restriction assumption of the variance of the error term ε in equation, effects across nested models or across different group cannot be easily compared (Allison 1999; Kohler, Karlson, and Holm 2011; Best and Wolf 2015; Breen, Karlson, and Holm 2013). For nonlinear cross-sectional models, the interpretation of predicted probabilities and related constructs are recommended (Long and Freese 2006, 157ff.). This option is not given for the fixed-effects model. The probability expression in equation (3) cannot be evaluated, as the unobserved heterogeneity vector $\boldsymbol{\alpha}$ is not estimated. Even if plausible values for $\boldsymbol{\alpha}$ are inserted in the equation, to conduct significance tests, one has to find plausible values for their variances and covariances with the other independent variables. Cameron and Trivedi (2009, 797) suggest for the binary logistic regression with fixed effects to interpret predicted probabilities of the estimation equation (8), which can be generalized to the multinomial case. However, although this circumvents the problem of finding a plausible conditional distribution for the unobserved heterogeneity $f_{\boldsymbol{\alpha}|\mathbf{x}}$, the object of interpretation is more unintuitive than with the odds ratio and logit effects. With this approach, one interprets the effects of a unit or marginal change in the independent variables at a specific time x_t on the probability that a specific time series of outcomes $(y_1,\ldots,y_T)$ is realized conditional on the probability of all permutations of the time

series. For realistic applications, any choice of the outcomes time series is arbitrary. Further, the interpretation of the conditional probability remains intuitive, as the permutation can only be understood as an analogue for the general tendency to choose each outcome. In sum, the odds ratio effects interpretation as shown above is the only viable option for the binary and multinomial fixed effects logistic regression.

6 Conclusion

This article introduces an implementation of multinomial logistic regression with fixed effects as derived by Chamberlain (1980). With this model it is possible to consistently estimate effects on multinomial categorical dependent variables, when time-invariant unobserved heterogeneity is present. In particular, time-invariant unobserved heterogeneity may be correlated with predictor variables. The implemented ado `femlogit` is applied to real data. In the first example with British election panel data, the effect of perceived distance in the left--right political dimension between a candidate and a voter on voting behavior is estimated. In the second example with multi-level data, the effect of smoking during pregnancy on pre-term vs. full term vs. post-term birth is analyzed. The specific advantage of the femlogit model with the first examples is that the effect of policy distance on vote intention is estimated net of all time-invariant voter characteristics that may affect vote intention, perceived policy distance, or both. Analogously, in the second example we can estimate the effect of smoking behavior on birth timing under control of all stable characteristic of the mother, which may be correlated with both smoking behavior and birth outcomes.

7 References

Abrevaya, Jason. 2006. "Estimating the Effect of Smoking on Birth Outcomes Using a Matched Panel Data Approach." *Journal of Applied Econometrics* 21 (4): 489–519. doi:10.1002/jae.851.

Allison, Paul D. 1999. "Comparing Logit and Probit Coefficients across Groups." *Sociological Methods & Research* 28 (2): 186–208. doi:10.1177/0049124192020004001.

———. 2009. *Fixed Effects Regression Models*. Los Angeles, CA: SAGE.

Best, Henning, and Christof Wolf. 2015. "Logistic Regression." In *Regression Analysis and Causal Inference*, edited by Henning Best and Christof Wolf, 57–82. London: SAGE.

Breen, Richard, Kristian Bernt Karlson, and Anders Holm. 2013. "Total, Direct, and Indirect Effects in Logit and Probit Models." *Sociological Methods & Research* 42 (2): 164–91. doi:10.1177/0049124113494572.

Brüderl, Josef, and Volker Ludwig. 2015. "Fixed-Effects Panel Regression." *Regression Analysis and Causal Inference*, 327–57. doi:10.1017/CBO9781107415324.004.

Cameron, A Colin, and Pravin K Trivedi. 2009. *Microeconometrics: Methods and Applications*. Cambridge and New York, NY: Cambridge University Press. doi:10.1007/s13398-014-0173-7.2.

Chamberlain, Gary. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 57 (1): 225–38.

Gangl, Markus. 2010. "Causal Inference in Sociological Research." *Annual Review of Sociology* 36 (1): 21–47. doi:10.1146/annurev.soc.012809.102702.

Gould, William, Jeffrey Pitblado, and Brian Poi. 2010. *Maximum Likelihood Estimation with Stata*. College Station, TX: Stata Press.

Heath, Anthony F, Roger M Jowell, John K Curtice, Jack A Brand, and James C Mitchell. 1993. *British General Election Panel Survey, 1987-1992*. Arbor and MI: Inter-university Consortium for Political and Social Research. doi:10.3886/ICPSR06451.v1.

Kessner, David, James Singer, Carolyn E Kalk, and Edward R Schlesinger. 1973. *Infant Death: An Analysis by Maternal Risk and Health Care*. Vol. 1. Contrasts in Health Status. Washington, DC: Institute of Medicine.

Kohler, Ulrich, Kristian Bernt Karlson, and Anders Holm. 2011. "Comparing Coefficients of Nested Nonlinear Probability Models." *Stata Journal* 11 (3): 420–38.

Lee, Myoung-Jae. 2002. *Panel Data Econometrics: Methods-of-Moments and Limited Dependent Variables*. San Diego, CA: Academic Press.

Long, J Scott, and Jeremy Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. 2nd ed. College Station, TX: Stata Press.

Pforr, Klaus. 2014. "Femlogit - Implementation of the Multinomial Logit Model with Fixed Effects." *The Stata Journal* 14 (4): 1–16.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata*. 3. ed. Vol. 2. College Station, TX: Stata Press.

Skrondal, Anders, and Sophia Rabe-Hesketh. 2003. "Multilevel Logistic Regression for Polytomous Data and Rankings." *Psychometrika*, no. 33: 267–87. doi:10.1007/BF02294801.

StataCorp Lp. 2009a. *Mata Reference Manual: Release 11*. College Station, TX: Stata Press.

———. 2009b. *Stata Reference Manual: Release 11*. College Station, TX: Stata Press.

———. 2009c. *Stata Structural Equation Modeling Reference Manual: Release 11*. College Station, TX: Stata Press.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nded. Vol. 58. Cambridge, MA: MIT Press. doi:10.1515/humr.2003.021.

World Health Organization. 2011. *ICD-10: International Statistical Classification of Diseases and Related Health Problems*. 10th revis. Geneva: World Health Organization.