Method Sheet - Analyzing the determinants of firms' location choices using count models

2023-10-12

Introduction

This document presents count model estimation methods and their application to STATA in the context of measuring and analyzing the determinants of firms' location choices in Tunisia. In particular, the aim is to analyze the role of mobile Internet broadband telecom infrastructures in firms' location choices.

There are two main methods of analyzing firm location in the economic literature: discrete choice models and count models. When there is a large number of zeros in the main outcome variable, which consists of the number of a certain event, and the choice set is large, count models should be preferred.

In applying a count model to the analysis of firms' location choices, the aim is to model the number of new firms in a particular geographical area and over a certain period as a function of the characteristics of that geographical area. So, when a firm is set up and chooses to locate somewhere, the attractiveness of an area will depend on agglomeration forces, but also on accessibility, or even labor market factors (human capital and labor costs). In developed countries, count models have been used to understand these factors (Jofre-Monseny, Marín-López, and Viladecans-Marsal 2011 ; Bhat, Paleti, and Singh 2014), and more specifically the role of fixed broadband Internet infrastructure in Ireland (McCoy et al. 2018) or France (Hasbi 2020).

Count models

Framework

Count models are based on a dependent variable which is count data:

- The dependent variable is a count (a non-negative integer): $y = 0, 1, 2, 3, 4...$
- The sample is concentrated on a few small discrete values.

The explained variable is discrete and non-negative. On the explanatory variable side, we want to study the factors affecting the mean number of the dependent variable. Returning to our research question, we use the number of new firms in a particular area in each period, which we model as a function of the characteristics of the geographical area.

Estimation using a Poisson model

Count models are commonly estimated using a Poisson model. This captures the discrete, non-negative nature of the data. Thus, the Poisson model predicts the number of occurrences of an event. The Poisson model states that the probability of the dependent variable Y being equal to a certain number y is:

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}$$

We're looking for the probability that Y is equal to a certain integer, which can be $0, 1, 2....$ This is the form of the Poisson distribution. For the Poisson model, μ is the intensity or rate parameter:

$$\mu = \exp(X_i^{'}\beta)$$

This is how the independent variables are introduced into the (X_i) model (they appear several times in the formula).

These coefficients can therefore be interpreted as follows: *A one-unit increase in x will increase/decrease the mean number of the dependent variable of the coefficient expressed as a percentage.*

Fixed-effects Poisson model

This model accounts for un-observable variables that do not change over time (*the fixed effects*) while using the Poisson distribution to predict counts i.e. the number of events occurring.

$$E[y_{it}|\alpha_i, X_i t] = \alpha_i \exp(X_{it}^{'}\beta) = \exp(\gamma_i + X_{it}^{'}\beta)$$

This equation specifies how the expected count y_{it} for an entity i at time t relates to some observed variables $X_i t$ and unobserved entity-specific effect γ_i. This equation says that the expected count is a function of some observed variables and an unobserved individual effect, both affecting the count multiplicatively.

When using the fixed-effect Poisson model, there is no problem of "incidental parameter problem". The Maximum Likelihood Estimation (MLE) remains consistent, just like in linear regression.

Properties and limitations of the Poisson model

- The **equidispersion property** of the Poisson distribution: equality of mean and variance :

$$E(y|x) = var(y|x) = \mu$$

This property is restrictive and often doesn't apply in practice, leading to **"overdispersion"** in the data. In such cases, we have to use the negative binomial model.

The *excess zeros problem* of the Poisson distribution: there are usually more zeros in the data than the Poisson model predicts. In these cases, use the zero-inflated Poisson model (or negative binomial).

Marginal effects for the Poisson model :

- The marginal effect of a variable on the average number of events is :

$$\delta E(y|x)/\delta x_j = \beta_j \exp(x_i^{'}\beta)$$

Interpretation of marginal effects: a one-unit increase in x will increase/decrease the mean number of the dependent variable of the marginal effect.

Negative binomial model

The negative binomial model is used with count data instead of the Poisson model if there is overdispersion in the data (Consul and Jain 1973). Unlike the Poisson model, the negative binomial model has the less restrictive property that the variance is not equal to the mean (μ) :

$$var(y|x) = \mu + \alpha\mu^2$$

Another form is $var(y|x) = \mu + \alpha\mu$, but this is less widely used. The negative binomial model also estimates the overdispersion parameter α.

Test de surdispersion :

- Estimate the negative binomial model, which includes the overdispersion parameter α, and test whether α is significantly different from zero.
- $H_0 : \alpha = 0$ or $H_a : \alpha \neq 0$

There are three possible cases: - when $\alpha = 0$, the Poisson model applies. - when $\alpha > 0$, overdispersion (often observed in real data). - when $\alpha < 0$, underdispersion (rare).

Incidence rate ratios (irr)

For the Poisson and negative binomial models, in addition to reporting the coefficients and marginal effects, we can also report the incidence rate ratios. The incidence rate ratios report exp(b) rather than b.

Then the interpretation of the incidence rate ratios is as follow : *irr=2 means that for each unit increase in x, the expected number of y will double.*

Risk of using Negative Binomial with fixed effects

Negative Binomial with Fixed Effects (NBFE) estimators should be used with great caution for a number of reasons. There is no true NBFE estimator (Allison and Waterman 2002). As Guimarães (2008) shows, the negative binomial conditional fixed effects model for count panel data does not control for individual fixed effects except under a quite specific set of assumptions.

It is even recommended that the Poisson approach with fixed effects be preferred to that of the NBFE estimator. The Poisson estimator (FEP) is totally robust to any type of relationship between the variance and the mean, so there is no need to 'correct' for overdispersion with FEP. It is simply necessary to calculate robust standard errors i.e. conditional mean needs to be correct. Furthermore, the question of whether there is overdispersion cannot be resolved with panel data by examining the raw data, or even the data conditional on x[1].

[1] https://www.statalist.org/forums/forum/general-stata-discussion/general/1539401-testin-overdispersion-in-negative-binomial

Application on Stata

Poisson with fixed effects model

```
xtset Year Location
```

Conditional fixed-effects model with robust standard error

```
xtpoisson y x i.a, fe vce(robust)
```

Si on veut reporter les incidence-rate ratios :

```
xtpoisson y x i.a, vce(robust) irr
```

Population-averaged model with robust standard errors

```
xtpoisson y x i.a, pa vce(robust)
```

Other command (ppmlhdfe) which implements Poisson pseudo-maximum likelihood regressions[2] (Correia, Guimarães, and Zylkin 2020).

```
ppmlhdfe y x, absorb(i.a) vce(robust)
```

Interpretations

Example with results of location choices

The estimated model:

$$Y_{it} = \alpha + \beta MBB_{it-2} + X_{it-2} + Z_{it-2} + \mu year_t + \eta_i + \epsilon_{it}$$

With:

- Y_{it} the count of new establishments creating in delegation i at time t
- MBB_{it-2}: the variable of interest which is a proxy of mobile broadband Internet quality in delegation i in time $t-2$. The main outcome is the number of antennas per people.
- X_{it-2}: a matrix of location characteristics for delegation i at time $t-2$ including agglomeration, competition and accessibility variables.
- Z_{it-2}: a matrix of labor market characteristics for municipality i at time t including human capital and labor cost proxies.
- μ: fixed effect capturing year specific effects.
- η_i: time invariant fixed effect controling for differences across delegations.
- ϵ_{it}: standard error clustered at the delegation level, capturing unobserved factors.

[2] http://scorreia.com/help/ppmlhdfe.html

Results on STATA

```
xtpoisson ncrea dens_4G_pop_2 r_dip_3_25_2 r_revenu_25_2 r_unemployment_25_2 nfirmdeleg_2 hhi_emploi_de
```

```
Conditional fixed-effects Poisson regression         Number of obs    =  1,032
Group variable: delegation                           Number of groups =    258

                                                     Obs per group:
                                                                  min =      4
                                                                  avg =    4.0
                                                                  max =      4

                                                     Wald chi2(6)     = 236.19
Log pseudolikelihood = -5235.9959                    Prob > chi2      = 0.0000

                                (Std. err. adjusted for clustering on delegation)
------------------------------------------------------------------------------
             |               Robust
       ncrea | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
 dens_4G_pop_2 |    .017881   .0048199     3.71   0.000     .0084344     .027328
  r_dip_3_25_2 |    1.20446   1.277017     0.94   0.346    -1.298449    3.707368
 r_revenu_25_2 |  -.0022138   .0002463    -8.99   0.000    -.0026966    -.001731
r_unemployment_2 |   2.420338   .4574821     5.29   0.000     1.523689    3.316986
 nfirmdeleg_2 |  -.0000673   .0000303    -2.23   0.026    -.0001266   -8.02e-06
 hhi_emploi_2 |   .1718548   .1218215     1.41   0.158    -.0669109    .4106205
------------------------------------------------------------------------------
```

Interpretation: We can reveal a positive impact of mobile broadband quality: when the density of 4G antennas per 10,000 inhabitants increases by one unit, the percentage change in the rate of new business creation increases by 1.8% ($e^{0.0179}$).

Population-averaged model with robust standard errors

```
xtpoisson ncrea dens_4G_pop_2 r_dip_3_25_2 r_revenu_25_2 r_unemployment_25_2 lpop_2 popdensity_2 dist_u
```

References

Allison, Paul D., and Richard P. Waterman. 2002. "7. Fixed-Effects Negative Binomial Regression Models." *Sociological Methodology* 32 (1): 247–65. https://doi.org/10.1111/1467-9531.00117.

Bhat, Chandra R., Rajesh Paleti, and Palvinder Singh. 2014. "A SPATIAL MULTIVARIATE COUNT MODEL FOR FIRM LOCATION DECISIONS." *Journal of Regional Science* 54 (3): 462–502. https://doi.org/10.1111/jors.12101.

Consul, P. C., and G. C. Jain. 1973. "A Generalization of the Poisson Distribution." *Technometrics* 15 (4): 791–99. https://doi.org/10.1080/00401706.1973.10489112.

Correia, Sergio, Paulo Guimarães, and Tom Zylkin. 2020. "Fast Poisson Estimation with High-Dimensional Fixed Effects." *The Stata Journal: Promoting Communications on Statistics and Stata* 20 (1): 95–115. https://doi.org/10.1177/1536867X20909691.

Guimarães, Paulo. 2008. "The Fixed Effects Negative Binomial Model Revisited." *Economics Letters* 99 (1): 63–66. https://doi.org/10.1016/j.econlet.2007.05.030.

Hasbi, Maude. 2020. "Impact of Very High-Speed Broadband on Company Creation and Entrepreneurship: Empirical Evidence." *Telecommunications Policy* 44 (3): 101873. https://doi.org/10.1016/j.telpol.2019.101873.

Jofre-Monseny, Jordi, Raquel Marín-López, and Elisabet Viladecans-Marsal. 2011. "The Mechanisms of Agglomeration: Evidence from the Effect of Inter-Industry Relations on the Location of New Firms." *Journal of Urban Economics* 70 (2-3): 61–74. https://doi.org/10.1016/j.jue.2011.05.002.

McCoy, Daire, Sean Lyons, Edgar Morgenroth, Donal Palcic, and Leonie Allen. 2018. "The Impact of Broadband and Other Infrastructure on the Location of New Business Establishments." *Journal of Regional Science* 58 (3): 509–34. https://doi.org/10.1111/jors.12376.