# Multiple Fixed Effects in Nonlinear Panel Data Models

## Theory and Evidence

Karyne B. Charbonneau

Princeton University

November 2012

**Abstract**

This paper considers the adaptability of nonlinear panel data models to multiple fixed effects. It is motivated by the gravity equation used in international trade, where influential papers such as Santos Silva and Tenreyro (2006) use nonlinear models with fixed effects for both importing and exporting countries. It is also relevant for other areas of microeconomics such as labor economics, where a wage equation might contain both worker and firm fixed effects, or industrial organization, where knowledge diffusion equations using patent data can include citing and cited country fixed effects. Econometric theory has mostly focused on the estimation of single fixed effects models. This paper investigates whether existing methods can be modified to eliminate multiple fixed effects for some specific models in which the incidental parameter problem has already been solved in the presence of a single fixed effect. We find that it is possible to generalize the conditional maximum likelihood approach of Rasch (1960, 1961) to include two fixed effects for the logit, the Poisson and the Negative binomial regression models considered by Hausman, Hall and Griliches (1984) as well as for the Gamma model. Surprisingly, Manski's (1987) maximum score estimator for binary response models cannot be adapted to the presence of two fixed effects. We also look at a multiplicative form model. In that case, it is possible to consistently estimate the parameters when there are two fixed effects with the use of a moment condition. Monte Carlo simulations show that the conditional logit estimator presented in this paper is less biased than other logit estimators without sacrificing on the precision. This superiority is emphasized in small samples. An application on trade data using both the logit and Poisson estimators further highlights the importance of properly accounting for two fixed effects. Indeed, estimates of the gravity model parameters produced by the method presented in this paper differ significantly from those obtained with the various estimators used in the trade literature.

# 1 Introduction

The so-called fixed effects have long been recognized as a key element of econometric modeling of panel data, and a significant literature now exists in econometric theory on the inclusion of fixed effects in both linear and nonlinear panel data models. The developed methods have also been put to use in various empirical studies. Many of these now include more than one fixed effect. However, econometric theory has mostly focused on single fixed effects. The present paper attempts to bridge part of this gap by looking at some specific nonlinear models. The empirical relevance is demonstrated using Monte Carlo simulations and an application to international trade data.

This paper is motivated by the fixed effects gravity equation models used in international trade.This specific area of economics is concerned with the estimation of the factors conducive to trade between countries. The importance of using fixed effects to control for country-specific characteristics has been emphasized in an influential paper by Anderson and VanWincoop (2003). Following this paper, many papers contributing to the gravity equation literature have included fixed effects in the estimation strategies. For example, Helpman, Melitz and Rubinstein (2008) and Santos Silva and Tenreyro (2006), developed nonlinear panel data models with fixed effects for both importing and exporting countries.

This paper investigates whether existing methods for eliminating fixed effects can be modified to eliminate multiple fixed effects. This is not only relevant for data consisting of country-pair, but also in a number of other areas of empirical micro-economics. For example, in a very influential paper, Abowd, Kramarz and Margolis (1999) used matched firm-employee data to study wage determinants of french workers. For such data sets one might want to allow for both firm and worker fixed effects. In a similar fashion, Aaronson, Barrow and Sander (2007) and Rivkin, Hanushek and Kain (2005) used matched data between students and teachers to study academic achievement. In both those papers multiple fixed effects were also used.

When studying "static" linear models, fixed effects do not generally cause any problem, since they can easily be differenced out to allow consistent estimation of the relevant parameters. However, when considering nonlinear panel data models, we encounter the well known incidental parameter problem identified by Neyman and Scott (1948).[1] This has motivated a rich literature on

[1]The incidental parameter problem refers to the fact that in nonlinear models with a fixed number of observations for each individual, the bias in the estimation of the fixed effects contaminates the estimates of the parameters of interest.

the estimation of single fixed effects nonlinear panel data models. The first model considered in the literature is the logit model studied in Rasch (1960,1961). Manski (1987) generalized this to develop a conditional maximum score estimator for binary response models that remains consistent under weak assumptions on the distribution of the errors. Hausman, Hall and Griliches (1984) used the relationship between the Poisson and multinomial distribution to solve the incidental parameter problem in the Poisson regression model (and Negative Binomial) in the presence of a single fixed effect. Like in the logit case, this results in a conditional likelihood approach that can be used to consistently estimate the parameters of interest.

With a more general approach to the problem, Hahn and Newey (2004) show that when $n$ and $T$ grow at the same rate, the fixed effects estimator is asymptotically biased and the asymptotic confidence intervals are wrong. They suggest two bias correction methods (the panel jackknife and the analytic bias correction).

The paper will proceed as follows; for each of the models mentioned above, to which we add Gamma and a more general multiplicative form model, we describe the estimation approach developed in the literature for one fixed effect and then try to generalize it to two. When it is not possible, we will give some intuition as to why the "usual" trick does not work. As a general rule of thumb, whenever we needed a pair of observations for the same individual to eliminate one fixed effect, we will need two "connected" pairs to get rid of two, in a very similar fashion as the difference-in-differences used in the linear models. We then proceed to showing the relevance of appropriately dealing with two fixed effects in nonlinear models using the estimators presented in this paper. To accomplish that, we first do Monte Carlo simulations for the logit model. Finally, we use data on trade flows between countries to test the logit and the Poisson estimators on the gravity equation.

Given the large number of empirical applications using multiple fixed effects, these methods are of broad applicability. Furthermore, we find that appropriately controlling for multiple fixed effects has substantial effect on the estimated parameters of interest relative to models without fixed effects or models inappropriately controlling for fixed effects.

# 2 Multiple Fixed Effects in Nonlinear Panel Data Models

## 2.1 Fixed Effects in a Logit Model

The first nonlinear model we shall consider is the simple and well documented logit model. There is a well-known application of the conditional maximum likelihood "trick" that allows us to solve the incidental parameter problem in a logit in the presence of one fixed effect. As we will see, it is possible to generalize this method to include two fixed effects. We begin by presenting the original solution, following somewhat closely that of Arellano and Honoré (2001), before moving on to two fixed effects.

For $T = 2$, suppose we have observations generated by:

$$y_{it} = 1\{x_{it}\beta + \alpha_i + \varepsilon_{it} \geq 0\} \quad i = 1, ..., n$$

where for all $i$ and $t$ the $\varepsilon_{it}$ are independent and have a logistic distribution conditional on $x$s and the individual fixed effect $\alpha$. This implies that we can express the following probability:

$$Pr(y_{i1} = 1 \mid x_{i1}, x_{i2}, \alpha_i) = \frac{\exp(x_{i1}\beta + \alpha_i)}{1 + \exp(x_{i1}\beta + \alpha_i)}. \tag{1}$$

It is then easy to show that the conditional likelihood will eliminate the fixed effect such that:

$$\begin{aligned} &Pr(y_{i1} = 1 \mid y_{i1}+y_{i2} = 1, x_{i1}, x_{i2}, \alpha_i) \\ &= \frac{Pr(y_{i1} = 1 \mid x_{i1}, x_{i2}, \alpha_i)Pr(y_{i2} = 0 \mid x_{i1}, x_{i2}, \alpha_i)}{Pr(y_{i1} = 1, y_{i2} = 0 \mid x_{i1}, x_{i2}, \alpha_i) + Pr(y_{i1} = 0, y_{i2} = 1 \mid x_{i1}, x_{i2}, \alpha_i)} \\ &= \frac{\exp(x_{i1} - x_{i2})\beta}{1 + \exp(x_{i1} - x_{i2})\beta}. \end{aligned} \tag{2}$$

We can then find an estimator for the parameter $\beta$ by applying this function to all pairs of observations for a given individual, and for all individuals. This can be generalized to the case where $T > 2$, and is easy enough to calculate. Note that we are conditioning on $y_{i1} + y_{i2} = 1$, which means that we are using the information contained in pairs of observations where the binary indicator changed. This important trick used here to eliminate the fixed effects is also the one used in Manski's (1987) maximum score estimator, which we will analyze in the next subsection. It is possible to get a likelihood function when $T > 2$, by conditioning on $\sum_{t=1}^{T} y_{it}$ to get the conditional distribution:

$$P\big(y_{i1}, \ldots, y_{it} \mid \sum_{t=1}^{T} y_{it}, x_{i1}, \ldots, x_{it}, \alpha_i\big) = \frac{\exp(\sum_{t=1}^{T} y_{it} x_{it} \beta)}{\sum_{(d_1,\ldots,d_t)\in B} \exp(\sum_{t=1}^{T} d_t x_{it} \beta)} \tag{3}$$

with $B$ being the set of all sequences of zeros and ones that have $\sum_{t=1}^{T} d_{it} = \sum_{t=1}^{T} y_{it}$ ( [6]). Note that this implies that $\sum_{t=1}^{T} y_{it}$ is a sufficient statistic for $\alpha_i$. We will see later that this sufficient statistic is also very useful in the Poisson model.

We now show that a similar trick can be applied in the case of two fixed effects in a logit model and provide an analogous result. Suppose the observations are now given by

$$y_{ij} = 1\{x_{ij}\beta + \mu_i + \alpha_j + \varepsilon_{ij} \geq 0\} \quad i = 1, \ldots, n, j = 1, \ldots, n \tag{4}$$

where $\varepsilon_{ij}$ follows a logit distribution.[2] Then, by applying the method used above to eliminate one fixed effect, we can write the following probabilities:[3]

$$Pr(y_{lj} = 1 \mid \mathbf{x}, \mu, \alpha, y_{lj} + y_{lk} = 1) = \frac{\exp\left[(x_{lj} - x_{lk})\beta + \alpha_j - \alpha_k\right]}{1 + \exp\left[(x_{lj} - x_{lk})\beta + \alpha_j - \alpha_k\right]} \tag{5}$$

and

$$Pr(y_{ij} = 1 \mid \mathbf{x}, \mu, \alpha, y_{ij} + y_{ik} = 1) = \frac{\exp\left[(x_{ij} - x_{ik})\beta + \alpha_j - \alpha_k\right]}{1 + \exp\left[(x_{ij} - x_{ik})\beta + \alpha_j - \alpha_k\right]}. \tag{6}$$

As can be seen, the two previous equations no longer depend on the $\mu$ fixed effects. However, they are still expressed in terms of the $\alpha$s. We now try to find a conditional probability that does not depend on the latter. First, we notice that we now have a logit with $(x_{ij} - x_{ik})$ as explanatory variable and $(\alpha_j - \alpha_k)$ as a fixed effect. We can therefore apply the trick a second time; so we compare it to another pair of observations with the same "fixed effect". Using both equations (5) and (6), and defining

$$\mathbf{c} \equiv \{y_{lj} + y_{lk} = 1, y_{ij} + y_{ik} = 1\}$$

[2]Here, and for all models presented, it is easier to imagine countries trading with each other. This can be seen as a panel where $n = T$. Then $y_{ij}$ refers to the observation involving the country of the $i$th row and the country of the $j$th column.

[3]Throughout the paper, $\mathbf{x}$ will refer to the vector of all $x$s.

we can now write the following conditional probability:

$$
\begin{aligned}
&Pr(y_{lj}=1 \mid \mathbf{x},\mu,\alpha, y_{lj}+y_{lk}=1, y_{ij}+y_{ik}=1, y_{ij}+y_{lj}=1) \\
&= \frac{Pr(y_{lj}=1, y_{ij}+y_{lj}=1 \mid \mathbf{x},\mu,\alpha,\mathbf{c})}{Pr(y_{ij}+y_{lj}=1 \mid \mathbf{x},\mu,\alpha,\mathbf{c})} \\
&= \frac{Pr(y_{lj}=1 \mid \mathbf{x},\mu,\alpha,\mathbf{c}) Pr(y_{ij}=0 \mid \mathbf{x},\mu,\alpha,\mathbf{c})}{Pr(y_{lj}=1, y_{ij}=0 \mid \mathbf{x},\mu,\alpha,\mathbf{c}) + Pr(y_{lj}=0, y_{ij}=1 \mid \mathbf{x},\mu,\alpha,\mathbf{c})} \\
&= \frac{\exp\left[(x_{lj}-x_{lk})\beta+\alpha_j-\alpha_k\right]}{\exp\left[(x_{lj}-x_{lk})\beta+\alpha_j-\alpha_k\right]+\exp\left[(x_{ij}-x_{ik})\beta+\alpha_j-\alpha_k\right]} \\
&= \frac{\exp\left[((x_{lj}-x_{lk})-(x_{ij}-x_{ik}))\beta\right]}{1+\exp\left[((x_{lj}-x_{lk})-(x_{ij}-x_{ik}))\beta\right]}. \qquad (7)
\end{aligned}
$$

The probability no longer depends on the fixed effects, hence allowing us to solve the incidental parameter problem in the presence of two fixed effects. Indeed, we could now write a conditional maximum likelihood function or apply the last expression to all quadruples of observations, just like with one fixed effect. The latter being easier to implement, the function to maximize is given by

$$
\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l,k\in Z_{ij}} \log\left(\frac{\exp\left[((x_{lj}-x_{lk})-(x_{ij}-x_{ik}))\beta\right]}{1+\exp\left[((x_{lj}-x_{lk})-(x_{ij}-x_{ik}))\beta\right]}\right). \qquad (8)
$$

where $Z_{ij}$ is the set of all the potential $k$ and $l$ that satisfy $y_{lj}+y_{lk}=1, y_{ij}+y_{ik}=1, y_{ij}+y_{lj}=1$ for the pair $ij$.

In the context of epidemiological studies, Hirji et al. (1987) show that you can use a similar recursive conditioning to eliminate what they call nuisance parameters and speed up computations. These parameters are not fixed effects and do not relate to the incidental parameters problem; they are simply normal covariates (like the $x$ variables in our model) that one needs to control for but for which the effect on the dependent variable is not of interest (for example, the constant). It is therefore interesting to see that such conditioning can also be used on parameters of a different nature, like fixed effects. Moreover, they do not present an economic application and do not compare the performance of their estimator to that of other estimators used in the economics literature, both things that we will look at in the next sections.

To assess the accuracy of this two fixed effects logit estimator and compare it to other logit estimators, we ran some Monte Carlo simulations. The results are presented in the third section.

We now move on to a related model where we achieve a different outcome.

## 2.2 Manski's maximum score estimator

Manski (1987) developed a consistent maximum score estimator for binary response models allowing for individual fixed or random effects in panel data. This estimator, unlike its predecessors[4] remains strongly consistent under very weak assumptions on the disturbances. We want to investigate the possibility of generalizing this estimator to the case where we have multiple fixed effects. The conditional maximum score estimator is similar in fashion to the estimator of the logit model. Indeed, it is also applied to a binary response model and uses pairs of observations for the same individual where the value of the indicator variable differs. However, unlike the logit conditional maximum likelihood, this estimator does not generalize to the case with two fixed effects, even under a stronger set of assumptions.

In Manski's original paper, the model has the form:

$$P(y_{it} = 1 \mid x_{i1}, x_{i2}, \alpha_i) = F_i(x_{it}\beta + \alpha_i) \quad t = 1, 2$$

That the distribution $F$ depends on $i$ is the first assumption in Manski (1987). It requires the disturbance to be stationary conditional on the identity of the panel member but does not restrict it to be the same across individuals.

Manski's key result resides in his first lemma:

**Lemma M 1.**

$$\begin{aligned}
x_{i2}\beta > x_{i1}\beta \quad &\Longleftrightarrow \quad P(y_{i2} = 1 \mid x_{i1}, x_{i2}, \alpha_i) > P(y_{i1} = 1 \mid x_{i1}, x_{i2}, \alpha_i) \\
x_{i2}\beta = x_{i1}\beta \quad &\Longleftrightarrow \quad P(y_{i2} = 1 \mid x_{i1}, x_{i2}, \alpha_i) = P(y_{i1} = 1 \mid x_{i1}, x_{i2}, \alpha_i) \\
x_{i2}\beta < x_{i1}\beta \quad &\Longleftrightarrow \quad P(y_{i2} = 1 \mid x_{i1}, x_{i2}, \alpha_i) < P(y_{i1} = 1 \mid x_{i1}, x_{i2}, \alpha_i)
\end{aligned} \tag{9}$$

So if we condition on $y_{i1} + y_{i2} = 1$, then we get:

$$P(y_{i2} = 1 \mid y_{i1} + y_{i2} = 1, x_{i1}, x_{i2}, \alpha_i) \begin{cases} > 1/2 & \text{if } (x_{i2} - x_{i1})\beta > 0 \\ = 1/2 & \text{if } (x_{i2} - x_{i1})\beta = 0 \\ < 1/2 & \text{if } (x_{i2} - x_{i1})\beta < 0 \end{cases}$$

[4] See for example Andersen (1970)

This is like Manski (1975) , so it is possible to use Maximum Score. This first lemma allows him to develop, under some identification conditions, a consistent estimator by maximizing for $b$ the sample analog of the following equation:

$$H(b) \equiv E[\text{sgn}((x_{i2} - x_{i1})b)(y_{i2} - y_{i1})] \tag{10}$$

for the observations where $y_{i1} \neq y_{i2}$.

Unfortunately, this approach cannot be generalized in such a way to generate an equivalent Lemma for the case of multiple fixed effects panel data models. Indeed, following a similar line of thoughts as for the logit case presented earlier, we would hope to adapt Lemma 1 by applying the same trick twice.

Suppose we have the following model:

$$P(y_{ij} = 1 \mid \mathbf{x}, \mu, \alpha) = F(x_{ij}\beta + \mu_i + \alpha_j) \quad i, j = 1, \ldots, n$$

We now restrict $F$ to be the same for all observations. In other words, all the disturbances are drawn from the same distribution. This is more restrictive than Manski's assumption, but still allows for an interesting range of models. We will show that even under this stricter set of assumptions, we cannot generalize this estimator to the case with two fixed effects. To do so we first apply the trick once to eliminate $\mu_i$ and get:

$$P(y_{ij} = 1 \mid y_{ij} + y_{ik} = 1, \mathbf{x}, \mu, \alpha) \begin{cases} > 1/2 & \text{if } (x_{ij} - x_{ik})\beta + \alpha_j - \alpha_k > 0 \\ = 1/2 & \text{if } (x_{ij} - x_{ik})\beta + \alpha_j - \alpha_k = 0 \\ < 1/2 & \text{if } (x_{ij} - x_{ik})\beta + \alpha_j - \alpha_k < 0 \end{cases}$$

This looks similar to the logit once we had used the conditioning a first time: explanatory variable $(x_{ij} - x_{ik})$ and fixed effect $\alpha_j - \alpha_k$. However, to apply the trick again, we would need $P(y_{ij} = 1 \mid y_{ij} + y_{ik} = 1, \mathbf{x}, \mu, \alpha)$ to have the form $F\big((x_{ij} - x_{ik})\beta + \alpha_j - \alpha_k\big)$ where $F$ is a CDF. Yet, this does not hold: we can't attest that this probability is always increasing. Therefore, we cannot apply Manski a second time: Manski's Maximum Score estimator cannot be adapted to the presence of two fixed effects, even under a stronger set of assumptions.

There is fundamental difference between this estimator and the logit that explains the opposite results. The logit can accommodate two fixed effects because using the known method once to deal with the first fixed effect gives us another logit, therefore allowing to apply said method a second time. This does not hold for Manski's maximum score estimator.

In general, models that can accommodate two fixed effects will have this feature: the model resulting from the use of the conditioning in one dimension is one for which we know the incidental parameter problem can be solved with the conditional approach. For example, conditioning once for the logit returns a logit. As we will see in the next subsection, conditioning once for the Poisson returns a logit (multinomial distribution). However, conditioning once for the maximum score returns a model with unknown properties.

### 2.3 Fixed Effects in a Poisson Model

We now turn to another model where the conditional maximum likelihood approach allows to deal with multiple fixed effects: the Poisson regression model. Here, as in the logit case, we can successfully solve the incidental parameter problem in the presence of two fixed effects. This is of particular interest since this model has become common in the trade application that motivates this study, namely the gravity equation literature.

Hausman, Hall and Griliches (1984) solved the incidental parameter problem for the Poisson model with one fixed effect. They used a conditional maximum likelihood approach to develop what is now called the fixed effect Poisson estimator. Lancaster (2002) proved that there really wasn't an incidental parameter problem in the Poisson model with one fixed effect. Indeed, he showed that in this case, the conditional and unconditional maximum likelihood were equivalent However, this equivalence cannot be established for a Poisson model with two fixed effects.[5] In fact, we show in the appendix that the incidental parameter problem remains in the case of two fixed effects (i.e the likelihood function is not maximized at the true parameter value). We therefore need a conditional approach, much like that of Hausman, Hall and Griliches (1984). Once again, this is especially important for trade applications, like the gravity model, that frequently use Poisson with two fixed effects. This section will go as follows; first we will recall the original fixed effect Poisson estimator.

[5] Lancaster (2002) showed the absence of an incidental parameter problem whenever the likelihood function could be factored into two orthogonal parts, one containing the parameters of interest and the other the incidental parameters. Such a decomposition cannot be done for the Poisson with two fixed effects.

Then, we will adapt it to deal with two fixed effects.

To illustrate the general idea behind conditional likelihood in the context of a Poisson model with one fixed effect, let us review the straightforward case where there are only two observations for each individual :

$$y_{it} \sim Pois(\exp(x_{it}\beta + \alpha_i)).$$

It follows from the relationship between Poisson and binomial distribution, that the distribution of $y_{i1}$ given that $y_{i1} + y_{i2} = K$ is given by

$$y_{i1} \sim bi\left(K, \frac{\exp(x_{i1}\beta)}{\exp(x_{i1}\beta) + \exp(x_{i2}\beta)}\right)$$

which does not involve the fixed effects, therefore allowing to consistently estimate $\beta$. With that in mind, we now present the fixed effect Poisson estimator of Hausman, Hall and Griliches (1984).

Suppose we have observations $y_{it}$ distributed Poisson with parameter $\lambda_{it} = \exp(x_{it}\beta + \alpha_i + \alpha_0)$ on $n$ individuals observed for $T$ periods. The parameters $\alpha_i$ represent individual fixed effects and $\alpha_0$ is the overall intercept so that we have $E(e^{\alpha_i}) = 1$. Therefore,

$$Pr(y_{it} \mid x_{it}, \alpha_i) = \frac{e^{-\lambda_{it}}(\lambda_{it})^{y_{it}}}{y_{it}!}. \tag{11}$$

The incidental parameter problem prevents us from consistently estimating the parameters in (11) by maximum likelihood. To solve this problem, Hausman, Hall and Griliches (1984) follow Andersen (1970, 1972) and condition on the sum $\sum_t y_{it}$. This is a sufficient statistic for $\alpha_i$. Just as in the simpler binomial case, it is well known that the distribution of $y_{it}$ conditional on $\sum_t y_{it}$ is a multinomial distribution [13]:

$$\begin{aligned} Pr(y_{i1}, \ldots, y_{iT} \mid \sum y_{it}) &= \frac{Pr\left(y_{i1}, \ldots, y_{i,T-1}, \sum_{t=1}^{T} y_{it} - \sum_{t=1}^{T-1} y_{it}\right)}{Pr(\sum y_{it})} \\ &= \frac{\frac{e^{-\sum_t \lambda_{it}} \prod_t \lambda_{it}^{y_{it}}}{\prod_t (y_{it}!)}}{\frac{e^{-\sum_t \lambda_{it}} \left(\sum_t \lambda_{it}\right)^{\sum_t y_{it}}}{\left(\sum_t y_{it}\right)!}} \\ &= \frac{\left(\sum_t y_{it}\right)!}{\prod_t (y_{it}!)} \prod_t \left[\frac{\lambda_{it}}{\sum_t \lambda_{it}}\right]^{y_{it}} \end{aligned} \tag{12}$$

We can simplify the term on the right as

$$\frac{e^{x_{it}\beta+\mu_i}}{\sum_t e^{x_{it}\beta+\mu_i}} = \frac{e^{x_{it}\beta}}{\sum_t e^{x_{it}\beta}}$$

which no longer depends on the fixed effects and can therefore be used to produce a likelihood function to consistently estimate the parameter $\beta$. Note the similarity with the logit specification developed earlier. We want to know if it is possible, as it was for the logit, to find a sufficient statistic for the panel data model with two fixed effects. This would allow us to write the conditional distribution of the $y_{it}$s as an expression that does not include the fixed effects. In other words, we want an equivalent to equation (12).

Now suppose we have a balanced panel of $n$ individuals where each observation is distributed Poisson in the following way:

$$y_{ij} \sim Pois(\exp(x_{ij}\beta + \mu_i + \alpha_j)) \tag{13}$$

where $\mu_i$ and $\alpha_j$ are individual fixed effects, $x_{ij}$ is a vector of explanatory variables and $\beta$ a vector of parameters to be estimated. We will define the mean of the distribution to be $\lambda_{ij} \equiv \exp(x_{ij}\beta + \mu_i + \alpha_j)$.

To gain some intuition, let us first look at the $2 \times 2$ case. We will then generalize to $n \times n$. In order to eliminate the incidental parameter problem related to the fixed effects, we need to know the following distribution:

$$y_{ij} \mid y_{ij} + y_{ik}, y_{lj} + y_{lk}, y_{ij} + y_{lj} \sim ? \tag{14}$$

For simplicity of notation, let us define:

$$\begin{aligned} y_{ij} &= m_0 \\ y_{ij} + y_{ik} &= m_1 \\ y_{lj} + y_{lk} &= m_2 \\ y_{ij} + y_{lj} &= m_3 \end{aligned} \tag{15}$$

Then, we can express the desired probability as:

$$
\begin{aligned}
Pr(y_{ij} = & m_0 \mid y_{ij} + y_{ik} = m_1, y_{lj} + y_{lk} = m_2, y_{ij} + y_{lj} = m_3) \qquad (16)\\
&= \frac{Pr(y_{ij} = m_0, y_{ik} = m_1 - m_0, y_{lj} = m_3 - m_0, y_{lk} = m_2 - m_3 + m_0)}{Pr(y_{ij} + y_{ik} = m_1, y_{lj} + y_{lk} = m_2, y_{ij} + y_{lj} = m_3)}\\
&= \frac{Pr(y_{ij} = m_0, y_{ik} = m_1 - m_0, y_{lj} = m_3 - m_0, y_{lk} = m_2 - m_3 + m_0)}{\sum_{t=\max\{0,-(m_2-m_3)\}}^{min\{m_1,m_3\}} Pr(y_{ij} = t, y_{ik} = m_1 - t, y_{lj} = m_3 - t, y_{lk} = m_2 - m_3 + t)}
\end{aligned}
$$

When a variable is distributed Poisson with mean $\lambda$ we know that the probability this variable takes a specific value $m$ is given by:

$$
f(m;\lambda) = \frac{\lambda^m e^{-\lambda}}{m!}
$$

Using this expression, let us write

$$
f_{ij} f_{ik} f_{lj} f_{lk} = \frac{\lambda_{ij}^{m_0} e^{-\lambda_{ij}}}{m_0!} \frac{\lambda_{ik}^{m_1-m_0} e^{-\lambda_{ik}}}{(m_1-m_0)!} \frac{\lambda_{lj}^{m_3-m_0} e^{-\lambda_{lj}}}{(m_3-m_0)!} \frac{\lambda_{lk}^{m_2-m_3+m_0} e^{-\lambda_{lk}}}{(m_2-m_3+m_0)!}
$$

and

$$
f_{ij}^{(t)} f_{ik}^{(t)} f_{lj}^{(t)} f_{lk}^{(t)} = \frac{\lambda_{ij}^{t} e^{-\lambda_{ij}}}{t!} \frac{\lambda_{ik}^{m_1-t} e^{-\lambda_{ik}}}{(m_1-t)!} \frac{\lambda_{lj}^{m_3-t} e^{-\lambda_{lj}}}{(m_3-t)!} \frac{\lambda_{lk}^{m_2-m_3+t} e^{-\lambda_{lk}}}{(m_2-m_3+t)!}
$$

Since the variables are independently distributed, we can now express the probability in (16) as

$$
\begin{aligned}
Pr(y_{ij} = & m_0 \mid y_{ij} + y_{ik} = m_1, y_{lj} + y_{lk} = m_2, y_{ij} + y_{lj} = m_3)\\
&= \frac{f_{ij} f_{ik} f_{lj} f_{lk}}{\sum_{t=\max\{0,-(m_2-m_3)\}}^{min\{m_1,m_3\}} f_{ij}^{(t)} f_{ik}^{(t)} f_{lj}^{(t)} f_{lk}^{(t)}}
\end{aligned}
$$

We invert to simplify the expression:

$$
\begin{aligned}
(Pr(y_{ij} = & m_0 \mid y_{ij} + y_{ik} = m_1, y_{lj} + y_{lk} = m_2, y_{ij} + y_{lj} = m_3))^{-1}\\
&= \frac{\sum_{t=\max\{0,-(m_2-m_3)\}}^{min\{m_1,m_3\}} f_{ij}^{(t)} f_{ik}^{(t)} f_{lj}^{(t)} f_{lk}^{(t)}}{f_{ij} f_{ik} f_{lj} f_{lk}}\\
&= \sum_{t=\max\{0,-(m_2-m_3)\}}^{min\{m_1,m_3\}} \frac{m_0!}{t!} \lambda_{ij}^{t-m_0} \frac{(m_1-m_0)!}{(m_1-t)!} \lambda_{ik}^{m_0-t} \frac{(m_3-m_0)!}{(m_3-t)!} \lambda_{lj}^{m_0-t} \frac{(m_2-m_3+m_0)!}{(m_2-m_3+t)!} \lambda_{lk}^{t-m_0}\\
&= \sum_{t=\max\{0,-(m_2-m_3)\}}^{min\{m_1,m_3\}} \frac{m_0!}{t!} \frac{(m_1-m_0)!}{(m_1-t)!} \left(\frac{\lambda_{ij}}{\lambda_{ik}}\right)^{t-m_0} \frac{(m_3-m_0)!}{(m_3-t)!} \frac{(m_2-m_3+m_0)!}{(m_2-m_3+t)!} \left(\frac{\lambda_{lj}}{\lambda_{lk}}\right)^{m_0-t}
\end{aligned}
$$

From the definition of the $\lambda$s we know that:

$$\frac{\lambda_{ij}}{\lambda_{ik}} = \frac{e^{x'_{ij}\beta+\mu_i+\alpha_j}}{e^{x'_{ik}\beta+\mu_i+\alpha_k}} = e^{(x_{ij}-x_{ik})'\beta+\alpha_j-\alpha_k}$$

and

$$\frac{\lambda_{lj}}{\lambda_{lk}} = \frac{e^{x'_{lj}\beta+\mu_l+\alpha_j}}{e^{x'_{lk}\beta+\mu_l+\alpha_k}} = e^{(x_{lj}-x_{lk})'\beta+\alpha_j-\alpha_k}$$

and therefore,

$$\frac{\lambda_{ij}}{\lambda_{ik}}\frac{\lambda_{lk}}{\lambda_{lj}} = e^{[(x_{lj}-x_{lk})'\beta-(x_{lj}-x_{lk})'\beta]}.$$

Finally we can write the desired probability as:

$$\begin{aligned}(Pr(y_{ij} = & m_0 \mid y_{ij}+y_{ik}=m_1, y_{lj}+y_{lk}=m_2, y_{ij}+y_{lj}=m_3))^{-1} \\ &= \sum_{t=\max\{0,-(m_2-m_3)\}}^{min\{m_1,m_3\}} \frac{m_0!}{t!}\frac{(m_1-m_0)!}{(m_1-t)!}\frac{(m_3-m_0)!}{(m_3-t)!}\frac{(m_2-m_3+m_0)!}{(m_2-m_3+t)!} \\ &\qquad\qquad \cdot \left(e^{[(x_{lj}-x_{lk})'\beta-(x_{lj}-x_{lk})'\beta]}\right)^{t-m_0}\end{aligned}$$

which does not depend on the fixed effects, eliminating the incidental parameter problem.

This was an example of the $2\times 2$ solution. Now let us look at the general $n\times n$ case. We are looking for a sufficient statistic for the fixed effects: a statistic such that when conditioned upon, the joint distribution of the $y_{ij}$ will not depend on the fixed effects $\mu_i$ and $\alpha_j$. We will show that this statistic is the vector of the sum of the columns $(\sum_i y_{ij})$ and the sum of the rows$(\sum_j y_{ij})$. Intuitively, the method goes as follows: we first apply Hausman, Hall and Griliches (1984) to eliminate $\mu_i$. We then find ourselves in a logit "world" with fixed effect $(\alpha_j-\alpha_k)$ and therefore can use Rasch to eliminate $(\alpha_j-\alpha_k)$.

First, let us write the conditional joint probability of the vector of all observations $Y$ as

$$Pr(Y \mid r,c,\alpha,\mu,x,\beta) = \frac{Pr(Y \mid \alpha,\mu,x,\beta)}{Pr(r,c \mid \alpha,\mu,x,\beta)} \tag{17}$$

with

$$Pr(r, c \mid \alpha, \mu, x, \beta) = \sum_{Y} Pr(r, c \mid Y, \alpha, \mu, x, \beta) \cdot Pr(Y \mid \alpha, \mu, x, \beta)$$
$$= \sum_{Y' \in Q} Pr(Y' \mid \alpha, \mu, x, \beta) \tag{18}$$

where $Q$ is the set of all possible distribution of $y_{ij}$ such that the sum of the rows is given by the vector $r$ and the sum of the columns by $c$. The vectors of fixed effects are given by $\alpha$ and $\mu$. Define $b_{ij} \equiv min(r_i - \sum_{k<j} y_{ik}, c_j - \sum_{k<i} y_{kj})$. Then we can express $Q$ as:

$$Q = \{0 \leq y_{ij} \leq b_{ij} \quad \forall i, j = 1, ..., n\}.$$

Using this expression, we write the probability as:

$$\frac{Pr(Y \mid \alpha, \mu, x, \beta)}{\sum_{Y' \in Q} Pr(Y' \mid \alpha, \mu, x, \beta)} = \frac{\frac{e^{-\sum_i \sum_j \lambda_{ij}} \prod_{i,j=1}^{n} \lambda_{ij}^{y_{ij}}}{\prod_{ij}^{n} (y_{ij}!)}}{\sum_{Y' \in Q} \frac{e^{-\sum_i \sum_j \lambda_{ij}} \prod_{i,j=1}^{n} \lambda_{ij}^{y'_{ij}}}{\prod_{ij}^{n} (y'_{ij}!)}} \tag{19}$$

This is the probability of getting the distribution $Y$ over the sum of probabilities of all possible distributions $Y'$ that have the same sum of rows and columns. This would allow us to eliminate both fixed effects, but the sum in the denominator would be impractical to implement empirically. Indeed, for numbers that we would find in realistic applications, such as those reflecting trade between countries, the number of possible $Y'$'s would be too large for computation. There is however an easy way around this. To accomplish our goal, it is sufficient to compare our matrix $Y$

to one alternative distribution of the $y_{ij}$'s with the same sum of rows and columns (here $Y'$):

$$\begin{aligned}
&\frac{Pr(Y \mid \alpha,\mu,x,\beta)}{Pr(Y \mid \alpha,\mu,x,\beta)+Pr(Y' \mid \alpha,\mu,x,\beta)} \\
&= \frac{\frac{e^{-\sum_i\sum_j \lambda_{ij}}\prod_{i,j=1}^{n}\lambda_{ij}^{y_{ij}}}{\prod_{ij}^{n}(y_{ij}!)}}{\frac{e^{-\sum_i\sum_j \lambda_{ij}}\prod_{i,j=1}^{n}\lambda_{ij}^{y_{ij}}}{\prod_{ij}^{n}(y_{ij}!)}+\frac{e^{-\sum_i\sum_j \lambda_{ij}}\prod_{i,j=1}^{n}\lambda_{ij}^{y'_{ij}}}{\prod_{ij}^{n}(y'_{ij}!)}} \\
&= \prod_{i,j=1}^{n}(y'_{ij}!)\frac{\prod_{i,j=1}^{n}\lambda_{ij}^{y_{ij}}}{\prod_{i,j=1}^{n}(y'_{ij}!)\prod_{i,j=1}^{n}\lambda_{ij}^{y_{ij}}+\prod_{i,j=1}^{n}(y_{ij}!)\prod_{i,j=1}^{n}\lambda_{ij}^{y'_{ij}}} \\
&= \prod_{i,j=1}^{n}(y'_{ij}!)\frac{\prod_{i,j=1}^{n}e^{(x_{ij}\beta+\mu_i+\alpha_j)y_{ij}}}{\prod_{i,j=1}^{n}(y'_{ij}!)\prod_{i,j=1}^{n}e^{(x_{ij}\beta+\mu_i+\alpha_j)y_{ij}}+\prod_{i,j=1}^{n}(y_{ij}!)\prod_{i,j=1}^{n}e^{(x_{ij}\beta+\mu_i+\alpha_j)y'_{ij}}} \\
&= \frac{1}{1+\frac{\prod_{i,j=1}^{n}(y_{ij}!)}{\prod_{i,j=1}^{n}(y'_{ij}!)}e^{\sum_i\sum_j(x_{ij}\beta+\mu_i+\alpha_j)(y'_{ij}-y_{ij})}} \\
&= \frac{1}{1+\frac{\prod_{i,j=1}^{n}(y_{ij}!)}{\prod_{i,j=1}^{n}(y'_{ij}!)}\left(e^{\sum_i\sum_j x_{ij}\beta(y'_{ij}-y_{ij})}\right)\left(e^{\sum_{i=1}^{n}\mu_i\sum_{j=1}^{n}(y'_{ij}-y_{ij})}\right)\left(e^{\sum_{j=1}^{n}\alpha_j\sum_{i=1}^{n}(y'_{ij}-y_{ij})}\right)} \\
&= \frac{1}{1+\frac{\prod_{i,j=1}^{n}(y_{ij}!)}{\prod_{i,j=1}^{n}(y'_{ij}!)}\left(e^{\sum_i\sum_j x_{ij}\beta(y'_{ij}-y_{ij})}\right)} \qquad (20)
\end{aligned}$$

The last equality results from the fact that $\sum_{j=1}^{n}(y'_{ij}-y_{ij}) = r_i - r_i = 0$ and $\sum_{i=1}^{n}(y'_{ij}-y_{ij}) = c_j - c_j = 0$. This likelihood function does not depend on the fixed effects. We can therefore apply it to $2\times 2$ matrices (much like in the logit case) to consistently estimate $\beta$. Thus, in a similar fashion to the logit, it is possible to use conditional maximum likelihood in a Poisson model to derive a consistent estimate of the parameter $\beta$ when we have two fixed effects, using as sufficient statistics the sum of rows and columns of the panel.[6]

To implement this estimator, select a random $l$ and $k$ for each observation $ij$ to compose a small $2\times 2$ matrix. Then, generate a second matrix ($Y'$) that has the same sum of columns and rows.

[6] Note that if the use of the Poisson maximum likelihood estimator for one fixed effect is not limited to observations distributed Poisson, this is not necessarily true for this two fixed effect equivalent. Indeed, if we know by consistency of maximum likelihood estimators that our function is maximized at the true value of $\beta$ when we have a Poisson distribution, we cannot prove consistency for other models (for example a multiplicative model such as the one illustrated in the next subsection). This is simply because equation (20) does not have an analytic solution for the value of $\beta$ that maximizes it.

For each pair $ij$, the procedure can be repeated $T$ times. Using this, minimize

$$\sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{j=1}^{n}\log\left(1+\frac{y_{ij}!y_{ik}!y_{lj}!y_{lk}!)}{y'_{ij}!y'_{ik}!y'_{lj}!y'_{lk}!}\left(e^{x_{ij}\beta(y'_{ij}-y_{ij})+x_{ik}\beta(y'_{ik}-y_{ik})+x_{lj}\beta(y'_{lj}-y_{lj})+x_{lk}\beta(y'_{lk}-y_{lk})}\right)\right)$$

We now show that this procedure can also be applied to other somewhat similar models.

### 2.4 Fixed Effects in Negative Binomial and Gamma Models

Some other models, more or less related to the Poisson regression model, can be adapted to two fixed effects using the same conditional likelihood approach seen in the previous subsection. It is the case for the Negative Binomial and Gamma models. Indeed, we show that if we have the following negative binomial distribution

$$y_{ij}\sim NB\big(e^{x_{ij}\beta},\frac{e^{\mu_i}}{(1+e^{\mu_i})}\frac{e^{\alpha_j}}{((1+e^{\alpha_j})}\big)$$

and can express the conditional expectation as

$$E[y_{ij}\mid x_{ij}]=\frac{e^{x_{ij}\beta+\mu_i+\alpha_j}}{1+e^{\mu_i}+e^{\alpha_j}}$$

then the sum of rows and the sum of columns are sufficient statistics for $\mu_i$ and $\alpha_j$. The proof is very similar to that of the Poisson case. Indeed, equations (17) and (18) still hold and comparing our matrix to an alternative $Y'$ we get the equivalent to equation (20). We can write:

$$\frac{Pr(Y\mid\alpha,\mu,x,\beta)}{Pr(Y\mid\alpha,\mu,x,\beta)+Pr(Y'\mid\alpha,\mu,x,\beta)}=\frac{A}{A+B}$$

where

$$A\equiv\prod_{i,j=1}^{n}\frac{\Gamma(y_{ij}+e^{x_{ij}\beta})}{\Gamma(e^{x_{ij}\beta})\Gamma(y_{ij}+1)}\left(\frac{1+e^{\mu_i}+e^{\alpha_j}}{(1+e^{\mu_i})(1+e^{\alpha_j})}\right)^{e^{x_{ij}\beta}}\left(\frac{e^{\mu_i+\alpha_j}}{(1+e^{\mu_i})(1+e^{\alpha_j})}\right)^{y_{ij}}$$

and

$$B\equiv\prod_{i,j=1}^{n}\frac{\Gamma(y'_{ij}+e^{x_{ij}\beta})}{\Gamma(e^{x_{ij}\beta})\Gamma(y'_{ij}+1)}\left(\frac{1+e^{\mu_i}+e^{\alpha_j}}{(1+e^{\mu_i})(1+e^{\alpha_j})}\right)^{e^{x_{ij}\beta}}\left(\frac{e^{\mu_i+\alpha_j}}{(1+e^{\mu_i})(1+e^{\alpha_j})}\right)^{y'_{ij}}$$

Therefore, we have

$$
\begin{aligned}
&\frac{Pr(Y \mid \alpha,\mu,x,\beta)}{Pr(Y \mid \alpha,\mu,x,\beta) + Pr(Y' \mid \alpha,\mu,x,\beta)} \\
&\quad = \frac{1}{1+\prod_{i,j=1}^{n} \frac{\Gamma(y'_{ij}+e^{x_{ij}\beta})}{\Gamma(y_{ij}+e^{x_{ij}\beta})} \frac{\Gamma(y_{ij}+1)}{\Gamma(y'_{ij}+1} \prod_i \left(\frac{e^{\mu_i}}{(1+e^{\mu_i})}\right)^{\sum_j (y'_{ij}-y_{ij})} \prod_j \left(\frac{e^{\alpha_j}}{(1+e^{\alpha_j})}\right)^{\sum_i (y'_{ij}-y_{ij})}} \\
&\quad = \frac{1}{1+\prod_{i,j=1}^{n} \frac{\Gamma(y'_{ij}+e^{x_{ij}\beta})}{\Gamma(y_{ij}+e^{x_{ij}\beta})} \frac{\Gamma(y_{ij}+1)}{\Gamma(y'_{ij}+1}} \qquad (21)
\end{aligned}
$$

We could then apply maximum likelihood following the same procedure as for the Poisson model to get consistent estimates of $\beta$.

Similarly, if we have a Gamma distribution such that

$$y_{ij} \sim Gamma(e^{x_{ij}\beta}, \frac{1}{e^{\mu_i}+e^{\alpha_j}})$$

with the following conditional expectation:

$$E[y_{ij} \mid x_{ij}] = \frac{e^{x_{ij}\beta}}{e^{\mu_i}+e^{\alpha_j}}$$

we can once again use the conditioning on the sum of rows and columns allows to eliminate the fixed effects and apply the same conditional maximum likelihood to consistently estimate $\beta$. We have:

$$\frac{Pr(Y \mid \alpha,\mu,x,\beta)}{Pr(Y \mid \alpha,\mu,x,\beta) + Pr(Y' \mid \alpha,\mu,x,\beta)} = \frac{C}{C+D}$$

where

$$C \equiv \prod_{i,j=1}^{n} (e^{\mu_i}+e^{\alpha_j})^{e^{e^{x_{ij}\beta}}} \frac{1}{\Gamma(e^{x_{ij}\beta})} y_{ij}^{e^{x_{ij}\beta}-1} e^{-y_{ij}(e^{\mu_i}+e^{\alpha_j})}$$

and

$$D \equiv \prod_{i,j=1}^{n} (e^{\mu_i}+e^{\alpha_j})^{e^{e^{x_{ij}\beta}}} \frac{1}{\Gamma(e^{x_{ij}\beta})} y_{ij}^{\prime e^{x_{ij}\beta}-1} e^{-y'_{ij}(e^{\mu_i}+e^{\alpha_j})}$$

Therefore,

$$\begin{aligned}
&\frac{Pr(Y \mid \alpha, \mu, x, \beta)}{Pr(Y \mid \alpha, \mu, x, \beta) + Pr(Y' \mid \alpha, \mu, x, \beta)} \\
&\quad = \frac{1}{1 + \prod_{i,j=1}^{n} \left(\frac{y_{ij}}{y'_{ij}}\right)^{e^{x_{ij}\beta}-1} e^{\sum_{ij}(y_{ij}-y'_{ij})(e^{\mu_i}+e^{\alpha_j})}} \\
&\quad = \frac{1}{1 + \prod_{i,j=1}^{n} \left(\frac{y_{ij}}{y'_{ij}}\right)^{e^{x_{ij}\beta}-1} e^{\sum_i e^{\mu_i}\sum_j (y_{ij}-y'_{ij}) + \sum_j e^{\alpha_j}\sum_i (y_{ij}-y'_{ij})}} \\
&\quad = \frac{1}{1 + \prod_{i,j=1}^{n} \left(\frac{y_{ij}}{y'_{ij}}\right)^{e^{x_{ij}\beta}-1}} \qquad (22)
\end{aligned}$$

We could then apply maximum likelihood following the same procedure as for the Poisson model to get consistent estimates of $\beta$. We conjecture that other homologous distribution could also be similarly adapted to the presence of two fixed effects.

### 2.5 Fixed Effects in Multiplicative Form

Finally, it is possible to deal with two fixed effects in the simpler, but useful, class of models in multiplicative form. Note that these models can be particularly pertinent when looking at gravity equations. In this case, we are looking for a moment condition that does not depend on the fixed effects. Much in the spirit of the difference-in-differences and the conditional likelihood, it is possible here to do pairwise comparisons and eliminate the fixed effects.

Suppose we have a balanced panel of $n$ individuals where each observation is given by:

$$E[y_{ij} \mid x_{ij}] = e^{x_{ij}\beta+\mu_i+\alpha_j} \qquad (23)$$

where $\mu_i$ and $\alpha_j$ are individual fixed effects, $x_{ij}$ is a vector of explanatory variables and $\beta$ a vector of parameters to be estimated. We define the residuals as:

$$\varepsilon_{ij} = y_{ij} - E[y_{ij} \mid x_{ij}].$$

We want to find a moment condition that does not depend on the fixed effects. By multiplying by $e^{-x_{ij}\beta}$, we can write

$$y_{ij}e^{-x_{ij}\beta} = e^{\mu_i+\alpha_j} + \varepsilon_{ij}e^{-x_{ij}\beta}. \qquad (24)$$

Similarly,

$$y_{ik} = e^{x_{ik}\beta+\mu_i+\alpha_k} + \varepsilon_{ik}$$

implies

$$e^{x_{ik}\beta} = y_{ik}e^{-\mu_i-\alpha_k} - \varepsilon_{ik}e^{-\mu_i-\alpha_k}. \tag{25}$$

Multiplying (24) by (25) we get:

$$\begin{aligned} y_{ij}e^{-x_{ij}\beta+x_{ik}\beta} =& e^{\mu_i+\alpha_j}(y_{ik}e^{-\mu_i-\alpha_k} - \varepsilon_{ik}e^{-\mu_i-\alpha_k}) + \varepsilon_{ij}e^{-x_{ij}\beta}(y_{ik}e^{-\mu_i-\alpha_k} - \varepsilon_{ik}e^{-\mu_i-\alpha_k}) \\ =& y_{ik}e^{\alpha_j-\alpha_k} - \varepsilon_{ik}e^{\alpha_j-\alpha_k} + \varepsilon_{ij}e^{-x_{ij}\beta-\mu_i-\alpha_k}(y_{ik} - \varepsilon_{ik}). \end{aligned} \tag{26}$$

Doing the same operations with $y_{lj}$ and $y_{lk}$, we get

$$y_{lj}e^{-x_{lj}\beta+x_{lk}\beta} = y_{lk}e^{\alpha_j-\alpha_k} - \varepsilon_{lk}e^{\alpha_j-\alpha_k} + \varepsilon_{lj}e^{-x_{lj}\beta-\mu_l-\alpha_k}(y_{lk} - \varepsilon_{lk}). \tag{27}$$

Combining (26) and (27), we get:

$$\begin{aligned} & y_{ik}y_{lj}e^{-x_{lj}\beta+x_{lk}\beta} - y_{lj}e^{-x_{lj}\beta+x_{lk}\beta}\varepsilon_{ik} + y_{lj}e^{-x_{lj}\beta+x_{lk}\beta-x_{ij}\beta-\mu_i-\alpha_j}\varepsilon_{ij}(y_{ik} - \varepsilon_{ik}) \\ & \quad = y_{lk}y_{ij}e^{-x_{ij}\beta+x_{ik}\beta} - y_{ij}e^{-x_{ij}\beta+x_{ik}\beta}\varepsilon_{lk} + y_{ij}e^{-x_{ij}\beta+x_{ik}\beta-x_{lj}\beta-\mu_l-\alpha_j}\varepsilon_{lj}(y_{lk} - \varepsilon_{lk}). \end{aligned}$$

Taking the conditional expectation, we get the following moment condition:

$$E[y_{ik}y_{lj}e^{-x_{lj}\beta+x_{lk}\beta} \mid \mathbf{x}] = E[y_{lk}y_{ij}e^{-x_{ij}\beta+x_{ik}\beta} \mid \mathbf{x}]$$

which does not depend on the fixed effects. This allows us to write the unconditional moment condition:

$$E\big[(y_{ik}y_{lj} - y_{lk}y_{ij}e^{(x_{ik}-x_{ij}+x_{lj}-x_{lk})\beta})(x_{ik} - x_{ij} + x_{lj} - x_{lk})\big] = 0 \tag{28}$$

We could then use equation (28) to develop a GMM consistent estimator for $\beta$, adding this model to the class including logit, Poisson, Negative Binomial, and Gamma, that can successfully produce consistent estimates in the presence of two fixed effects.

# 3 Monte Carlo Simulations

In this section, we present Monte Carlo evidence to support the multiple fixed effects estimators developed in this paper. For this, we focus on the logit estimator given by the maximization of equation (8). The simulations will compare that estimator to a regular logit, ignoring the fixed effects, and a logit estimating all the fixed effects (putting in dummies). Recall that this last estimator is subject to the incidental parameter problem.

To account for different possible features of the data, this comparison will be made for four different designs. All of these designs are applied to the estimation of the following model:

$$y_{ij} = 1\{x_{ij}'\beta + \mu_i + \alpha_j + \varepsilon_{ij} \geq 0\} \quad i = 1, \ldots, n, j = 1, \ldots, n$$

where $x_{ij}$ is a vector of five explanatory variables[7] drawn from a standard normal distribution and the error term $\varepsilon_{ij}$ is drawn from a logistic distribution. The first design has no fixed effects ($\mu_i = \alpha_j = 0 \quad \forall i, j$). The second design has fixed effects drawn from a standard normal distributions uncorrelated with the explanatory variables. In both these first two designs, $\beta_1 = 1$ and $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$. The third design has fixed effects correlated with the first explanatory variables. More specifically, $x_1 = rndn + \alpha + \eta$ where $rndn$ is a standard normal and $\alpha$ and $\eta$ are the same fixed effects as in the second design. To illustrate how this affects the coefficient on this variable differently, we now also have $\beta_2 = 1$. For these simulations to stay as close as possible to a data application, such as the one presented in the next section, these first three designs do not have resampling of the $x$s in each replication. However, since it is more common in Monte Carlo studies to have resampling, the fourth and last design replicates the second design, but with resampling of the data. Each design is estimated for $n = 136$ and for a sample half that size. These are both standard size ranges for trade studies. Whenever fixed effects are estimated, the coefficients are truncated in order to insure convergence. The results from a 1000 replications are given in Tables 1 through 4. For each estimator considered, we report the median bias, the median absolute deviation (MAD), the mean bias, and the root mean squared error (RMSE) for all five coefficient estimates.

Using the median absolute deviation as the measure of the precision and looking at the median

[7]We chose to keep five variables to have simulations as close as possible to the application of the next section.

Table 1: Design 1

| | | Median bias | MAD | Mean bias | RMSE |
|---|---|---|---|---|---|
| *Full sample:* $n = 136$ | | | | | |
| Logit | $\beta_1$ | 0.0001 | 0.0107 | 0.0002 | 0.0155 |
| | $\beta_2$ | -0.0004 | 0.0068 | -0.0003 | 0.0103 |
| | $\beta_3$ | -0.0002 | 0.0077 | -0.0002 | 0.0108 |
| | $\beta_4$ | -0.0002 | 0.0074 | -0.0001 | 0.0109 |
| | $\beta_5$ | 0.0003 | 0.0072 | 0.0003 | 0.0109 |
| Logit FE | $\beta_1$ | 0.0187 | 0.0209 | 0.0186 | 0.0290 |
| | $\beta_2$ | -0.0012 | 0.0126 | -0.0003 | 0.0186 |
| | $\beta_3$ | 0.0007 | 0.0126 | 0.0005 | 0.0194 |
| | $\beta_4$ | 0.0003 | 0.0122 | -0.0001 | 0.0190 |
| | $\beta_5$ | 0.0007 | 0.0130 | 0.0001 | 0.0185 |
| Conditional Logit | $\beta_1$ | 0.0007 | 0.0159 | 0.0011 | 0.0226 |
| | $\beta_2$ | -0.0008 | 0.0128 | -0.0002 | 0.0188 |
| | $\beta_3$ | 0.0008 | 0.0129 | 0.0005 | 0.0193 |
| | $\beta_4$ | 0.0002 | 0.0118 | -0.0002 | 0.0189 |
| | $\beta_5$ | 0.0006 | 0.0129 | 0.0001 | 0.0184 |
| *Small sample:* $n = 68$ | | | | | |
| Logit | $\beta_1$ | 0.0027 | 0.0208 | 0.0025 | 0.0306 |
| | $\beta_2$ | -0.0006 | 0.0134 | 0.0000 | 0.0212 |
| | $\beta_3$ | -0.0007 | 0.0136 | -0.0006 | 0.0207 |
| | $\beta_4$ | -0.0005 | 0.0148 | 0.0003 | 0.0218 |
| | $\beta_5$ | -0.0005 | 0.0148 | -0.0003 | 0.0220 |
| Logit FE | $\beta_1$ | 0.0392 | 0.0420 | 0.0392 | 0.0604 |
| | $\beta_2$ | 0.0020 | 0.0268 | 0.0006 | 0.0393 |
| | $\beta_3$ | 0.0005 | 0.0254 | 0.0002 | 0.0377 |
| | $\beta_4$ | 0.0004 | 0.0253 | 0.0005 | 0.0392 |
| | $\beta_5$ | -0.0021 | 0.0256 | -0.0017 | 0.0380 |
| Conditional Logit | $\beta_1$ | 0.0031 | 0.303 | 0.0038 | 0.0459 |
| | $\beta_2$ | 0.0027 | 0.0272 | 0.0005 | 0.0386 |
| | $\beta_3$ | -0.0002 | 0.0276 | 0.0005 | 0.0376 |
| | $\beta_4$ | 0.0010 | 0.0251 | 0.0013 | 0.0388 |
| | $\beta_5$ | -0.0024 | 0.0252 | -0.0016 | 0.0373 |

Design 1 has no fixed effects and the true coefficients are as follows: $\beta_1 = 1$ and $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

Table 2: Design 2

| | | Median bias | MAD | Mean bias | RMSE |
|---|---|---|---|---|---|
| *Full sample:* $n = 136$ | | | | | |
| Logit | $\beta_1$ | -0.2448 | 0.2448 | -0.2455 | 0.2490 |
| | $\beta_2$ | 0.0032 | 0.0222 | 0.0014 | 0.0332 |
| | $\beta_3$ | 0.0018 | 0.0222 | 0.0003 | 0.0321 |
| | $\beta_4$ | 0.0006 | 0.0218 | 0.0003 | 0.0324 |
| | $\beta_5$ | 0.0005 | 0.0217 | 0.0002 | 0.0327 |
| Logit FE | $\beta_1$ | 0.0201 | 0.0225 | 0.0207 | 0.0320 |
| | $\beta_2$ | -0.0005 | 0.0134 | -0.0004 | 0.0204 |
| | $\beta_3$ | -0.0009 | 0.0138 | -0.0010 | 0.0204 |
| | $\beta_4$ | 0.0003 | 0.0140 | -0.0005 | 0.0200 |
| | $\beta_5$ | 0.0003 | 0.0137 | 0.0007 | 0.0208 |
| Conditional Logit | $\beta_1$ | 0.0019 | 0.0174 | 0.0021 | 0.0252 |
| | $\beta_2$ | -0.0001 | 0.0139 | -0.0002 | 0.0209 |
| | $\beta_3$ | -0.0012 | 0.0139 | -0.0011 | 0.0207 |
| | $\beta_4$ | 0.0000 | 0.0138 | -0.0006 | 0.0199 |
| | $\beta_5$ | 0.0001 | 0.0142 | 0.0005 | 0.0210 |
| *Small sample:* $n = 68$ | | | | | |
| Logit | $\beta_1$ | -0.2408 | 0.2408 | -0.2414 | 0.2492 |
| | $\beta_2$ | -0.0007 | 0.0337 | 0.0001 | 0.0495 |
| | $\beta_3$ | -0.0024 | 0.0319 | -0.0018 | 0.0466 |
| | $\beta_4$ | 0.0005 | 0.0336 | 0.0004 | 0.0492 |
| | $\beta_5$ | -0.0024 | 0.0341 | -0.0018 | 0.0434 |
| Logit FE | $\beta_1$ | 0.0403 | 0.0439 | 0.0402 | 0.0635 |
| | $\beta_2$ | 0.0006 | 0.0287 | 0.0008 | 0.0427 |
| | $\beta_3$ | -0.0002 | 0.0284 | 0.0008 | 0.0410 |
| | $\beta_4$ | 0.0003 | 0.0297 | -0.0007 | 0.0429 |
| | $\beta_5$ | -0.0034 | 0.0283 | -0.0018 | 0.0434 |
| Conditional Logit | $\beta_1$ | 0.0005 | 0.0325 | 0.0019 | 0.0493 |
| | $\beta_2$ | 0.0000 | 0.0302 | 0.0013 | 0.0426 |
| | $\beta_3$ | 0.0006 | 0.0276 | 0.0013 | 0.0401 |
| | $\beta_4$ | -0.0022 | 0.0298 | -0.0008 | 0.0437 |
| | $\beta_5$ | -0.0029 | 0.0279 | -0.0020 | 0.0432 |

Design 2 has fixed effects drawn from a random normal uncorrelated with the $x$'s and the true coefficients are as follows: $\beta_1 = 1$ and $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

Table 3: Design 3

| | | Median bias | MAD | Mean bias | RMSE |
|---|---|---|---|---|---|
| *Full sample:* $n = 136$ | | | | | |
| Logit | $\beta_1$ | 0.4986 | 0.4986 | 0.4987 | 0.4992 |
| | $\beta_2$ | -0.0917 | 0.0917 | -0.0915 | 0.0929 |
| | $\beta_3$ | -0.0112 | 0.0121 | -0.0114 | 0.0166 |
| | $\beta_4$ | 0.0016 | 0.0087 | 0.0016 | 0.0127 |
| | $\beta_5$ | 0.0083 | 0.0105 | 0.0083 | 0.0155 |
| Logit FE | $\beta_1$ | 0.0266 | 0.0283 | 0.0256 | 0.0384 |
| | $\beta_2$ | 0.0241 | 0.0257 | 0.0244 | 0.0366 |
| | $\beta_3$ | -0.0013 | 0.0168 | -0.0016 | 0.0243 |
| | $\beta_4$ | -0.0007 | 0.0152 | -0.0006 | 0.0236 |
| | $\beta_5$ | -0.0002 | 0.0159 | -0.0001 | 0.0242 |
| Conditional Logit | $\beta_1$ | 0.0025 | 0.0202 | 0.0014 | 0.0306 |
| | $\beta_2$ | 0.0002 | 0.0196 | 0.0006 | 0.0296 |
| | $\beta_3$ | -0.0007 | 0.0172 | -0.0012 | 0.0250 |
| | $\beta_4$ | -.0003 | 0.0163 | 0.0000 | 0.0249 |
| | $\beta_5$ | -0.0005 | 0.0164 | 0.0005 | 0.0251 |
| *Small sample:* $n = 68$ | | | | | |
| Logit | $\beta_1$ | 0.4918 | 0.4918 | 0.4951 | 0.4974 |
| | $\beta_2$ | -0.0902 | 0.0902 | -0.0899 | 0.0972 |
| | $\beta_3$ | -0.0102 | 0.0203 | -0.0119 | 0.0315 |
| | $\beta_4$ | 0.0008 | 0.0211 | 0.0011 | 0.0294 |
| | $\beta_5$ | 0.0102 | 0.0218 | 0.0108 | 0.0316 |
| Logit FE | $\beta_1$ | 0.0491 | 0.0521 | 0.0504 | 0.0769 |
| | $\beta_2$ | 0.0552 | 0.0585 | 0.0538 | 0.0807 |
| | $\beta_3$ | -0.0027 | 0.0349 | -0.0010 | 0.0512 |
| | $\beta_4$ | 0.0014 | 0.0342 | 0.0005 | 0.0503 |
| | $\beta_5$ | -0.0023 | 0.0346 | -0.0007 | 0.0513 |
| Conditional Logit | $\beta_1$ | -0.0011 | 0.0416 | 0.0011 | 0.0613 |
| | $\beta_2$ | 0.0042 | 0.0434 | 0.0045 | 0.0632 |
| | $\beta_3$ | -0.0024 | 0.0358 | 0.0000 | 0.0523 |
| | $\beta_4$ | 0.0020 | 0.0367 | 0.0004 | 0.0522 |
| | $\beta_5$ | -0.0041 | 0.0334 | -0.0017 | 0.0517 |

Design 3 has fixed effects correlated with $x_1$ but not with the other explanatory variables and the true coefficients are as follows: $\beta_1 = \beta_2 = 1$ and $\beta_3 = \beta_4 = \beta_5 = 0$.

Table 4: Design 4

| | | Median bias | MAD | Mean bias | RMSE |
|---|---|---|---|---|---|
| *Full sample:* $n = 136$ | | | | | |
| Logit | $\beta_1$ | -0.2455 | 0.2455 | -0.2447 | 0.2479 |
| | $\beta_2$ | -0.0009 | 0.0221 | -0.0008 | 0.0323 |
| | $\beta_3$ | 0.0010 | 0.0218 | 0.0003 | 0.0328 |
| | $\beta_4$ | -0.0003 | 0.0229 | -0.0003 | 0.0323 |
| | $\beta_5$ | 0.0011 | 0.0221 | 0.0009 | 0.0326 |
| Logit FE | $\beta_1$ | 0.0200 | 0.0221 | 0.0199 | 0.0309 |
| | $\beta_2$ | -0.0003 | 0.0128 | -0.0003 | 0.0198 |
| | $\beta_3$ | 0.0006 | 0.0137 | 0.0000 | 0.0208 |
| | $\beta_4$ | 0.0021 | 0.0143 | 0.0016 | 0.0207 |
| | $\beta_5$ | 0.0005 | 0.0126 | 0.0008 | 0.0204 |
| Conditional Logit | $\beta_1$ | 0.0015 | 0.0163 | 0.0008 | 0.0242 |
| | $\beta_2$ | 0.0008 | 0.0134 | -0.0001 | 0.0202 |
| | $\beta_3$ | 0.0004 | 0.0144 | -0.0002 | 0.0212 |
| | $\beta_4$ | 0.0016 | 0.0142 | 0.0014 | 0.0207 |
| | $\beta_5$ | 0.0005 | 0.0138 | 0.0006 | 0.0206 |
| *Small sample:* $n = 68$ | | | | | |
| Logit | $\beta_1$ | -0.2447 | 0.2447 | -0.2430 | 0.2493 |
| | $\beta_2$ | -0.0013 | 0.0327 | 0.0002 | 0.0487 |
| | $\beta_3$ | 0.0011 | 0.0350 | 0.0010 | 0.0501 |
| | $\beta_4$ | -0.0015 | 0.0341 | -0.0012 | 0.0501 |
| | $\beta_5$ | 0.0045 | 0.0330 | 0.0019 | 0.0469 |
| Logit FE | $\beta_1$ | 0.0403 | 0.0449 | 0.0410 | 0.0637 |
| | $\beta_2$ | 0.0021 | 0.0306 | 0.0021 | 0.0438 |
| | $\beta_3$ | -0.0012 | 0.0275 | 0.0003 | 0.0404 |
| | $\beta_4$ | 0.0017 | 0.0274 | 0.0005 | 0.0409 |
| | $\beta_5$ | 0.0026 | 0.0288 | 0.0020 | 0.0431 |
| Conditional Logit | $\beta_1$ | -0.0011 | 0.0340 | 0.0016 | 0.0490 |
| | $\beta_2$ | 0.0012 | 0.0304 | 0.0016 | 0.0436 |
| | $\beta_3$ | 0.0009 | 0.0275 | 0.0010 | 0.0402 |
| | $\beta_4$ | 0.0004 | 0.0260 | 0.0007 | 0.0403 |
| | $\beta_5$ | 0.0033 | 0.0283 | 0.0023 | 0.0429 |

Design 4 has fixed effects drawn from a random normal uncorrelated with the $x$'s, there is resampling of the $x$'s and the true coefficients are as follows: $\beta_1 = 1$ and $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$.

bias [8] we see that for all designs, the conditional logit presented in this paper has smaller bias without sacrificing on the precision. As expected, sample size matters, both for the bias and the precision. Indeed, the median absolute deviation is approximately twice as large in the small sample for all three estimators in all designs. Moreover, while for the logit estimating the fixed effects, cutting the sample by half about doubles the median bias on the positive coefficient in all designs, it has no significant impact on the median bias for the conditional logit. Except in the first design, which has no fixed effects, the regular logit is severely biased, especially when the fixed effects are correlated with the explanatory variable. Design 1 shows that wrongly assuming that there are fixed effects when there are non can lead to biased estimates when using a logit estimating fixed effects, but not when using the conditional logit presented in this paper. Comparing the second and third designs, we observe that when the fixed effects are correlated with one of the explanatory variables, not only does it increase the bias on the coefficient of that variable for the logit estimating fixed effects, but it also causes bias on the other positive coefficient. Indeed, we see that $\beta_1$ has a median bias of 0.0266 compared to 0.0201 and $\beta_2$ a median bias of 0.0241 (in design 3). This is amplified in the small sample. Still comparing design 2 and 3, we further observe that the median bias remains similarly small for the conditional logit, showing that this estimator is equally capable of dealing with correlated and uncorrelated fixed effects. Finally, comparing designs 2 and 4, we see that resampling the $x$s in each replication has little to no effect on the results.

In short, these Monte Carlo simulations confirm that conditional logit presented in this paper is less biased, as precise, and more robust to different fixed effects than other logit estimators. Some preliminary simulations suggest similar results for the other estimators of section 2. We now move on to applying some of these estimators to trade data.

# 4 Application: Gravity equation and the extensive margins of trade

Understanding how different trade barriers influence trade flows is key when one wants to study the impact of distance, trade agreements, and other trade frictions. To do that, economists have been using the gravity equation for over 50 years. As Bernard et al. (2007) put it, "the gravity equation for bilateral trade flows is one of the most successful empirical relationships in international

[8] We are always mostly discussing results for the positive coefficient, since that is where the action takes place.

economics". The gravity equation was first applied to aggregate trade. As its name suggests, it was initially motivated by the Newtonian theory of gravitation (bilateral trade should be positively related to the size of countries, as measured by their GDP, and negatively related to their distance). It now has a plethora of microeconomic foundations. More recent work has emphasized the role of extensive margin adjustments in understanding the variations of aggregate trade flows[9] and has derived gravity equations for these extensive margin adjustments (see for example Bernard, Redding and Schott (2011) and Mayer, Melitz and Ottaviano (2012) ).

### 4.1 Data

For this application, we use the CEPII data (both the BACI and Gravity datasets) for 2005. This trade database is widely used, as in Keith Head and Thierry Mayer's work. We illustrate the estimation of two-way fixed effects using data on bilateral trade and including importer and exporter fixed effects for a single year. The year 2005 was chosen based on data availability but results are similar for other years. After merging the bilateral trade data with data on country characteristics such as distance, common border or colonial status, we obtain a balanced panel of 211 countries that account for the majority of world trade.

Table 5: Summary Statistics

| *Panel A* | | | | |
|---|---|---|---|---|
| | Trade | Trade1 | Trade3 | Trade5 |
| % zeros | 44.1729 | 51.2999 | 58.0749 | 61.7513 |
| *Panel B* | | | | |
| | Range | Mean | Median | Std. Dev. |
| pn if Trade=1 | [1,4941] | 273.738 | 20 | 647.315 |
| dni | [6,208] | 117.237 | 120 | 50.5608 |
| dnj | [8,208] | 117.237 | 114 | 55.9579 |

*pn* is the number of products (defined at the HS 6-digit level) traded.
$Trade = 1$ if $pn > 0$, otherwise 0.
$Trade1 = 1$ if $pn > 1$, otherwise 0.
$Trade3 = 1$if $pn3$, otherwise 0.
$Trade5 = 1$ if $pn > 5$, otherwise 0.
$dni(j)$: number of countries the importer (exporter) imports from (exports to).

[9] Trade frictions impact aggregate trade flows both through the amount each firm exports (the intensive margins) and the number of firms exporting (the extensive margin). Note that the extensive margin can also refer to the number of products exported.

Among these 211 countries, about 44% of unilateral trade flows are zeros. This is line with what Helpman, Melitz and Rubinstein find for a sample of 158 countries. If we restrict the definition positive trade for flows of more than one product, we see that the number of zeros exceeds 50% (and more for three or five products). This restriction is used as a robustness check in the logit estimation. This prominence of zeros in international trade further motivates the use of econometric models that can adequately deal with zeros.

Among countries that do trade, we see that the product number ranges from 1 to 4,941[10] , with a mean of 274 products, but a median of 20. Approximately 12% of positive trade flows are only composed of 1 product while only 26% of trade flows have more than 150 products. Finally, countries import from and export to 117 other countries on average.

## 4.2 Logit

There are many "zeros and ones" relationships in trade and the logit is very widely used. In the context of the gravity equation literature, the logit is most commonly used to study the extensive margins of trade in heterogeneous firms models. In an influential paper, Helpman, Melitz and Rubinstein (2008) try to improve on traditional estimates of gravity equation by accounting for both firm heterogeneity (in a Melitz (2003) framework) and the frequently forgotten zero trade flows. To do this, they use a two-stage procedure, where the first stage consists in estimating the probability that a country trades with another. Although they use a probit with importer and exporter fixed effects, we can argue that a logit would be more appropriate. Indeed, a probit with multiple fixed effects suffers from the incidental parameter problem and therefore the estimates are biased. In a paper estimating the Chaney (2008) model with French firm-level data, Crozet and Koenig (2010) also use the probability of exporting as a first stage in their empirical strategy. More specifically, they run a logit with firm and import country-year fixed effects to disentangle the elasticity of trade barriers on the intensive and extensive margins.

The following application shows that papers like Helpman, Melitz and Rubinstein (2008) and Crozet and Koenig (2010), as well as any other that uses a similar framework, should use the conditional logit to properly account for the multiple fixed effects. Indeed, estimates can differ

[10]There are 5,017 HS 6-digit products. Note also that, not surprisingly, the two countries that trade the largest number of products (4,941) are the US and Canada.

significantly. We write the probability of country $j$ exporting to country $i$ as

$$\text{Prob}[Trade_{ij}] = \beta_0+\beta_1 \ln(D_{ij}) + \beta_2 Border_{ij} + \beta_3 Legal_{ij} + \beta_4 Language_{ij} + \beta_5 Colony_{ij} + \beta_6 Currency_{ij} + \beta_7 RTA_{ij} + \mu_i + \alpha_j + \varepsilon_{ij} \tag{29}$$

where $D_{ij}$ is the simple distance between country $i$'s and country $j$'s most populated cities, $Border_{ij}$ is a dummy that takes the value 1 if $i$ and $j$ share a border, $Legal_{ij}$ is a dummy that takes the value 1 if the two countries have the same legal system,$Language_{ij}$ is a dummy that takes the value 1 if $i$ and $j$ have the same official language, $Colony_{ij}$ is a dummy that takes the value 1 if $i$ and $j$ were ever in a colonial relationship, $Currency_{ij}$ is a dummy that takes the value 1 if the two countries use the same currency, $RTA_{ij}$ is a dummy that takes the value 1 if $i$ and $j$ are in a regional trade agreement, and, finally, $\mu_i$ and $\alpha_j$ are respectively importer and exporter fixed effects. The results are presented in Table 6.

Table 6: Logit results (benchmark)

| Variables | OLS with FE | Logit | Logit with FE | Conditional Logit |
|---|---|---|---|---|
| Distance | -0.1116 | -0.6491 | -1.2146 | -0.2282 |
| | (0.0028) | (0.0169) | (0.0329) | (0.0461) |
| Border | -0.0765 | 0.0620 | -0.4194 | 1.4751 |
| | (0.0159) | (0.1606) | (0.2424) | (0.2743) |
| Legal | 0.0314 | 0.4098 | 0.2662 | 0.2989 |
| | (0.0044 ) | (0.0247) | (0.0422) | (0.0650) |
| Language | 0.0756 | -0.4588 | 0.7870 | 0.9452 |
| | (0.0056) | (0.0283) | (0.0597) | (0.0824) |
| Colonial ties | 0.0377 | 3.5425 | 0.8699 | 1.9481 |
| | (0.0149) | (0.3251) | (0.4086) | (0.5377) |
| Currency | 0.0540 | -0.5732 | 0.5458 | 0.0331 |
| | (0.0198) | (0.1386) | (0.2055) | (0.0066) |
| RTA | -0.0959 | 2.2365 | 1.3113 | 1.4140 |
| | (0.0082) | (0.1120) | (0.1529) | (0.1788) |

These results are for Prob[$Trade = 1$] where $Trade = 1$ when the number of products traded is greater than 0. OLS with FE refers to a simple linear probability model with fixed effects. The results in the second column are for the regular logit ignoring the fixed effects, the results in the third column are for the logit estimating all the fixed effects and finally, the results in the fourth column are for the estimator presented in this paper.

Standard errors clustered at $ij$ level (allowing for importer and exporter correlation).

The results for $\beta_1$ [11] differ greatly between estimators, more than what is expected when looking

[11] In what follows we mostly discuss results for the coefficient on distance, because it is generally the most talked

at the Monte Carlo simulations. Both estimators being relatively precise, this difference is puzzling, but suggests that distance might have a smaller impact on the probability of exporting than what traditional estimates indicate. Of course, caution should be exercised when comparing the results from this application with those of the Monte Carlo simulations since it is not clear how the data relates to the distributional assumptions made in the latter. Note that the conditional logit estimated effect of distance on the probability of exporting is closer to that produced by the linear probability model than that of the logit estimating the fixed effects.

Generally, the estimated effects of the other variables on the probability of exporting differ for the conditional logit and the logit estimating the fixed effects, but this difference, unlike that for distance, is much closer to what the Monte Carlo simulations suggested. One notable exception is the effect of a common border. Indeed, like the probit in Helpman, Melitz and Rubinstein (2008), the logit with fixed effects produces a surprising negative effect of sharing a border on the probability of trading. However, the conditional logit's positive estimate suggests that this result might be due to the inconsistency of the estimator. Note that the conditional logit is the only one of the four estimators where all estimated coefficients match their expected sign.

As a robustness check and to make sure that none of these estimators were strongly influenced by one outlier country, they were each calculated dropping each country in succession. The results are all very similar. Defining the dependent variable as "0" or "1" can give a heavy weight to relatively small trade flows: a country pair trading one product gets the same dependent variable value as a country pair trading 1,000 products. To test the robustness of the estimators to that issue, we recalculate all of them using three different definitions of positive trade (more than one product, more than three, and more than five). The results are presented in Table 7.

As predicted, considering trade as positive only when the number of products exported is greater than one increases the impact of distance on the probability of exporting. Accounting for a rough measure of trade flow size suggests that to trade many products, countries do have to be closer. The effect of distance becomes more and more negative as we increase the number of zeros by changing the $Trade$ variable. This is true for all estimators except the regular logit. This could be because, as the Monte Carlo simulations illustrate, the bigger the coefficient (in absolute value) the bigger the bias. Though respective coefficients for both the conditional logit

about trade barrier. *Border* is also commonly discussed, but as it is not significant in our estimations, it is of lesser interest.

Table 7: Logit results (robustness check)

| | Variables | OLS | Logit | Logit with FE | Conditional Logit |
|---|---|---|---|---|---|
| *Panel A: Trade1* | Distance | -0.1185 | -0.6389 | -1.4553 | -0.2502 |
| | | (0.0028) | (0.0163) | (0.0371) | (0.0492) |
| | Border | -0.0517 | 0.3069 | -0.1719 | 1.6626 |
| | | (0.0170) | (0.1542) | (0.2532) | (0.3545) |
| | Legal | 0.0225 | 0.3283 | 0.2721 | 0.3055 |
| | | (0.0043 ) | (0.0242) | (0.0440) | (0.0731) |
| | Language | 0.0945 | -0.4194 | 0.9819 | 1.1447 |
| | | (0.0054) | (0.0284) | (0.0638) | (0.0942) |
| | Colonial ties | 0.0491 | 3.5133 | 1.1530 | 2.2784 |
| | | (0.0155) | (0.2759) | (0.3703) | (0.5831) |
| | Currency | 0.0334 | -0.6519 | 0.6564 | 0.0372 |
| | | (0.0205) | (0.1315) | (0.2062) | (0.0086) |
| | RTA | -0.0524 | 2.1052 | 1.2354 | 1.6153 |
| | | (0.0084) | (0.0931) | (0.1402) | (0.1675) |
| *Panel B: Trade3* | Distance | -0.1207 | -0.6184 | -1.7427 | -0.2751 |
| | | (0.0028) | (0.0159) | (0.0428) | (0.0505) |
| | Border | -0.0142 | 0.5755 | 0.2205 | 1.7963 |
| | | (0.0178) | (0.1512) | (0.2772) | (0.4537) |
| | Legal | 0.0196 | 0.2644 | 0.3879 | 0.3772 |
| | | (0.0042 ) | (0.0243) | (0.0472) | (0.0833) |
| | Language | 0.1027 | -0.3710 | 1.0679 | 1.2772 |
| | | (0.0052) | (0.0290) | (0.0683) | (0.0989) |
| | Colonial ties | 0.0517 | 3.1438 | 1.1043 | 2.1905 |
| | | (0.0160) | (0.3251) | (0.3315) | (0.5520) |
| | Currency | 0.0235 | -0.5948 | 1.0211 | 0.0407 |
| | | (0.0208) | (0.1271) | (0.2249) | (0.0085) |
| | RTA | 0.0065 | 2.0604 | 1.2449 | 1.8341 |
| | | (0.0086) | (0.0812) | (0.1389) | (0.1673) |
| *Panel C: Trade5* | Distance | -0.1165 | -0.5965 | -1.8686 | -0.2788 |
| | | (0.0028) | (0.0158) | (0.0463) | (0.0535) |
| | Border | 0.0100 | 0.6822 | 0.4508 | 1.9881 |
| | | (0.0179) | (0.1460) | (0.2817) | (0.4265) |
| | Legal | 0.0151 | 0.2328 | 0.4026 | 0.3838 |
| | | (0.0041 ) | (0.0246) | (0.0495) | (0.0928) |
| | Language | 0.1104 | 1.2253 | 0.7870 | 1.4356 |
| | | (0.0051) | (0.0296) | (0.0728) | (0.0898) |
| | Colonial ties | 0.0538 | 3.0461 | 1.0601 | 1.9623 |
| | | (0.0165) | (0.1843) | (0.3205) | (0.7479) |
| | Currency | 0.0166 | -0.5604 | 1.2330 | 0.0372 |
| | | (0.0208) | (0.1260) | (0.2449) | (0.0092) |
| | RTA | 0.0448 | 2.0147 | 1.2473 | 1.9366 |
| | | (0.0086) | (0.0747) | (0.1405) | (0.1840) |

These results are for Prob[$Trade1 = 1$] where $Trade1 = 1$ when the number of products traded is greater than 1. Similarly for Trade3 (greater than 3) and Trade5 (greater than 5).Standard errors clustered at $ij$ level.

and the logit estimating fixed effects move in the same direction, they are getting farther apart, thus emphasizing our original concern about their large difference. As a final robustness check, we have done the same calculations with different measures of distance (i.e distance between capital cities, distance weighted by population, CES distance weighted by population). It does not affect the results significantly.

The large difference between estimators, especially between the conditional logit and the logit estimating the fixed effects, shows the importance of properly accounting for multiple fixed effects. On the one hand, all the results suggest that the impact of distance on the probability of exporting could be smaller than what we thought. On the other hand, however, this large difference can be cause for concern. Indeed, its magnitude is not in line with the Monte Carlo simulations. Therefore, it might indicate that the model is misspecified. If, for example, different countries had different $\beta_1$ or if distance was in fact interacted with something else, then both estimators would give an average $\beta_1$. Since each gives different weights to the same observations, this could explain why the estimates differ so much. It could then imply that the true $\beta_1$ is quite different from all the estimates. As a preliminary check, we relaxed the functional form assumptions on distance by using non-parametric dummies for quartiles of the distance distribution. The results did not suggest any problem and were very much in line with the original results presented in Tables 6 and 7. Be that as it may, and whether or not the model is misspecified, this application illustrates the pertinence of computing the conditional logit estimator for two fixed effects.

### 4.3 Poisson

The Poisson maximum likelihood estimator is very widely used in international trade, particularly so in the gravity equation literature since Santos Silva and Tenreyro (2006)'s influential paper. Although often applied to a continuous variable (for example trade flows), we focus here on the count data application. Indeed, while the Poisson maximum likelihood estimator with one fixed effect can be implemented for a continuous dependent variable, the same cannot be said for the conditional Poisson with two fixed effects developed in this paper. This does not mean that there isn't a wide variety of count-data cases where we can use it.

In the gravity equation literature, the study of the extensive margin provides a good example of count data to which we can apply a Poisson estimator. Hummels and Klenow (2005) use the number of products as a measure of the extensive margin of trade and find that it accounts for about 60 percent of the greater exports of large economies. However, they do use a count-data model but rather the log of the extensive margin in a linear regression. Bernard et al. (2007) also use the number of products as part of the measure of the extensive margin of trade (the other part being the number of firms exporting) in their multiproduct firm model. They show that adjustments on the product margin have a great impact on aggregate trade flows, thus highlighting the importance of properly measuring it. When firm-level data is available, one can also use count-data models on the number of destination countries for each firm to estimate destination-specific fixed cost of exporting, as in Eaton, Kortum and Kramarz (2011) .

The following application shows that papers like Hummels and Klenow (2005) and Bernard et al. (2007), as well as any other that uses a similar framework, should use the conditional Poisson to properly account for the multiple fixed effects. Indeed, estimates can differ significantly. We estimate the following model:

$$\begin{aligned} Pn_{ij} \sim Poisson(\exp(\beta_0 + \beta_1 \ln(D_{ij}) + \beta_2 Border_{ij} + \beta_3 Legal_{ij} + \beta_4 Language_{ij} \\ + \beta_5 Colony_{ij} + \beta_6 Currency_{ij} + \beta_7 RTA_{ij} + \mu_i + \alpha_j) \end{aligned} \tag{30}$$

where $Pn_{ij}$ is the number of 6-digit HS products county $j$ exports to country $i$ and the other variables are as defined before. The results are presented in Table 8.

Once again, properly accounting for the multiple fixed effects by using the conditional Poisson estimator developed in this paper produces significantly different estimates of the gravity equation. Results are in line with the literature, though perhaps a little smaller in absolute value than what is found in papers like Helpman, Melitz and Rubinstein (2008) and Santos Silva and Tenreyro (2006). That might be due to the larger number of countries used here or on the count data specification (these other papers look at continuous trade flows). Focusing on the distance coefficient, we see that the two OLS estimates and the Poisson follow the same ordering and relative size generally found in the literature (see for example Santos Silva and Tenreyro (2006)). The significantly smaller distance elasticity found with the Poisson Pseudo-Maximum-Likelihood approach advertised by Santos Silva and Tenreyro (2006), has been the subject of some debate. Table 8 indicates that this might be due to the use of a Poisson estimator that does not appropriately deal with importer and exporter fixed effects. Indeed, the coefficient found with the conditional Poisson estimator is much closer to

Table 8: Poisson results

| Estimator | OLS FE | OLS FE | Poisson | Poisson FE | Conditional Poisson |
|---|---|---|---|---|---|
| Dependent variable | $\ln(Pn_{ij})$ | $\ln(1+Pn_{ij})$ | $Pn_{ij}$ | $Pn_{ij}$ | $Pn_{ij}$ |
| Distance | -0.9214 | -0.6232 | -0.4106 | -0.5528 | -0.7845 |
| | (0.0132) | (0.0102) | (0.0199) | (0.0165) | (0.0047) |
| Border | 0.5009 | 0.7206 | 0.3079 | -0.1741 | 0.5231 |
| | (0.0736) | (0.0760) | (0.0715) | (0.0612) | (0.1874) |
| Legal | 0.1598 | 0.0850 | -0.0419 | 0.2113 | 0.2406 |
| | (0.0165) | (0.0130 ) | (0.0346) | (0.0202) | (0.0194) |
| Language | 0.6414 | 0.5201 | -0.2032 | 0.3387 | 0.8116 |
| | (0.0247) | (0.0183) | (0.0434) | (0.0353) | (0.0898) |
| Colonial ties | 0.6824 | 0.6869 | 1.4207 | 0.4615 | 0.6485 |
| | (0.0617) | (0.0699) | (0.0749) | (0.0553) | (0.1409) |
| Currency | 0.2070 | -0.1512 | 0.2030 | -0.2331 | 0.0580 |
| | (0.0830) | (0.0737) | (0.0916) | (0.0728) | (0.0061) |
| RTA | 0.3919 | 0.9574 | 1.2591 | 0.2363 | 0.3615 |
| | (0.0296) | (0.0294) | (0.0456) | (0.0368) | (0.0370) |

Note that the OLS in the first column does not include the "zero" trade flows.
Standard errors clustered at $ij$ level (allowing for importer and exporter correlation).

more standard results. The same holds for the coefficients on the other variables.

Since the Poisson estimate can be sensitive to the distance measure used, as a robustness check we computed all estimators for different measures of distance. All results were very similar.

# 5 Conclusion

This paper looked at estimators of nonlinear panel data models with multiple fixed effects. There is an abundance of empirical methods applying two fixed effects in nonlinear models. However, current estimators are subject to the incidental parameters problem. Although many methods have been developed to address this problem in models with a single fixed effect, very little has been done for the cases with two or more fixed effects. Attempting to fill this important gap, we developed methods to appropriately deal with two fixed effects for a broad class of nonlinear models.

In general, we still use the conditional maximum likelihood based on Rasch (1960, 1961). If with one fixed effect it suffices to condition on the sum of the observations in one dimension (typically, for one individual, the sum of $y_{it}$ over time), with two fixed effects we condition in both dimensions (for one importer $i$, the sum of $y_{ij}$ for all exporters $j$; for one exporter $j$, the sum of $y_{ij}$ over all importers $i$). This approach allows us to consistently estimate the parameters of interest for the logit, Poisson, Negative binomial and Gamma models with two fixed effects. However, we found that if the same conditioning can be used on Manski's maximum score estimator with a single fixed effect, the method fails when there are two. That estimator cannot be generalized to the case of two fixed effects. For the more general multiplicative form model with two fixed effects, we used a slightly different approach. We developed a moment condition that does not rely on either of the two fixed effects to give us a consistent estimator.

We showed that the conditioning method that is the core of this paper performs well in recovering the true parameters in Monte Carlo studies. Indeed, we found that the conditional logit presented in this paper is less biased, as precise, and more robust to different fixed effects than other logit estimators. We also showed that this same procedure yields quite different estimated coefficients from methods subject to the incidental parameters problem in applications with actual trade data. That is true when applying both the logit and Poisson models to gravity-type equations with importer and exporter fixed effects.The coefficients estimated by our procedure are typically closer to those estimated using OLS.

The methods developed in this paper have broad applicability and significant impact when used. Our applications highlight the importance of appropriately controlling for multiple fixed effects in nonlinear panel data models for recovering the parameters of the underlying relationships of interest.

The work is however not over. We still need to investigate the asymptotic properties of the estimators developed in this paper. For this, we plan to use the results for panels with one fixed effects where $n$ and $T$ grow at the same rate. We are also working on the censored regression model for which the one fixed effect case was solved by Honoré (1992) and Honoré and Powell (1994) . Finally, we want to expand the data application to a broader range of models.

## References


Aaronson, D., L. Barrow, and W. Sander (2007): "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics*, 25(1), 95–135.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High wage workers and high wage firms," *Econometrica*, 67(2), 251–333.

Andersen, E. (1970): "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society*, 32, 283–301.

——— (1972): "The Numerical Solution of a Set of Conditional Estimation Equations," *Journal of the Royal Statistical Society*, 34, 42–54.

Anderson, J. E., and E. van Wincoop (2003): "Gravity with Gravitas: A Solution to the Border Puzzle," *The American Economic Review*, 93(1), 170–192.

Arellano, M., and B. E. Honoré (2001): "Panel Data Models: Some Recent Developments," *Handbook of Econometrics*, 5(53), 3229–3296.

Bernard, A. B., J. B. Jensen, S. J. Redding, and P. K. Schott (2007): "Firms in International Trade," *NBER working paper*.

Bernard, A. B., S. J. Redding, and P. K. Schott (2011): "Multiproduct Firms and Trade Liberalization," *The Quaterly Journal of Economics*, 126(3), 1271–1318.

Chaney, T. (2008): "Distorted Gravity: The Intensive and Extensive Margins of International Trade," *The American Economic Review*, 98(4), 1707–1721.

Crozet, M., and P. Koenig (2010): "Structural gravity equations with intensive and extensive margins," *Canadian Journal of Economics*, 43(1).

Eaton, J., S. Kortum, and F. Kramarz (2011): "An Anatomy of International Trade: Evidence from French Firms," *Econometrica*, 79(5), 1453–1498.

Hahn, J., and W. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72(4), 1295–1319.

Hausman, J., B. H. Hall, and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents-RD Relationship," *Econometrica*, 52(4), 909–938.

Helpman, E., M. Melitz, and Y. Rubinstein (2008): "Estimating Trade Flows:Trading Partners and Trading Volumes," *The Quaterly Journal of Economics*, CXXIII(2), 441–187.

Hirji, K. F., C. R. Mehta, and N. R. Patel (1987): "Computing Distributions for Exact Logistic Regression," *Journal of the American Statistical Association*, 82(400), 1110–1117.

Honoré, B. E. (1992): "Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60(3), 533–565.

Honoré, B. E., and J. L. Powell (1994): "Pairwise difference estimators of censored and truncated regression models," *Journal of Econometrics*, 64, 241–278.

Hummels, D., and P. J. Klenow (2005): "The Variety and Quality of a Nations' Exports," *The American Economic Review*, 95(3), 704–723.

Lancaster, T. (2002): "Orthogonal Parameters and Panel Data," *The Review of Economic Studies*, 69(3), 647–666.

Manski, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3(3), 205–228.

——— (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55(2), 357–362.

Melitz, M., T. Mayer, and G. I. P. Ottaviano (2012): "Market Size, Competition, and the Product Mix of Exporters," *Working Paper*.

Melitz, M. J. (2003): "The Impact of Trade on Intra-Industry Reallocation and Aggregate Industry Productivity," *Econometrica*, 71, 1695–1725.

Neyman, J., and E. Scott (1948): "Consistent estimates based on partially consistent observations," *Econometrica*, 16, 1–32.

Rasch, G. (1960): "Probabilistic Models for Some Intelligence and Attainment Tests," *Denmarks Paedagogiske Institut, Copenhagen.*

——— (1961): "On the general laws and the meaning of measurement in psychology," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistis and Probability*, 4.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005): "Teachers, Schools, and Academic Achievement," *Econometrica*, 73(2), 417–158.

Silva, J. M. C. S., and S. Tenreyro (2006): "The Log of Gravity," *The Review of Economics and Statistics*, 88(4), 641–658.

# A Appendix

We show here the existence of the incidental parameter problem for the simple $2 \times 2$ model where the fixed effects sum up to zero. In this case, the log-likelihood function is given by:

$$\begin{aligned}\ell(\alpha,\eta,b) =& y_{11}(b_0 + x_{11}b_1 + \alpha + \eta) - \exp(b_0 + x_{11}b_1 + \alpha + \eta) + y_{12}(b_0 + x_{12}b_1 + \alpha - \eta) \\ & - \exp(b_0 + x_{12}b_1 + \alpha - \eta) + y_{21}(b_0 + x_{21}b_1 - \alpha + \eta) - \exp(b_0 + x_{21}b_1 - \alpha + \eta) \\ & + y_{22}(b_0 + x_{22}b_1 - \alpha - \eta) - exp(b_0 + x_{22}b_1 - \alpha - \eta)\end{aligned}$$

We first maximize with respect to the fixed effects to find the two first order conditions:

$$\begin{aligned}\frac{\partial \ell(\cdot)}{\partial \alpha} =& - \exp(\alpha)\big[\exp(\eta)\exp(b_0 + x_{11}b_1) + \exp(-\eta)\exp(b_0 + x_{12}b_1)\big] \\ & + \exp(-\alpha)\big[\exp(\eta)\exp(b_0 + x_{21}b_1) + \exp(-\eta)\exp(b_0 + x_{22}b_1)\big] + (y_{11} + y_{12} - y_{21} - y_{22})\end{aligned} \tag{31}$$

$$\begin{aligned}\frac{\partial \ell(\cdot)}{\partial \eta} =& - \exp(\eta)\big[\exp(\alpha)\exp(b_0 + x_{11}b_1) + \exp(-\alpha)\exp(b_0 + x_{21}b_1)\big] \\ & + \exp(-\eta)\big[\exp(\alpha)\exp(b_0 + x_{12}b_1) + \exp(-\alpha)\exp(b_0 + x_{22}b_1)\big] + (y_{11} - y_{12} + y_{21} - y_{22})\end{aligned} \tag{32}$$

Adding (31) and (32) we get

$$2y_{11} - 2y_{22} - 2\exp(\alpha)\exp(\eta)\exp(b_0 + x_{11}b_1) + 2\exp(-\alpha)\exp(-\eta)\exp(b_0 + x_{22}b_1) = 0$$

which can be rewritten as

$$-\exp(\alpha)^2\exp(\eta)^2\exp(b_0 + x_{11}b_1) + (y_{11} - y_{22})\exp(\alpha)\exp(\eta) + \exp(b_0 + x_{22}b_1) = 0$$

This allows us to solve for $\exp(\alpha)\exp(\eta)$ to get:

$$\exp(\alpha)\exp(\eta) = \frac{(y_{22} - y_{11}) \pm \sqrt{(y_{11} - y_{22})^2 + 4\exp(b_0 + x_{11}b_1)\exp(b_0 + x_{22}b_1)}}{-2\exp(b_0 + x_{11}b_1)} \tag{33}$$

Similarly we can get

$$\exp(-\alpha)\exp(-\eta) = \frac{(y_{22}-y_{11}) \pm \sqrt{(y_{11}-y_{22})^2 + 4\exp(b_0+x_{11}b_1)\exp(b_0+x_{22}b_1)}}{2\exp(b_0+x_{22}b_1)} \tag{34}$$

Because both equations (33) and (34) must be positive, we can determine the sign in front of the square root. The term in the square root is larger than $(y_{22}-y_{11})$. Therefore, the square root must be subtracted from the numerator of equation (33) and added in equation (34).

Now, if we subtract (32) from (31), we get

$$2y_{12} - y_{21} - 2\exp(\alpha)\exp(-\eta)\exp(b_0+x_{12}b_1) + \exp(-\alpha)\exp(\eta)\exp(b_0+x_{21}b_1) = 0$$

which can be rewritten as

$$-\exp(\alpha)^2\exp(-\eta)^2\exp(b_0+x_{12}b_1) + (y_{12}-y_{21})\exp(\alpha)\exp(\eta) + \exp(b_0+x_{21}b_1) = 0$$

This in turn allows us to write

$$\exp(\alpha)\exp(-\eta) = \frac{(y_{21}-y_{12}) \pm \sqrt{(y_{12}-y_{21})^2 + 4\exp(b_0+x_{12}b_1)\exp(b_0+x_{21}b_1)}}{-2\exp(b_0+x_{12}b_1)} \tag{35}$$

and

$$\exp(-\alpha)\exp(\eta) = \frac{(y_{21}-y_{12}) \pm \sqrt{(y_{12}-y_{21})^2 + 4\exp(b_0+x_{12}b_1)\exp(b_0+x_{21}b_1)}}{2\exp(b_0+x_{21}b_1)} \tag{36}$$

We can do a similar analysis as for equations (33) and (34) to determine the unknown sign in the numerator of equations (35) and (36). It will reveal that the square root must be subtracted in the numerator of equation (35) and added in equation (36).

Now taking the derivative of the objective function with respect to $b_1$ we find:

$$\begin{aligned}\frac{\partial \ell(\cdot)}{\partial b_1} =& y_{11}x_{11} + y_{12}x_{12} + y_{21}x_{21} + y_{22}x_{22} \\ &- x_{11}\exp(b_0+x_{11}b_1)\exp(\alpha)\exp(\eta) - x_{12}\exp(b_0+x_{12}b_1)\exp(\alpha)\exp(-\eta) \\ &- x_{21}\exp(b_0+x_{21}b_1)\exp(-\alpha)\exp(\eta) - x_{22}\exp(b_0+x_{22}b_1)\exp(-\alpha)\exp(-\eta)\end{aligned}$$

Using the envelope theorem and substituting in equations (33)−(36) we can write:

$$
\begin{aligned}
\frac{\partial \ell(\cdot)}{\partial b} =& y_{11}x_{11} + y_{12}x_{12} + y_{21}x_{21} + y_{22}x_{22} \\
&+ \frac{1}{2}x_{11}\big[(y_{22} - y_{11}) - \sqrt{(y_{11} - y_{22})^2 + 4\exp(b_0 + x_{11}b_1)\exp(b_0 + x_{22}b_1)}\big] \\
&+ \frac{1}{2}x_{12}\big[(y_{21} - y_{12}) - \sqrt{(y_{12} - y_{21})^2 + 4\exp(b_0 + x_{12}b_1)\exp(b_0 + x_{21}b_1)}\big] \\
&- \frac{1}{2}x_{21}\big[(y_{21} - y_{12}) + \sqrt{(y_{12} - y_{21})^2 + 4\exp(b_0 + x_{12}b_1)\exp(b_0 + x_{21}b_1)}\big] \\
&- \frac{1}{2}x_{22}\big[(y_{22} - y_{11}) + \sqrt{(y_{11} - y_{22})^2 + 4\exp(b_0 + x_{11}b_1)\exp(b_0 + x_{22}b_1)}\big] \qquad (37)
\end{aligned}
$$

To show that this expression is not maximized at the true $\beta_1$, we will first show that if we could simply replace the $y$s by their expected value, equation (37) would indeed be equal to zero. We will then look at the expected value of equation (37) to find that Jensen's inequality causes us to either overestimate or underestimate the parameter of interest, depending on some values of the $x$s.

Let us evaluate equation (37) when replacing the $y$s by their expected value[12] and assuming $b_1$ is the true value. To do that, first define:

$$
\begin{aligned}
sqrt_{11-22} &\equiv \sqrt{(\exp(b_0 + x_{11}b_1)\exp(\alpha)\exp(\eta) + \exp(b_0 + x_{22}b_1)\exp(-\alpha)\exp(-\eta))^2}\big] \\
sqrt_{12-21} &\equiv \sqrt{(\exp(b_0 + x_{12}b_1)\exp(\alpha)\exp(\eta) + \exp(b_0 + x_{21}b_1)\exp(-\alpha)\exp(-\eta))^2}\big]
\end{aligned}
$$

We then have:

$$
\begin{aligned}
\frac{\partial \ell(\cdot)}{\partial b} =& \exp(b_0 + x_{11}b_1)\exp(\alpha)\exp(\eta)x_{11} + \exp(b_0 + x_{12}b_1)\exp(\alpha)\exp(-\eta)x_{12} \\
&+ \exp(b_0 + x_{21}b_1)\exp(-\alpha)\exp(\eta)x_{21} + \exp(b_0 + x_{22}b_1)\exp(-\alpha)\exp(-\eta)x_{22} \\
&+ \frac{1}{2}x_{11}\big[(\exp(b_0 + x_{22}b_1)\exp(-\alpha)\exp(-\eta) - \exp(b_0 + x_{11}b_1)\exp(\alpha)\exp(\eta)) - sqrt_{11-22} \\
&+ \frac{1}{2}x_{12}\big[(\exp(b_0 + x_{21}b_1)\exp(-\alpha)\exp(-\eta) - \exp(b_0 + x_{12}b_1)\exp(\alpha)\exp(\eta)) - sqrt_{12-21} \\
&- \frac{1}{2}x_{21}\big[(\exp(b_0 + x_{21}b_1)\exp(-\alpha)\exp(-\eta) - \exp(b_0 + x_{12}b_1)\exp(\alpha)\exp(\eta)) + sqrt_{12-21} \\
&- \frac{1}{2}x_{22}\big[(\exp(b_0 + x_{22}b_1)\exp(-\alpha)\exp(-\eta) - \exp(b_0 + x_{11}b_1)\exp(\alpha)\exp(\eta)) + sqrt_{11-22}
\end{aligned}
$$

[12] e.g. $y_{11} = \exp(\alpha)\exp(\eta)\exp(b_0 + x_{11}b_1)$

This simplifies to

$$
\begin{aligned}
\frac{\partial \ell(\cdot)}{\partial b} =& \exp(b_0 + x_{11}b_1)\exp(\alpha)\exp(\eta)x_{11} + \exp(b_0 + x_{12}b_1)\exp(\alpha)\exp(-\eta)x_{12} \\
&+ \exp(b_0 + x_{21}b_1)\exp(-\alpha)\exp(\eta)x_{21} + \exp(b_0 + x_{22}b_1)\exp(-\alpha)\exp(-\eta)x_{22} \\
&- \big[\exp(b_0 + x_{11}b_1)\exp(\alpha)\exp(\eta)x_{11} + \exp(b_0 + x_{12}b_1)\exp(\alpha)\exp(-\eta)x_{12} \\
&+ \exp(b_0 + x_{21}b_1)\exp(-\alpha)\exp(\eta)x_{21} + \exp(b_0 + x_{22}b_1)\exp(-\alpha)\exp(-\eta)x_{22}\big] \\
=&0
\end{aligned}
$$

Hence, if the expectation of the function in (37) was equivalent to evaluating the function at the expectation, the likelihood would be maximized at the true parameter value. However, we know by Jensen's inequality that this is not true[13]. Indeed, if $E[\frac{1}{2}x_{11}(y_{22} - y_{11})] = \frac{1}{2}x_{11}E[y_{22} - y_{11}]$, we have

$$
\begin{aligned}
&E[\frac{1}{2}x_{11}\sqrt{(y_{11} - y_{22})^2 + 4\exp(b_0 + x_{11}b_1)\exp(b_0 + x_{22}b_1)}] \\
&\qquad \geq \frac{1}{2}x_{11}\sqrt{(E[y_{11} - y_{22}])^2 + 4\exp(b_0 + x_{11}b_1)\exp(b_0 + x_{22}b_1)}
\end{aligned}
\tag{38}
$$

if $x_{11}$ is positive, and,

$$
\begin{aligned}
&E[\frac{1}{2}x_{11}\sqrt{(y_{11} - y_{22})^2 + 4\exp(b_0 + x_{11}b_1)\exp(b_0 + x_{22}b_1)}] \\
&\qquad \leq \frac{1}{2}x_{11}\sqrt{(E[y_{11} - y_{22}])^2 + 4\exp(b_0 + x_{11}b_1)\exp(b_0 + x_{22}b_1)}
\end{aligned}
\tag{39}
$$

if $x_{11}$ is negative. Rewriting equation (37) as

$$
\begin{aligned}
\frac{\partial \ell(\cdot)}{\partial b} =& y_{11}x_{11} + y_{12}x_{12} + y_{21}x_{21} + y_{22}x_{22} \\
&+ \frac{1}{2}(x_{11} - x_{22})(y_{22} - y_{11}) - \frac{1}{2}(x_{11} + x_{22})\sqrt{(y_{11} - y_{22})^2 + 4\exp(b_0 + x_{11}b_1)\exp(b_0 + x_{22}b_1)} \\
&+ \frac{1}{2}(x_{12} - x_{21})(y_{21} - y_{12}) - \frac{1}{2}(x_{12} + x_{21})\sqrt{(y_{12} - y_{21})^2 + 4\exp(b_0 + x_{12}b_1)\exp(b_0 + x_{21}b_1)}
\end{aligned}
$$

and using the results from equations (38) and (39), we can conclude that

[13] It is not true in general at least, but could hold for some value of $x$ or for some very specific distribution of the $x$'s.

- if $(x_{11} + x_{22} > 0)$ and $(x_{12} + x_{21} > 0)$ then

$$E[\frac{\partial \ell(\cdot)}{\partial b_1}|_{\beta}] < 0$$

- if $(x_{11} + x_{22} < 0)$ and $(x_{12} + x_{21} < 0)$ then

$$E[\frac{\partial \ell(\cdot)}{\partial b_1}|_{\beta}] > 0$$

- but if $(x_{11} + x_{22} < 0)$ and $(x_{12} + x_{21} > 0)$ or $(x_{11} + x_{22} > 0)$ and $(x_{12} + x_{21} < 0)$ then it depends on the relative size of the different variables and parameters.

This shows that a Poisson model with two fixed effects does suffer from the incidental parameter problem and confirms the necessity of using the conditional estimator developed in the present paper.