PPMLHDFE: Fast Poisson Estimation with High Dimensional Fixed Effects

Paulo Guimarães

2020 Portuguese Stata Conference

Introduction

- Poisson regression is the standard approach to model count data
- alternative for multiplicative models where the dependent variable is nonnegative
- only assumption required for consistency is the correct specification of the conditional mean of the dependent variable
- Poisson regression vs Poisson pseudo maximum likelihood (PPML) regression

Advantages of PPML

- dependent variable with nonnegative values
- no need to specify a distribution for the dependent variable
- natural way to deal with zero values on the dependent variable
- Unlike log linear OLS, it is robust to heteroskedasticidity

Why is OLS sometimes preferred?

- sometimes researchers resort to log-linear regressions in contexts where PPML would be better justified
- one reason is ability to estimate linear regressions with multiple fixed effects
- Stata users are familiar with the user-written package *reghdfe*
- *reghdfe* (Sergio Correia) is the state-of-the-art tool for estimation of linear regression models with HDFE
- But PPML with HDFE can be implemented with (almost) the same ease as linear regression with HDFE

Generalized Linear Models

- GLMs are a class of regression models based on the exponential family of distributions (Nelder,1972)
- GLMs include popular nonlinear regression models such as logit, probit, cloglog, and Poisson
- the exponential family is given by

$$f_y(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

where a(.), b(.), and c(.), are specific functions and ϕ and θ are parameters

Generalized Linear Models (cont.)

- for these models

$$E(y) = \mu = b'(\theta)$$

  and

$$V(y) = b''(\theta)a(\phi).$$

- given a set of n independent observations, each indexed by i, the expected value can be related to a set of covariates ($\mathbf{x}_i$) by means of a link function g(.). More specifically it is assumed that

$$E(y_i) = \mu_i = g^{-1}(\mathbf{x}_i\beta),$$

  and the likelihood for the GLM may be written as

$$L(\theta, \phi; y_1, y_2, ..., y_n) = \prod_{i=1}^{n} exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$$

Estimation

- application of the Gauss-Newton algorithm with the expected Hessian leads to the following updating equation:

  $$\beta^{(r)} = \left(\mathbf{X}'\mathbf{W}^{(r-1)}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{(r-1)}\mathbf{z}^{(r-1)},$$

  where $\mathbf{X}$ is the design matrix of explanatory variables, $\mathbf{W}^{(r-1)}$ is a weighting matrix, $\mathbf{z}^{(r-1)}$ is a transformation of the dependent variable, and r is an index for iteration obtained by recursive application of weighted least squares
- this approach is known as Iteratively Reweighted Least Squares (IRLS)

The Poisson regression model

- for Poisson regression we have

$$E(y_i) = \mu_i = \exp(\mathbf{x}_i \beta)$$

and the regression weights to implement IRLS simplify to

$$\mathbf{W}^{(r-1)} = \text{diag}\left\{\exp(\mathbf{x}_i \beta^{(r-1)})\right\}$$

while the dependent variable for the intermediary regression becomes

$$z_i^{(r-1)} = \left\{\frac{y - \exp(\mathbf{x}_i \beta^{(r-1)})}{\exp(\mathbf{x}_i \beta^{(r-1)})} + \mathbf{x}_i \beta^{(r-1)}\right\}$$

Dealing with HDFE

- $\mathbf{X}$ may contain a large number of fixed effects that render the direct calculation of $(\mathbf{X}'\mathbf{W}^{(r-1)}\mathbf{X})$ impractical, if not impossible
- the solution is to use an alternative updating formula that estimates only the coefficients of the non-fixed effect covariates (say, δ)
- we can rely on the FWL theorem to expurgate the fixed effects and use the following updating equation:

$$\delta^{(r)} = \left(\widetilde{\mathbf{X}}'\mathbf{W}^{(r-1)}\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}'\mathbf{W}^{(r-1)}\widetilde{\mathbf{z}}^{(r-1)},$$

where $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{z}}$ are weighted within-transformed versions of the main covariate matrix $\mathbf{X}$ and working dependent variable $\mathbf{z}$, respectively

Existence of MLE

- MLE for Poisson regression may not exist and algorithms may be unable to converge or converge to incorrect estimates
- problem identified by Santos Silva and Tenreyro (2010)
- Correia Guimaraes and Zylkin (2018) discuss the necessary and sufficient conditions for the existence of estimates in a wide class of GLM models
- CGZ show that for the case of Poisson regression it is always possible to find MLE estimates if some observations are dropped from the sample
- these observations are called separated observations because they do not convey relevant information for the estimation process and can be safely discarded
- CGZ propose a method to identify separated observations that will succeed even in the presence of HDFEs

ppmlhdfe

- ppmlhdfe - Poisson pseudo-likelihood regression with multiple levels of fixed effects
- authored by Sergio Correia, Paulo Guimaraes and Tom Zylkin
- requires the installation of the latest versions of *ftools* and *reghdfe*

ppmlhdfe (cont)

- same flexibility as *reghdfe* allowing for multiple fixed effects and interactions
- allows weights, multi-way clustered standard errors, and count model specific options such as *exposure* and *irr*
- takes great care to verify the existence of maximum-likelihood estimates

Accelerating HDFE-IRLS

- *ppmlhdfe* directly embeds the Mata routines of *reghdfe*
- we within-transform (or *partial out*) the original untransformed variables **z** and **X** in the first IRLS iteration only and progressively update these variables
- the criterion for the inner loops of *reghdfe* becomes tighter as we approach convergence
- in practice, these innovations can reduce the total number of calls to *reghdfe* by 50% or more

Final Notes

- dedicated github website: ppmlhdfe
- (forthcoming) article in Stata Journal describing command usage
- the approach could be easily extended to any other model from the GLM family