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)

June 28, 2019

North American Summer Meeting of the Econometric Society

Motivation: why should we use generalized linear models?

- Practitioners often prefer least squares when seemingly better alternatives exist. Example:
    - 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 & Mullahy 2001, Santos Silva & Tenreyro 2006, etc.)
    - 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 (Wooldridge 1999, Fernandez-Val and Weidner 2016), Weidner and Zylkin 2019)
    - Works with non-count variables (Gourieroux et al 1984)
    - Practical estimator through IRLS and alternating projections (Guimarães 2014, Correia 2017, Zylkin et al 2018)
- However, there was another obstacle we did not anticipate:
    - Our implementation often 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 1973, 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, Geyer 2008, Clarkson and Jenrich 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.

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

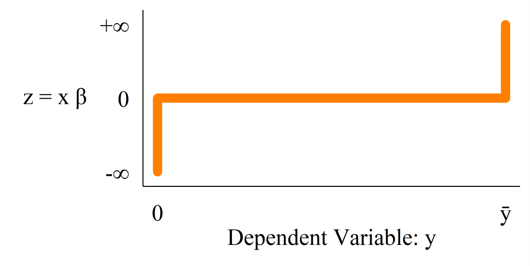
Consider the class of GLMs defined by exponential log-likelihood functions:

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

- 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 through this talk.
- 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.

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 (3/3)

- 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 solution.
- Observations where $z_i \neq 0$ will have a perfect fit.
- If $\mathcal{L}_i$ is unbounded above, conditions are slightly more complex; 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: $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$
    - Example: $\hat{\beta} = \lim_{a \to \infty} a + 3$
    - We are mostly interested in the non-infinite component

Proposition 3: Addressing non-existence

- Given a $\mathcal{L}_i$ bounded above, ML solution in the extended reals will always exist.
- Given vector z identifying all instances of non-existence, if we first drop perfectly predicted observations (and resulting perfect 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 Jenrich 1991)
    - Slow under non-existence; often fails as "close to 0" is data specific.
2. Solve a modified simplex algorithm (Clarkson and Jenrich 1991)
    - Cannot handle fixed effects or other partialled-out covariates
3. Analytically solve computational geometry problem (Geyer 2008), or use eigenvalues of Fischer information matrix (Eck and Geyer 2018).
    - Extremely slow and complex (Geyer 2008); might not converge (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; potentially allowing for fixed effects
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 `ppmhdfe` package
- 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 non-sensical results (Zorn 2005, Gelman 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 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 (?)

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