Verifying the existence of ML estimates for GLMs

Sergio Correia (Federal Reserve Board)
Paulo Guimarães (Banco de Portugal, CEFUP, and IZA)
Thomas Zylkin (Robins School of Business, University of Richmond)

July 12, 2019

STATA Conference — University of Chicago

paper: https://arxiv.org/abs/1903.01633

examples: https://github.com/sergiocorreia/ppmlhdfe/blob/master/guides/

Motivation: why should we use generalized linear models?

- Practitioners often prefer least squares when seemingly better alternatives exist. Examples:
    - Linear probability model instead of logit/probit
    - Log transformations instead of Poisson
- This comes with several disadvantages:
    - Inconsistent estimates under heteroskedasticity due to Jensen's inequality; bias can be quite severe (Manning and Mullahy 2001; Santos Silva and Tenreyro 2006; Nichols 2010)
    - Linear models might lead to a wrong support: predicted probabilities outside [0-1], log(0), etc.

Digression: genesis of this paper

- We wanted to run pseudo-ML poisson regressions with fixed effects:
    - Paulo: $\log(1 + wages)$
    - Tom: $\log(1 + trade)$
    - Sergio: $\log(1 + credit)$
- Should have been feasible:
    - No incidental parameters problem in many standard panel settings (Wooldridge 1999; Fernández-Val and Weidner 2016; Weidner and Zylkin 2019)
    - Works with non-count variables (Gourieroux, Monfort, and Trognon 1984)
    - Practical estimator through IRLS and alternating projections (Guimarães 2014; Correia 2017; Larch et al. 2019)
- However, there was another obstacle we did not anticipate:
    - Our implementation sometimes failed to converge, or converged to incorrect solutions.
    - Problem was aggravated when working with many levels of fixed effects (our intended goal)

How can maximum likelihood estimates *not* exist?

Consider a Poisson regression on a simple dataset without constant:

- Log-likelihood: $\mathcal{L}(\beta) = \sum[y_i(x_i\beta) - \exp(x_i\beta) - \log(y_i!)]$
- FOC: $\sum x_i[y_i - \exp(x_i\beta)] = 0$

y	x
0	1
0	1
0	0
1	0
2	0
3	0

- In this example, the FOC becomes $\exp(\beta) = 0$, maximized only at infinity!
    - Note that at infinity the first two observations are fit perfectly, with $\mathcal{L}_i = 0$
- More generally, non-existence can arise from any **linear combination of regressors** including fixed effects.

Existing literature

- Non-existence conditions have been independently (re)discovered multiple times:
  - Log-linear frequency table models (Haberman 1974)
  - Binary choice (Silvapulle 1981; Albert and Anderson 1984)
  - GLM sufficient–but–not–necessary conditions (Wedderburn 1976; Santos Silva and Tenreyro 2010)
  - GLM (Verbeek 1989; Geyer 1990, 2009; Clarkson and Jennrich 1991 - all three unaware of each other).
- Most researchers still unaware of problem outside of binary choice models; no textbook mentions as of 2019.
  - Software implementations either fail to converge or inconspicuously converge to wrong results.

Our contribution

1. Derive existence conditions for a broader class of models than in existing work
    - Including Gamma PML, Inverse Gaussian PML
2. Clarify how to correct for non-existence of *some* parameters.
    - Finite components of β can be consistently estimated; inference is possible
3. Introduce a novel and easy-to-implement algorithm that detects and corrects for non-existence
    - Particularly useful with high-dimensional fixed effects and partialled-out covariates.
    - Can be implemented with run–of–the-mill tools.
    - programmed in our new HDFE PPML command `ppmlhdfe` (Correia, Guimarães, and Zylkin 2019)

Proposition 1: non-existence conditions (1/4)

Consider the class of GLMs defined by the following log-likelihood function:

$$\mathcal{L} = \sum_i \mathcal{L}_i = \sum_i \left[a(\phi)\, y_i\, \theta_i - a(\phi)\, b(\theta_i) + c(y_i, \phi)\right]$$

- a, b, and c are known functions; ϕ is a scale parameter
- $\theta_i = \theta(x_i\beta)$ is the canonical link function; where $\theta' > 0$
- $y_i \geq 0$ is an outcome variable. Potentially $y \leq \bar{y}$ as in logit/probit but for simplicity we'll ignore this for the most part.
- Its conditional mean is $\mu_i = E[y_i|x_i] = b'(\theta_i)$
- Assume for simplicity that regressors X have full column rank.
- Assume that $\mathcal{L}_i$ has a finite upper bound (***rules out e.g. log link Gamma PML***)

ML solution for β will **not** exist iff there is a non-zero vector γ such that:

$$x_i\gamma = z_i \begin{cases} \leq 0 & \text{if } y_i = 0 \\ = 0 & \text{if } 0 < y_i < \bar{y} \\ \geq 0 & \text{if } y_i = \bar{y} \end{cases}$$

Proposition 1: non-existence conditions (2/4)

ML solution for β will **not** exist iff there is a non-zero vector γ such that:

$$x_i\gamma = z_i \begin{cases} \leq 0 & \text{if } y_i = 0 \\ = 0 & \text{if } 0 < y_i < \bar{y} \\ \geq 0 & \text{if } y_i = \bar{y} \end{cases}$$

Intuition If $\exists$ a linear combination of regressors $z_i = x_i\gamma$ satisfying these conditions, then

$$\frac{d\mathcal{L}(\beta + k\gamma^*)}{dk} = \sum_{y_i=0} \alpha_i \left[-b'(\theta_i)\right] \theta' z_i + \sum_{y_i=\bar{y}} \alpha_i \left[\bar{y} - b'(\theta_i)\right] \theta' z_i > 0,$$

for any $k > 0$, which implies we can always increase the objective function by searching in the direction described by γ^*.

Proposition 1: non-existence conditions (3/4)

ML solution for β will **not** exist iff there is a non-zero vector γ such that:

$$x_i\gamma = z_i \begin{cases} \leq 0 & \text{if } y_i = 0 \\ = 0 & \text{if } 0 < y_i < \bar{y} \\ \geq 0 & \text{if } y_i = \bar{y} \end{cases}$$

Poisson PML example For PPML, $\bar{y} = \infty$, and only the first two conditions matter

$$\frac{d\mathcal{L}(\beta + k\gamma^*)}{dk} = \sum_{y_i=0} - \exp\left(x_i\beta + kz_i\right) z_i + \sum_{y_i>0} \left[y_i - \exp\left(x_i\beta\right)\right] z_i > 0,$$

Note the second term is 0 and the first term is positive and asymptotically decreasing towards 0 as $k \to \infty$ (finite solution for β not possible!)

Proposition 1: non-existence conditions (4/4)

- Linear combination z is a “certificate of non-existence”: hard to obtain, but can be used to verify non-existence
  - If we add z to the regressor set, its associated FOC will not have a finite solution.
- Observations where $z_i \neq 0$ will be perfectly predicted 0’s and $\bar{y}$’s
- If $\mathcal{L}_i$ is unbounded above, conditions are more complex (and ultimately less innocuous)
  - See proposition 2 of the paper.

Addressing non-existence

- As in perfect collinearity, first look for specification problems:
    - In a Poisson wage regression, did we add "unemployment benefits" as covariate?
    - In a Poisson trade regression, did we add an "is embargoed?" indicator?
- If no specification problems, it's due to sampling error
- Solution: allow estimates to take values in the extended reals: $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$
    - Permits solutions like this: $\hat{\beta}_1 = \lim_{a \to \infty} a + 3$, $\hat{\beta}_2 = \lim_{a \to \infty} a + 2$, $\hat{\beta}_3 = 1.5$
    - We are mostly interested in the non-infinite components: $\hat{\beta}_1 - \hat{\beta}_2 = 1$, $\hat{\beta}_3 = 1.5$
    - Can show "separated" observations drop out of FOC's for finite $\hat{\beta}$'s (including that of $\hat{\beta}_1 - \hat{\beta}_2$)

Proposition 3: Addressing non-existence

- Given a $\mathcal{L}_i$ bounded above, a unique ML solution in the extended reals will always exist.
- Given a z identifying all instances of non-existence, if we first drop perfectly predicted observations (and resulting perfectly collinear variables) ML solution **in the reals** will always exist.
  - It will consistently estimate the non-infinite components of β, allowing for inference on them (proposition 3d)
  - We can recover infinite components by regressing z against x.

Obtaining z: Existing Alternatives

1. Drop boundary observations with $\mathcal{L}_i$ close to 0 (Clarkson and Jennrich 1991)
    - Slow under non-existence; often fails as "close to 0" is data specific.
2. Solve a modified simplex algorithm (Clarkson and Jennrich 1991)
    - Cannot handle fixed effects or other high-dimensional covariates
3. Analytically solve computational geometry problem (Geyer 2009), or use eigenvalues of Fischer information matrix (Eck and Geyer 2018).
    - Extremely slow and complex (Geyer 2009); requires full working with full information matrix (Eck and Geyer 2018); cannot handle fixed effects (both).

None works well with fixed effects!

Obtaining z: Iterative Rectifier (our algorithm)

1. Define a working dependent variable $z_i = \mathbb{1}_{y_i=0}$
2. Given an arbitrarily large integer K, set weights $w_i = \begin{cases} 1 & \text{if } y_i = 0 \\ K & \text{if } y_i > 0 \end{cases}$
3. (Weighted least squares) Regress z on X with weights w (fixed effects no problem!)
4. Stop if all $\hat{z}_i \geq 0$
5. Else, update $z_i = max(\hat{z}_i, 0)$ and repeat from step 3

- Steps 2-3 are the "weighting method" of solving least squares with equality constraints (Stewart 1997); step 5 is a "rectifier" that enforces a positive dependent variable
- Proofs in proposition 4 and appendix
- Stata implementation in our `ppmlhdfe` package ; examples at our github
- Convergence usually achieved in a few iterations, but choosing weights too large could lead to numerical instability.

Other existing approaches

- Naïve approach: drop the regressors causing non-existence and proceed as usual
  - Leads to nonsensical results (Zorn 2005; Gelman et al. 2008)
- Penalize estimates beyond plausible values (Firth regression, Bayesian aproach)
  - "For Poisson regression and other models with the logarithmic link, we would not often expect effects larger than 5 on the logarithmic scale" (Gelman et al 2008)
  - Not a ML estimator
  - Many datasets (e.g. in trade) can have plausible effects way beyond 5.
- Solutions specific to binary choice discussed in Konis (2007)

Comparison of solutions

Method	Advantages	Concerns
1. Drop regressors	-	Nonsensical
2. Drop $\mu_i < \varepsilon$ observations	Simple	Fails often: ε is data dependent
3. Bayesian: penalize $\mu_i < \varepsilon$	It's Bayesian	It's Bayesian. ε is data dependent
4. Modified simplex	Fast for small k	Slow for large k Can't handle FEs
5. Directions of recession	Exact answer "at infinity"	Complex, very slow (?) Can't handle FEs
6. Iterative rectifier	Simple works well with large k and FEs	Numerical accuracy (?)

Example (1/3)

y	x1	x2
0	2	-1
0	-1	2
0	0	0
1	0	0
2	5	-10
3	6	-12

- The first $y = 0$ value in this data set is “separated” by the linear combination $z = 2x_1 + x_2$.
- In theory, the coefficients for x_1 and x_2 are both infinite, but we can still obtain a finite estimate for the transformed parameter $\beta_1 - 2\beta_2$
    - Math + interpretation are analogous to the case of perfect collinearity

Current workhorse Stata commands like `poisson` and `ppml` either fail to converge or give incorrect estimates.

- `poisson` does not converge.
- `ppml` recognizes there is a problem, but incorrectly attributes it to the regressor x_1:

```
. ppml y x1 x2

note: checking the existence of the estimates

Number of regressors excluded to ensure that the estimates exist: 1
Excluded regressors:  x1
Number of observations excluded: 0
```

y	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
x2	-.2050602	.0794416	-2.58	0.010	-.3607629	-.0493576
_cons	-1.359994	.8990047	-1.51	0.130	-3.122011	.4020225

Example (3/3)

Here is an example of how `ppmlhdfe` handles this situation. The `sep(ir)` option specifies we want to use our "IR" algorithm.

```
. ppmlhdfe y x1 x2, sep(ir)
(ReLU method dropped 1 separated observation in 1 iterations)
note: 1 variable omitted because of collinearity: x2
PPML regression                                   No. of obs      =          5
                                                  Residual df     =          3
                                                  Wald chi2(1)    =       6.04
Deviance             =  1.993162063               Prob > chi2     =     0.0140
Log pseudolikelihood = -4.799356454               Pseudo R2       =     0.3506

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |    .356474   .1450006     2.46   0.014      .072278    .6406699
          x2 |          0  (omitted)
       _cons |  -1.049287   .8213881    -1.28   0.201    -2.659178    .5606042
------------------------------------------------------------------------------
```

There are lots of other options as well (can use simplex method instead, can ask `ppmlhdfe` to compute the contents of z). Read more here.

Conclusion

Non-existence of estimates:

- Affects a broad class of GLMs beyond just binary choice models
- Poorly understood (no textbook mentions); not addressed in statistical packages
- Leads practitioners to stay with least squares despite limitations

This paper:

- Presents non-existence conditions for a broad class of GLMs
- Discusses how to address non-existence: drop perfectly predicted observations, then proceed as normal
- Introduces an algorithm for detecting and addressing non-existence that is conceptually simple, easy-to-implement, and allows for fixed effects

New “fast” FE-PPML command `ppmlhdfe` incorporates our methods: `ssc install ppmlhdfe`

References i

Albert, A., and J. A. Anderson. 1984. "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika* 71 (1): 1–10. https://doi.org/10.2307/2336390.

Clarkson, Douglas B., and Robert I. Jennrich. 1991. "Computing Extended Maximum Likelihood Estimates for Linear Parameter Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 53 (2): 417–26. http://www.jstor.org/stable/2345752.

Correia, Sergio. 2017. "Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator." Unpublished manuscript.

Correia, Sergio, Paulo Guimarães, and Thomas Zylkin. 2019. "Ppmlhdfe: Fast Poisson Estimation with High-Dimensional Data." *arXiv Preprint arXiv:1903.01690*, April.

Eck, Daniel J, and Charles J Geyer. 2018. "Computationally Efficient Likelihood Inference in Exponential Families When the Maximum Likelihood Estimator Does Not Exist." *arXiv Preprint arXiv:1803.11240*.

Fernández-Val, Iván, and Martin Weidner. 2016. "Individual and Time Effects in Nonlinear Panel Models with Large N, T." *Journal of Econometrics* 192 (1): 291–312. https://doi.org/10.1016/j.jeconom.2015.12.014.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *Annals of Applied Statistics* 2 (4): 1360–83. https://doi.org/10.1214/08-AOAS191.

Geyer, Charles J. 1990. "Likelihood and Exponential Families." PhD thesis, University of Washington.

———. 2009. "Likelihood Inference in Exponential Families and Directions of Recession." *Electronic Journal of Statistics* 3: 259–89.

Gourieroux, Author C, A Monfort, and A Trognon. 1984. "Pseudo Maximum Likelihood Methods: Theory." *Econometrica* 52 (3): 681–700.

Guimarães, Paulo. 2014. "POI2HDFE: Stata Module to Estimate a Poisson Regression with Two High-Dimensional Fixed Effects." Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s457777.html.

Haberman, Shelby J. 1974. *The Analysis of Frequency Data*. Vol. 4. University of Chicago Press.

Konis, Kjell. 2007. "Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models." PhD thesis, University of Oxford.

Larch, Mario, Joschka Wanner, Yoto V. Yotov, and Thomas Zylkin. 2019. "Currency Unions and Trade: A Ppml Re-Assessment with High-Dimensional Fixed Effects." *Oxford Bulletin of Economics and Statistics*. https://doi.org/10.1111/obes.12283.

Manning, Willard G., and John Mullahy. 2001. "Estimating Log Models: To Transform or Not to Transform?" *Journal of Health Economics* 20 (4): 461–94. https://doi.org/10.1016/S0167-6296(01)00086-8.

Nichols, Austin. 2010. "Regression for Nonnegative Skewed Dependent Variables." In *BOS10 Stata Conference*, 2:15–16. Stata Users Group.

Santos Silva, J. M. C., and Silvana Tenreyro. 2006. "The Log of Gravity." *Review of Economics and Statistics* 88 (4): 641–58. https://doi.org/10.1162/rest.88.4.641.

Santos Silva, Joao M C, and Silvana Tenreyro. 2010. "On the Existence of the Maximum Likelihood Estimates in Poisson Regression." *Economics Letters* 107 (2): 310–12. https://doi.org/10.1016/j.econlet.2010.02.020.

Silvapulle, Mervyn J. 1981. "On the Existence of Maximum Likelihood Estimators for the Binomial Response Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 43 (3): 310–13. http://www.jstor.org/stable/2984941.

Stewart, Gilbert W. 1997. "On the Weighting Method for Least Squares Problems with Linear Equality Constraints." *BIT Numerical Mathematics* 37 (4): 961–67.

Verbeek, Albert. 1989. "The Compactification of Generalized Linear Models." In *Statistical Modelling*, 314–27. Springer.

Wedderburn, R. W. M. 1976. “On the Existence and Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models.” *Biometrika* 63 (1): 27–32. https://doi.org/10.1093/biomet/63.1.27.

Weidner, Martin, and Thomas Zylkin. 2019. “Bias and Consistency in Three-Way Gravity Models.” Unpublished manuscript.

Wooldridge, Jeffrey M. 1999. “Distribution-Free Estimation of Some Nonlinear Panel Data Models.” *Journal of Econometrics* 90 (1): 77–97.

Zorn, Christopher. 2005. “A Solution to Separation in Binary Response Models.” *Political Analysis* 13 (2): 157–70. https://doi.org/10.1093/pan/mpi009.