Handling Sparse Non-negative Data in Finance*

Agostino Capponi† Zhaonan Qu‡

March 21, 2025

Latest Version Here

Abstract

Recent studies recommend Poisson and other pseudo maximum likelihood (PML) estimators over log-linear regression to model count or non-negative count-like dependent variables in finance. However, estimator selection is often ad hoc, with Poisson used as a rule of thumb. We show this approach is suboptimal and propose a general class of moment estimators, deriving the optimal one by balancing bias-variance trade-offs while accounting for heteroskedasticity and sparsity. A simple cross-validation procedure selects the best estimator. Applied to corporate defaults, credit ratings, patents, and real estate investments, our method reveals that the optimal choice varies significantly and often diverges from Poisson.

Keywords: Count-like Data; Heteroskedasticity; Sparsity; Pseudo Maximum Likelihood Estimation; Bias-variance Trade-off

*We are grateful for stimulating conversations with Charles-Albert Lehalle, Han Hong, Garud Iyengar, Yanguang Liu, and Johan Ugander, and to Ron Bekkerman and Davide Proserpio for sharing the dataset on residential real estate investment, as well as for technical support from Moody's Corporation for access to their Default and Recovery Database (DRD).

†Columbia University, Department of Industrial Engineering and Operations Research

‡Columbia University, Data Science Institute

1 Introduction

Data with non-negative outcome (dependent) variables are ubiquitous in applications across a diverse range of disciplines, including economics (Hausman et al., 1984; Card, 2001), finance (Coles et al., 2006; Hirshleifer et al., 2012; Cohn et al., 2022), health services (Dow and Norton, 2003), epidemiology (Holford, 1980; Böhning et al., 1999), political science (King, 1989), and sociology (Boulton and Williford, 2018). Specific examples include count data, such as corporate defaults or patents granted, and continuous count-like data, such as health expenditures or sales revenue. A common approach to modeling non-negative data, particularly in finance and economics, is log-linear regression, i.e., regressing a log transformation of the outcome variable on explanatory variables (He and Tian, 2013; Akey and Appel, 2021). To deal with log of zeros, this approach typically shifts the outcome by 1 before applying the log transformation. An alternative approach is the pseudo maximum likelihood (PML) estimation based on count models, most often the Poisson regression (Gourieroux et al., 1984a; Wooldridge, 1999).[1]

Despite the popularity of log-linear regression, there has been increasing evidence that Poisson PML may be a better approach to modeling non-negative data. For example, Cohn et al. (2022); Chen and Roth (2024) raise the issue of lack of meaningful interpretation of log-linear estimates, while King (1988); Silva and Tenreyro (2006, 2011); Cohn et al. (2022); Mullahy and Norton (2022) demonstrate the fragility of log-linear regression in the presence of *heteroskedasticity* or *sparsity*, which are two prominent features of count and non-negative data in finance. As a solution, many works recommend using Poisson PML as a more robust alternative. This development has had profound impacts on empirical research in corporate finance, especially technological innovation, as effect estimates could change significantly when switching to a Poisson specification, which is being widely accepted in recent empirical research (Sautner et al., 2023; Addoum et al., 2023; Hollingsworth et al., 2024).

[1] In the econometrics literature, some works use the term "quasi-maximum likelihood" (White, 1982; Wooldridge, 1999). In this paper, we follow the terminology of Gourieroux et al. (1984a) and Silva and Tenreyro (2006) and use "pseudo maximum likelihood".

However, is Poisson PML *always* the best option when handling non-negative data? Previous works have studied this question by either focusing on the heteroskedasticity or the sparsity of the data. For example, the efficiency properties of Poisson and other PML estimators under heteroskedasticity are well-known and depend on the *dispersion* of data (Gourieroux et al., 1984b; Wooldridge, 1999). On the other hand, when data exhibits high levels of sparsity, which may stem from censoring or the infrequent occurrence of events, Silva and Tenreyro (2006, 2011) provide simulation evidence on the robustness of Poisson PML compared to log-linear regression and other PML estimators, while zero-inflated (Lambert, 1992) and hurdle models (Mullahy, 1986) have also been suggested and studied. Crucially, non-negative data across numerous finance and economics applications often display both sparsity and heteroskedasticity, as also demonstrated in our empirical analysis in Section 5. Despite this, research works that concurrently address both characteristics are uncommon, which drives the central question of our paper: What is the most effective method for managing non-negative data where both sparsity and heteroskedasticity are significant factors?

In this work, we show that it is imperative to simultaneously take into account sparsity and heteroskedasticity when modeling non-negative data, rather than treating these features in isolation. More precisely, we show that the choice of an estimator depends on the relative magnitude of the two features, which creates a bias-variance trade-off. To precisely characterize this trade-off, we propose a novel class of moment estimators that encompasses traditional PML estimators, such as Poisson PML, as special cases. These estimators place heterogeneous weights on observations in the estimating equations. On one hand, heteroskedasticity typically encourages larger weights on samples with smaller conditional means to improve efficiency. On the other hand, the presence of many zeros discourages such weighting to reduce bias. As a result, there exists an estimator that optimally balances these two aspects of the data. Fig. 1 illustrates how the optimal estimator depends on different levels of sparsity and heteroskedasticity in a simulation setting (see Section 4 for details). It is worth noting that this trade-off is a unique feature of data with sparse non-negative

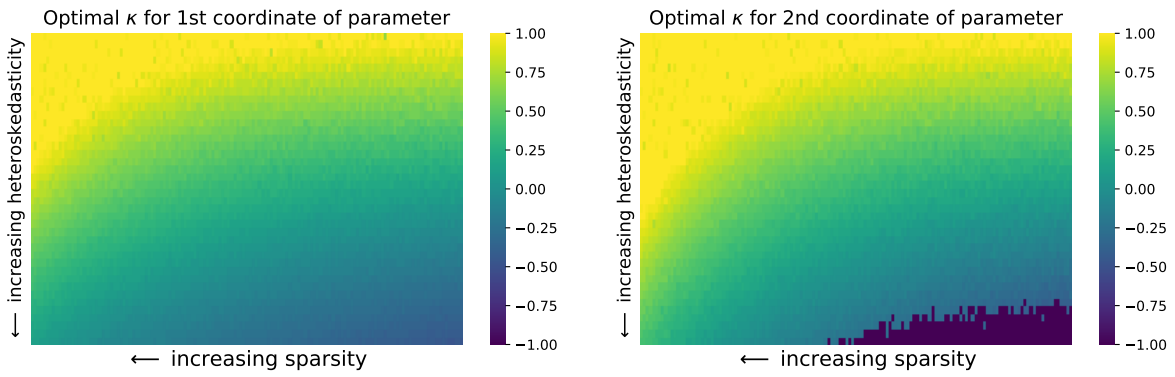


Figure 1: Phase transition of the optimal estimator for different levels of heteroskedasticity and sparsity. Estimators are indexed by a parameter taking values between -1 and 1, which controls the relative weight the moment estimator assigns to samples with larger conditional means vs. smaller conditional means. Poisson PML corresponds to a value of 0, non-linear least squares corresponds to 1, while gamma PML corresponds to -1. When heteroskedasticity is low relative to sparsity, a larger value is optimal. Poisson is only optimal for a small region in the figures, while non-linear squares and Poisson can also be optimal.

outcome variables, and is distinct from the conventional bias-variance trade-off in statistics and machine learning that relies on the notion of model complexity.

To guide empirical researchers in determining the optimal choice of estimator, we propose a simple data-driven cross-validation procedure. We apply our framework across diverse finance datasets, including those related to corporate defaults, credit ratings, corporate patents, and real estate investments. Our findings reveal that the optimal estimator varies significantly with the dataset, and is not always the Poisson PML estimator. Furthermore, the choice of estimator can lead to markedly different parameter estimates, which calls into question the prevailing practice of applying a single estimator, such as Poisson, universally across all applications. Notably, although non-linear least squares has been avoided in practice due to the conventional wisdom that it is inefficient (Silva and Tenreyro, 2006), our framework indicates that it may be the preferred method when heteroskedasticity is mild compared to sparsity. Indeed, this proves to be true for some of the datasets we consider. Ultimately, we advocate for a systematic approach to modeling sparse non-negative data that jointly takes into account salient features of the data, rather than a one-size-fits-all

approach commonly recommended in existing literature.

The rest of the paper is organized as follows. In Section 2, we provide the econometric background on non-negative data modeling, and introduce the formal setup of the paper, along with discussions of key concepts and objects, including dispersion and asymmetric censoring. In Section 3, we propose a class of moment estimators that we call generalized pseudo maximum likelihood estimators. We characterize the bias and variance of these estimators in the presence of heteroskedasticity and asymmetric censoring. These results provide a framework to determine the most appropriate estimator from the class of generalized pseudo maximum likelihood estimators. In Section 4, we conduct numerical experiments to study the bias-variance trade-off behavior of generalized PML estimators under excess sparsity and heteroskedasticity. In Section 5, we apply our framework to datasets in finance. Readers primarily interested in the empirical implications of our paper, especially in finance, may skip directly to Sections 4 and 5.

2 Econometrics of Nonnegative Data

As mentioned previously, non-negative data in finance often exhibits two prominent features: sparse and heteroskedastic outcomes. *Heteroskedasticity* arises naturally when the outcome variable is constrained to be non-negative. Intuitively, samples with larger conditional means are less constrained by the non-negativity condition, hence tend to have higher variance. This property is closely related to the *dispersion* of data. Meanwhile, count-like data also has distributions that are often right-skewed and exhibits *sparsity*, i.e., having many zero responses (Greene, 1994). These characteristics can result from intrinsic sparsity structures in the data generating process, such as the rare nature of the underlying events or network structures. For example, financial networks that encode contractual relationships or interactions between entities are typically sparse and exhibit a core-periphery structure (Craig and Von Peter, 2014). Sparse outcomes also arise naturally when the data is censored

near zero. For example, in economic trade, measurements of large trade volumes are usually more accurate than those of small trade volumes, which may be rounded down to zero.

One of the most popular approaches to modeling data with non-negative dependent variables is log-linear regression, i.e., assuming that the logarithm of the response variable Y is linear in the covariates X. To deal with zero values, this approach typically uses the transformation $\log(1+Y)$, resulting in the specification

$$\tilde{Y} := \log(1+Y) = \theta^T X + \epsilon.$$

Prominent examples include the estimation of international trade data and panel data with non-negative outcomes, such as earnings (Card, 2001), and count and count-like data in finance, such as number of corporate patents granted (Hirshleifer et al., 2012) and firms' toxic waste release volumes (Akey and Appel, 2021). However, previous works have raised some important drawbacks of the log-linear regression approach. For example, Cohn et al. (2022) demonstrate that log-linear models do not always produce estimates with economically meaningful interpretations, and can even result in the wrong sign in expectation. Chen and Roth (2024) similarly show that treatment effect estimates based on log-linear transformations are not invariant under change of units and therefore should not be interpreted as percentages. More importantly, as demonstrated in the series of works by Silva and Tenreyro (2006, 2010, 2011, 2022), log-linear regression is not robust against heteroskedastic and sparse data. Similar issues with other non-linear transformations of non-negative outcome variables, such as the inverse hyperbolic sine,

A popular alternative to transformation of non-negative outcomes is *pseudo maximum likelihood* (PML) estimation (White, 1982; Gourieroux et al., 1984a,b; Wooldridge, 1999). PML has been widely applied to model data when the true likelihood is difficult to specify or when one is uncertain whether the data distribution is correctly specified, e.g., in generalized linear models (McCullagh, 2019). The Poisson PML (PPML) is a popular choice in practice

for modeling non-negative outcomes, especially count data (Wooldridge, 1999; Cameron and Trivedi, 2013). When $Y \in \mathbb{R}_+$ is integer valued, a natural modeling approach is to assume that Y conditional on X has a Poisson distribution with rate parameter $\lambda(X, \theta)$ that is log-linear in X with parameter θ, i.e., $\lambda(X, \theta) = \exp(\theta^T X)$. The likelihood is given by

$$\mathbb{P}(Y = y \mid X) = \frac{(\lambda(X, \theta))^y \cdot \exp(-\lambda(X, \theta))}{y!} = \frac{e^{y \cdot \theta^T X} \cdot \exp(-e^{\theta^T X})}{y!}.$$

The maximum likelihood estimator (MLE) $\hat{\theta}$ of θ, given n independent and identically distributed (i.i.d.) samples $\{y_i, x_i\}_{i=1}^n$, is then obtained as the maximizer of the log-likelihood

$$\max_{\theta} \sum_i y_i \cdot \theta^T x_i - \exp(\theta^T x_i) \tag{1}$$

modulo constants. When the true distribution of Y given X is not Poisson, the PML approach continues to maximize (1). A seminal result of Gourieroux et al. (1984a) shows that, even if Y is not Poisson (or even integer valued), the PML estimator which maximizes (1) is still consistent as long as the conditional mean of Y given X is log-linear, i.e., $\mathbb{E}(Y \mid X) = \exp(\theta^T X)$. Importantly, the consistency of PPML holds even if the data is heteroskedastic, i.e., $\mathrm{Var}(Y \mid X)$ depends on X. In contrast, the log-linear regression is consistent only if the data is homoskedastic (Manning and Mullahy, 2001).[2]

Although Poisson pseudo maximum likelihood is consistent under heteroskedasticity, it is only *efficient* when the conditional variance of Y is equal to the conditional mean of Y, i.e., $\mathrm{Var}(Y \mid X) = \mathbb{E}(Y \mid X)$. When the conditional variance of Y does not scale linearly with the conditional mean, giving rise to the so-called over-dispersion and under-dispersion phenomena (Zorn, 1998), there exist other consistent PML estimators that are more efficient.

[2]The distinction between PPML and log-linear regression lies in the ordering of expectation and logarithm operations. Whereas log-linear regression assumes that the conditional expectation of the logarithm of Y (or $Y + 1$) is linear in X, i.e., $\mathbb{E}(\log Y \mid X) = \beta^T X$, PPML assumes that the logarithm of the conditional expectation of Y is linear in X, i.e., $\log(\mathbb{E}(Y \mid X)) = \theta^T X$. The implications of this distinction, including the robustness of PPML to heteroskedasticity, are manifested through convexity and Jensen's inequality, and are discussed extensively in Silva and Tenreyro (2006, 2022).

For example, the gamma PML (GPML) is efficient when $\text{Var}(Y \mid X) = \mathbb{E}^2(Y \mid X)$, while the closely related negative binomial PML is efficient when $\text{Var}(Y \mid X) = \mathbb{E}(Y \mid X) + a\mathbb{E}^2(Y \mid X)$. The non-linear least squares (NLS) estimator can also be viewed as a PML estimator, and is efficient when $\text{Var}(Y \mid X)$ is constant.

While the efficiency properties of PML estimators are well-documented, their behaviors when used to model non-negative data with many zero outcomes is less clear. Silva and Tenreyro (2006, 2011) provide some simulation evidence on the robustness of Poisson PML to sparsity resulting from rounding errors or censoring near zero. They therefore recommend the Poisson PML over other PML estimators on account of its robustness under sparsity and heteroskedasticity. Meanwhile, the general consensus among existing works is that NLS performs poorly under strong heteroskedasticity, and is often avoided in practice for this reason. However, when the level of sparsity in the data is high, our findings in Sections 4 and 5 suggest that NLS can actually outperform Poisson and other PML estimators in terms of mean squared error (MSE), even under equidispersion, when PPML is efficient. These findings seem to contradict conventional wisdom, and raise some important questions on PML estimators. In particular, although their behaviors with respect to heteroskedasticity and sparsity have been studied separately before, it remains unclear how they compare with each other when both features are considered. When modeling data with non-negative outcomes in practice, the answers to such questions can help us decide which modeling approach to use that best handles the two features of the data.

2.1 Our Contributions and Key Takeways

We establish a bias-variance trade-off framework to analyze estimation methods for non-negative outcome data. On one hand, PML estimation on data with many zero outcomes leads to bias due to unreliable observations with small conditional means; on the other hand, heteroskedasticity leads to efficiency loss due to noisy observations with large conditional means. These two features therefore create a tension regarding how to optimally utilize the

information provided by samples. To formalize this intuition, we propose a general class of estimators that encompasses pseudo maximum likelihood estimators commonly used in practice such as Poisson PML, gamma PML, and NLS. These *generalized PML estimators* place heterogeneous weights on each observation in their estimation equations, with some favoring samples with small conditional means (such as the gamma PML) while others favoring samples with large conditional means (such as the NLS). We study the bias and asymptotic variance of estimators under different levels of heteroskedasticity, sparsity, and sample weights. Intuitively, under strong over-dispersion, estimators that place more weight on samples with smaller conditional means are more efficient. On the other hand, when there are many zeros, estimators that place more weight on samples with larger conditional means can reduce bias. As a result, there exists an estimator that optimally balances these two aspects of the data. It is worth noting that this trade-off is a unique feature of the generalized PML estimators for data with sparse non-negative outcome variables, and is distinct from the conventional bias-variance trade-off in statistics and machine learning that relies on the notion of model complexity.

We find that, depending on the data at hand, some generalized PML estimators are more appropriate than others. When data exhibits sparsity, the potential bias that it induces can outweigh any efficiency gain from using a generalized PML that up-wieghts samples with small conditional means. Conversely, when heteroskedasticity is severe, it is often more appropriate to use such a generalized PML, even if doing so may increase the bias due to excess sparsity. We therefore recommend determining the most appropriate estimator that takes into account the optimal trade-off between heteroskedasticity and excess sparsity. We propose a simple cross validation procedure to achieve this task, demonstrating its usefulness when modeling default and credit ratings data. We find that a generalized PML estimator close to the Poisson is optimal when modeling industry level defaults. On the other hand, for the data on credit ratings, the non-linear least squares (NLS) in fact provides significantly better fit, which suggests that the dispersion of data is mild relative to its sparsity. Given that

NLS is generally considered inefficient when modeling non-negative outcomes, our finding highlights the importance of assessing model fitness on the specific data at hand, instead of following a general rule and always use a single fixed estimator.

Our study thus provides a principled approach to modeling data with non-negative outcome variables that unifies existing pseudo maximum likelihood approaches, i.e., our framework includes as special cases the existing PML estimators, such as Poisson, gamma, and NLS; these estimators are often applied in an ad hoc manner, whereas our method provides a principled way to determine the best estimator to use. In settings where log-linear regression or Poisson pseudo maximum likelihood are often applied, such as technological innovation and corporate finance (Hirshleifer et al., 2012; Cohn et al., 2022), using a simple cross validation procedure, one can in practice determine the generalized PML estimator that best describes the data with non-negative outcomes.

In the rest of this section, we introduce the key elements of our bias-variance trade-off framework. We start with a discussion of pseudo maximum likelihood estimation in the context of heteroskedastic non-negative data and dispersion in Section 2.2. In Section 2.3, we discuss measurement errors and asymmetric censoring that results in excess zeros. In Section 2.4, we propose a unified perspective on pseudo maximum likelihood estimation by incorporating them into a general class of moment estimators. This perspective provides intuition for the bias-variance trade-off results in Section 3.

2.2 Heteroskedasticity and Pseudo Maximum Likelihood

A popular instance of data with non-negative outcome variables is economic trade data between countries or geographical locations. We start our discussions based on this instance, although they apply more generally to non-negative data arising in other domains. In Silva and Tenreyro (2006), the authors focus on a class of multiplicative trade models, often

referred to as "gravity models" (Anderson, 1979; Anderson and Van Wincoop, 2003):

$$Y_{ij} = \alpha_0 G_i^{\alpha_1} G_j^{\alpha_2} D_{ij}^{\alpha_3} \eta \equiv \exp(\theta^T X_{ij}) \cdot \eta, \tag{2}$$

where Y_{ij} is the trade volume between countries i and j, G_i is the GDP of country i, and D_{ij} is the distance between the two countries. η is a multiplicative error that accounts for the randomness in data that results in the imperfect fit $Y_{ij} \neq \alpha_0 G_i^{\alpha_1} G_j^{\alpha_2} D_{ij}^{\alpha_3}$, with $\mathbb{E}(\eta \mid X) = 1$. Assuming for now that $Y_{ij} > 0$, a common approach to estimating this model is to run an OLS regression of the following log-linearized model

$$\log(Y) = \theta^T X + \log \eta.$$

The validity of this approach relies on the standard requirement that $\log \eta$ is uncorrelated with X. However, $\mathbb{E}(\log \eta \mid X) \neq \log \mathbb{E}(\eta \mid X) = 0$ in general due to Jensen's inequality. The quantity $\mathbb{E}(\log \eta \mid X)$ depends on X because the non-linear logarithm transformation applied to η induces dependence of $\mathbb{E}(\log \eta \mid X)$ on higher moments of the conditional distribution of η on X, not just its mean, unless η is of a particular form. Therefore, the log-linear OLS regression approach can be severely biased, which has been confirmed in experiments and theoretical analysis in Silva and Tenreyro (2006); Cohn et al. (2022).

The failure of log-linear OLS regression for the gravity model (2) is closely related to the problem of heteroskedasticity intrinsic in data with non-negative outcomes. Suppose we would like to express the data in the following general additive form:

$$Y = f(X, \theta) + \varepsilon. \tag{3}$$

With $f(X, \theta) = \exp(\theta^T X)$, it is connected to the multiplicative gravity model (2) through

$$\varepsilon = (\eta - 1) \cdot \exp(\theta^T X). \tag{4}$$

The conditional variance $\text{Var}(\varepsilon \mid X)$ therefore will in general depend on X, even if the conditional variance $\text{Var}(\eta \mid X)$ does not, resulting in heteroskedasticity. This feature is intrinsic to non-negative data because, as mentioned in the introduction, when $f(X, \theta) > 0$ is small, in order to guarantee $Y = f(X, \theta) + \varepsilon \geq 0$, the dispersion of ϵ is restricted, while larger values of $\exp(\theta^T X)$ allows larger dispersion of ε.

In the presence of heteroskedasticity, a popular alternative to estimating the non-linear model (3) is **pseudo maximum likelihood** (PML). This approach does not assume knowledge of the distribution of ε, only the correct specification of $\mathbb{E}(Y \mid X) = f(X, \theta)$. Under this condition, PML then minimizes a "pseudo log-likelihood", which is constructed assuming a specified probability distribution for ε. For example, the non-linear least squares estimator, which solves the problem

$$\min_{\theta} \sum_{i} (y_i - f(x_i, \theta))^2, \tag{5}$$

can be understood as a PML estimator that assumes a normal distribution of ε. While the resulting estimator achieves consistent estimation of θ under regularity conditions (Gourieroux et al., 1984a), it is only efficient when ε is *homoskedastic*, i.e., $\text{Var}(\varepsilon \mid X)$ does not depend on X. If in addition we had knowledge of the functional form of the conditional variance $\text{Var}(\varepsilon \mid X)$, we can obtain the efficient weighted NLS estimator. However, such information is often not available in practice. In this case, the Poisson pseudo maximum likelihood (PPML) estimator is usually preferable to the NLS. It constructs the pseudo log-likelihood by assuming that $Y \mid X$ has a Poisson distribution with mean parameter $f(X, \theta)$. Even when Y is not necessarily integer-valued, consistency can be achieved given the correct specification of $f(X, \theta)$. Recall that Poisson distributions have equal mean and variance, so that the Poisson PML approach can be understood as assuming $\text{Var}(\varepsilon \mid X) = \mathbb{E}(\varepsilon \mid X)$. In fact, it is efficient among all PML estimators when the condition $\text{Var}(\varepsilon \mid X) = \mathbb{E}(\varepsilon \mid X)$ is satisfied. Given our discussion on the intrinsic heteroskedasticity of non-negative data,

we see that Poisson PML is preferable to NLS partly because it implicitly accounts for the heteroskedasticity with a conditional variance that *increases* with the conditional mean. Indeed, despite its ability to achieve consistency under correct specification of $f(X, \theta)$, the NLS is known to perform very poorly in practice for non-negative outcome data compared to PPML under strong heteroskedasticity (Silva and Tenreyro, 2006), due to its efficiency loss in this setting.

An additional advantage of PPML is that its pseudo log-likelihood function

$$\sum_i y_i \cdot \log f(x_i, \theta) - f(x_i, \theta) \tag{6}$$

is concave in θ whenever $f(X, \theta)$ is log-concave. Examples include the exponential function $\exp(\theta^T X)$ and the Gaussian. In contrast, the NLS objective (5) is usually non-convex for non-linear $f(X, \theta)$. As a result, it may have multiple local maxima, and standard non-convex optimization packages may only be able to find local maximizers instead of the true NLS estimator. Besides heteroskedasticity, this distinction between NLS and Poisson PML in terms of the difficulty of solving the corresponding optimization problems likely also contributes to the poor performance of NLS estimators observed in practice, although it is less studied than the effects of heteroskedasticity.

Besides Poisson PML, other PML estimators based on different distributions are also popular in practice, especially when the so-called over-dispersion or under-dispersion is present in the data (Greene, 1994; Zorn, 1998). Recall that Poisson PML implicitly assumes, and is efficient under, the condition of equal mean and variance. In many cases, this requirement is not satisfied by data, which appears to be more (or less) dispersed. More precisely, we may have

$$\mathrm{Var}(Y \mid X) = a\mathbb{E}(Y \mid X) + \mathbb{E}^{\alpha}(Y \mid X). \tag{7}$$

For $\alpha > 1$, the data is said to be over-dispersed, while for $a = 0$ and $0 \leq \alpha < 1$, the

data is said to be under-dispersed. For $\alpha \neq 1$, there exist other PML estimators that are more efficient than the PPML. For example, the gamma PML (GPML) is efficient when $\alpha = 2$, while the closely related negative binomial PML is efficient when $a = 1$ and $\alpha = 2$ (Gourieroux et al., 1984a). In practice, when we do not know the variance, previous works have provided methods to determine α using statistical tests based on auxiliary regressions (Park, 1966; Manning and Mullahy, 2001) and tests for over-dispersion (Cameron and Trivedi, 1990).

It may appear that under heteroskedasticity, we should always use a PML estimator that best captures the conditional variance as a function of the conditional mean, by determining the value of α. However, that is only half of the story. In simulations, Silva and Tenreyro (2006) observe that the estimator resulting from maximizing the PPML objective consistently outperforms other PML estimators, even when $\text{Var}(\varepsilon \mid X) = \mathbb{E}(\varepsilon \mid X)$ fails. In addition to heteroskedasticity, their simulations introduce censoring and rounding to the data, producing many observations whose outcomes are zero, which is another feature of non-negative outcome variables often observed in practice. It turns out that this mechanism of censoring or rounding is important in the modeling of non-negative data, and explains why Poisson PML may be preferable even in the presence of over-dispersion. We therefore turn to it next.

2.3 Excess Sparsity and Asymmetric Censoring

Another prominent feature of data with non-negative outcomes is the presence of many zeros. In the international trade example, economic, political or geographical factors could impact the trade between two countries, resulting in small or zero trade volumes between them. Indeed, economic trade data typically contain many zero observations of Y_{ij} in (2), i.e., no trade between pairs of countries or small trade volumes that are often *rounded* to zero. More generally, a large number of zero observations is common in non-negative outcome data in many applications (Greene, 1994). We will refer to this feature as **excess sparsity**. This feature is related, but distinct from, the heteroskedasticity and dispersion of data discussed

in the preceding section. Previous works have proposed several methods to explain and model this phenomenon. We now discuss two that are relevant to our analysis, starting from a modified Poisson process.

Suppose that given explanatory variables X, the non-negative outcome variable Y follows a Poisson distribution with mean parameter $\exp(\theta_0^T X)$. Then the probability of $Y = 0$ is given by

$$\mathbb{P}(Y = 0 \mid X) = \exp(-\exp(\theta_0^T X)).$$

Suppose, on the other hand, that the fraction of zeros observed in data is not consistent with the predictions of this Poisson model, which in a large sample is close to $\int_X \mathbb{P}(Y = 0 \mid X)dX$. In this case, we need to modify the data-generating process to account for this discrepancy. A popular proposal is the two-part or **hurdle model** (Mullahy, 1986; King, 1989). It assumes a two-part generating process of Y given X. An independent binomial random event determines whether Y is set to zero. If not, the realization of Y is determined according to a *truncated* distribution that takes on strictly positive values. In the simplest case, instead of assuming $\mathbb{P}(Y = 0 \mid X) = \exp(-\exp(\theta_0^T X))$ as in a Poisson model, we assume

$$\mathbb{P}(Y = 0 \mid X) = \exp(-\beta \exp(\theta_0^T X)),$$

which can be understood as the probability of a Poisson variable with mean parameter $\beta \exp(\theta_0^T X)$ being equal to 0, while the probability for $Y > 0$ follows a truncated Poisson with mean parameter $\exp(\theta_0^T X)$, i.e.,

$$P(y \mid x) = \begin{cases} \exp(-\beta \exp(\theta_0^T x)) & y = 0 \\ \frac{1-\exp(-\beta \exp(\theta_0^T x))}{1-\exp(-\exp(\theta_0^T x))} \cdot \frac{e^{-\exp(\theta_0^T x)} \exp^y(\theta_0^T x)}{y!} & y > 0 \end{cases}.$$

When $\beta = 1$, we recover the original Poisson model. When $\beta < 1$, this modified data generating process will result in more sparsity than the standard Poisson model, and similarly

for $\beta > 1$. More generally, we can use non-Poisson based distributions for both the binomial zero event variable and the positive outcome variable. This hurdle model has been widely used in practice for non-negative data with excess sparsity (Greene, 1994; Zorn, 1998).

We can also understand the data generating process in the hurdle model equivalently as the following two-step censoring model. For example, when $\beta < 1$, in the first step, a variable Y is drawn according to a Poisson distribution with mean parameter $\exp(\theta_0^T X)$. If $Y > 0$, then with probability

$$\frac{\exp(-\beta \exp(\theta_0^T X)) - \exp(-\exp(\theta_0^T X))}{1 - \exp(-\exp(\theta_0^T X))},$$

Y is censored to 0. We can verify that the resulting variable is zero with probability $\exp(-\beta \exp(\theta_0^T X))$, and has the same distribution as the truncated Poisson when it is positive. This censoring process is closely related to the **zero-inflation** model of Lambert (1992), which is another common (though less popular) approach to handle excess zeros. In Lambert (1992), a standard Poisson model is generated with mean parameter $\exp(\theta_0^T X)$. Then with probability

$$\frac{1}{1 + \exp(\beta \theta_0^T X)},$$

the observation is censored to zero. The main difference with the hurdle model is in the censoring probabilities, which is modeled by a logistic function for the zero-inflated Poisson. The choice between two-part models and sample selection models have been studied by Manning et al. (1987) and Leung and Yu (1996), among others. See also Zorn (1998) for a discussion of the hurdle model and zero-inflated model.

In both the hurdle and the zero-inflation models, the probability of being censored to zero depends on X. For example, in the zero-inflation model with $\beta > 0$, the censoring probability $\frac{1}{1+\exp(\beta \theta_0^T X)}$ *decreases* with $\exp(\theta_0^T X)$, the conditional mean. Similarly, in the hurdle model, the probability $\exp(-\beta \exp(\theta_0^T X))$ of observing a zero response variable decreases rapidly in $\exp(\theta_0^T X)$. This feature aligns with how excess zeros appear in applications. In many

survey and census data, such as economic trade data, a small outcome or response variable is often more likely to be rounded down to zero than a large outcome, which is less prone to measurement errors (Frankel, 1997). This ***asymmetric* censoring** feature of the data turns out to be important in our bias-variance trade-off analysis, because its presence calls for the use of PML estimators that may otherwise be inefficient under heteroskedasticity. For example, in the presence of many zeros as a result of asymmetric censoring, Silva and Tenreyro (2006, 2011) demonstrate in simulations that PPML is generally more robust than the gamma PML, even under over-dispersion with $\text{Var}(Y \mid X) = \mathbb{E}^2(Y \mid X)$, when GPML is the most efficient PML.

In our paper, in order to mathematically model the asymmetric censoring that results in sparse outcome variables, we consider the following DGP:

$$\begin{aligned} Y^{true} &= f(X,\theta) + \varepsilon \\ Y^{observed} &= \begin{cases} Y^{true} & \text{with probability } 1 - P(X,\theta_0) \\ 0 & \text{with probability } P(X,\theta_0), \end{cases} \end{aligned} \tag{8}$$

where the censoring probability that we are primarily interested in is

$$P(X,\theta_0) = \exp(-\beta f(X,\theta)) \text{ or } 1/(1 + f^{\beta}(X,\theta)). \tag{9}$$

As we will see, this censoring mechanism will create a bias for a continuum of (generalized) PML estimators indexed by a parameter κ that increases while the variance decreases in κ, resulting in an optimal choice of estimator given asymmetric censoring and heteroskedasticity.

2.4 A Unified Perspective on Pseudo Maximum Likelihood

Having introduced heteroskedasticity and asymmetric censoring, we now turn to a class of pseudo maximum likelihood estimators that motivate the main results in this paper.

Although PMLs are generally based on pseudo likelihoods constructed from qualitatively different distributions such as Poisson and normal, there is another way to view them that reveals their closer connections through the first order conditions of their respective optimization problems. In particular, the following PML are all consistent estimators of θ_0 when i.i.d. data is generated from the model $Y = \exp(\theta_0^T X) + \varepsilon$ with $\mathbb{E}(\varepsilon \mid X) = 0$ (see for example Gourieroux et al. (1984b)):

(1) Non-linear least squares

$$\min_\theta \sum_i (y_i - \exp(\theta^T x_i))^2 \Rightarrow \sum_i (y_i - \exp(\theta^T x_i)) \exp(\theta^T x_i) x_i = 0,$$

which is asymptotically optimal when $\mathrm{Var}(Y \mid X) = 1$.

(2) Poisson PML

$$\max_\theta \sum_i y_i \theta^T x_i - \exp(\theta^T x_i) \Leftrightarrow \sum_i (y_i - \exp(\theta^T x_i)) x_i = 0,$$

which is optimal when $\mathrm{Var}(Y \mid X) = \exp(\theta_0^T X)$.

(3) gamma PML

$$\max_\theta \sum_i -\theta^T x_i - y_i \exp(-\theta^T x_i) \Leftrightarrow \sum_i (y_i - \exp(\theta^T x_i)) \exp(-\theta^T x_i) x_i = 0,$$

which is optimal when $\mathrm{Var}(Y \mid X) = \exp(2\theta_0^T X) = (E(Y \mid X))^2$.

(4) Negative binomial PML

$$\max_\theta \sum_i y_i \theta^T x_i - (\frac{1}{b} + y_i) \log(1 + b \exp(\theta^T x_i)) \Leftrightarrow \sum_i x_i \frac{y_i - \exp(\theta^T x_i)}{1 + b \exp(\theta^T x_i)} = 0,$$

which is optimal when $\mathrm{Var}(Y \mid X) = \exp(\theta_0^T X) + a \exp(2\theta_0^T X)$. Note that the first order

conditions of the above PML estimators are all of the form

$$\sum_i (y_i - \exp(\theta^T x_i))(a + \exp(\theta^T x_i))^\kappa x_i = 0. \tag{10}$$

NLS, Poisson, and gamma PML corresponds to $a = 0$ and $\kappa = 1, 0, -1$ respectively, while negative binomial corresponds to $a = 1$ and $\kappa = -1$. Therefore, the PML estimators can be understood as solving a system of equations that balance the *observed* outcomes y_i with the *expected* outcomes $\exp(\theta^T x_i)$, weighted by the covariates x_i and an estimator-specific term $(a + \exp(\theta^T x_i))^\kappa$ for different values of a, κ.

The system of equations (10) provides a unified perspective on PML estimators for data with non-negative response. Importantly, it provides intuition on the behavior of PML estimators with respect to heteroskedasticity and excess zeros due to censoring, which was already present in the work of Silva and Tenreyro (2006). For example, it is well-known that the NLS is very inefficient under strong heteroskedasticity. We can understand this inefficiency in terms of its weight $(a + \exp(\theta^T x_i))^\kappa = \exp(\theta^T x_i)$ on the samples: NLS places larger weights on samples with larger conditional mean $\exp(\theta^T x_i)$. However, such samples also tend to be noisier when $\mathrm{Var}(Y \mid X)$ increases with $\exp(\theta_0^T X)$. In this regard, the gamma PML estimator, which places weight $\exp(-\theta^T X)$ on samples, relies more on samples with small conditional means $\exp(\theta^T X)$, which are less noisy, leading to efficiency gains under strong heteroskedasticity. The Poisson PML has uniform weight $(a + \exp(\theta^T x_i))^\kappa \equiv 1$ on samples, and therefore places less weight on large samples than the NLS and less weight on small samples than the gamma PML, leading to efficiency gain under mild heteroskedasticity. The negative binomial, with $a = 1$ and $\kappa = -1$, has similar weights on samples with large conditional mean $\exp(\theta^T X)$ as the gamma PML, and similar weights on samples with small $\exp(\theta^T X)$ as the Poisson PML.

In general, under stronger heteroskedasticity, i.e., larger α in $\mathrm{Var}(Y \mid X) = \mathbb{E}^\alpha(Y \mid X)$, PMLs which place smaller weights on samples with large conditional mean $\mathbb{E}(Y \mid X)$, i.e.,

smaller κ in (10), tend to be more efficient. On the other hand, when there is asymmetric censoring of the data, where smaller samples are more likely to be affected by measurement errors or censoring, the *opposite* is true. This is because such censoring affects the samples with small $\mathbb{E}(Y \mid X)$ more, and a PML estimator that places more weight on small $\mathbb{E}(Y \mid X)$, i.e., smaller κ, will suffer from a more severe bias.

3 Generalized PML Estimators

The above discussions suggest that there is an intrinsic dilemma when applying pseudo maximum likelihood type methods to data with non-negative outcomes. On one hand, heteroskedasticity necessitates smaller weights on *larger* (noisier) samples to improve efficiency. On the other hand, asymmetric censoring requires smaller weights on *smaller* (more likely to be censored) samples to reduce bias. In the former case, PMLs with small κ such as gamma PML and negative binomial are generally preferred, while in the latter case PMLs with larger κ such as NLS are generally preferred. Silva and Tenreyro (2006) argue that the Poisson PML is a reasonable compromise between NLS and gamma PML when both heteroskedasticity and asymmetric censoring are present. However, depending on the relative severity of the two problems, there may be other estimators that can better balance between bias and variance. Our work is precisely motivated by this consideration.

In this work, we propose and analyze the following general class of Z-estimators (Amemiya, 1985) with estimating equations

$$\frac{1}{n}\sum_i \psi(y_i, x_i, \theta) = \frac{1}{n}\sum_i (y_i - \exp(\theta^T x_i))(a + \exp(\theta^T x_i))^\kappa x_i = 0, \tag{11}$$

where $\kappa \in \mathbb{R}$. This class of estimators includes as special cases the standard PML estimators. When $a = 0$ and $\kappa = 0$, we recover the Poisson PML. When $a = 0, \kappa = 1$, we recover NLS. When $a = 0, \kappa = -1$ we recover the gamma PML. When $a = 1, \kappa = -1$, we recover the negative binomial PML. For non-integer values of κ, the resulting estimator does not arise

from pseudo maximum likelihood estimation based on a particular distribution, but their estimating equations generalize those of the standard PML estimators. For this reason, we refer to them as **generalized PML estimators**[3].

For the class of Z-estimators solving (11), we next study their behaviors under heteroskedasticity of the form

$$\mathrm{Var}(Y \mid X) = \mathbb{E}^{\alpha}(Y \mid X) = \exp(\alpha \theta^T X)$$

and asymmetric censoring of the form (8) with probability

$$P(X, \theta_0) = \exp(-(\tau \exp(\theta_0^T X))^{\beta}) \text{ or } 1/(1 + (\tau \exp(\theta_0^T X))^{\beta}).$$

As we will see, depending on the magnitudes of heteroskedasticity and excess sparsity, quantified by α and (τ, β) respectively, generalized PML estimators with different values of κ will perform differently. Given a particular configuration of problem parameters, we analyze the bias and variance of the estimators, which allows us to characterize the optimal choice of a, κ in terms of the mean squared error (MSE). These results provide guidance in practice on how to model data with non-negative outcome variables that are tailored to the particular setting at hand, which could feature more heteroskedasticity or excess sparsity.

3.1 Generalized PML Estimators

In this section, we study the bias-variance trade-off and computation of generalized PML estimators. As a reminder, given data with non-negative response variable Y and covariates X, our framework assumes that the data is generated i.i.d. according to

$$Y_i = \begin{cases} \exp(\theta_0^T X_i) + \varepsilon_i & \text{with probability } 1 - P(\theta_0, X_i) \\ 0 & \text{with probability } P(\theta_0, X_i), \end{cases}$$

[3]When $\kappa > 0$, the estimators are better understood as M-estimators that maximize a non-concave sample average objective, especially when some covariates are strictly positive or negative. See Section 3.4

which is potentially heteroskedastic and asymmetrically censored:

$$\begin{aligned} \mathbb{E}(\varepsilon_i \mid X_i) &= 0 \\ V(\exp(\varepsilon_i) \mid X_i) &= \exp^{\alpha}(\theta_0^T X_i) = \exp(\alpha \theta_0^T X_i), \alpha \in [0,2] \\ P(X_i, \theta_0) &= \exp(-(\tau \exp(\theta_0^T X_i))^{\beta}) \text{ or } 1/(1+(\tau \exp(\theta_0^T X_i))^{\beta}). \end{aligned} \tag{12}$$

We assume for now that $\tau = 1$ (see Section 4 for a discussion of τ). For this data, we consider a family of Z-estimators defined in (11) that include the classic PML estimators. For now, we focus on the case with $a = 0$ and $\kappa \in [-1, 1]$, resulting in the estimating equations

$$\frac{1}{n}\sum_i (Y_i - \exp(\theta^T X_i)) \cdot \exp^{\kappa}(\theta^T X_i) \cdot X_i = 0, \quad \kappa \in [-1,1]. \tag{13}$$

For each pair of values (α, β) (or (α, τ)), we are interested in the behavior of generalized PML estimators, and characterizing the κ that minimizes the MSE.

3.2 Asymptotic Bias and Variance

We first provide the general formulae for the bias and asymptotic variance for the estimators defined by (13). Under the censoring model (8), the bias does not have an exact formula in general. Instead, we provide an approximation formula based on Taylor expansions.

Theorem 1. *The bias of generalized PML estimators can be approximated by*

$$\hat{\theta} - \theta_0 = (A^T A)^{-1} A b + o_p(1), \tag{14}$$

where, with $P(\theta_0, x)$ the censoring probability in (8)*, A and b are given by*

$$\begin{aligned} A &= \int (P(\theta_0, x)\kappa - 1) \cdot x x^T p(x) \exp((\kappa+1)\theta_0^T x) dx, \\ b &= \int P(\theta_0, x) \cdot x^T p(x) \exp((\kappa+1)\theta_0^T x) dx. \end{aligned} \tag{15}$$

We can use Theorem 1 to estimate the bias of generalized PML estimators under censoring. For example, if $\kappa = -1$, which corresponds to Gamma PML, we have

$$A = \int (-1 - P(\theta_0, x)) \cdot xx^T p(x) dx,$$
$$b = \int P(\theta_0, x) \cdot xp(x) dx,$$

which can be used to explicitly quantify the bias given a particular $P(\theta_0, x)$, such as $P(\theta_0, x) = \frac{1}{\exp(\tau\theta_0^T x)+1}$, as in the zero-inflation model of Lambert (1992). In the absence of censoring, i.e., $P(\theta_0, x) \equiv 0$, Theorem 1 implies that generalized PML estimators are all consistent.

Under standard regularity conditions, the limit of Z-estimators converges to the solution $\tilde{\theta}_0$ to the population limit of their estimating equations, which for (13) is given by

$$\mathbb{E}(\psi(y, x, \theta)) = \mathbb{E}(y - \exp(\theta^T x)) \cdot \exp(\kappa\theta^T x) \cdot x = 0,$$

In the absence of censoring, we have

$$\mathbb{E}(y - \exp(\theta^T x)) \cdot \exp(\kappa\theta^T x) \cdot x = \mathbb{E}_x(\exp(\theta_0^T x) - \exp(\theta^T x)) \cdot \exp(\kappa\theta^T x) \cdot x,$$

which evaluates to 0 when $\theta = \theta_0$, ensuring the consistency of generalized PML estimators. However, in the presence of censoring with probability $P(\theta_0, x)$, the conditions that characterize the limit of the generalized PML estimators become

$$\mathbb{E}_x \left\{(1 - P(\theta_0, x)) \cdot \exp(\theta_0^T x) - \exp(\theta^T x)\right\} \cdot \exp(\kappa\theta^T x) \cdot x = 0,$$

which is no longer solved by $\theta = \theta_0$. We use $\tilde{\theta}_0$ to differentiate from the true parameter θ_0 when censoring results in bias.

Note that the terms A and b do not depend on the heteroskedasticity parameter α and only on κ and the censoring probability $P(\theta_0, X)$. This is expected because the censoring

probabilities are defined in terms of the conditional mean $\exp(\theta_0^T X)$ only, which does not depend on α, which controls the variance of ε. If we assume an alternative censoring mechanism, where the censoring probability depends on the *realized* outcome $\exp(\theta_0^T X) + \varepsilon$, then the bias will depend on α as well. On the other hand, as shown in Theorem 2, the asymptotic variance will depend on both α and κ, as well as $P(\theta_0, X)$.

The quality of the approximation formula in Theorem 1 is an important consideration in practice. Intuitively, since the Taylor expansions used in the derivation of Theorem 1 rely on the small magnitude of $\exp(\theta_0^T x)$, the formula (14) provides a good approximation of the bias when the censoring probability and κ result in integrals with small densities on x with large $\exp(\theta_0^T x)$. In Section 4, we provide simulation evidence that the approximation formula we derive is very close to the true bias in a wide range of settings.

Next, we compute the asymptotic variance of the generalized PML estimators, taking into account their bias that results from the censoring model (8).

Theorem 2. *The asymptotic distribution of generalized PML estimators is given by* $\sqrt{n}(\theta - \tilde{\theta}_0) \to_d \mathcal{N}(0, J^{-1} I J^{-1})$*, where*

$$
\begin{aligned}
I &= \mathbb{E}_x x x^T \exp(2\kappa\theta^T x) \cdot \left[P(\theta_0, x) \left(\exp(\theta^T x) \right)^2 + (1 - P(\theta_0, x)) \left(\exp(\theta_0^T x) - \exp(\theta^T x) \right)^2 \exp(\alpha \theta_0^T x) \right] \\
J &= \mathbb{E}_x \left[- \left((1 - P(\theta_0, x)) \cdot \exp(\theta_0^T x) - \exp(\theta^T x) \right) \cdot \exp(\kappa \theta^T x) \cdot \kappa x x^T + \exp(\kappa \theta^T x) \exp(\theta^T x) x x^T \right],
\end{aligned}
\tag{16}
$$

evaluated at $\theta = \tilde{\theta}_0$*, which solves the following equations that are the population limit of the estimating equations* (11)*:*

$$
\mathbb{E}_x \left\{ (1 - P(\theta_0, x)) \cdot \exp(\theta_0^T x) - \exp(\theta^T x) \right\} \cdot \exp(\kappa \theta^T x) \cdot x = 0. \tag{17}
$$

Theorem 2 relies on the classic result on the asymptotic variance of Z-estimators of the

form $1/n \sum_i \psi(y_i, x_i, \theta)$, which is given by $J^{-1} I J^{-1}$ with

$$I = \mathbb{E}_x \mathbb{E}\left[\psi(y, x, \tilde{\theta}_0)\psi(y, x, \tilde{\theta}_0)^T \mid x\right],$$
$$J = \mathbb{E}_x \mathbb{E}\left[-\frac{\partial \psi(y, x, \tilde{\theta}_0)}{\partial \theta} \mid x\right].$$

Note that we need to evaluate I, J at the limit $\tilde{\theta}_0$ of the generalized PML estimators, which under the censoring model solves (17). In practice, we can form the plug-in estimator of the variance by replacing θ with the generalized PML estimator $\hat{\theta}_0$ solving (13) and replacing the integral with the sample average in (16). Note that unlike the bias, the asymptotic variance depends on all three components of our framework: α, which governs heteroskedasticity, $P(\theta_0, x)$, which describes the censoring model, and κ, which determines the generalized PML estimator.

3.3 Bias-Variance Trade-off

Having established the bias and asymptotic variance of generalized PML estimators under the heteroskedasticity model (12) and the censoring model (8), we now turn to the bias-variance trade-off under heteroskedasticity and censoring. To be concrete, we focus on the NLS ($\kappa = 1$) and Poisson PML ($\kappa = 0$) estimators in the one-dimensional setting, but our analysis can be extended to the general case. We have the following formula for the bias.

Corollary 1. Assume that the parameter θ is one-dimensional. Near θ_0, the biases of NLS and Poisson PML estimators are approximated by

$$\tilde{\theta}_{PPML} - \theta_0 = -\frac{\int p(X) X \cdot P(\theta_0, X) \cdot \exp(\theta_0^T X) dX}{\int p(X) X^2 \cdot \exp(\theta_0^T X) dX},$$
$$\tilde{\theta}_{NLS} - \theta_0 = -\frac{\int p(X) X \cdot P(\theta_0, X) \cdot \exp(2\theta_0^T X) dX}{\int p(X) X^2 \left(1 + P(\theta_0, X)\right) \exp(2\theta_0^T X) dX}.$$

Comparing the bias of PPML and NLS, we can see that the NLS is more robust to

censoring toward zero. Consider for example the censoring model where

$$P(\theta_0, X) = \begin{cases} 1 & \exp(\theta_0^T X) \leq c \\ 0 & \text{otherwise.} \end{cases}$$

This threshold censoring model rounds down samples with bounded conditional mean to zero, and can be viewed as the limit $\beta \to \infty$ of the censoring models in (12), with $c = 1$. When $c \leq 1$, we have

$$\begin{aligned} &\int p(X)X \cdot P(\theta_0, X) \cdot \exp(\theta_0^T X)dX = \int_{\exp(\theta_0^T X)\leq c} p(X)X \cdot \exp(\theta_0^T X)dX \\ &\geq \int_{\exp(2\theta_0^T X)\leq c} p(X)X \cdot \exp(\theta_0^T X)dX = \int p(X)X \cdot P(\theta_0, X) \cdot \exp(2\theta_0^T X)dX, \end{aligned}$$

assuming that $X \in \mathbb{R}_+$, while

$$\int p(X)X^2 \cdot \exp(\theta_0^T X)dX \leq \int p(X)X^2 \exp(2\theta_0^T X)dX \leq \int p(X)X^2 \left(1 + P(\theta_0, X)\right) \exp(2\theta_0^T X)dX$$

as long as the tail of $p(X)$ decays fast enough. Corollary 1 then implies that $|\tilde{\theta}_{NLS} - \theta_0| \leq |\tilde{\theta}_{PPML} - \theta_0|$.

In the presence of censoring, the asymptotic variance of PPML and NLS can also be computed following Theorem 2. For PPML, we can modify I, J as follows:

$$I_{PPML} = \mathbb{E}_x \mathbb{E}\left[(y - \exp(\theta_0^T x))^2 \cdot xx^T \mid x\right] = \mathbb{E}_x xx^T \left[\exp(\alpha\theta_0^T x) + (\exp(2\theta^T x) - \exp(\alpha\theta_0^T x))P(\theta_0, x)\right],$$

where we note that the first term is the usual variance term for the DGP without censoring. Interestingly, if $\alpha = 2$, the variance of the censoring model coincides with that of the uncensored model, regardless of the actual censoring mechanism. When $\alpha = 0$, the variance of NLS is smaller than that of PPML. As α increases from 0 to 2, the variance of NLS starts to increase, while the variance of PPML initially decreases and then increases, but always

staying below that of the NLS. These behaviors lead to a particular value of $\overline{\alpha}$ where the MSE of NLS starts to dominate that of PPML. Therefore, below this level $\overline{\alpha}$ of heteroskedasticity, NLS should in fact be preferred than PPML, even though it may have larger variance.

More generally, if we consider the two dimensional plane where the x axis is α, which quantifies heteroskedasticity, and the y axis is τ, which quantifies asymmetric censoring towards zero in (9), we should expect to observe a **phase transition**, where in part of the region, NLS will be preferable, while in the other part of the region, PPML will be preferable. In Section 4, we demonstrate with extensive numerical experiments that such a phase transition indeed occurs. See, for example, Fig. 6.

Lastly, we remark on the behavior of the negative binomial (NB) compared to the class of generalized PML estimators in (13). Recall the estimating equations of NB PML:

$$\sum_i (y_i - \exp(\theta^T x_i))(1 + \exp(\theta^T x_i))^{-1} x_i = 0.$$

Compared to the Gamma PML, whose estimating equations are

$$\sum_i (y_i - \exp(\theta^T x_i))(\exp(\theta^T x_i))^{-1} x_i = 0,$$

NB PML is similarly robust to high degrees of heteroskedasticity and large outliers. However, since it does not place overly large weights on samples with small $\exp(\theta^T x_i)$, it has the additional advantage of being more robust to excess sparsity than the Gamma PML. In fact, for samples near zero, it places similar weights as the Poisson PML. Therefore, NB PML should be preferable when both heteroskedasticity and excess sparsity are prominent. We can similarly perform a trade-off analysis for generalized PML estimators of the form

$$\sum_i (y_i - \exp(\theta^T x_i))(1 + \exp(\theta^T x_i))^{-\kappa} x_i = 0.$$

The bias-variance trade-off that we consider in this paper has important implications

on how we should handle data with non-negative outcomes in practice. Although Poisson PML and negative binomial PML are popular choices in practice, our results suggest a more principled approach to select the optimal (generalized) PML estimator given the magnitudes of heteroskedasticity and excess sparsity of the data at hand. These characteristics can in principle be determined from data, using for example statistical tests (Park, 1966; Mullahy, 1986), although we leave a systematic study of this topic to future work. The ability to distinguish between heteroskedasticity and excess sparsity is also important, as they could lead to similar dispersion behaviors in the data, but require opposite types of PML estimators. Zorn (1998) discusses the effect of zero inflation and hurdle models on the dispersion of data. The presence of excess zeros could result in over-dispersion of the data when it is generated by zero inflation, while hurdle process can result in either over or under dispersion. This property is important since in the presence of zero inflation-type censoring, we should use generalized PMLs with larger κ, while if we mistakenly attribute the excessive zeros to over-dispersion, we would be misled to use generalized PMLs with smaller κ.

Next, we discuss the computational considerations of generalized PML estimators. We formulate optimization problems that give rise to the generalized PML estimating equations as first order conditions, which provides a straightforward method for computing the generalized PML estimators given data.

3.4 Computational Considerations for Generalized PML Estimators

Now we turn our attention to computing generalized PML estimators for general $\kappa \in [0, 1]$. Recall the family of generalized PML estimators that we consider

$$\sum_i \left\{ y_i - \exp(\theta^T x_i) \right\} \cdot \exp(\kappa \theta^T x_i) \cdot x_i = 0. \tag{18}$$

An important practical question is how one can solve it for arbitrary κ. In this section, we discuss one approach, which finds local optimizers of optimization problems whose first order

conditions correspond to the estimating equations (18).

For $\kappa = -1$ and 0, the objectives are those of the familiar gamma and Poisson PML:

$$\max_{\theta} \sum_i -y_i \exp(-\theta^T x_i) - \theta^T x_i$$
$$\max_{\theta} \sum_i -\exp(\theta^T x_i) + y_i \theta^T x_i,$$

which are both concave in θ. More generally, we can verify that when $\kappa \in (-1, 0) \cup (0, 1)$, (18) is the first order condition of the problem

$$\max_{\theta} \sum_i \frac{1}{\kappa} y_i \exp(\kappa \theta^T x_i) - \frac{1}{\kappa + 1} \exp((\kappa + 1)\theta^T x_i). \tag{19}$$

For $\kappa \in (-1, 0)$, (19) is concave in θ, which is easy to solve. On the other hand, for $\kappa \in (0, 1)$, (19) is no longer concave in θ, since the term $\frac{1}{\kappa} y_i \exp(\kappa \theta^T x_i)$ is *convex* in θ and the term $-\frac{1}{\kappa+1} \exp((\kappa+1)\theta^T x_i)$ is concave in θ. In this case, we can still use non-convex optimization software to find local maxima, such as `scipy`'s `optimize` module or the Gauss-Newton method (Gratton et al., 2007). In our simulations, we find that the `scipy.optimize` module with `Nelder-Mead` method has comparable accuracy as `CVX` when $\kappa < 0$, and behaves well for $\kappa > 0$. Note also that the global maximizer of (19) is bounded, so that its first order condition is necessary for optimality. Moreover, as long as x_i can take both positive and negative values, the solution to the first order condition is also bounded. To see this point more clearly, consider the first order condition

$$\sum_i \left\{ y_i - \exp(\theta^T x_i) \right\} \cdot \exp(\kappa \theta^T x_i) \cdot x_i = 0.$$

When $-1 < \kappa < 0$, we can verify that with $\theta \to \pm\infty$ the left hand side cannot attain 0, regardless of the signs of x_i. However, when $\kappa > 0$, if $x_i \geq 0$ for all i or $x_i \leq 0$ for all i, then the left hand side can attain a value of zero when $\theta \to -\infty$ or ∞. However, when some x_i is positive and some x_i is negative, the left hand side cannot attain 0 at infinity. We

can therefore better understand the generalized PML estimators as solutions to (19) when $\kappa > 0$. Note also that when $\kappa = 1$, both optimization problems

$$\max_{\theta} y_i \exp(\theta^T x_i) - \frac{1}{2}\exp(2\theta^T x_i) \Longleftrightarrow \min_{\theta}(y_i - \exp(\theta^T x_i))^2$$

yield the same NLS estimator, by expanding the square in the second objective and removing constant terms. Intuitively, large values of y_i will make the NLS objective "less" concave, since the convex term $y_i \exp(\kappa\theta^T x_i)$ will be large. On the other hand, many zero y_i's will result in the concave term $-\frac{1}{2}\exp(2\theta^T x_i)$ dominating, leading to an objective that is closer to being concave. This argument provides some intuition on why NLS could actually be preferable to the Poisson PML.

Lastly, note that compared to $\kappa = -1$ (gamma PML), it is preferable to use $\kappa \in (-1, 0)$ since the objective $\max_\theta \frac{1}{\kappa} y_i \exp(\kappa\theta^T x_i) - \frac{1}{\kappa+1}\exp((\kappa+1)\theta^T x_i)$ will still retain strict concavity when there is a large number of zeros. In comparison, the gamma PML has objective

$$\max_{\theta} \sum_i -y_i \exp(-\theta^T x_i) - \theta^T x_i,$$

which could become close to linear (have a "flat" loss function) when many y_i are zero, resulting in poor solution quality of the optimization problem. This feature may partially explain the poor performance of gamma PML in the presence of many zeros, even under strong over-dispersion.

4 Simulation Studies

In this section, we verify our theoretical findings in this paper, and demonstrate the ubiquity of the bias-variance trade-off for data with non-negative outcomes in simulation studies with different data generating processes.

4.1 Simulation Design

Our simulation design starts with the setup in Silva and Tenreyro (2006), combined with the censoring model that we consider in this paper. First, we generate a 2-dimensional covariate X with its first coordinate X_1 drawn from a standard normal distribution and second coordinate X_2 drawn from the uniform distribution on $[0, 1]$. Given X, the outcome variable is generated according to the model

$$Y_i = \begin{cases} \exp(\theta_0^T X_i) \cdot \eta_i & \text{with probability } 1 - P(\theta_0, X_i) \\ 0 & \text{with probability } P(\theta_0, X_i), \end{cases}$$

where the parameter of interest $\theta_0 = [1, 1]$ and the multiplicative error η_i is log-normal, i.e., $\log \eta_i$ is normal with zero mean. The variance of η_i is set to be $\exp^{\alpha-2}(\theta_0^T X_i)$, so that the conditional variance of $\exp(\theta_0^T X_i) \cdot \eta_i$ is equal to $\exp^{\alpha}(\theta_0^T X_i)$. The censoring probability $P(\theta_0, X_i)$ is given by

$$P(X_i, \theta_0) = \exp(-(\tau \exp(\theta_0^T X_i))^{\beta}) \text{ or } 1/(1 + (\tau \exp(\theta_0^T X_i))^{\beta}),$$

which can be viewed as continuous approximations to the step function. Compared to the double exponential specification, the power decay specification provides a heavier tail, but their qualitative behaviors with respect to β and τ are similar. β regulates the contrast in censoring probabilities between samples with small or large conditional means. As β increases from 1, samples with $\tau \exp(\theta_0^T X_i) < 1$ become increasingly susceptible to censoring, while samples with $\tau \exp(\theta_0^T X_i) \geq 1$ become less affected. τ regulates the point of "discontinuity". As τ increases from 0, the range of $\exp(\theta_0^T X_i)$ for which censoring is significant becomes smaller and more concentrated around 0, resulting in overall *less* sparsity.

We consider the family of generalized PML estimators defined by

$$\frac{1}{n}\sum_i (Y_i - \exp(\theta^T X_i)) \cdot \exp^\kappa(\theta^T X_i) \cdot X_i = 0, \quad \kappa \in [-1, 1],$$

and study their bias and variance. A practical consideration is how to obtain the generalized PML estimators for $\kappa \notin \{-1, 0, 1\}$. See Section 3.4 on how to form optimization problems whose first order conditions recover the estimating equations (13).

4.2 Accuracy of Bias Formula

In this section, we investigate the accuracy of the asymptotic bias and variance formulae derived in Section 3. Since they are based on the population estimating equations

$$\mathbb{E}_x \left\{ (1 - P(\theta_0, x)) \cdot \exp(\theta_0^T x) - \exp(\theta^T x) \right\} \cdot \exp(\kappa \theta^T x) \cdot x = 0,$$

we first verify through simulations that they accurately characterize the expectation of the PML estimators when data is subject to censoring. For each replication, we compute PML estimators $\hat{\theta}_\kappa$ based on 1000 i.i.d. samples, and average over 1000 replications to simulate the expectation $\mathbb{E}\hat{\theta}_\kappa$ of the PML estimators. Then we plug $\mathbb{E}\hat{\theta}_\kappa$ into θ in the above equation, and evaluate the equations by replacing $\mathbb{E}_x$ with average over 10000 samples of x. We find that the moment equations evaluate to similar values near zero compared to the reference case without censoring.

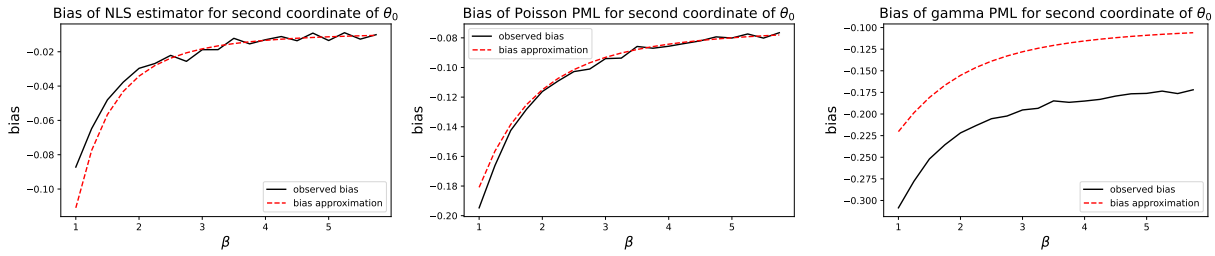


Figure 2: Comparison of simulated and analytic bias with censoring probability $\frac{1}{1+\exp(\beta\theta_0^T x)}$.

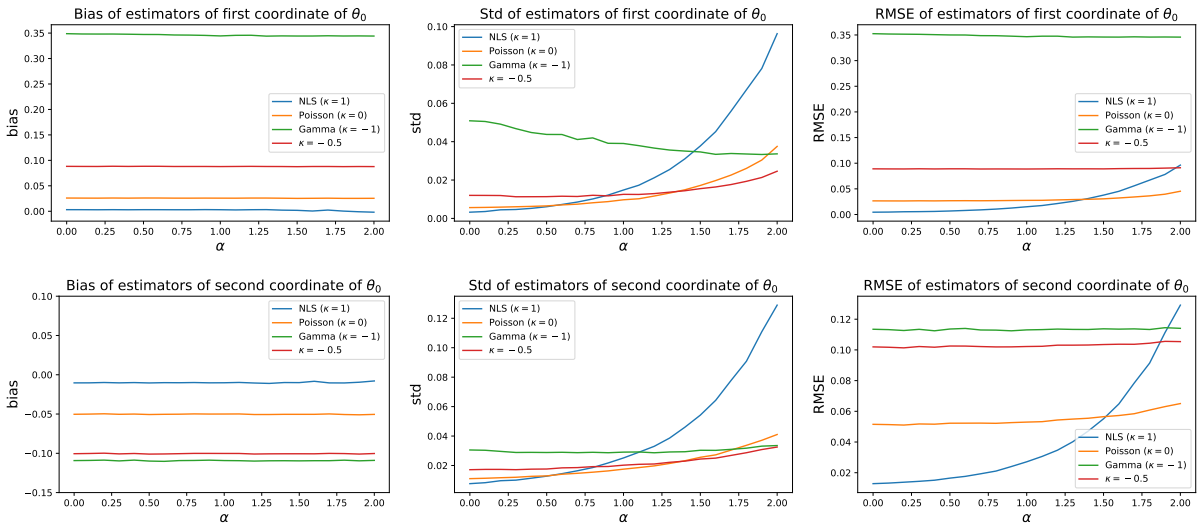


Figure 3: Bias, standard error, and RMSE of generalized PML estimators as a function of the heteroskedasticity parameter α and censoring probability $\frac{1}{1+(\tau \exp(\theta_0^T x))^\beta}$ with $\beta = 2$ and $\tau = 2$. Results are calculated based on 5000 replications of data with sample size 3000.

Next, we investigate the accuracy of our approximation formula (14) for bias in Fig. 2. We find that (14) is quite accurate for NLS and Poisson PML, particularly for $\beta > 1$, but less accurate for the gamma PML, which remains to be understood further.

4.3 Bias-Variance Trade-off

In this part of the simulation studies, we investigate the various manifestations of the bias-variance trade-off, by holding subsets of the tuple $(\alpha, \tau, \beta, \kappa)$ in our framework fixed and varying the rest.

First, in Fig. 3 we plot the bias, standard deviation, and root mean squared error (RMSE) of generalized PML estimators as a function of the degree of heteroskedasticity $\alpha \in [0, 2]$, for fixed $\beta = 2, \tau = 2$. We find that bias is decreasing in κ, and is essentially flat in α, which is expected since the bias approximation formula does not depend on α. Moreover, the standard deviation increases with α for most estimators, but those with larger κ experience faster increase of the standard deviation, as they are less efficient under heteroskedasticity. As a result, when α grows, for the first coordinate of θ_0, the standard deviation of NLS starts

to be dominated by that of the Poisson at $\alpha \approx 0.6$, which in turn is lower bounded by the standard deviation of the generalized PML estimator with $\kappa = -0.5$ starting $\alpha \approx 1.35$ and lower bounded by the gamma PML starting $\alpha \approx 1.5$. Interestingly, the standard deviation of gamma PML ($\kappa = -1$) is *larger* than that of $\kappa = -0.5$ even for $\alpha = 2$, when in the absence of asymmetric censoring gamma PML is most efficient. This is due to censoring having an effect on the asymptotic variance.

If we ignore the presence of bias due to asymmetric censoring, we may think that when heteroskedasticity level $\alpha \in [0.6, 1.35]$, Poisson PML should be preferable, which in fact is what has been recommended as a rule of thumb in practice. However, the RMSE comparisons in Fig. 3 suggest that the NLS in fact dominates the Poisson (and gamma) for all $\alpha \leq 1.35$, due to its smaller bias. Moreover, the efficiency gain of gamma PML and $\kappa = -0.5$ for large α is overwhelmed by their large biases, making the Poisson PML the best option for $\alpha > 1.35$. Similar observations hold for the second coordinate of θ_0. These results suggest that in the presence of sparsity, conventional consensus on the superiority of Poisson PML may be incomplete and misguided, and that the bias plays an important role in a more systematic analysis. As further evidence, in Silva and Tenreyro (2006), the authors conclude that in a similar simulation setup, the Poisson PML outperforms the other estimators in terms of bias and RMSE for a variety of specifications of heteroskedasticity as well as data with many zeros resulting from rounding. However, in their Table 1, we see that the NLS estimator in fact outperforms other estimators when the degree of heteroskedasticity $\alpha < 1$. This observation is especially salient when y_i has rounding error, i.e., small values of y_i are rounded down to 0. When heteroskedasticity intensifies to $\alpha = 2$, NLS is dominated by Poisson PML. Our simulation results in Fig. 3 suggest that, under the common asymmetric censoring model that results in excessive sparsity, the RMSE of NLS continues to dominate that of the Poisson PML beyond $\alpha = 1$, with a transition threshold of $\alpha \approx 1.35$ to 1.5.

Next, in Fig. 4, we plot the bias, variance, and RMSE of generalized PML estimators as a function of τ in the censoring probability $\frac{1}{1+(\tau \exp(\theta_0^T x))^\beta}$ with $\beta = 2$ and $\alpha = 1$. Recall that

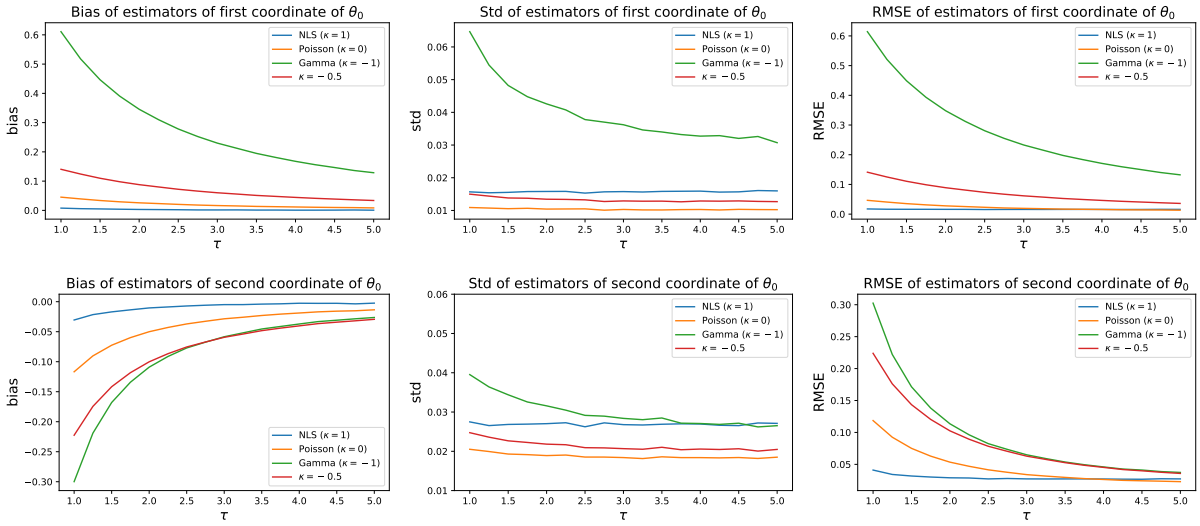


Figure 4: Biases of PML estimators under censoring probability $\frac{1}{1+(\tau \exp(\theta_0^T x))^\beta}$ and $\alpha = 1, \beta = 2$. Results are calculated based on 5000 replications of data with sample size 3000.

increasing τ reduces overall sparsity of the data, so we should expect the bias of estimators to decrease, which is indeed the case. Moreover, the bias plots show that NLS is relatively insensitive to the level of sparsity in the data, whereas PML estimators with negative κ are quite sensitive to small values of τ, when censoring is strong and samples with small conditional means are less reliable. The variance of Poisson is the smallest among all generalized PML estimators, since $\alpha = 1$. The RMSE plots again demonstrate the bias-variance trade-off, with NLS having the smallest RMSE for $\tau \leq 4$, while Poisson starts to have smaller RMSE for $\tau > 4$, when the bias from sparsity becomes less dominant. These results again suggest that in the presence of excess sparsity, contrary to popular belief, the Poisson PML is not necessarily the best choice, even when it is expected to be the most efficient. The optimal choice depends on the degree of censoring.

Our next set of simulation results in Fig. 5 present the bias, standard deviation, and RMSE of generalized PML estimators with $\kappa \in [-1, 1]$, under censoring probability $\frac{1}{1+(\tau \exp(\theta_0^T x))^\beta}$ with $\tau = 1$, $\beta = 2$, and $\alpha \in [0, 1, 2]$,. We find that the bias decreases as κ increases from -1 to 1, as the corresponding generalized PML estimators place increasing weights on larger samples thus reducing bias. Similarly, the standard deviation of estimators decreases then

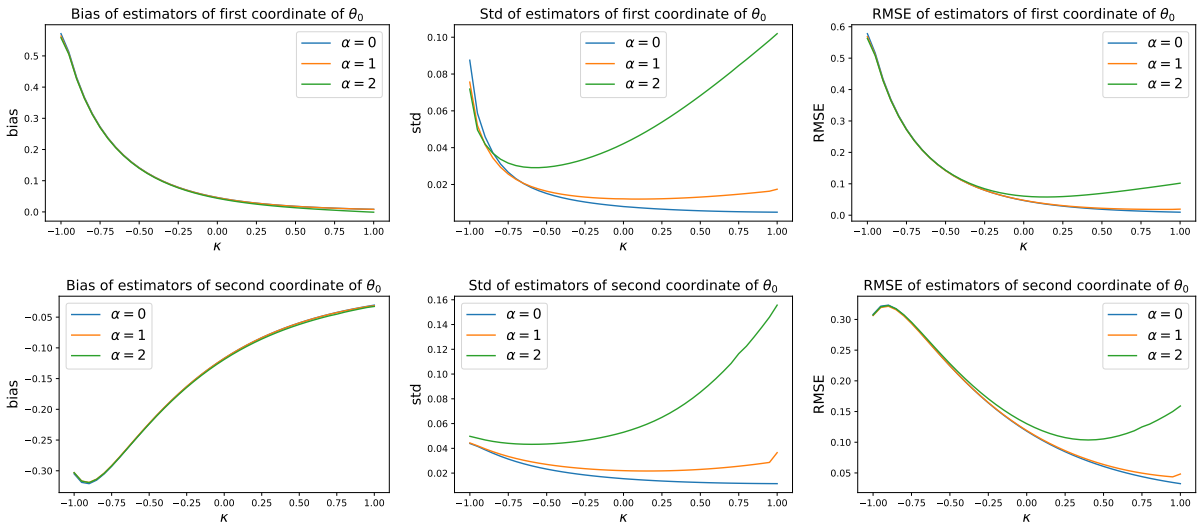


Figure 5: Biases of PML estimators under censoring probability $\frac{1}{1+\exp(\beta\theta_0^T x)}$ and $\beta = 2$.

increases with κ, resulting in a choice of κ with the smallest variance. However, this κ with the smallest standard deviation is generally different from that in the case without censoring. For example, when $\alpha = 2$, we know that the gamma PML ($\kappa = -1$) is the most efficient estimator in the absence of censoring. However, in the presence of asymmetric censoring, the standard deviation is minimized at around $\kappa \approx -0.6$ for both coordinates of θ_0. This observation is in line with our theoretical result on the asymptotic variance, which depends on the censoring parameters. Lastly, the RMSE plots confirm our intuition and quantitative results on the bias-variance trade-off. There is an optimal κ with the smallest RMSE, which depends on the degree of heteroskedasticity α and sparsity τ. For example, although Poisson PML ($\kappa = 0$) is widely used practice especially when heteroskedasticity is not negligible (e.g., $\alpha = 1$), in the presence of moderate censoring ($\tau = 1$ and $\beta = 2$), NLS ($\kappa = 1$) in fact has lower RMSE and should be preferred.

We summarize the preceding three sets of simulation results in the phase transition plot in Fig. 6. The x axis indexes the level of sparsity τ while the y axis indexes the degree of heteroskedasticity α. For each pair (α, τ), we plot the κ corresponding to the optimal generalized PML estimator. We see that as heteroskedasticity decreases or sparsity increases,

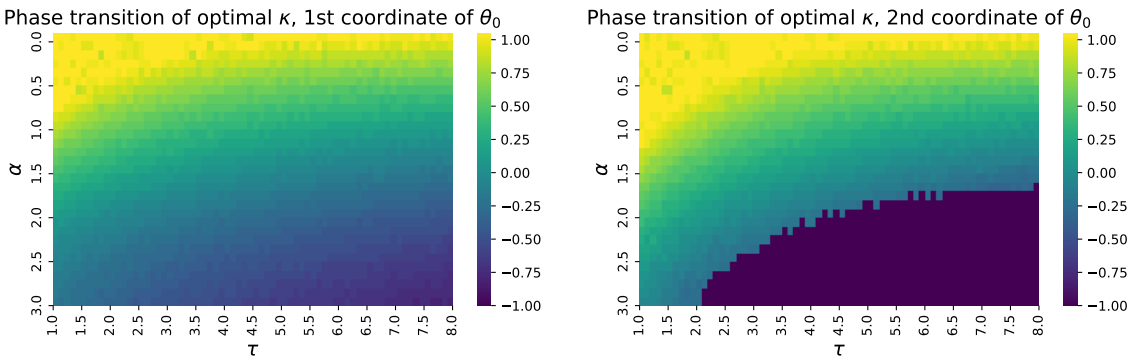


Figure 6: Phase transition of the optimal κ for different levels of heteroskedasticity (α) and sparsity (τ). Results based on simulated data with true parameter $\theta_0 = (1, 1)$.

the optimal κ decreases, and vice versa. Therefore, we advocate for a more systematic modeling of non-negative dependent variables based on assessing the relative importance of the two aspects of data. In order to determine α and τ, one could consider performing tests along the lines of Park (1966); Mullahy (1986) or Cameron and Trivedi (1990). In this paper, we propose an alternative, simpler method based on cross validation, which we describe in detail in the next section, where we apply our framework to a collection of datasets in finance to demonstrate its usefulness.

5 Analysis of four Financial Datasets

In this section, we illustrate the proposed generalized PML approach using data on corporate defaults, credit ratings, and corporate patents. For the first two datasets, we are interested in modeling the default rate and credit ratings for a collection of industries defined in the Moody's 35 Industry Categories[4]. In the third dataset, we model the number of corporate patents granted using the dataset provided by Cohn et al. (2022), who replicate the study of Hirshleifer et al. (2012) on the effect of over-confident CEOs on corporate innovation. We demonstrate that by varying the parameter κ in (13), the resulting generalized PML

[4] https://www.moodys.com/sites/products/ProductAttachments/MCO_35%20Industry%20Categories.pdf

estimators exhibit different behaviors. In particular, we can employ a simple cross validation procedure to choose the optimal generalized PML estimator that could significantly outperform standard PML estimators in terms of out of sample RMSE. We further discuss how we can leverage hypothesis testing to determine the appropriate parameters. Our results in this section demonstrate that the proposed framework provides a practical approach to modeling data with non-negative outcome variables.

5.1 Data

We construct our datasets on corporate default and credit rating from several sources. For the outcome variable, we leverage data from the Moody's Default and Recovery Database (DRD), which provides a detailed history of firm defaults across a broad collection of industries. To construct the outcome variable, we aggregate the total number of default events across all firms in each of the Moody's 35 Industry Categories for each month between January 2019 and December 2023. Each observation will consist of the monthly default number for a particular industry category and particular month-year, which is a sparse non-negative variable. Default events as defined in the Moody's database include bankruptcy, missed payments, distressed exchange, and others. Fig. 7 provides a histogram of the number of monthly default events. The majority of industry-year/month pairs have zero defaults, as expected. A large outlier of 35 events is recorded for February 2023 in the banking industry. These events correspond to the series of bank failures and bankruptcies in early 2023. The ratings outcome is constructed in a similar way: we compute the monthly number of firms in each of Moody's Industry Categories that are rated investment grade, i.e., Baa3 or higher. We then normalize the counts by their standard deviations in each month. Distribution of this non-negative outcome variable is also described in the histogram in Fig. 7.

To construct the predictors for default rates, we follow the setup in Duffie et al. (2007) and consider the following macroeconomic and firm-specific factors:

- Average firm's distance to default, constructed based on Vassalou and Xing (2004);

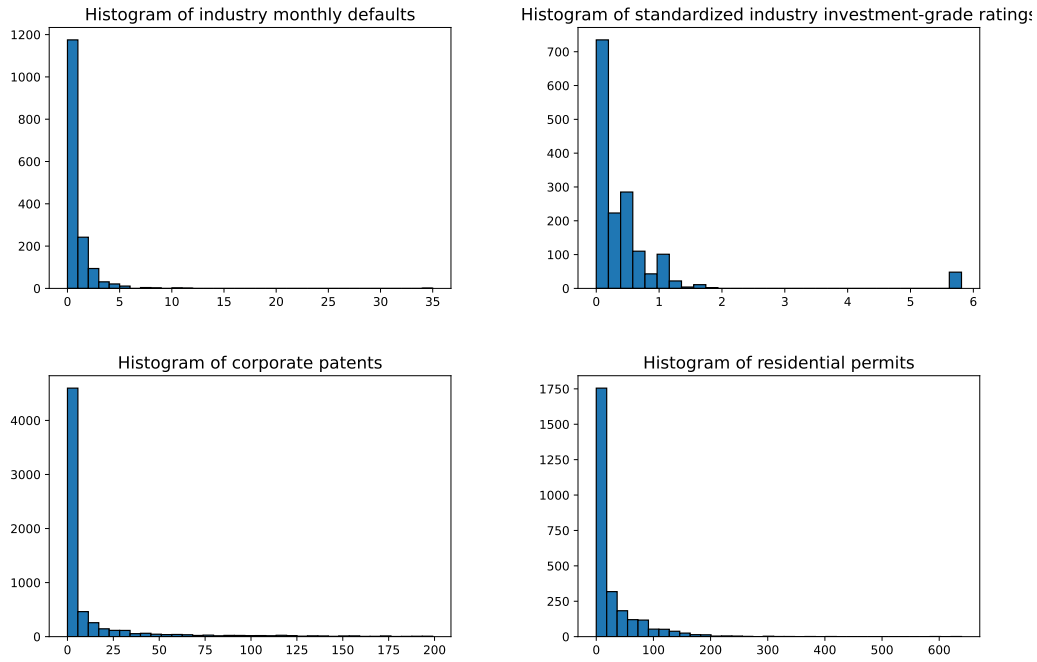


Figure 7: Upper left: Histogram of monthly defaults in Moody's 35 Industrial Categories between January 2020 and December 2023. A single large outlier of 35 events is recorded for February 2023 in the banking industry. Upper right: Histogram of monthly average ratings, using a numerical conversion of the Moody's ratings where smaller values correspond to higher ratings. Lower left: Replication dataset of Hirshleifer et al. (2012) constructed by Cohn et al. (2022) of corporate patents, truncated to samples with $y \leq 200$. Lower right: Residential permits based on data used in Bekkerman et al. (2023).

- Monthly industry average of trailing 1-year stock return;
- The 3-month Treasury bill rate;
- Trailing 1-year return on the S&P 500 index.

The T-bill data is obtained from the Federal Reserve Bank of St. Louis, the trailing 1-year return on the S&P 500 index is obtained from CRSP. Monthly industry average of trailing 1-year stock return is obtained from CRSP/Compustat. To calculate firm's distance to default, we combine several data sources available on WRDS and apply the iterative algorithm proposed by Vassalou and Xing (2004). In order to match stock returns data from CRSP with the Moody's 35 Industry Categories, we use the 3-digit NAICS code in the Moody's dataset and the first 3 digits of the 6-digit NAICS code in the CRSP data.

Similarly, for the credit rating outcome, we use the following predictors constructed from CRSP/Compustat data that have been found to have predictive power (Hirk et al., 2022):

- Total debt to total asset ratio;
- Total debt to earnings before interest, taxes, depreciation, and amortization ratio;
- Retained earnings divided by total assets;
- Long-term debt to assets ratio;
- Earnings before interest and taxes divided by total assets

5.2 Predicting Industry Level Default Rates and Credit Ratings

In this section, we investigate the performance of generalized PML estimators when they are used to predict default rates and credit ratings at the industry sector level. To this end, we split the samples 80%–20% randomly into train and test sets. For default rates, each sample consists of the number of defaults in a particular Moody's 35 Industry Category in each year-month (outcome variable), along with that category's average distance to default, average trailing 1 year stock return, as well as the 3 month T-bill rate and trailing 1 year S&P 500 returns (covariates). For credit ratings, each sample consists of the normalized number of investment-grade firms for each industry category in each year-month (outcome variable), along with the indicators of solvency and financial soundness (covariates).

We estimate generalized PML estimators with $\kappa \in [-1, 1]$ on the train set, and report the test set RMSEs averaged over 2000 random splittings of the data. When $\kappa \neq -1, 0, 1$, the corresponding generalized PML estimator is obtained by solving (19) using the `scipy.optimize` module. The average out of sample RMSEs for default rates are summarized in Fig. 8. We observe that the RMSE is optimized at $\kappa = 0.10$, with an average value of 0.435. Therefore, for the purpose of predicting corporate defaults in this setting, the conventional choice of Poisson PML estimator with $\kappa = 0$ is close to the optimal choice. However, this is not

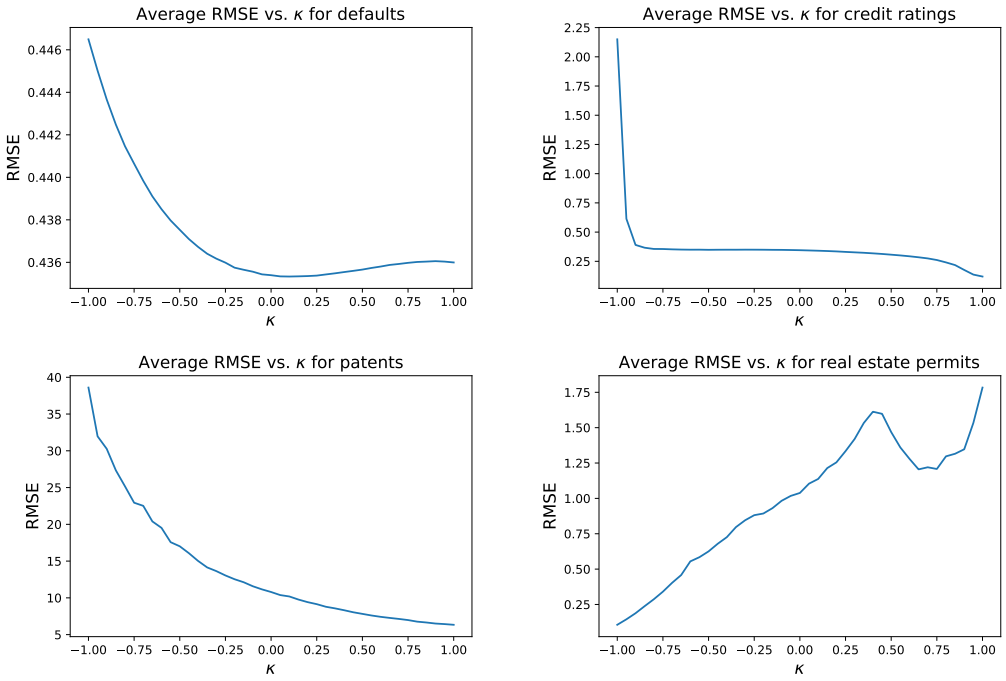


Figure 8: Average out of sample RMSEs of generalized PML estimators on corporate defaults (top left), credit ratings data (top right), corporate patents (bottom left), and residential permits (bottom right).

always the case for other datasets and applications, as demonstrated by the results in Fig. 8 for credit ratings, and it is important to determine the optimal generalized PML estimator in a principled manner, e.g., using the cross validation procedure we propose. Moreover, for the default data, the RMSEs of generalized PML estimators with $\kappa \geq 0$ tend to outperform those with $\kappa < 0$. This provides evidence that the excess sparsity feature of the default data dominates efficiency considerations, leading us to favor estimators with smaller weights on samples with small means. On the other hand, the results for credit rating prediction, summarized in Fig. 8, suggest that NLS ($\kappa = 1$) has a significantly better fit of the credit rating data compared to the Poisson or other generalized PML estimators. This demonstrates that, although NLS is generally considered inefficient when modeling non-negative outcome data, it could still be preferable in some cases, particularly under excess sparsity. To assess the degree of heteroskedasticity, in Fig. 9 we provide box plots of the outcome variable y_i as a function of the conditional mean $\exp(\hat{\theta}^T x_i)$ predicted by the estimated models. We see that

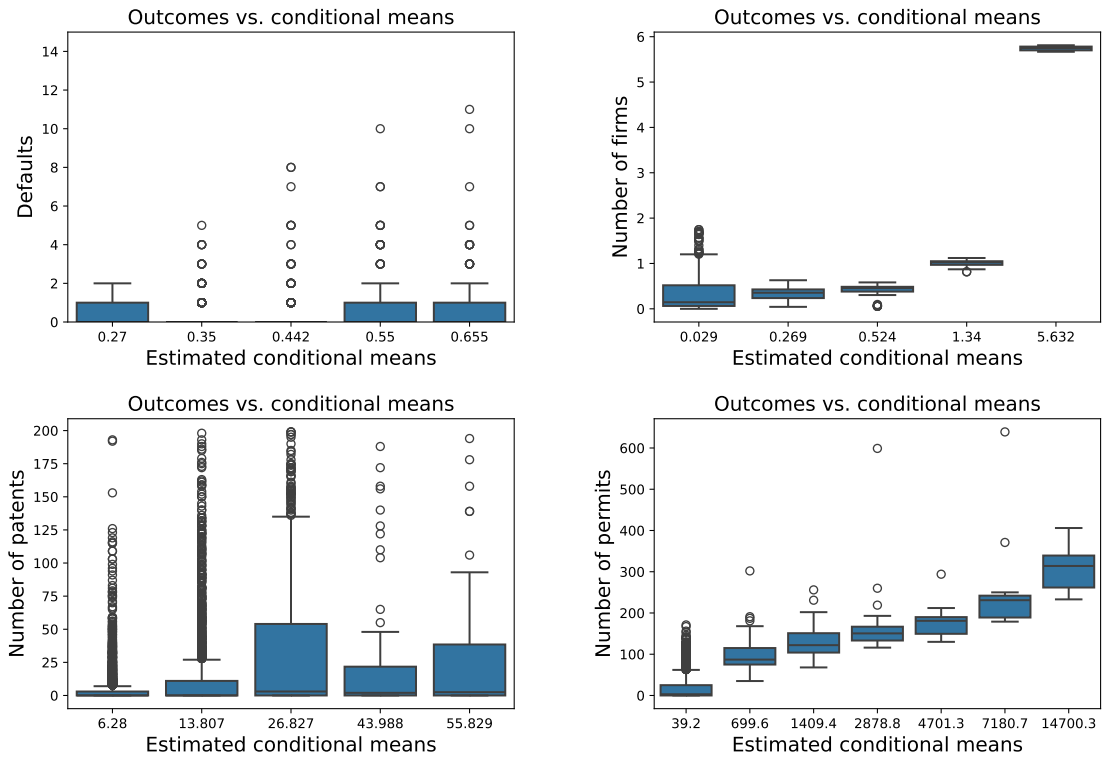


Figure 9: Box plots of outcomes vs. estimated conditional means for the default, credit rating, patent, and real estate data. For ratings and patents data, heteroskedasticity is relatively mild, which justifies the use of NLS ($\kappa = 1$). On the other hand, heteroskedasticity is significant for real estate permits data, and moderate for defaults data, which explains the optimal choices of $\kappa = -1$ and $\kappa = 0.1$, respectively.

although the defaults data exhibits slightly higher sparsity, as illustrated by the histogram Fig. 7, its degree of heteroskedasticity is considerably higher than that of the ratings data, which justifies the superior RMSE of NLS over that of the Poisson PML on the ratings data.

Our results in this section demonstrate that the out of sample RMSE can be used as a general cross validation metric to determine the optimal κ when modeling data with non-negative outcomes. For example, when estimating a gravity model on trade data, we can generate a plot similar to that in Fig. 8, and use it to assess the relative performance of the family estimators with $\kappa \in [-1, 1]$. In settings with strong heteroskedasticity, an estimator with $\kappa \leq 0$ should generally be preferred, while in settings with excess sparsity, an estimator with $\kappa \geq 0$ will be selected. We conclude by noting that other procedures to select κ may be more appropriate depending on the modeling task. For example, we may determine the

optimal κ based on the bias-variance trade-off analysis in this paper. To determine the degree α of heteroskedasticity, we can employ hypothesis tests discussed in Manning and Mullahy (2001); Silva and Tenreyro (2006). To determine the degree β of asymmetric censoring, we can consider estimating a zero-inflated model (Lambert, 1992) and use the estimated parameter of zero-inflation. We leave a systematic study of this proposal to future works.

6 Conclusion

In this paper, we show that existing approaches to modeling non-negative count-like data in finance may not adequately capture two prominent features of the data: heteroskedasticity and sparsity. We establish framework under which to study the trade-off the two features. We propose a novel class of estimators we call generalized pseudo maximum likelihood (PML) estimators, which includes the classic PML estimators such as the Poisson PML and non-linear least squares (NLS). Given a particular dataset with non-negative outcomes, we propose a simple cross validation procedure that enables the researcher to select the estimator that best fits the data. We demonstrate with extensive numerical studies the validity and appeal of our approach. In particular, although NLS (and other generalized PML estimators with $\kappa > 0$) is generally considered to be inefficient under heteroskedasticity and over-dispersion and thus avoided in practice, we find in our numerical and empirical studies that it could be preferable for some datasets when there is strong sparsity and the associated optimization problem is solved appropriately. Our work therefore calls for a more systematic assessment of count-like data in finance and provides a principled approach to the modeling of such data in practice.

References

Addoum, J. M., Ng, D. T., and Ortiz-Bobea, A. (2023). Temperature shocks and industry earnings news. *Journal of Financial Economics*, 150(1):1–45.

Akey, P. and Appel, I. (2021). The limits of limited liability: Evidence from industrial pollution. *The Journal of Finance*, 76(1):5–55.

Amemiya, T. (1985). Advanced econometrics. *Harvard University Press*, 2:153–161.

Anderson, J. E. (1979). A theoretical foundation for the gravity equation. *American Economic Review*, 69(1):106–116.

Anderson, J. E. and Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, 93(1):170–192.

Bekkerman, R., Cohen, M. C., Kung, E., Maiden, J., and Proserpio, D. (2023). The effect of short-term rentals on residential investment. *Marketing Science*, 42(4):819–834.

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999). The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 162(2):195–209.

Boulton, A. J. and Williford, A. (2018). Analyzing skewed continuous outcomes with many zeros: A tutorial for social work and youth prevention science researchers. *Journal of the Society for Social Work and Research*, 9(4):721–740.

Cameron, A. C. and Trivedi, P. K. (1990). Regression-based tests for overdispersion in the poisson model. *Journal of econometrics*, 46(3):347–364.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression Analysis of Count Data*. Number 53. Cambridge university press.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160.

Chen, J. and Roth, J. (2024). Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics*, 139(2):891–936.

Cohn, J. B., Liu, Z., and Wardlaw, M. I. (2022). Count (and count-like) data in finance. *Journal of Financial Economics*, 146(2):529–551.

Coles, J. L., Daniel, N. D., and Naveen, L. (2006). Managerial incentives and risk-taking. *Journal of financial Economics*, 79(2):431–468.

Craig, B. and Von Peter, G. (2014). Interbank tiering and money center banks. *Journal of Financial Intermediation*, 23(3):322–347.

Dow, W. H. and Norton, E. C. (2003). Choosing between and interpreting the heckit and two-part models for corner solutions. *Health Services and Outcomes Research Methodology*, 4:5–18.

Duffie, D., Saita, L., and Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, 83(3):635–665.

Frankel, J. A. (1997). Regional trading blocs in the world economic system. *Institute for International Economics*.

Gourieroux, C., Monfort, A., and Trognon, A. (1984a). Pseudo maximum likelihood methods: Applications to poisson models. *Econometrica: Journal of the Econometric Society*, pages 701–720.

Gourieroux, C., Monfort, A., and Trognon, A. (1984b). Pseudo maximum likelihood methods: Theory. *Econometrica: journal of the Econometric Society*, pages 681–700.

Gratton, S., Lawless, A. S., and Nichols, N. K. (2007). Approximate gauss–newton methods for nonlinear least squares problems. *SIAM Journal on Optimization*, 18(1):106–132.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models.

Hausman, J., Hall, B. H., and Griliches, Z. (1984). Econometric models for count data with an application to the patents-r & d relationship. *Econometrica: Journal of the Econometric Society*, pages 909–938.

He, J. J. and Tian, X. (2013). The dark side of analyst coverage: The case of innovation. *Journal of Financial Economics*, 109(3):856–878.

Hirk, R., Vana, L., and Hornik, K. (2022). A corporate credit rating model with autoregressive errors. *Journal of Empirical Finance*, 69:224–240.

Hirshleifer, D., Low, A., and Teoh, S. H. (2012). Are overconfident ceos better innovators? *The Journal of Finance*, 67(4):1457–1498.

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics*, 36(2):299–305.

Hollingsworth, A., Karbownik, K., Thomasson, M. A., and Wray, A. (2024). The gift of a lifetime: The hospital, modern medicine, and mortality. *American Economic Review*, 114(7):2201–2238.

King, G. (1988). Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential poisson regression model. *American Journal of Political Science*, pages 838–863.

King, G. (1989). Event count models for international relations: Generalizations and applications. *International Studies Quarterly*, 33(2):123–147.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Leung, S. F. and Yu, S. (1996). On the choice between sample selection and two-part models. *Journal of Econometrics*, 72(1-2):197–229.

Manning, W. G., Duan, N., and Rogers, W. H. (1987). Monte carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics*, 35(1):59–82.

Manning, W. G. and Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20(4):461–494.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365.

Mullahy, J. and Norton, E. C. (2022). Why transform y? a critical assessment of dependent-variable transformations in regression models for skewed and sometimes-zero outcomes. Technical report, National Bureau of Economic Research.

Park, R. E. (1966). Estimation with heteroscedastic error terms. *Econometrica*, 34(4).

Sautner, Z., Van Lent, L., Vilkov, G., and Zhang, R. (2023). Firm-level climate change exposure. *The Journal of Finance*, 78(3):1449–1498.

Silva, J. S. and Tenreyro, S. (2006). The log of gravity. *Review of Economics and Statistics*, 88(4):641–658.

Silva, J. S. and Tenreyro, S. (2010). On the existence of the maximum likelihood estimates in poisson regression. *Economics Letters*, 107(2):310–312.

Silva, J. S. and Tenreyro, S. (2011). Further simulation evidence on the performance of the poisson pseudo-maximum likelihood estimator. *Economics Letters*, 112(2):220–222.

Silva, J. S. and Tenreyro, S. (2022). The log of gravity at 15. *Portuguese Economic Journal*, 21(3):423–437.

Vassalou, M. and Xing, Y. (2004). Default risk in equity returns. *The Journal of Finance*, 59(2):831–868.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25.

Wooldridge, J. M. (1999). Quasi-likelihood methods for count data. *Handbook of applied econometrics volume 2: Microeconomics*, pages 321–368.

Zorn, C. J. (1998). An analytic and empirical examination of zero-inflated and hurdle poisson specifications. *Sociological Methods & Research*, 26(3):368–400.

A Introduction to Pseudo Maximum Likelihood

We begin our discussions of pseudo maximum likelihood by introducing the seminal works of Silva and Tenreyro (Silva and Tenreyro, 2006, 2011), who popularized the use of Poisson pseudo maximum likelihood estimation for non-negative outcome data in economics.

A.1 The Empirical Success of Poisson Pseudo Maximum Likelihood

In Silva and Tenreyro (2006), the authors focus on a class of multiplicative trade models

$$Y_{ij} = \alpha_0 G_i^{\alpha_1} G_j^{\alpha_2} D_{ij}^{\alpha_3} \eta \equiv \exp(\beta^T X) \cdot \eta, \tag{20}$$

where Y_{ij} is trade volume, G_i is GDP of country i, and D_{ij} is distance between two countries. In the presence of heteroskedasticity, they demonstrate that log-linearized models can be severely biased, whereas PPML is robust. The reason behind the bias is that in the log-linear specification

$$\log(1 + Y) = \beta^T X + \log \eta,$$

of the trade equation, where η is heteroskedastic, e.g., $\mathrm{Var}(\eta) = f(X)$, the logarithm of η is correlated with X as well and $\mathbb{E}(\log \eta \mid X) \neq 0$, violating the requirement for consistency of OLS estimator based on the log-linear model.

Moreover, an important feature of trade data is that there could be many zero observations of Y_{ij}, i.e., no trade between pairs of countries or small trade volumes that are often *rounded* to zero. This can also be interpreted as $D_{ij} = +\infty$. In the presence of many zeros, Silva and Tenreyro (2011) demonstrate with extensive simulations that PPML is more robust compared to log-linear models, as well as other pseudo maximum likelihood methods.

Given the specification $Y = \exp(\theta^T X) + \epsilon$ with $Y \geq 0$ and $\mathbb{E}(\epsilon \mid X) = 0$, here is the intuition why the model is inherently heteroskedastic. When $\exp(\theta^T X)$ is small, in order

to guarantee $Y \geq 0$, the dispersion of ϵ around 0 must be very small, whereas for larger $\exp(\theta^T X)$, the variance of ϵ is allowed to be larger. In other words, the heteroskedasticity of the model is inherent in the specification $Y = \exp(\theta^T X) + \epsilon$.

Assuming $\text{Var}(\epsilon \mid X) \propto \mathbb{E}(\epsilon \mid X) = \exp(\theta^T X)$, we can consider the efficient *weighted* nonlinear least squares estimator

$$\min_{\theta} \sum_i \exp(-\theta^T x_i)(y_i - \exp(\theta^T x_i))^2$$

which has first order equations $\sum_i \exp(\theta^T x_i)x_i - x_i \exp(-\theta^T x_i)y_i^2 = 0$, which is asymptotically equivalent to the first order conditions of the PPML objective

$$\sum_i (y_i - \exp(\theta^T x_i))x_i = 0.$$

In practice, when we do not know the variance, the unweighted NLS is used. Silva and Tenreyro (2006) observe that the estimator resulting from maximizing the PPML objective is much better compared to the NLS estimator, even when $\text{Var}(\epsilon \mid X) \propto \mathbb{E}(\epsilon \mid X)$ fails, e.g., $\text{Var}(\epsilon \mid X) \propto \mathbb{E}(\epsilon \mid X)^2$. To summarize, Silva and Tenreyro (2006) observe that PPML has superior performance across a variety of patterns of heteroskedasticity as well as in settings with many zero responses, whereas alternative estimators such as NLS and log-linear OLS only perform well when the specification of variance is correct.

Remark 3. **Poisson Approach for General Response.** Given (Y, X) with $Y \in \mathbb{R}$, we can also apply the transformation $Y \to \exp(Y)$ to obtain a positive response variable. Given the transformed data, we can then perform pseudo Poisson maximum likelihood, an consistency is guaranteed as long as $\mathbb{E}[\exp(Y) \mid X] = \exp(\theta^T X)$. When is this condition satisfied and how does this approach compare to the usual OLS regression? The requirement is then $\mathbb{E}(\exp(r) \mid X] = 1$ instead of $\mathbb{E}[r \mid X] = 0$ if $Y = X^T\beta + r$

Remark 4. **Connections to GLM.** The consistency of PPML given log-linear conditional

expectations is a special case of the general result about generalized linear models[5]:

Given $\mathbb{E}(y \mid x) = \sigma(\theta^T x)$ for some monotonic σ, there exists a convex loss function, the "matching" or "surrogate" loss, whose minimizer is a consistent estimator of θ. This convex loss function can be written as

$$r(\theta; x, y) = \int_0^{\theta^T x} (\sigma(z) - y) dz$$

and an also be represented as a Bregman divergence. For PPML, σ is the exponential, and we can check that r is the negative of the pseudo log-likelihood $y_i \cdot \theta^T x_i - \exp(\theta^T x_i)$. This connection to GLM suggests that the general robustness properties of PPML extends to other GLMs as well. Moreover, for different applications we may choose the most appropriate σ depending on the data and assumptions.

Remark 5. The connection to GLM does not fully answer the question of why PPML is suitable for non-negative response data. In particular, other GLM models can also be used for non-negative responses, such as the negative binomial. But in simulations and in practice, as observed by Silva and Tenreyro (2006, 2011); Gourieroux et al. (1984a,b), PPML outperforms alternatives under heteroskedasticity and sparse data (many $y_i = 0$).

A.2 Gourieroux et al. (1984a,b)

Consider the linear exponential family with density

$$l(u, m) = \exp(A(m) + B(u) + C(m)u)$$

with $A(m), B(u)$ are scalar and $C(m)$ is vector, and m is the mean of the distribution.

Besides the standard Poisson model, they also consider the Poisson model with specification

[5]See section 8.5 of https://www.cs.ox.ac.uk/people/varun.kanade/teaching/CLT-MT2022/lectures/CLT.pdf

error, which seems to be equivalent to the doubly stochastic Poisson models:

$$\mathbb{P}(y_i \mid x_i) = \int \frac{\exp(-\exp(\theta^T x_i + \delta_i))}{y_i} \exp(y_i(\theta^T x_i + \delta_i)) g(\delta_i) d\delta_i$$

In other words, for each observation the Poisson parameter is assume to be *random* and of the form $\theta^T x_i + \delta_i$. When $\exp(\delta_i)$ has a Gamma distribution $\gamma(\alpha, \beta)$, y_i has a negative binomial distribution, and one can obtain the ML estimator of θ. However, if $\exp(\delta_i)$ does not have a Gamma distribution, the MLE is inconsistent, but the PPML estimator is consistent for all distributions of $\exp(\delta_i)$ having second moment.

B Generalized PML Estimators for General Conditional Mean Models

Instead of assuming $f(X, \theta) = \exp(\theta^T X)$, we can consider the more general specification $f(X, \theta) = f(\theta^T X)$. Then the general first order conditions of the NLS, Poisson, Gamma, and negative binomial PMLs, are given by Gourieroux et al. (1984b):

$$\begin{aligned}
\text{NLS:} \quad & \sum_i (f'(\theta^T x_i) - y_i f'(\theta^T x_i)/f(\theta^T x_i)) \cdot f(\theta^T x_i) \cdot x_i = 0 \\
\text{Poisson:} \quad & \sum_i (f'(\theta^T x_i) - y_i f'(\theta^T x_i)/f(\theta^T x_i)) \cdot x_i = 0 \\
\text{Gamma:} \quad & \sum_i (f'(\theta^T x_i) - y_i f'(\theta^T x_i)/f(\theta^T x_i)) \cdot \frac{1}{f(\theta^T x_i)} \cdot x_i = 0 \\
\text{Negative Bionomial:} \quad & \sum_i (f'(\theta^T x_i) - y_i f'(\theta^T x_i)/f(\theta^T x_i)) \cdot \frac{1}{1 + f(\theta^T x_i)} \cdot x_i = 0.
\end{aligned}$$

More generally, we can consider a continuous family of estimators indexed by $\kappa \in \mathbb{R}$ that solve the estimating equations

$$\sum_i (f'(\theta^T x_i) - y_i f'(\theta^T x_i)/f(\theta^T x_i)) \cdot f^\kappa(\theta^T x_i) \cdot x_i = 0, \tag{21}$$

as well as

$$\sum_i (f'(\theta^T x_i) - y_i f'(\theta^T x_i)/f(\theta^T x_i)) \cdot (1 + f(\theta^T x_i))^\kappa \cdot x_i = 0. \tag{22}$$

When $\kappa \geq 0$, the resulting estimator is robust to measurement errors that bias towards 0, while when $\kappa \leq 0$, the resulting estimator is robust to measurement errors that bias towards ∞. As a result, the Poisson estimating equation $\sum_i (f'(\theta^T x_i) - y_i f'(\theta^T x_i)/f(\theta^T x_i)) \cdot x_i = 0$ corresponding to $\kappa = 0$ is a reasonable compromise that is robust to both types of measurement errors. However, depending on how the conditional variance varies with the conditional mean. If the variance increases, then NLS maybe be robust to bias towards 0, but not to heteroskedasticity. However, based on observations if

$$V(Y \mid X) = \mathbb{E}[Y_i \mid X_i],$$

NLS may still be ok, but if

$$V(Y \mid X) = \mathbb{E}^2[Y_i \mid X_i],$$

NLS performs very poorly. So it could be possible that when V does not grow too fast, the benefit of NLS that are robust to small outliers can outweigh the disadvantage of not being robust to heteroskedastic errors. For example, in the simulation results in Table 1 of Silva and Tenreyro (2006), the NLS performs the best even when $V(Y \mid X) = \mathbb{E}[Y_i \mid X_i]$ and there are rounding errors to 0. But its performance significantly deteriorates when $V(Y \mid X) = \mathbb{E}^2[Y_i \mid X_i]$.

If we consider a graph where the x axis is how variance increases with conditional mean, and the y axis is how much measurement bias towards 0 is, we will observe a phase transition where in some regions each of the estimators will be preferable.

Can also consider the Negative Binomial

$$\sum_i (y_i - \exp(\theta^T x_i))(1 + \exp(\theta^T x_i))^{-1} x_i = 0.$$

Compared with Gamma,

$$\sum_i (y_i - \exp(\theta^T x_i))(\exp(\theta^T x_i))^{-1} x_i = 0,$$

NB is similarly robust to high degrees of heteroskedasticity and large outliers. However it has the additional advantage of being robust to small outliers, just like the Poisson PML. Similarly, if we do not want to downweight large samples as much, an estimator satisfying

$$\sum_i (y_i - \exp(\theta^T x_i))(1 + \exp(\theta^T x_i))^{-\alpha} x_i = 0$$

for some $\alpha \in (0, 1)$ will be closer in behavior to Poisson for larger samples.

Therefore, we may understand the pseudo ML estimators in terms of how they weight observations using the value of the conditional mean $f(\theta^T X)$. There are at least three sources of mis-specification/measurement errors that can impact the performance of PML estimators:

1. Heteroskedasticity, most commonly variance increasing with mean. Estimators that **downweight** large samples generally have better performance;

2. Measurement bias towards 0, i.e., small responses being recorded as 0. Estimators that **upweight** large samples generally have better performance;

3. Measurement bias towards ∞, i.e., large responses being distorted more. For example, if a measurement error of the form $(1 + \alpha)Y_i$ results in large samples being distorted more. Estimators that **downweight** large samples generally have better performance. This type of measurement bias could be behind data where there are some large outliers.

We therefore see that depending on the severity of each issue, it is reasonable to expect

different PML estimators can perform differently. We therefore propose to jointly analyze the trade-offs of these issues. The key components of our analysis will be the degree of heteroskedasticity and measurement bias towards 0 (and ∞). Given these parameters of the problem, there will be a particular PML that is optimal, and our goal is to characterize that PML.

C Proofs

C.1 Proof of Theorem 1

Proof. First note that without censoring,

$$\begin{aligned}\mathbb{E}(y - \exp(\theta^T x)) \cdot \exp(\kappa\theta^T x) \cdot x &= \mathbb{E}(\exp(\theta_0^T x + \epsilon) - \exp(\theta^T x)) \cdot \exp(\kappa\theta^T x) \cdot x \\ &= \mathbb{E}(\exp(\theta_0^T x) - \exp(\theta^T x)) \cdot \exp(\kappa\theta^T x) \cdot x\end{aligned}$$

using $\mathbb{E}[\exp(\epsilon) \mid x] = 1$. Therefore when $\theta = \theta_0$, the above expectation is zero.

On the other hand, under censoring with probability $P(\theta_0, x)$:

$$\begin{aligned}Y_i^{true} &= \exp(\theta_0^T X_i + \epsilon_i) \\ Y_i &= \begin{cases} Y_i^{true} & 1 - P(\theta_0, X_i) \\ 0 & P(\theta_0, X_i), \end{cases}\end{aligned}$$

we have

$$\mathbb{E}(y - \exp(\theta^T x)) \cdot \exp(\kappa\theta^T x) \cdot x = \mathbb{E}_x \mathbb{E}\left[(y - \exp(\theta^T x)) \mid x\right] \cdot \exp(\kappa\theta^T x) \cdot x$$

and

$$\begin{aligned}\mathbb{E}\left[(y-\exp(\theta^T x)) \mid x\right] &= (1-P(\theta_0,x))\cdot(\exp(\theta_0^T x)-\exp(\theta^T x)) + P(\theta_0,x)(-\exp(\theta^T x)) \\ &= (1-P(\theta_0,x))\cdot\exp(\theta_0^T x)-\exp(\theta^T x)\end{aligned}$$

so that the moment conditions become

$$\mathbb{E}_x\left\{(1-P(\theta_0,x))\cdot\exp(\theta_0^T x)-\exp(\theta^T x)\right\}\cdot\exp(\kappa\theta^T x)\cdot x = 0, \tag{23}$$

and we denote the solution to these moment conditions $\tilde{\theta}_0$.

Barring explicit solutions, we can use Taylor expansion to estimate bias as follows: expanding $\exp(\theta^T X)-\exp(\theta_0^T X)$ around θ_0 as

$$\begin{aligned}\exp(\theta^T X) &\approx \exp(\theta_0^T X)+\exp(\theta_0^T X)(\theta-\theta_0)^T X \\ \exp(\kappa\theta^T X) &\approx \exp(\kappa\theta_0^T X)+\exp(\kappa\theta_0^T X)(\theta-\theta_0)^T\kappa X\end{aligned}$$

we have

$$\begin{aligned}&\mathbb{E}_x\left\{(1-P(\theta_0,x))\cdot\exp(\theta_0^T x)-\exp(\theta^T x)\right\}\cdot\exp(\kappa\theta^T x)\cdot x = \\ &\int_x\left((1-P(\theta_0,x))\cdot\exp(\theta_0^T x)-\exp(\theta^T x)\right)\cdot\exp(\kappa\theta^T x)\cdot xp(x)dx \approx \\ &\int_x\left((1-P(\theta_0,x))\cdot\exp(\theta_0^T x)-\exp(\theta_0^T x)-\exp(\theta_0^T x)\cdot(\theta-\theta_0)^T x\right)\cdot \\ &\left(\exp(\kappa\theta_0^T x)+\exp(\kappa\theta_0^T x)(\theta-\theta_0)^T\kappa x\right)\cdot xp(x)dx = \\ &\int\left((1-P(\theta_0,x))-1-(\theta-\theta_0)^T x\right)\cdot\left(1+(\theta-\theta_0)^T\kappa x\right)\cdot xp(x)\exp((\kappa+1)\theta_0^T x) \approx \\ &\int\left(-P(\theta_0,x)-(\theta-\theta_0)^T x+P(\theta_0,x)\cdot(\theta-\theta_0)^T\kappa x\right)\cdot xp(x)\exp((\kappa+1)\theta_0^T x)\end{aligned}$$

which implies

$$\int (P(\theta_0, x)\kappa - 1)(\theta - \theta_0)^T x \cdot xp(x) \exp((\kappa+1)\theta_0^T x) = \int P(\theta_0, x) \cdot xp(x) \exp((\kappa+1)\theta_0^T x)$$

or

$$(\theta - \theta_0)^T \cdot \int (P(\theta_0, x)\kappa - 1) x x^T p(x) \exp((\kappa+1)\theta_0^T x) = \int P(\theta_0, x) \cdot x^T p(x) \exp((\kappa+1)\theta_0^T x)$$

so that

$$\theta - \theta_0 = (A^T A)^{-1} A b$$

where

$$A = \int (P(\theta_0, x)\kappa - 1) x x^T p(x) \exp((\kappa+1)\theta_0^T x)$$
$$b = \int P(\theta_0, x) \cdot xp(x) \exp((\kappa+1)\theta_0^T x).$$

□

C.2 Proof of Theorem 2

Proof. Recall that the generalized PML estimators $\hat{\theta}$ (omitting dependence on κ) defined in (11) converge to the limit $\tilde{\theta}_0$, which solves

$$\mathbb{E}_x \left\{ (1 - P(\theta_0, x)) \cdot \exp(\theta_0^T x) - \exp(\theta^T x) \right\} \cdot \exp(\kappa \theta^T x) \cdot x = 0.$$

Under standard regularity conditions, the asymptotic variance of the Z-estimators that solve

$$\frac{1}{n} \sum_i \psi(y_i, x_i, \theta) = 0$$

is given by $J^{-1}IJ^{-1}$, where

$$I = \mathbb{E}_x \mathbb{E}\left[\psi(y, x, \tilde{\theta}_0)\psi(y, x, \tilde{\theta}_0)^T \mid x\right]$$
$$J = \mathbb{E}_x \mathbb{E}\left[-\frac{\partial \psi(y, x, \tilde{\theta}_0)}{\partial \theta} \mid x\right].$$

We have, using $V[\exp(\varepsilon_i) \mid X_i] = \exp^{\alpha}(\theta_0^T X_i) = \exp(\alpha\theta_0^T X_i)$,

$$\begin{aligned}
I &= \mathbb{E}_x \mathbb{E}\left[\left((y - \exp(\theta^T x)) \cdot \exp(\kappa\theta^T x) \cdot x\right)\left((y - \exp(\theta^T x)) \cdot \exp(\kappa\theta^T x) \cdot x\right)^T \mid x\right] \\
&= \mathbb{E}_x xx^T \exp(2\kappa\theta^T x) \cdot \mathbb{E}\left[\left(y - \exp(\theta^T x)\right)^2 \mid x\right] \\
&= \mathbb{E}_x xx^T \exp(2\kappa\theta^T x) \cdot \mathbb{E}\left[P(\theta_0, x)\left(\exp(\theta^T x)\right)^2 + (1 - P(\theta_0, x))\left(\exp(\theta_0^T x + \varepsilon) - \exp(\theta^T x)\right)^2 \mid x\right] \\
&= \mathbb{E}_x xx^T \exp(2\kappa\theta^T x) \cdot \mathbb{E}\left[P(\theta_0, x)\left(\exp(\theta^T x)\right)^2 + (1 - P(\theta_0, x))\left(\exp(\theta_0^T x) - \exp(\theta^T x)\right)^2 \exp(\alpha\theta_0^T x) \mid x\right],
\end{aligned} \tag{24}$$

and

$$\begin{aligned}
J &= \mathbb{E}_x \mathbb{E}\left[-(y - \exp(\theta^T x)) \cdot \exp(\kappa\theta^T x) \cdot \kappa xx^T + \exp(\kappa\theta^T x)\exp(\theta^T x)xx^T \mid x\right] \\
&= \mathbb{E}_x \mathbb{E}\left[-\left((1 - P(\theta_0, x)) \cdot \exp(\theta_0^T x) - \exp(\theta^T x)\right) \cdot \exp(\kappa\theta^T x) \cdot \kappa xx^T + \exp(\kappa\theta^T x)\exp(\theta^T x)xx^T \mid x\right].
\end{aligned}$$

Combining these calculations, we have

$$\sqrt{n}(\hat{\theta} - \tilde{\theta}_0) \to_d \mathcal{N}(0, J^{-1}IJ^{-1}),$$

where

$$\begin{aligned}
I &= \mathbb{E}_x xx^T \exp(2\kappa\theta^T x) \cdot \left[P(\theta_0, x)\left(\exp(\theta^T x)\right)^2 + (1 - P(\theta_0, x))\left(\exp(\theta_0^T x) - \exp(\theta^T x)\right)^2 \exp(\alpha\theta_0^T x)\right] \\
J &= \mathbb{E}_x \left[-\left((1 - P(\theta_0, x)) \cdot \exp(\theta_0^T x) - \exp(\theta^T x)\right) \cdot \exp(\kappa\theta^T x) \cdot \kappa xx^T + \exp(\kappa\theta^T x)\exp(\theta^T x)xx^T\right],
\end{aligned}$$

evaluated at $\theta = \tilde{\theta}_0$. □

The variance of various pseudo ML estimators without censoring has been characterized by Gourieroux et al. (1984b). Here we leverage their result to compute the variance of PPML and NLS.

Lemma 1. The asymptotic variance of PPML and NLS are given by

$$V_{NLS} = \mathbb{E}_x \left[\exp(2\theta_0^T x)\right]^{-1} \cdot \left(\mathbb{E}_x (x \exp(\theta_0^T x))(x \exp(\theta_0^T x))^T \cdot V\left[y \mid x\right]\right) \cdot \mathbb{E}_x \left[\exp(2\theta_0^T x)\right]^{-1}$$
$$V_{PPML} = \mathbb{E}_x \left[xx^T \exp(\theta_0^T x)\right]^{-1} \cdot \left(\mathbb{E}_x xx^T V\left[y \mid x\right]\right) \cdot \mathbb{E}_x \left[xx^T \exp(\theta_0^T x)\right]^{-1}$$

Proof. The general formula for the asymptotic variance of PML estimators is given by $J^{-1}IJ^{-1}$, where

$$J = \mathbb{E}_x \mathbb{E}\left[-\frac{\partial^2 \varphi(y, x, \theta_0)}{\partial \theta \partial \theta^T} \mid x\right]$$
$$I = \mathbb{E}_x \mathbb{E}\left[\frac{\partial \varphi(y, x, \theta_0)}{\partial \theta} \frac{\partial \varphi(y, x, \theta_0)}{\partial \theta}^T \mid x\right].$$

Recall that for NLS,

$$\varphi_{NLS}(y, x, \theta_0) = (\exp(\theta_0^T x) - y)^2,$$

so that

$$\begin{aligned}
J_{NLS} &= 2\mathbb{E}_x\mathbb{E}\left[-x^2\exp(2\theta_0^T x) - (\exp(\theta_0^T x) + x^2\exp(\theta_0^T x))(\exp(\theta_0^T x) - y) \mid x\right] \\
&= 2\mathbb{E}_x\mathbb{E}\left[-\exp(2\theta_0^T x) + \exp(\theta_0^T x)y - x^2 y\exp(\theta_0^T x) \mid x\right] \\
&= 2\mathbb{E}_x\left[-\exp(2\theta_0^T x)\right] \\
I_{NLS} &= 4\mathbb{E}_x\mathbb{E}\left[\left(x\exp(\theta_0^T x)(\exp(\theta_0^T x) - y)\right)\cdot\left(x\exp(\theta_0^T x)(\exp(\theta_0^T x) - y)\right)^T \mid x\right] \\
&= 4\mathbb{E}_x(x\exp(\theta_0^T x))(x\exp(\theta_0^T x))^T\mathbb{E}\left[\left((\exp(\theta_0^T x) - y)\right)^2 \mid x\right] \\
&= 4\mathbb{E}_x(x\exp(\theta_0^T x))(x\exp(\theta_0^T x))^T\mathbb{E}\left[\epsilon^2 \mid x\right] \\
&= 4\mathbb{E}_x(x\exp(\theta_0^T x))(x\exp(\theta_0^T x))^T\cdot V\left[y \mid x\right]
\end{aligned}$$

and the asymptotic variance is given by

$$\mathbb{E}_x\left[\exp(2\theta_0^T x)\right]^{-1}\cdot\left(\mathbb{E}_x(x\exp(\theta_0^T x))(x\exp(\theta_0^T x))^T\cdot V\left[y \mid x\right]\right)\cdot\mathbb{E}_x\left[\exp(2\theta_0^T x)\right]^{-1}.$$

Similarly, for PPML, we have

$$\varphi_{PPML}(y, x, \theta_0) = y\cdot\theta_0^T x - \exp(\theta_0^T x),$$

so that

$$\begin{aligned}
J_{PPML} &= \mathbb{E}_x\left[xx^T\exp(\theta_0^T x)\right] \\
I_{PPML} &= \mathbb{E}_x\mathbb{E}\left[(y - \exp(\theta_0^T x))^2\cdot xx^T \mid x\right] \\
&= \mathbb{E}_x xx^T\mathbb{E}\left[(\epsilon)^2\cdot \mid x\right] \\
&= \mathbb{E}_x xx^T V\left[y \mid x\right]
\end{aligned}$$

and the asymptotic variance is given by

$$J^{-1}IJ^{-1} = \mathbb{E}_x \left[xx^T \exp(\theta_0^T x) \right]^{-1} \cdot \left(\mathbb{E}_x xx^T V \left[y \mid x \right] \right) \cdot \mathbb{E}_x \left[xx^T \exp(\theta_0^T x) \right]^{-1}.$$

□