University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 321

Testing the Binomial Fixed Effects Logit Model; with an Application to Female Labor Supply

Rainer Winkelmann and Lin Xu

April 2019

Testing the binomial fixed effects logit model; with an application to female labor supply*

Rainer Winkelmann
University of Zurich

Lin Xu
University of Zurich

April 2019

Abstract

Regression models for proportions are frequently encountered in applied work. The conditional expectation is bound between 0 and 1 and, therefore, must be non-linear which requires non-standard panel data extensions. The quasi-maximum likelihood estimator of Papke and Wooldridge (1996) suffers from the incidental parameters problem when including fixed effects. In this paper, we re-consider the binomial panel logit model with fixed effects (Machado, 2004). We show that the conditional maximum likelihood estimator is very easy to implement using standard software. We investigate the properties of the estimator under misspecification and derive a new test for overdispersion in the binomial fixed effects logit model. Models and test are applied in a study of contracted work-time percentage, measured as proportion of full-time work, for women in Switzerland.

Keywords: proportions data, unobserved heterogeneity, conditional maximum likelihood, overdispersion

JEL classification: C23, J21.

* Department of Economics, University of Zurich, E-mail: rainer.winkelmann@econ.uzh.ch and lin.xu@econ.uzh.ch.
We are grateful to Damian Kozbur, Johannes Kunz, Kevin Staub and Tiemen Woutersen for helpful discussions. The data used in the application have been provided by the Swiss Household Panel. We gratefully acknowledge financial support by the Swiss National Science Foundation.

1 Introduction

After half a century of research on econometric models for limited dependent variables (Maddala, 1983, Wooldridge, 2002), it remains the case that only a small portion deals with proportions data, and even a smaller one with panel models for such proportions. Among the latter, Machado (2004) proposes the binomial fixed effects logit model, Papke and Wooldridge (2008) a correlated random effects probit quasi-likelihood estimator and Ramalho et al. (2016) a class of exponential GMM estimators.

And yet, proportions and related types of data are regularly encountered in applied econometric work. Often, they correspond to the number of "successes" in a sequence of Bernoulli trials, such as homicide- or unemployment rates, or the fraction of days absent from work during a work week. Also, variety scores (e.g., the number of applicable items in a general health questionnaire), bounded count data, as well as ratings, share the key feature of discreteness and the existence of an upper and lower bound for the outcome.

In all these cases, the binomial model with a logit function for the expected proportion provides a natural starting point for modelling. For the fixed effects setting, Machado (2004) shows that the incidental parameters problem can be overcome by a conditional maximum likelihood (CML) estimator, much like it is the case for the binary response logit model. She also provides Monte Carlo evidence indicating that the dummy variables (DV) approach is subject to an upward bias that is decreasing in both the length of the panel, T, and in the number of Bernoulli trials, K. For $T > 5$ and $K > 5$, CML and DV approaches yield quite comparable results (Machado, 2004).

This paper advances this earlier work in three important directions: First, we show how the binomial logit fixed effects estimator can be implemented using any off-the-shelf statistical software that includes a conditional logit routine. Second, we study the properties of the CML and DV estimators for the case that the binomial distributional assumption fails. The leading example is that of overdispersion, as it originates from unobserved heterogeneity or dependence among the Bernoulli trials. Since the CML estimator is not a pseudo ML estimator in the sense of Gourieroux, Monfort and Trognon (1984), it does not possess formal robustness properties. We therefore investigate the extent of bias in a series of simulation experiments. Third and finally, we derive and implement a new test for the binomial assumption, i.e., a test for the hypothesis of no overdispersion, as existing tests (e.g. Dean, 1992) cannot be applied because the fixed effects are not estimated by the CML estimator.

To illustrate the proposed methods, we conduct a study of the determinants of women's work behavior, as measured by the contracted work-time percentage, where 0 means no work and 1 means full-time work. The binomial logit estimates indicate that having children is associated with substantially reduced work-time percentage, ceteris paribus. Perhaps more surprisingly, having a partner makes the effect more pronounced, whereas speaking French reduces it. Without accounting for fixed effects, the work-time percentage difference between mothers and non-mothers is underestimated, indicating a positive selection into motherhood.

2 Model and estimation

A proper panel model for proportions $y_{it} \in [0, 1]$ must overcome two challenges. First, the model should observe the restricted support of the outcome, as well as being able to handle data clustering at the end points. For instance, the log-odds transformation (see Berkson's minimum chi-square method in Maddala, 1983), $\log[y_{it}/(1 - y_{it})]$ is not defined for $y_{it} = 0$ or $y_{it} = 1$. Another method facing the same limitation is beta regression, which is flexible for fitting continuous proportional data but cannot give predictions at the boundaries with positive probability. Second, direct control for unobserved time-invariant individual heterogeneity (that may or may not be correlated with the regressors) using a dummy for each cross-sectional unit is subject to the incidental parameters problem, giving inconsistent and severely biased estimation of structural parameters when the length of panel T is fixed.

Machado (2004) addresses these two issues by proposing a conditional maximum likelihood estimator for the binomial fixed effects logit model.

Assumption 1

Let $Y_{it} = Ky_{it}$ where K is a known integer, and

$$y_{it} \in \left\{0, \frac{1}{K}, \frac{2}{K}, \dots, 1\right\}$$

such that

$$Y_{it}|p_{it} \sim \text{binomial}(K, p_{it}), \qquad i = 1, \dots, N; \qquad t = 1, \dots, T \tag{1}$$

Here, K is the number of "trials", $Y_{it} = Ky_{it}$ is the "number of successes", and y_{it} is the proportion, or fraction of successes for observation unit i in period t.

Assumption 2

Let the expected proportion depend on covariates x_{it} and an individual specific effect α_i as follows:

$$\mathrm{E}(y_{it}|x_{it},\alpha_i) = p_{it} = \frac{\exp(x'_{it}\beta+\alpha_i)}{1+\exp(x'_{it}\beta+\alpha_i)} \equiv \Lambda_{it} \tag{2}$$

x_{it} and α_i can be correlated.

Assumption 3

Observations are independent between individuals and, conditional on group effects α_i, serially uncorrelated.

The objective of the analysis is estimation of β. Under Assumptions 1-3, the joint binomial density for $Y_{i1}, Y_{i2}, \ldots, Y_{iT}$ conditional on $\sum_t Y_{it}$ is given by (see Machado, 2004)

$$f\left(Y_{i1}, Y_{i2}, \ldots, Y_{iT}|\sum_t Y_{it}\right) = \frac{\Pi_t\binom{K}{y_{it}}p_{it}^{Y_{it}}(1-p_{it})^{K-Y_{it}}}{\sum_{q\in Q_i}\Pi_t\binom{K}{q_t}p_{it}^{q_t}(1-p_{it})^{K-q_t}} = \frac{\exp(\sum_t Y_{it}x'_{it}\beta)\Pi_t\binom{K}{y_{it}}}{\sum_{q\in Q_i}\exp(\sum_t q_t x'_{it}\beta)\Pi_t\binom{K}{q_t}} \tag{3}$$

where $Q_i = \{q = (q_1, q_2, \ldots, q_T)|q_t \in \{0, 1, 2, \ldots, K\}, \sum_t q_t = \sum_t Y_{it}\}$. The conditional binomial approach eliminates the fixed effects α_i which appear in the numerator and denominator with same power. Observations for which $\sum_t Y_{it} = 0$ or $\sum_t Y_{it} = 1$ have a conditional probability of 1 and do not contribute to estimation of β. For proportion data, such outcomes tend to be much less prevalent than for binary outcomes.

2.1 An alternative implementation

Next, we show how the binomial logit fixed effects estimator can be implemented using any off-the-shelf statistical software with a conditional logit routine. Our approach to estimation exploits the fact that the binomial distribution arises as the sum of K independent Bernoulli trials, which technically leads to an equivalence of two estimators, one based on a binomial log-likelihood function and the other based on a Bernoulli log-likelihood for an expanded dataset. Consider first the case of a pure cross-section. In this case, α_i is not identified and we drop it for the moment from expression (2).

For the expanded dataset, one simply generates a sequence of K copies for each i, keeping the regressors unchanged, where the proportion y_i is replaced by a sequence of 0/1 indicator variables

d_{ij} in arbitrary order such that

$$\sum_{j=1}^{K} d_{ij} = Ky_i \tag{4}$$

It follows that d_{ij} and y_i have the same CEF:

$$\mathrm{E}(y_i|x_i) = \mathrm{E}\left(\frac{\sum_{j=1}^{K} d_{ij}}{K}\middle| x_i\right) = \mathrm{E}(d_{ij}|x_i) \tag{5}$$

Suppose that $\mathrm{E}(y_i|x_i) = \Lambda_i$ is of logistic form. Then, the (Bernoulli) log-likelihood function of the expanded dataset is given by

$$\begin{aligned} \log L &= \sum_{i}^{N}\sum_{j}^{K} d_{ij}\log(\Lambda_i) + (1-d_{ij})\log(1-\Lambda_i) \\ &= \sum_{i}^{N} Y_i \log(\Lambda_i) + (K-Y_i)\log(1-\Lambda_i) \end{aligned}$$

and we obtain the following first derivative, or score function,

$$\begin{aligned} \frac{\partial \log L}{\partial \beta} &= K\sum_{i}^{N}\left(\frac{y_i}{\Lambda_i} - \frac{1-y_i}{1-\Lambda_i}\right)\Lambda_i' x_i \\ &= K\sum_{i}^{N}\left(y_i - \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}\right)x_i \end{aligned} \tag{6}$$

This first-order condition is identical to the score function of the binomial log-likelihood and differs from that of the Bernoulli quasi-log-likelihood (Papke and Wooldridge, 1996) only by a proportionality constant. Hence, the three estimators are identical in a cross-section.

For panel data, the binomial probability p_{it} has conditional expectation as defined in (2). In the expanded model, every y_{it} is replaced by K binary variables d_{ijt}, such that $\sum_j d_{ijt} = Ky_{it} = Y_{it}$. Conditional on $\sum_t \sum_j d_{ijt} = \sum_t Ky_{it}$, $i = 1, 2, \ldots, N$, the conditional density function for each individual i can be written as follows

$$f\left(\{d_{ijt}\}|\sum_t\sum_j d_{ijt}\right) = \frac{\Pi_t\Pi_j p_{it}^{d_{ijt}}(1-p_{it})^{1-d_{ijt}}}{\sum_{s\in S_i}\Pi_t\Pi_j p_{it}^{s_{jt}}(1-p_{it})^{K-s_{jt}}} = \frac{\exp(\sum_t\sum_j d_{ijt}x_{it}'\beta)}{\sum_{s\in S_i}\exp(\sum_t\sum_j s_{jt}x_{it}'\beta)} \tag{7}$$

where $S_i = \{(s_{11}, s_{12}, \ldots, s_{1T}, s_{21}, \ldots, s_{KT})|s_{jt} \in \{0,1\}, \sum_t\sum_j s_{jt} = \sum_t\sum_j d_{ijt})\}$

Compared with equation (3), the number of s such that $\{s|\sum_j s_{ijt} = q_{it}\}$ is $\binom{K}{q_t}$ for given q. Equation (7) is therefore basically the same as equation (3), except for the term $\Pi_t\binom{K}{y_{it}}$ in the numerator. But this term does not depend on any parameter and thus drops out of the first-order condition for the maximum of the log-likelihood function. Specifically, the conditional Bernoulli log-likelihood function can be written as:

$$\log L = \sum_i \left[\sum_t \sum_j d_{ijt} x'_{it}\beta - \log(\sum_{s\in S_i} \exp(\sum_t \sum_j s_{jt} x'_{it}\beta)) \right] \tag{8}$$

with first derivative

$$\frac{\partial \log L}{\partial \beta} = \sum_i \left[\sum_t K y_{it} x'_{it} - \frac{\sum_{s\in S_i} \exp(\sum_t \sum_j s_{jt} x'_{it}\beta) \sum_t \sum_j s_{jt} x'_{it}}{\sum_{s\in S_i} \exp(\sum_t \sum_j s_{jt} x'_{it}\beta)} \right] \tag{9}$$

which is the same as in the conditional binomial model, up to an additive constant, and will give the same consistent estimator of β, after elimination of the fixed effects.

2.2 Overdispersion

Departures from the binomial proportions model can take a number of forms. A first one is a violation of the independence assumption of the underlying Bernoulli trials. Positive dependence, or contagion, among the sequence of Bernoulli trials causes overdispersion, a conditional variance exceeding the one implied by the binomial model equal to $Kp_{it}(1-p_{it})$. Another violation is "unobserved heterogeneity", where p_{it} is no longer a constant but rather a random variable, say $\tilde{p}_{it}$. Marginalizing over $\tilde{p}_{it}$ then leads to a mixture model that is characterized by overdispersion as well. Depending on the distribution of $\tilde{p}_{it}$, proportions can for example have a u-shaped probability function even conditional on α_i and x_{it}, i.e., probability mass stacked at the endpoints of 0 and 1, which is never the case for a binomial distribution that has either a single mode, or two adjacent modes.

A prominent example for a continuous mixture is the beta-binomial model, where

$$\tilde{p}_{it} \sim \text{beta}(u_{it}, v_{it}) \tag{10}$$

and

$$u_{it} = \phi\Lambda(x_{it}\beta + \alpha_i), v_{it} = \phi(1 - \Lambda(x_{it}\beta + \alpha_i))$$

It is easy to show that the thus obtained beta-binomial distribution has expectation $K\Lambda_{it}$ and variance

$$\text{Var}(Y_i|K, \Lambda_{it}, \phi) = K\Lambda_{it}(1 - \Lambda_{it})\left(1 + \frac{K-1}{\phi+1}\right) \tag{11}$$

Thus, the variance of the beta-binomial model is proportional to that of the binomial model. The degree of overdispersion increases in K, the number of trials, and it decreases in the parameter ϕ. The binomial variance is obtained for $K = 1$, or in the limit, for $\phi \to \infty$, which also means that $\text{Var}(\tilde{p}_{it}) \to 0$.

In general, fixed effects conditional maximum likelihood estimators are not consistent if the underlying model is misspecified. The reason is that the first-order condition is no longer a moment condition for the mean, but rather a function of the conditional probabilities (Gourieroux et al., 1984). However, it might still be the case that the CML estimator works satisfactorily as long as the degree of overdispersion, and hence the departure from the binomial assumption, is not too large. We will explore this type of robustness in a series of simulation experiments.

2.3 Simulation study

We conduct simulation experiments for two different data generating processes (DGPs): one, where the binomial assumption is satisfied, and a second one, based on the beta-binomial model, where overdispersion is present. Unobserved time-invariant individual heterogeneity is positively correlated with the regressor in both cases. The degree of overdispersion is varied from 10 to 100 percent.

Both set-ups use the same logit conditional expectation function with a single regressor

$$\text{E}(y_{it}|x_{it}, \alpha_i) = \Lambda(\beta_0 + \beta_1 x_{it} + \alpha_i) = \frac{\exp(\beta_0 + \beta_1 x_{it} + \alpha_i)}{1 + \exp(\beta_0 + \beta_1 x_{it} + \alpha_i)} \quad , \tag{12}$$

where $\beta_0 = 0$, $\beta_1 = 2$ and the size of the cross-section is either $N = 100$ or $N = 500$. The time dimension increases from $T = 2$, $T = 5$ to $T = 10$.

The regressor x_{it} is drawn from a uniform distribution with support $[-1, 1]$ and has therefore a mean of 0 and a variance of 1/3. Draws are independent both across individuals and over time. We make a correlated random effects assumption:

$$\alpha_i = \sqrt{T}\bar{x}_i + \varepsilon_i, \tag{13}$$

where $\varepsilon_i \sim N(0,1)$. It follows that the correlation between α_i and $\bar{x}_i$ is 0.5, a substantial amount.

Once the mean is given, the dependent variable is obtained by generating pseudo random numbers from either a binomial or a beta-binomial distribution. Specifically, we first draw integer random numbers from a (beta) binomial distribution with parameters K and $\Lambda(x_{it}\beta_1+\alpha_i)$, and then divide the result by the number of categories K. e.g.:

$$y_{it} = \frac{Ky_{it}}{K}, Ky_{it} \sim \text{binomial}(K, p_{it}), p_{it} = \Lambda(\beta_0 + \beta_1 x_{it} + \alpha_i) \tag{14}$$

K is exogenously determined. In our case, we set K to 2, 5 or 10. For $K=2$, $K \times y_{it}$ can be 0, 1, or 2, with corresponding fractions of $y_{it} = 0, 0.5$, or 1, respectively; if $K=10$, y_{it} takes on one-digit decimals: 0, 0.1, 0.2,..., 1.

The theoretical predictions are clear. Ignoring the presence of the individual specific component and estimating the marginal, pooled model will have two effects:

- β_1 is upward biased due to the positive correlation between x_{it} and α_i.
- β_1 is downward biased due to omitted heterogeneity. In the probit model, there is a closed form expression for this bias (Wooldridge, 2002). In the logit model, it needs to be computed numerically, but the direction is the same.

Which one of the two biases is larger is an empirical matter. The DV estimator, on the other hand, suffers from the standard upward incidental parameters bias (Abrevaya, 1997).

– – – – – – – – – – Table 1 about here – – – – – – – – – –

Table 1 shows the simulation results based on 1000 replications, for a sample size of $N = 100$. The mean and standard deviation of estimated coefficients across replications are reported. The three estimators are referred to as Blogit CML, Blogit DV, and pooled logit respectively. The first row of each sub-panel (for $K = 2$, $K = 5$ and $K = 10$) gives the results for the DGP without overdispersion, i.e. results where the dependent variable has a binomial distribution conditional on x_{it} and α_i.

As Machado (2004), we find that the Blogit CML model estimates the true structural slope parameter very well even for small samples. There is a 2% upward bias for $T = K = 2$ that vanishes quickly as either T or K increase. The sampling variability inmcreases not only in T but also in K, albeit at a less than $\sqrt{K}$ rate. The Blogit DV estimators have a larger bias and a larger standard error, and hence a higher mean squared error, in all settings. It becomes small as T and K increase. For instance, for $T = 10$ and $K = 10$, the mean Blogit DV estimate is 2.025, whereas the mean Blogit CML estimate is 2.000. On the other hand, the pooled logit estimator has no tendency to converge. Depending on parameterization of the DGP, it overestimates or underestimates the true parameter $\beta_1 = 2$. In Table 2, the simulations are repeated for a larger sample, $N = 500$ instead of $N = 100$, but the qualitative conclusions remain unchanged.

Beta-binomial DGP

Simulations from the beta-binomial model add a further step: instead of directly obtaining binomial responses with conditional (on x_{it} and α_i) success probability $p_{it} = \Lambda(\beta_0 + \beta_1 x_{it} + \alpha_i)$, $\tilde{p}_{it}$ is now drawn from a beta distribution with mean p_{it}:

$$\tilde{p}_{it} \sim \text{beta}(\phi\Lambda(\beta_0 + \beta_1 x_{it} + \alpha_i), \phi(1 - \Lambda(\beta_0 + \beta_1 x_{it} + \alpha_i)))$$

From (11), we know that the multiplicative variance inflation factor depends on both K and ϕ. To keep the degree of overdispersion the same for $K = 2, 5, 10$, we adjust ϕ accordingly. For example, for 10% overdispersion and $K = 2$, we have $1 + (K - 1)/(\phi + 1) = 1.1$, so $\phi = 9$.

– – – – – – – – – – Table 2 about here – – – – – – – – – –

From Table 1 (for $N = 100$) and Table 2 (for $N = 500$), three key patterns emerge. First, overdispersion leads to an upward bias of both the Blogit CML and the Blogit DV estimators. The bias increases in the amount of overdispersion. Second, the Blogit CML estimator always dominates the Blogit DV estimator, both in terms of bias and standard error. This result did already hold for the binomial case, and it persists in the presence of overdispersion. Third, for a given degree of overdispersion, the bias is decreasing in T as well as in K.

The overall conclusion is that the Blogit CML estimator maintains a rather good performance even if the binomial model is misspecified, as long as the degree of overdispersion is modest, or else, as long as T and / or K are large. To take the two polar cases, for $N = 100$, if $K = T = 2$ the mean estimate with 10% overdispersion is 2.1, a 5% upward bias. For $K = T = 10$ the mean estimate with 100% overdispersion is 2.025, a 1.25% upward bias.

3 A test for overdispersion

Existing methods to test the binomial assumption for y_{it}, e.g. Dean's (1992) score test, require estimates of the expectation Λ_{it} under the null hypothesis of the binomial model. However the Blogit CML approach does not estimate α_i, so Λ_{it} is not complete. To ascertain the validity of the Blogit CML model assumption, we construct a novel statistic test as follows.

We observe panel data (y_{it}, x_{it}), $i = 1, \ldots, N$, $t = 1, \ldots, T$, where $T \geq 2$. Consider the case $T = 2$ first. The response variable $y_{it} \in \{0, \frac{1}{K}, \frac{2}{K}, \ldots, 1\}$ can be at the endpoint, zero or one, with increment $1/K$. To test whether non-binomial variance dispersion exists, a binary random variable M_{it} is drawn from a Bernoulli distribution with mean y_{it}, $M_{it} \sim \text{Bernoulli}(y_{it}), \forall t = 1, 2$. Equivalently, we can draw a random element from the d_{ij}-sequence defined in (4). The basic idea of the test will be a comparison between the variances of the differences $Y_{i1} - Y_{i2}$ and $M_{i1} - M_{i2}$. If the underlying probabilities p_{it} are the same across periods, then the outcomes $Y_{i1}, \ldots, Y_{iT}$ can be regarded as random draws from i.i.d binomial distributions and the variance of $Y_{i1} - Y_{i2}$ should be equal to the sum of binomial variances, under assumptions A1 and A3. On the other hand, the Bernoulli draws from the same distributions have standard variances. If there is over- or under-dispersion, the variance of $Y_{i1} - Y_{i2}$ will be larger or smaller than the variance calculated from Bernoulli draws. Since we do not observe p_{it}, the probability of success for the binomial draws is estimated using y_{it}.

More specifically, consider the variable

$$z_i = \frac{(Y_{i1} - Y_{i2})^2 - K(M_{i1} - M_{i2})^2}{K(K-1)} \tag{15}$$

Conditional on y_{i1}, y_{i2},

$$\begin{aligned} \mathrm{E}[(M_{i1} - M_{i2})^2 | y_{i1}, y_{i2}] &= y_{i1}(1 - y_{i1}) + y_{i2}(1 - y_{i2}) + (y_{i1} - y_{i2})^2 \\ &= y_{i1} + y_{i2} - 2y_{i1}y_{i2} \end{aligned}$$

Therefore, under A1, A2 and A3, the expectation of z_i is given by

$$\mathrm{E}(z_i) = \frac{1}{K(K-1)} \left[\mathrm{Var}(Y_{i1}) + \mathrm{Var}(Y_{i2}) + (\mathrm{E}Y_{i1} - \mathrm{E}Y_{i2})^2 - K(\Lambda_{i1} + \Lambda_{i2} - 2\Lambda_{i1}\Lambda_{i2}) \right]$$

Under the binomial assumption, $\mathrm{Var}(Y_{it}) = K\Lambda_{it}(1 - \Lambda_{it})$, and it follows that

$$\begin{aligned} \mathrm{E}(z_i) = &\frac{1}{K(K-1)} \left[K\Lambda_{i1}(1 - \Lambda_{i1}) + K\Lambda_{i2}(1 - \Lambda_{i2}) + K^2(\Lambda_{i1} - \Lambda_{i2})^2 \right. \\ &\left. -K[\Lambda_{i1}(1 - \Lambda_{i1}) + \Lambda_{i2}(1 - \Lambda_{i2}) + (\Lambda_{i1} - \Lambda_{i2})^2] \right] \\ = &(\Lambda_{i1} - \Lambda_{i2})^2. \end{aligned} \tag{16}$$

Hence, the expected value of z_i is zero under the null hypothesis of binomial dispersion as long as $\Lambda_{i1} = \Lambda_{i2}$. From (2), this is the case for observations for which the regressors x_{it} do not change over time, i.e., $x_{i1} = x_{i2}$. With overdispersion, $\mathrm{E}(z_i)$ is greater than zero, with underdispersion smaller than zero.

3.1 Case I : a test for discrete covariates

Define the set of individuals with the same expectations over time, $A = \{i : \Lambda_{i1} = \Lambda_{i2}\}$, for which $E(z_i|i \in A) = 0$ holds. With time invariant fixed effect α_i, the set A is equal to $\{i : x_{i1} = x_{i2}\}$. It is feasible to find such a set A if all covariates are finite discrete variables. The test term for discrete x_{it} is defined as

$$\tau_A = \hat{E}(z_i|i \in A) = \frac{\sum_{i \in A} z_i}{|A|}, \tag{17}$$

where $|A|$ represents the number of elements in A. Under H_0, $\tau_A \xrightarrow{p} 0$. Further, by the central limit theorem (CLT), the statistic τ_A converges to a normal distribution,

$$\sqrt{|A|}(\tau_A - 0) \xrightarrow{d} N(0, \sigma_A^2), \tag{18}$$

where $\sigma_A^2 = \mathrm{Var}(z_i|i \in A)$. In practice, σ_A^2 is replaced by the sample variance $\hat{\sigma}_A^2$. So we reject the binomial distribution assumption at the $\alpha\%$ significance level if $\left|\frac{\tau_A}{\hat{\sigma}_A/\sqrt{|A|}}\right| \geq Z_{1-\frac{\alpha}{2}}$.

Note that individuals in the set A do not contribute to the estimation of the Blogit CML model, since x_{it} are canceled out as fixed effects. Nonetheless, they are needed for generating our dispersion test. This non-parametric method to build a test is similar to finding proper cell estimators in matching theory, but likewise faces the curse of dimensionality. It is hard to find the set A when the dimension of x_{it} becomes larger. If $|A|$ shrinks, the convergence rate $\sqrt{|A|}$ will decrease and the estimator τ_A will converge more slowly.

3.2 Case II: a kernel test for continuous covariates

The set $A = \{i : \Lambda_{i1} = \Lambda_{i2}\}$ is empty or very small when Λ_{i1} and Λ_{i2} are continuous. A more general method uses a kernel estimator for the conditional expectation $\mathrm{E}(z_i|\Lambda_{i1} - \Lambda_{i2} = 0)$. The main idea is to put more weight on individuals with smaller $|\Lambda_{i1} - \Lambda_{i2}|$. Since we do not observe the underlying expectations Λ_{it} directly, we find the set A according to observables x_{it}. Under the assumption of a single scalar regressor and time-invariant unobserved heterogeneity, we can decompose the conditional expectation (16) by a Taylor expansion at x_{i2},

$$\begin{aligned}(\Lambda_{i1}-\Lambda_{i2})^2 &= [\Lambda(x_{i1}\beta+\alpha_i)-\Lambda(x_{i2}\beta+\alpha_i)]^2\\ &= [\Lambda'(x_{i2}\beta+\alpha_i)\beta\,(x_{i1}-x_{i2})+\frac{\Lambda''(x_{i2}\beta+\alpha_i)}{2!}\beta^2\,(x_{i1}-x_{i2})^2+o((x_{i1}-x_{i2})^2)]^2\\ &= [\Lambda'(x_{i2}\beta+\alpha_i)\beta\,(x_{i1}-x_{i2})]^2+o(\beta^2(x_{i1}-x_{i2})^2),\end{aligned}$$

Denote $\Delta_i = (x_{i1}-x_{i2})\beta$,

$$E(z_i|\Lambda_{i1}-\Lambda_{i2}) = (\Lambda_{i1}-\Lambda_{i2})^2 = (\Lambda'_{i2}\Delta_i)^2+o(\Delta_i^2).$$

As the fixed effect α_i is canceled out, an alternative conditional expectation function is given by Δ_i,

$$\tau(\Delta) = E(z_i|\Delta_i=\Delta, X_i) = (\Lambda'_{i2}\Delta)^2.$$

Then, under the binomial assumption,

$$E(z_i|\Lambda_{i1}-\Lambda_{i2}=0) = \tau(0) = 0.$$

The result generalizes to a vector-valued x, in which case $\Delta_i = (x_{i1}-x_{i2})'\beta$.

The next step is to build a kernel estimator for $\tau(0)$. One conditional moment estimator for $\tau(\Delta)$ is $\hat{\tau}(\Delta) = \frac{\sum_{i=1}^N K(\frac{\Delta_i-\Delta}{h})z_i}{\sum_{i=1}^N K(\frac{\Delta_i-\Delta}{h})}$, where h is the kernel bandwidth for Δ_i and $K(\frac{\Delta_i-\Delta}{h})$ is the kernel function. For a given sample, Δ_i needs to be replaced by $\hat{\Delta}_i = (x_{i1}-x_{i2})\hat{\beta}$, where $\hat{\beta}$ is estimated. We can use the Blogit CML estimator for estimation, as it is consistent under the binomial null hypothesis. We construct a local estimate $\hat{\tau}$ for the object of interest $\tau(0)$ (see Pagan and Ullah, 1999):

$$\hat{\tau} = \frac{\sum_{i=1}^N K(\frac{\hat{\Delta}_i}{h})z_i}{\sum_{i=1}^N K(\frac{\hat{\Delta}_i}{h})} = \sum_{i=1}^N w_{ni}z_i, \quad w_{ni} = \frac{K(\frac{\hat{\Delta}_i}{h})}{\sum_{i=1}^N K(\frac{\hat{\Delta}_i}{h})},$$

We chose the Gaussian function $K(\frac{\hat{\Delta}_i}{h}) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{(\hat{\Delta}_i/h)^2}{2})$ for simplicity.

Asymptotic property and kernel bandwidth choice

Let $f = f(\Delta=0)$ denote the continuous density function of the random variable Δ at point 0. The kernel density estimator $\hat{f}$ for f is

$$\hat{f} = \sum_{i=1}^N \frac{K(\frac{\hat{\Delta}_i}{h})}{nh}.$$

In addition, rewrite z_i as the sum of its conditional expectation $E(z_i|\Delta_i) = \tau(\Delta_i)$ and an error term u_i, such that

$$z_i = \tau(\Delta_i) + u_i = (\Lambda'_{i2}\Delta_i)^2 + u_i$$

where $\mathrm{E}(u_i|\Delta_i, X_i) = 0$ and $\mathrm{Var}(u_i|\Delta_i, X_i) = \sigma^2$.

The estimator $\hat{\tau}$ is a combination of $\hat{f}$ and z_i

$$\hat{\tau} = \frac{\sum_{i=1}^{N} \frac{1}{nh} K(\frac{\hat{\Delta}_i}{h}) z_i}{\sum_{i=1}^{N} \frac{1}{nh} K(\frac{\hat{\Delta}_i}{h})} = \frac{1}{\hat{f}} \sum_{i=1}^{N} \frac{1}{nh} K\left(\frac{\hat{\Delta}_i}{h}\right) z_i = \frac{1}{\hat{f}} \sum_{i=1}^{N} \frac{1}{nh} K\left(\frac{\hat{\Delta}_i}{h}\right) (\Lambda'_{i2}\Delta_i)^2 + u_i.$$

The expectation of $\hat{\tau}$ is

$$\begin{aligned}
\mathrm{E}(\hat{\tau}) &= E(\frac{1}{\hat{f}} \sum_{i=1}^{N} \frac{1}{nh} K(\frac{\hat{\Delta}_i}{h})(\Lambda'_{i2}\Delta_i)^2 + \frac{1}{\hat{f}} \sum_{i=1}^{N} \frac{1}{nh} K(\frac{\hat{\Delta}_i}{h}) u_i) \\
&= \int\int \frac{1}{h\hat{f}} K(\nu)(\Lambda'_{i2})^2 (h\nu)^2 f(h\nu, \Lambda_{i2}) h d\nu d\Lambda_{i2} + \hat{E}(u_i|\hat{\Delta} = 0), \quad \text{where we replace } \Delta = h\nu \\
&= h^2 \int\int K(\nu)(\nu)^2 (\Lambda'_{i2})^2 \frac{f(\nu, \Lambda_{i2})}{\hat{f}} d\nu d\Lambda_{i2} \\
&= h^2 \mu_2 E[(\Lambda'_{i2})^2 | \Delta = 0], \quad \text{where } \mu_2 = \int K(\nu)(\nu)^2 d\nu
\end{aligned}$$

We therefore obtain a bias

$$\mathrm{Bias}(\hat{\tau}) = \mathrm{E}(\hat{\tau}) - \tau(0) = \mathrm{E}(\hat{\tau}) = h^2 \mu_2 E[(\Lambda'_{i2})^2 | \Delta = 0], \tag{19}$$

that is proportional to h^2.

To guarantee consistency of the estimator $\hat{\tau}_n$, convergence of the mean square error to zero is required. The MSE is equal to $\mathrm{MSE}(\hat{\tau}) = \mathrm{Bias}(\hat{\tau})^2 + \mathrm{Var}(\hat{\tau})$. So the bias for τ_n should decrease to zero, as n increases:

$$h^2 \longrightarrow 0, \quad \text{as } n \longrightarrow \infty. \tag{20}$$

Besides the convergence condition for bias, we also consider the asymptotic performance of the variance of $\hat{\tau}$. Using a result on the variance of conditional expectations from Pagan and Ulah (1999), we obtain:

$$\mathrm{Var}(\hat{\tau}) = \frac{\sigma^2}{nhf} \int K^2(\nu) d\nu, \quad \mathrm{Var}(\hat{\tau}) \propto \frac{1}{nh}. \quad \text{If } n \longrightarrow \infty, \quad \frac{1}{nh} \longrightarrow 0. \tag{21}$$

To make sure that the MSE converges at the fastest speed, bias^2 and variance should converge at the same rate: $h^4 \propto \frac{1}{nh}$. Otherwise, the slower speed dominates the convergence rate. Thus, h is of order $h \propto n^{-\frac{1}{5}}$ and by the central limit theorem,

$$\sqrt{nh}(\hat{\tau} - E(\hat{\tau})) \xrightarrow{d} N(0, f^{-1}\sigma^2 \int K^2(\nu)d\nu) \tag{22}$$

Here $\sigma^2 = \text{Var}(z_i^2|\Delta = 0)$, with the same definition in the discrete case (eq 18). In practice, the approximate bias is calculated by $\hat{\text{E}}(\hat{\tau}) = \sum_{i=1}^{N} w_{ni}(y_{i2}(1 - y_{i2})\hat{\Delta})^2$, σ^2 is replaced by $\hat{\sigma}^2 = \sum w_{ni}(z_i - \tau(\hat{\Delta}_i))^2$ and $\widehat{\text{Var}}(\hat{\tau}) = \frac{\sigma^2}{\hat{f}^2}\frac{\sum_{i=1}^{N} K^2(\nu)}{n^2h^2}$. Hence, $\left|\frac{\hat{\tau} - \hat{\text{E}}(\hat{\tau})}{\sqrt{\widehat{\text{Var}}(\hat{\tau})}}\right|$ can be used as a t-test.

To shrink the bias, in other words, to undersmooth, h should be $h \propto o(n^{-\frac{1}{5}})$. But this involves a trade-off between bias and variance, i.e., smaller bandwidth decreases the convergence rate and decreases the asymptotic variance, making rejection much harder under the alternative assumptions. By Silver's rule of thumb, the bandwidth for the Gaussian kernel function is set to $h = 0.9\sigma_{\Delta_i} n^{-\frac{1}{5}}$, where σ_{Δ_i} is the standard deviation of Δ_i. In practice, we obtain the sample standard deviation $\hat{\sigma}_{\hat{\Delta}_i}$ at first and divide $\hat{\Delta}_i$ by $\hat{\sigma}_{\hat{\Delta}_i}$. Hence, the standard deviation of $\frac{\hat{\Delta}_i}{\sigma_{\hat{\Delta}_i}}$ is 1. The standardized $\frac{\hat{\Delta}_i}{\sigma_{\hat{\Delta}_i}}$ is the argument of the Gaussian kernel function, $k(\frac{\hat{\Delta}_i}{h}) = k(\frac{\hat{\Delta}_i/\sigma_{\hat{\Delta}_i}}{h/\sigma_{\hat{\Delta}_i}}) = k(\frac{\hat{\Delta}_i/\sigma_{\hat{\Delta}_i}}{h'})$. So the bandwidth can be simplified to $h' = 0.9n^{-\frac{1}{5}}$.

Sign of test

Under the alternative hypothesis, the dependence between binary outcomes d_{itj} is not equal to 0, $\text{corr}(d_{itj}, d_{itj'}) \neq 0, \forall j, j' = 1, ...K, j \neq j'$. The dependence can be measured by the dispersion degree η in the variance function,

$$\text{Var}(Y_i) = K\Lambda_{it}(1 - \Lambda_{it})(1 + \eta) \tag{23}$$

For the beta-binomial model, η is equal to $\eta = \frac{K-1}{\phi+1} > 0$. If $\eta < 0$, Y_i is under-dispersed relative to the binomial model. When $\eta = 0$, Y_i has a the binomial distribution. Overdispersion describes positive correlations at off-diagonal entries in the variance matrix of $d_{it}, \forall i, t$, while underdispersion is generated from negative correlations. If y_{it} does not follow H_0, the test statistic will be statistically significant with positive sign in overdispersed samples and with negative sign for underdispersed ones. Underdispersed data are hard to simulate by usual methods, but prevalent in empirical data. One example of an underdispersion DGP can be found in Ahn and Chen (1995). Here, we focus on over-dispersed data as are obtained from the beta-binomial distribution.

3.3 Case III: conditional moment test for multiple periods

The test can be extended to multiple time periods. With $T = 2$, there is a single moment condition for $E(z_i|\Delta_i = 0)$ that can be tested. For $T > 2$, one possibility is to combine $T - 1$ such moment conditions into a single test statistic.

In the discrete case, let $g_{i,t} = \mathbb{1}(x_{i,t} = x_{i,t+1})z_{i,t}$,

$$g_{i,t} = \mathbb{1}(x_{i,t} = x_{i,t+1})\frac{(Y_{i,t} - Y_{i,t+1})^2 - K(M_{i,t} - M_{i,t+1})^2}{K(K-1)}, \quad t = 1, ..., T-1$$

As we derived before, $E(g_{i,t}) = 0$

In the continuous case, moment conditions are

$$g_{i,t} = \frac{K(\frac{\hat{\Delta}_{i,t}}{h})(z_{i,t} - \tau(\hat{\Delta}_{i,t}))}{\sum_{i=1}^{N} \frac{1}{n} K(\frac{\hat{\Delta}_{i,t}}{h})}, \quad t = 1, ..., T-1$$

where $\hat{\Delta}_{it} = (x_{it} - x_{i,t+1})\hat{\beta}$ and $\tau(\hat{\Delta}_{i,t}) = (y_{i,t+1}(1 - y_{i,t+1})\hat{\Delta}_{it})^2$. Under the null hypothesis, $\mathrm{E}(g_{i,t}) = 0$.

These moment conditions can be written in matrix form for individual $i = 1, ..., N$ as :

$$g_i = \begin{pmatrix} g_{i,1} \\ \cdots \\ g_{i,T-1} \end{pmatrix}, \quad \text{and the sample mean is } \bar{g}_n = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} g_{i,1} \\ \cdots \\ \frac{1}{n}\sum_{i=1}^{n} g_{i,T-1} \end{pmatrix}.$$

Let $\hat{S}$ denote the sample variance:

$$\hat{S} = \frac{1}{n}\sum_{i=1}^{n} g_i g_i' - \bar{g}_n \bar{g}_n',$$

Since $\bar{g}_n \xrightarrow{p} \mathrm{E}(g_i) = 0$, a test statistic is given by

$$J = n \cdot \bar{g}_n' \hat{S}^{-1} \bar{g}_n = (\sqrt{n} \cdot \bar{g}_n)' \hat{S}^{-1} (\sqrt{n} \cdot \bar{g}_n)$$

Theory: $\sqrt{n} \cdot \bar{g}_n \xrightarrow{d} N(0, S)$ and $\hat{S} \xrightarrow{p} S$, then $J \xrightarrow{d} \chi^2_{T-1}$.

This chi-square test rejects the binomial distribution assumption at the α% significance level if $J \geq \chi^2_\alpha(T-1)$.

3.4 Simulation study

We conduct simulation experiments to examine the performance of these tests under two scenarios. In the first setting, explanatory variables are discrete (in fact, there is a single binary regressor, to keep things as simple as possible), while the explanatory variable is continuous in the second. The remaining aspects of the DGP regarding fixed effects, expectation functions and parameters setting are kept the same as those in section 2.3.

− − − − − − − − − Table 3 about here − − − − − − − − −

Table 3 presents rejection rates, the probability that our test rejects the binomial assumption, over 1000 replications for a binary x_{it}. x_{it} is either 0 or 1 with equal probability of 50%. In this case, the probability of $x_{i1} = x_{i2}$ is also 50%, which means that on average half of the observations will be in the set A of individuals with the same expectations over time and thus informative for computing the test statistic. As before, simulations are conducted for $T = 2, 5, 10$ and for $K = 2, 5, 10$.

The first row of each sub-panel shows results for 0% overdispersion, i.e. for sampling from a binomial DGP applies. In this case, the rejection rates are equivalent to the proportion of type-I errors and ideally should be close to the nominal size of the test, in this case 5%. We find that this is mostly the case. The smallest empirical rejection rate is 3.7 percent, the largest one is 9%. If anything, there is overall a slight tendency to overreject in these small samples, more so in the multivariate ($T > 2$) version of the test than in the scalar version ($T = 2$).

In the current set-up, where we test the validity of a specific model assumption, we would be more worried about under- than overrejection. Also, the power properties are very important in this context. The lower part of each subpanels shows the rejection rates for misspecified binomial models, and hence the power. Reassuringly, we find that the test has some power already against the alternative of rather modest overdispersion (10%), in particular for $N = 500$ and $T = 10$, where close to 30% of wrong null hypotheses are rejected. As the dispersion degree increases, the power of the test also grows, and it reaches 100% for DGPs where overdispersion, the number of observations and the number of time periods are large.

− − − − − − − − − Table 4 about here − − − − − − − − −

In Table 4, we show the results for the kernel weighted test statistics appropriate for continuous variables. x_{it} is drawn from a uniform distribution between -1 and 1, with mean 0 and variance 1/3. The general patterns regarding type-I errors and power of the tests are mostly similar to those of Table 3. As in Table 3, the power of the test tends to decrease in K, for a given overall degree of overdispersion, but this tendency is more uniform in the continuous version of the test. It indicates that the power of the test reacts differently to the two parameters driving overdispersion, and in particular that it is more sensitive to increases in ϕ rather than K.

The combined results from our simulation experiments are re-assuring: on one hand, modest amounts of overdispersion cause only minor bias of the BLogit CML estimator; on the other hand, the test we derive has good power properties against medium or high-dispersion alternatives to the binomial assumption.

4 Application to labor supply

In this illustrative application, we consider the interplay between fertility and female labor supply in the context of Switzerland, using data from the Swiss Household Panel (SHP) for the years 2012-2016. The SHP is an ongoing longitudinal survey of households and people living in Switzerland that covers a large range of topics on living conditions, both objective and subjective, including work, fertility and health. We restrict the analysis to women aged 25-45, who participated in the survey at least twice during the five-year period. This gives us a sample of 5,854 person-year observations for 1,712 different women.

There is a long literature on female labor supply (see e.g. Mroz, 1987) that has mostly focussed on the binary participation decision, i.e. the extensive margin, or on annual hours of work. In the Swiss context, it is more natural to model the work-time percentage, which is a number between 0 and 100%. These work-time percentages are written into contracts and also advertised in job vacancies. For instance, 60 percent work-time means that the worker works the equivalent of 3 days per week and also is paid only 60% of a full-time salary. In practice, the large majority of agreed-upon work-time percentages are multiples of 10%. Figure 1 shows the distribution of work-time percentages in our sample. The relative frequency of zeros is 14.4%, meaning that the estimated participation rate in our sample for this age group is 85.6%, a number very close to the offical numbers published by the Federal Statical Office (BfS, 2016).

The graph reveals one frequently noted "puzzle" of female labor supply in Switzerland: while the participation rate of women is one of the highest among OECD countries, there is a large prevalence of part-time work, varying from very small amounts to more substantial ones, but the proper full-time rate (say 100% or 90%) is actually not that high (here 21%). As a consequence quite substantial male-female gaps emerge over the life-cycle regarding earnings, eligiblity for retirement benefits and career development more generally. Clearly, the move to part time work for women with children provides a main explanation for this pattern. The Box-plots in Figure 2 show, for our data, how the median work-time percentage drops from 80 percent or higher for those aged 30 or below to 50 percent for women in their early 40s. Figure 3 documents how the probability of having at least one child increases with age. Around 10 percent of Swiss women remain childless.

Table 5 provides some descriptive statistics (means and standard deviations) for both the dependent and the explanatory variables used in the following empirical analysis. The average work-time percentage is 55%, with a standard deviation of 0.34. We have re-coded the work-time percentage as a strict multiple of 0.1, by moving the few intermittent values to the decile below. Hence, we can treat 10-times the work-time percentage as a binomial variable with outcomes $0, 1, \ldots, 10$. Under the binomial assumption, the standard deviation for a fraction with a mean of 0.56 is equal to $\sqrt{0.56(1-0.56)/10} = 0.157$, substantially below the observed standard deviation of 0.346. Hence, there is evidence of over-dispersion at the marginal level. However, this does not necessarily invalidate the key assumption of the binomial distribution *conditional* on covariates and individual specific fixed effects.

– – – – – – – – – – Table 5 about here – – – – – – – – – –

Women have an average age of 36.3 years and 63.1 percent report having at least one child in the year they are surveyed. For 58.4 percent of person-year observations, there is a partner present in the household. The health status is captured by a 5-point scale for self assessed health, where 0 means "not well at all" and 4 means "very well". We treat it as a cardinal scale for simplicity, and also abstract from its potential endogeneity to working or having children. Finally, we include information on language region. There is quite a bit of evidence that work-norms differ between the French and the German speaking part of Switzerland, with some stigma attached to working mothers, in particular for the first years of the child and full-time work, in the German-speaking part of Switzerland (65% or our sample) but much less so in the French-speaking part (29% of our sample) (Steinhauer, 2018).

Our final estimation model includes four year effects, age-squared (the linear age term is dropped; alternatively, one could identify the linear age effect by setting a second year effect equal to zero), indicators for the presence of a child and partner, and the health variable. Since language region is mostly constant over time, it is near-collinear with the fixed effects when applying the BLogit CML or Blogit DV estimators, and we therefore only include its interaction with the child-indicator variable.

Results are given in Table 6. The first column shows the estimated coefficients from the BLogit CML and the second that from the BLogit DV model. The last two columns add corresponding (binary) logit models for the extensive margin model (work yes/no), again using alternatively the CML or DV estimators. Standard errors are clustered at the individual level throughout.

As is the case for the binary logit model with fixed effects, DV estimation of the binomial model is subject to the perfect prediction problem (see e.g. Kunz, Staub and Winkelmann, 2018). Outcomes for women, whose work-time percentage is either zero or one in each year are perfectly predicted, meaning that the associated dummy coefficient will tend to minus or plus infinity, respectively. For the Blogit CML, perfectly prediction formally does not arise as the α_i's are not estimated. However, all such observations have mechanically a log-conditional likelihood contribution of zero and thus do not contribute to estimation of β as well. To use the same estimation sample everywhere, we drop all perfectly predicted outcomes, leading to a final sample size of 4,661 person-year observations for the work-time percentage model.

− − − − − − − − − − Table 6 about here − − − − − − − − − −

When interpreting magnitudes, we note the recent suggestion by Kemp and Santos Silva (2016) to focus on expected (semi-) elasticities. These can be estimated without knowledge of α_i and are thus very suitable for our conditional maximum likelihood approach. For the binomial proportion model with $\mathrm{E}(y_{it}|x_{it}, \alpha_i) = \Lambda_{it}$, we get

$$\partial \log E(y_{it}|x_{it}, \alpha_i)/\partial x_{it} = \beta(1 - \Lambda_{it})$$

A good estimator of the overall mean of Λ_{it} is the sample mean of the outcome, $\bar{\Lambda} = \bar{y} = 0.55$, so that the CML estimators $\hat{\beta}$ can be multiplied by 0.45 to obtain average semi elasticities with respect to changes in the associated covariate. From columns (1) and (2) of Table 6, we find a large negative association between having a child and the amount of work. The point estimate of the main effect is about -2, which means that not having a child increases the expected work-time percentage by about 90 percent, i.e., by a factor of close to two. This effect is highly statistically significant, as are two of the three interaction effects: having a child reduces the work-time percentage more if a partner is present than otherwise, suggesting the presence of pecuniary motives for work, or needing to "make ends meet". The labor supply response of women to having children is about half as large for French speaking women as it is for German speakers, corroborating the social norm results found in the earlier literature.

In this application, the BLogit CML and the Blogit DV results are very similar. The DV results are always a bit larger in absolute value, but the difference never exceeds five percent. This resonates with our simulation results, because both T and K are both relatively large. Nevertheless, and perhaps surprisingly, the joint test for the binomial assumption derived in section 3.3. indicates a clear rejection (test value of 37.7 with a χ^2 5% critical value of 9.5). However, we know from the simulation results (Tables 1 and 2) that even with 50% overdispersion, the bias of the Blogit CML is very small for $K = 10$ and $T = 5$, a setting similar to the current application. At the same time, the probability of rejecting the wrong H0 is very close to 1 (see Table 4). On a practical note, the CML estimator can be computed much faster than the DV estimator, by a factor of about 10 in our case. Of course, this problem would be exacerbated in applications with more cross-sectional units, to the point where computation of the Blogit DV estimator may become unfeasible.

In the last two columns of Table 6, we allow for a comparison with results from a more conventional binary logit extensive margin estimator. A first point to note is that the sample becomes much smaller, since all observations with variation in the positive range only, i.e., percentages between 10% and 100%, are now coded as "1" and thus become perfectly predicted. Their variation does not contribute to estimation, the usable sample size drops by 3/4, and the standard errors of the estimated coefficients increase accordingly. We had to drop the Italian times children interaction, as it could not be estimated in the reduced sample.

Second, we note that the estimated coefficients tend to be substantially larger. This needs to be relativized, though, as the implied expected semi-elasticities for the probability of work are obtained using a much smaller factor, as $(1 - \bar{y})$ is equal to the non-participation rate of 0.145 in this case. In terms of statistical significance, we find that the health and partner coefficients were not statistically in the work-time percentage model, but they are in the participation model. In terms of point estimate, the interaction between French and children just offsets the main effect of having at least one child, meaning that there is no difference in participation probabilities for

French speaking mothers and non-mothers, although some labor supply responsiveness was found in the work-time percentage model for the combined extensive and intensive margin effect. In terms of estimation method, and in contrast to the work-time percentage model, the participation model suffers from a massive incidental parameters bias, since the point estimates for the DV estimator exceed those of the CML estimator by fifty percent on average.

5 Concluding remarks

Machado (2004) introduce the fixed effects binomial model as a method for proportions or discrete bounded outcomes more generally. However, there are no existing results showing that the conditional binomial logit maximum likelihood estimator is robust to misspecification. In this paper, we focus on the consequences of overdispersion as it originates for instance from neglected unobserved heterogeneity. We show in simulations experiments that the Blogit CML estimator maintains a rather good performance even if the binomial model is misspecified as long as one of three conditions is met: either is the degree of overdispersion modest, or else, the length of the panel T or the number of Bernoulli trials K must be large.

We then derive a test of the null hypothesis that the binomial assumption is valid, against the alternative hypothesis of overdispersion. The test computes the variance of within-individual outcome differences. For the subset of observations with regressors that do not change over time, the mean difference is zero (or close to zero if regressors do not differ "too much") and it is possible to compute variances with and without the binomial assumption that do not depend on the fixed effects. This is essential as fixed effects are not estimated by the Blogit CML estimator. Our simulation experiments show that the test has good power properties under the alternative of medium or large degrees of overdispersion. But these are exactly the case, where the bias of the Blogit CML estimator becomes noticeable.

Proportions data are ubiquitous in empirical economic research. We study in our empirical application an outcome related to women's work behavior, namely the contracted work-time percentage. In our sample of mid-aged women obtained from the Swiss Household Panel, 65% of all women report working part-time, i.e. a percentage between 10 and 90%. The empirical analysis using the fixed effects binomial logit estimator estimates substantially different work-time percentages for mothers and non-mothers. Having a partner makes the effect more pronounced, whereas speaking French reduces it. We show how these coefficients can be interpreted in terms of expected semi-elasticities even if the fixed effects are not estimated. In comparison to the fixed-effects logit estimation for the participation model, much fewer observations are lost in the work-time percentage model due to perfect prediction, contributing to its overall much more precise estimation of the model parameters .

In future work, we will consider alternative estimators that could be pursued if the binomial null hypothesis is rejected. If the logit conditional expectation function is to be kept, a binomial logit correlated random effects model is a possible approach. Such a model would explicitly account for overdispersion, by assuming for instance that unobserved heterogeneity follows a normal distribution with mean depending on the regressors.

6 References

Abrevaya, J. (1997) The equivalence of two estimators of the fixed-effects logit model, *Economics Letters* 55(1), 4143.

Ahn, H., and J.J. Chen (1995) Generation of over-dispersed and under-dispersed binomial variates, *Journal of Computational and Graphical Statistics*, 4(1), 55-64.

Berkson, J. (1944) Application of the logistic function to bio-assay, *Journal of the American Statistical Association* 39, 357-365.

BfS Bundesamt für Statistik (2016) Arbeitsmarktindikatoren der Schweiz, Neuchâtel, Switzerland.

Chamberlain, G. (1980) Analysis of covariance with qualitative data, *Review of Economic Studies*, 47, 225-238.

Dean, C.B. (1992) Testing for overdispersion in Poisson and binomial regression models, *Journal of the American Statistical Association*, 87(418), 451-457.

Gourieroux, C., A. Monfort and A. Trognon (1984) Pseudo maximum likelihood methods: Theory, *Econometrica: Journal of the Econometric Society*, 681-700.

Kemp, G. and J. Santos Silva (2016) Partial effects in fixed-effects models, United Kingdom Stata Users' Group Meetings 2016.

Kunz, J., K. Staub and R. Winkelmann (2018) Predicting fixed effects in panel probit models, York

Health, Econometrics and Data Group Working Paper #18/23.

Machado, M.P. (2004) A consistent estimator for the binomial distribution in the presence of "incidental parameters": an application to patent data, *Journal of Econometrics*, 119(1): 73-98.

Maddala, G.S. (1983) *Limited-dependent and qualitative variables in econometrics*, Cambridge university press.

Mundlak, Y. (1978) On the Pooling of Time Series and Cross Section Data, *Econometrica*, 46 (1), 69-85.

Mroz, T.A. (1987) The Sensitiviy of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, *Econometrica*, 55, 765-799.

Pagan, A. and A. Ullah (1999) *Nonparametric econometrics*, Cambridge University Press.

Papke, L.E. and J.M. Wooldridge (1996) Econometric methods for fractional response variables with an application to 401(k) plan participation rates, *Journal of Applied Econometrics* 11, 619-632

Papke, L.E. and J.M. Wooldridge (2008) Panel data methods for fractional response variables with an application to test pass rates, *Journal of Econometrics* 145, 121-133.

Ramalho, E.A., J.J.S Ramalho and L.M.S. Coelho (2016) Exponential regression of fractional-response fixed-effects models with an application to firm capital structure, *Journal of Econometric Methods*, 7(1).

Steinhauer, A. (2018) Working Moms, Childlessness, And Female Identity, CEPR Discussion Paper Nr. 12929.

Wooldridge, J.M. (2002) Econometric Analysis of Cross Section and Panel Data, MIT Press.

Table 1: Simulation Results N=100

N=100	overdispersion	T=2			T=5			T=10		
		Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit
K=2	0%	2.049	2.880	2.242	2.003	2.280	1.986	2.006	2.134	1.877
		(0.419)	(0.621)	(0.255)	(0.178)	(0.211)	(0.145)	(0.118)	(0.128)	(0.101)
	10%	2.100	2.968	2.225	2.036	2.320	1.988	2.010	2.138	1.876
		(0.419)	(0.632)	(0.263)	(0.192)	(0.227)	(0.156)	(0.121)	(0.131)	(0.105)
	50%	2.516	3.674	1.898	2.179	2.484	1.908	2.073	2.204	1.937
		(0.822)	(1.283)	(0.399)	(0.360)	(0.429)	(0.279)	(0.246)	(0.266)	(0.222)
	overdispersion	T=2			T=5			T=10		
		Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit
K=5	0%	2.012	2.279	2.243	2.000	2.101	1.990	2.002	2.051	1.872
		(0.233)	(0.272)	(0.190)	(0.111)	(0.118)	(0.104)	(0.073)	(0.075)	(0.076)
	10%	2.033	2.303	2.239	2.013	2.115	1.990	2.010	2.059	1.875
		(0.252)	(0.295)	(0.184)	(0.120)	(0.128)	(0.115)	(0.079)	(0.081)	(0.078)
	50%	2.139	2.431	2.205	2.054	2.159	1.971	2.032	2.081	1.882
		(0.300)	(0.354)	(0.214)	(0.139)	(0.149)	(0.122)	(0.086)	(0.088)	(0.083)
	100%	2.320	2.657	2.140	2.111	2.219	1.958	2.052	2.102	1.898
		(0.402)	(0.485)	(0.253)	(0.176)	(0.188)	(0.147)	(0.116)	(0.119)	(0.104)
	overdispersion	T=2			T=5			T=10		
		Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit
K=10	0%	2.005	2.129	2.241	2.001	2.050	1.989	2.000	2.025	1.871
		(0.157)	(0.169)	(0.153)	(0.078)	(0.081)	(0.091)	(0.052)	(0.052)	(0.063)
	10%	2.017	2.142	2.237	2.004	2.053	1.986	2.002	2.026	1.868
		(0.174)	(0.187)	(0.158)	(0.082)	(0.084)	(0.091)	(0.055)	(0.055)	(0.063)
	50%	2.065	2.195	2.241	2.026	2.076	1.984	2.012	2.036	1.869
		(0.202)	(0.219)	(0.175)	(0.098)	(0.101)	(0.100)	(0.064)	(0.065)	(0.072)
	100%	2.139	2.275	2.214	2.051	2.102	1.979	2.025	2.049	1.876
		(0.239)	(0.260)	(0.183)	(0.109)	(0.113)	(0.106)	(0.074)	(0.075)	(0.076)

Results for 1000 Monte Carlo replications; Standard deviations in parentheses. In each period, the number of observation is 100. $x_{it} \sim U[-1,1]$. $\beta_1 = 2$ and $\alpha_i = \sqrt{T}\bar{x}_i + N(0,1)$. Overdispersion factor $\frac{k-1}{\phi+1}$ represents dispersion degree of variance. Overdispersion degree 0% is generated by binomial distribution and postive dispersion degree is generated by a beta-binomial DGP.

Table 2: Simulation Results N=500

N=500	overdispersion	T=2			T=5			T=10		
		Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit
K=2	0%	2.013	2.826	2.234	2.002	2.278	1.987	2.002	2.130	1.870
		(0.170)	(0.254)	(0.112)	(0.082)	(0.097)	(0.068)	(0.053)	(0.057)	(0.046)
	10%	2.078	2.936	2.214	2.026	2.307	1.975	2.011	2.139	1.870
		(0.195)	(0.293)	(0.120)	(0.084)	(0.099)	(0.070)	(0.058)	(0.062)	(0.050)
	50%	2.361	3.424	1.849	2.142	2.438	1.881	2.062	2.191	1.926
		(0.311)	(0.489)	(0.175)	(0.151)	(0.180)	(0.120)	(0.102)	(0.110)	(0.092)
	overdispersion	T=2			T=5			T=10		
		Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit
K=5	0%	2.003	2.268	2.232	2.001	2.102	1.986	2.000	2.049	1.867
		(0.102)	(0.119)	(0.078)	(0.051)	(0.054)	(0.050)	(0.033)	(0.034)	(0.033)
	10%	2.032	2.302	2.237	2.007	2.109	1.984	2.006	2.055	1.868
		(0.108)	(0.126)	(0.088)	(0.052)	(0.055)	(0.051)	(0.035)	(0.036)	(0.034)
	50%	2.143	2.435	2.204	2.052	2.157	1.973	2.023	2.073	1.871
		(0.132)	(0.156)	(0.090)	(0.063)	(0.068)	(0.055)	(0.041)	(0.043)	(0.040)
	100%	2.289	2.617	2.121	2.103	2.211	1.951	2.050	2.100	1.895
		(0.168)	(0.203)	(0.107)	(0.080)	(0.085)	(0.063)	(0.049)	(0.051)	(0.046)
	overdispersion	T=2			T=5			T=10		
		Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit	Blogit CML	Blogit DV	Pooled logit
K=10	0%	1.997	2.121	2.235	1.999	2.049	1.984	2.000	2.024	1.867
		(0.071)	(0.076)	(0.067)	(0.034)	(0.035)	(0.042)	(0.023)	(0.023)	(0.029)
	10%	2.014	2.140	2.233	2.003	2.053	1.986	2.002	2.026	1.869
		(0.075)	(0.081)	(0.073)	(0.038)	(0.039)	(0.043)	(0.025)	(0.025)	(0.031)
	50%	2.066	2.195	2.228	2.023	2.073	1.980	2.013	2.037	1.868
		(0.092)	(0.099)	(0.076)	(0.043)	(0.044)	(0.044)	(0.029)	(0.029)	(0.033)
	100%	2.134	2.269	2.206	2.050	2.100	1.974	2.024	2.048	1.871
		(0.108)	(0.117)	(0.083)	(0.050)	(0.051)	(0.048)	(0.035)	(0.035)	(0.035)

Notes: see Table 1

Table 3: Simulation Results for Rejection Rates when x is discrete

N=100	overdispersion	T=2	T=5	T=10	N=500	overdispersion	T=2	T=5	T=10
K=2	0%	0.059	0.058	0.09	K=2	0%	0.053	0.037	0.056
	10%	0.071	0.097	0.133		10%	0.155	0.236	0.293
	50%	0.483	0.408	0.325		50%	0.991	0.937	0.859
K=5	0%	0.054	0.073	0.075	K=5	0%	0.05	0.063	0.049
	10%	0.061	0.06	0.099		10%	0.094	0.128	0.148
	50%	0.218	0.343	0.449		50%	0.866	0.996	1
	100%	0.54	0.735	0.849		100%	1	1	1
K=10	0%	0.049	0.063	0.078	K=10	0%	0.047	0.048	0.061
	10%	0.049	0.060	0.086		10%	0.084	0.108	0.122
	50%	0.181	0.315	0.445		50%	0.809	0.987	1
	100%	0.468	0.815	0.907		100%	0.999	1	1

Notes: x_{it} is binary variable with 50% probability equal to 0 or 1. The rest DGP is same as in Table 1; The null hypothesis is that of binomial dispersion.

Table 4: Simulation Results for Rejection Rates when x is continuous

N=100	overdispersion	T=2	T=5	T=10	N=500	overdispersion	T=2	T=5	T=10
K=2	0%	0.053	0.063	0.097	K=2	0%	0.062	0.047	0.056
	10%	0.07	0.095	0.185		10%	0.138	0.223	0.394
	50%	0.424	0.684	0.899		50%	0.948	1	1
K=5	0%	0.05	0.053	0.078	K=5	0%	0.042	0.05	0.057
	10%	0.044	0.063	0.110		10%	0.083	0.105	0.190
	50%	0.239	0.512	0.805		50%	0.772	0.992	1
	100%	0.608	0.958	0.999		100%	0.996	1	1
K=10	0%	0.041	0.062	0.091	K=10	0%	0.052	0.056	0.043
	10%	0.041	0.057	0.086		10%	0.073	0.094	0.179
	50%	0.190	0.402	0.673		50%	0.677	0.988	1
	100%	0.482	0.908	0.997		100%	0.991	1	1

Note: see Table 1 DGP.

Table 5. Descriptive statistics ($NT = 5,854$)

	mean	std. dev.
Work-time percentage	0.557	0.346
Age	36.30	6.01
Children (yes=1)	0.631	0.482
Partner (yes=1)	0.584	0.492
Self-rated health	3.114	0.610
Years of schooling	14.43	3.09
French speaking (yes=1)	0.293	0.455
Italian speaking (yes=1)	0.043	0.204

Source: Swiss Household Panel 2012-2016, own calculations.

Table 6. Determinants of female labor supply (SHP 2012-2016)

	Work-time Percentage		Work (yes/no)	
	Blogit CML	Blogit DV	Logit CML	Logit DV
Age squared	-0.001	-0.001	0.005	0.007
	(0.002)	(0.002)	(0.005)	(0.008)
Self-rated health	0.071	0.073	0.448	0.642
	(0.038)	(0.039)	(0.153)	(0.220)
Partner (yes=1)	0.322	0.333	1.371	1.927
	(0.253)	(0.260)	(0.658)	(0.952)
Children (yes=1)	-2.097	-2.160	-1.975	-2.771
	(0.276)	(0.286)	(0.956)	(1.381)
Children × Partner	-0.824	-0.848	-2.216	-3.188
	(0.260)	(0.268)	(1.056)	(1.510)
Children × French	1.152	1.194	1.876	2.777
	(0.405)	(0.418)	(1.089)	(1.619)
Children × Italian	-0.247	-0.266		
	(0.720)	(0.754)		
Year 2013	0.143	0.147	-0.018	-0.018
	(0.147)	(0.151)	(0.465)	(0.647)
Year 2014	0.246	0.253	-0.147	-0.265
	(0.278)	(0.285)	(0.820)	(1.143)
Year 2015	0.338	0.347	-0.247	-0.408
	(0.412)	(0.423)	(1.205)	(1.677)
Year 2016	0.387	0.397	-0.449	-0.717
	(0.545)	(0.560)	(1.577)	(2.204)
Number of person-years	4,661	4,661	1,071	1,071
Number of persons	1,334	1,334	295	295
Log pseudolikelihood	-23,183.6	-1,838.3	-358.8	-595.9
Fixed effects	yes	yes	yes	yes

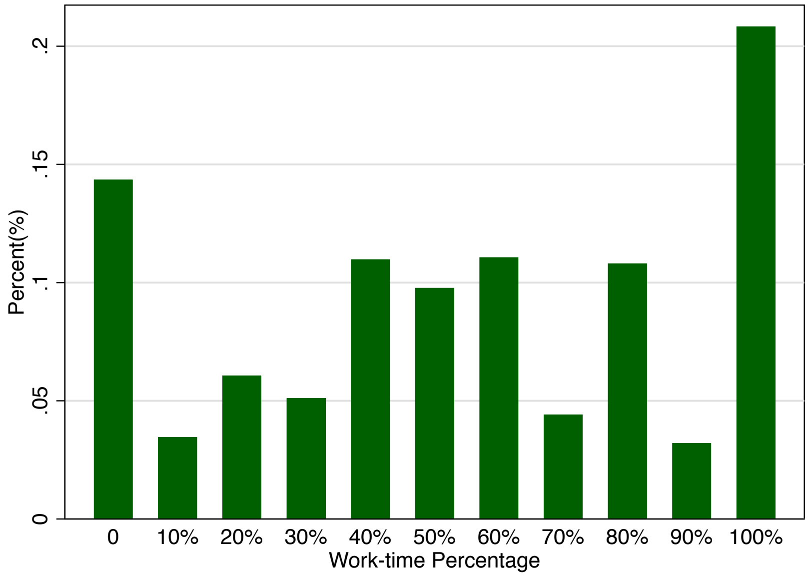

Figure 1: Distribution of Work-time percentage in 2016

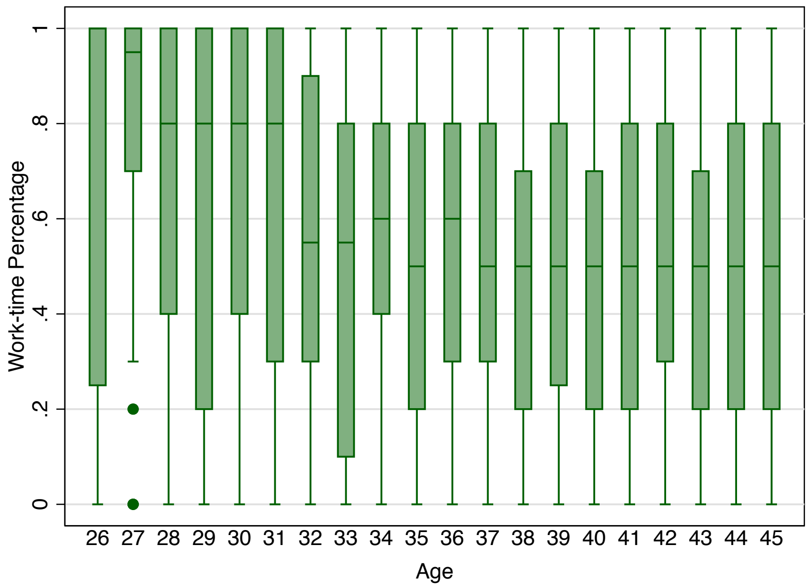

Figure 2: work-time percentage by age in 2016

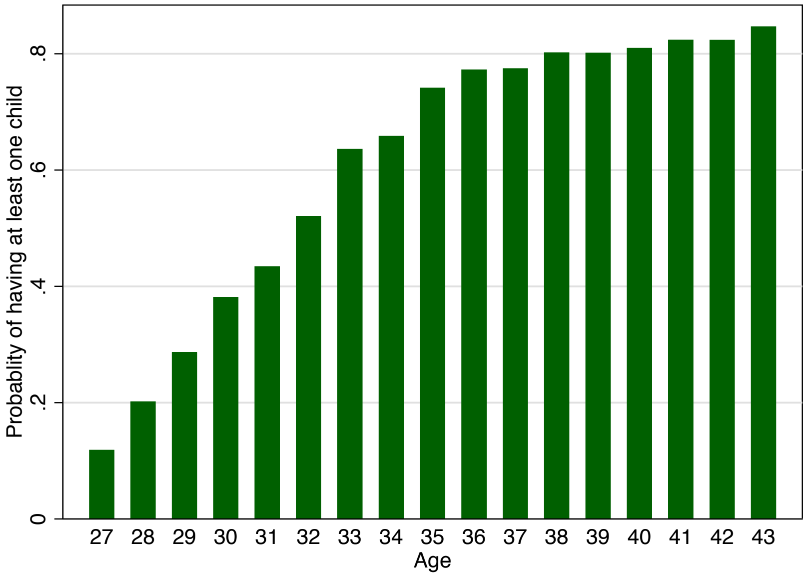

Figure 3: Probability of having at least one child by age