Discussion Papers in Economics

Misspecified Exponential Regressions: Estimation, Interpretation, and Average Marginal Effects

By

J.M.C. Santos Silva
(University of Surrey)

&

Rainer Winkelmann
(University of Zurich).

DP 01/24

School of Economics
University of Surrey
Guildford
Surrey GU2 7XH, UK
Telephone +44 (0)1483 689380
Facsimile +44 (0)1483 689548
Web https://www.surrey.ac.uk/school-economics
ISSN: 1749-5075

Misspecified exponential regressions: Estimation, interpretation, and average marginal effects*

J.M.C. Santos Silva† Rainer Winkelmann‡

4 February 2024

Abstract

Exponential regressions are frequently used when outcomes are non-negative. They are attractive because they are easy to interpret and to estimate, using pseudo maximum likelihood (PML). However, the validity of these methods depends on the correct specification of the conditional expectation, and little is known regarding their properties when the conditional expectation is misspecified. We show that PML estimators of misspecified exponential models provide optimal approximations to the conditional expectation, in a weighted mean squared error sense, and we give conditions under which their Poisson PML estimator identifies average marginal effects.

JEL classification codes: C13, C21, C25, C51.

Key words: Optimal approximations; Poisson regression; Pseudo maximum likelihood; Stein's Lemma.

*We are very grateful to John Mullahy, Dennis Novy, Vitaliy Oryshchenko, Paulo Parente, Silvana Tenreyro, and Tom Zylkin for their comments and suggestions on earlier versions of this work. The usual disclaimer applies.

INTRODUCTION

The pseudo maximum likelihood (PML) estimators introduced by Gourieroux, Monfort and Trognon (1984) owe their popularity to the fact that, even if other distributional assumptions are violated, they remain consistent as long as the models correctly specify the conditional expectation function. However, in practice, it is likely that models will suffer from some degree of misspecification, and relatively little is known about the properties of PML estimators when the conditional expectation is misspecified.

We consider PML estimators for working models that assume that the conditional mean is equal to the exponential of a linear combination of regressors. We call them linear index function exponential models, or LIFE models in short. If correctly specified, the slope parameters estimate the elasticities or semi-elasticities of the mean outcome with respect to the regressor. Under misspecification, this is no longer the case, and we are interested in the interpretation of these estimators when the assumed conditional expectation is incorrect.

We show that PML estimators for misspecified exponential regressions provide best LIFE predictors in a minimum weighted mean squared error (MSE) sense. Furthermore, asymptotically, they provide optimal LIFE approximations to the conditional expectation, also in a minimum weighted MSE sense. Therefore, PML estimators of exponential regression models are interpretable, even when the models misspecify the conditional expectation.[1] In contrast, as far as we are aware, estimators for other models with positive conditional expectation, such as the Tobit (Tobin, 1958) or zero-inflated models (Mullahy, 1986), do not have a clear interpretation when the models are misspecified. We also give conditions under which the Poisson PML (PPML) estimator for the exponential model identifies the average marginal effects (AMEs) of the regressors, or the average treatment effect in the case of a binary regressor, and present numerical illustrations of the main results.

[1] As is well known, ordinary least squares provides the best *linear* approximation, in a minimum MSE sense, to the conditional expectation (see, e.g., Goldberger, 1991, p. 52)

SET UP AND NOTATION

Suppose that a sample $\{(y_i, x_i)\}_{i=1}^{n}$ is available and that the specified working model is

$$y_i = \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}) + u_i = \exp(x_i'\beta) + u_i, \tag{1}$$

where y_i is the outcome of interest, $x_i = (1, x_{1i}, \ldots, x_{ki})'$ is a vector of regressors, $\beta = (\beta_0, \ldots, \beta_k)'$ is a conformable vector of parameters, and u_i is the error term. LIFE models such as (1) are commonly used to describe counts (e.g., Winkelmann, 2008) and other data for which the conditional mean is positive, being especially popular in health economics (e.g., Manning and Mullahy, 2001) and in the estimation of gravity equations for trade, migration, investment, and other flows (e.g., Santos Silva and Tenreyro, 2006); Santos Silva and Tenreyro (2022) briefly survey the use of LIFE models in economics and related areas.

The popularity of LIFE models is partially due to the fact that important economic models such as the Cobb-Douglas production function and the gravity equation are of this form, but it is also due to the availability of suitable PML estimators that are valid under very mild assumptions. Indeed, as shown by Gourieroux, Monfort and Trognon (1984), when

$$E[y_i|x_i] = \exp\left(x_i'\beta^0\right), \tag{2}$$

the PML estimators of β in (1) will identify β^0, the parameters of the conditional expectation, even if other aspects of the distribution are misspecified.

Letting $w(x_i, b)$ denote a strictly positive weight that depends on x_i and on a vector of parameters b, the PML estimators of β in (1) satisfy moment conditions of the form

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \exp\left(x_i'\hat{\beta}_w\right)\right)\frac{\exp\left(x_i'\hat{\beta}_w\right)}{w\left(x_i, \hat{\beta}_w\right)}x_i = 0, \tag{3}$$

where $\hat{\beta}_w$ is the PML estimator of β obtained with weights $w\left(x_i, \hat{\beta}_w\right)$, whose form depends on the estimator. Indeed, (3) coincides with first-order conditions of the nonlinear least squares (NLS) estimator when $w(x_i, b) = 1$, with those of the PPML estimator when $w(x_i, b) = \exp(x_i'b)$, and with those of the gamma PML estimator when $w(x_i, b) = \exp(2x_i'b)$.[2]

Although the robustness of the PML estimators makes them very attractive, in practice many factors conspire against the validity of (2), the form of the conditional mean assumed by the working model in (1). The literature provides many examples of models with positive conditional expectations that do not have the exponential form assumed in (1). Examples are the zero-inflated and the hurdle models introduced by Mullahy (1986), the Tobit model (Tobin, 1958), and the beta-binomial regression model for counts with an explicit upper bound (Heckman and Willis, 1977).[3] The misspecification of the linear index can result from measurement error (e.g., Kukush, Schneeweis, and Wolf, 2004), omitted variables (e.g., Gail et al., 1984, and Drake and McQuarrie, 1995), and coefficient heterogeneity (Breinlich, Novy, and Santos Silva, 2024), to name just a few possibilities. Next we consider the interpretation of PML estimators for (1) when (2) does not hold.

PML ESTIMATORS OF MISSPECIFIED EXPONENTIAL MODELS

Exponential models as approximations

The seminal papers of White (1981, 1982) consider the properties of estimators of misspecified non-linear models. Specifically, White (1981) establishes conditions under which the NLS estimator of a misspecified model asymptotically provides the best approximation, in a minimum MSE sense, to the conditional expectation function. This result, however, does not extend to other PML estimators of (1). In turn, White (1982) shows that maximum likelihood estimators of misspecified models identify the parameters that minimize the Kullback-Leibler (1951) divergence between the model and the true distribution. Again, this result is not useful when the purpose is simply to estimate the conditional mean, or an approximation to it, because minimizing the Kullback-Leibler

[2]Naturally, other consistent estimators can obtained by choosing different forms of $w(x_i, b)$.

[3]The Tobit model is designed to deal with censoring, but it is often used to model uncensored observations with a mass-point as zero. Indeed, that was the kind of data used by Tobin (1958) to illustrate the application of his method.

divergence takes into account other aspects of the conditional distribution, such as the variance.

To understand the properties of PML estimators of (1) when the conditional mean is misspecified, we write $\hat{\beta}_w$ as the following self-consistent weighted least squares estimator

$$\hat{\beta}_w = \arg\min_b \frac{1}{n}\sum_{i=1}^{n}\frac{\left(y_i - \exp\left(x_i'b\right)\right)^2}{w\left(x_i, \hat{\beta}_w\right)}, \tag{4}$$

whose first-order conditions are given by (3).[4] Self consistent estimators are functions of the data and of the estimators themselves (see, e.g., Efron, 1967, Powell, 1986, Flury and Tarpey, 2006, and Peng, 2012), and can be implemented using a self-consistent algorithm that starts with a suitable set of estimates and iteratively updates them until convergence; that is, until the estimates "confirm themselves" (see Flury and Tarpey, 2006).

Equation (4) shows that, even when (2) does not hold, the PML estimators of β in (1) have a clear interpretation: they minimize the weighted sum of the squared residuals $\left(y_i - \exp\left(x_i'b\right)\right)$ and, in this sense, they deliver the best predictor of y_i of the form $\exp(x_i'b)$. That is, they are the best LIFE predictors in a minimum weighted MSE sense.

A more interesting insight is obtained by considering the probability limit or pseudo-true value of $\hat{\beta}_w$, which we denote by β_w^*. Following Cox (1961), β_w^* is such that

$$E_x\left[\left(\mu(x_i) - \exp\left(x_i'\beta_w^*\right)\right)\frac{\exp\left(x_i'\beta_w^*\right)}{w\left(x_i, \beta_w^*\right)}x_i\right] = 0, \tag{5}$$

where we have used the law of iterated expectations to replace y_i with $E[y_i|x_i] = \mu(x_i)$. Assuming that it is possible to interchange integration and differentiation, β_w^* can also be defined as

$$\beta_w^* = \arg\min_b E_x\left[\frac{\left(\mu(x_i) - \exp\left(x_i'b\right)\right)^2}{w\left(x_i, \beta_w^*\right)}\right]. \tag{6}$$

The definition of β_w^* in (6) shows that, asymptotically, the PML estimators of (1) minimize the weighted mean square of the difference $\left(\mu(x_i) - \exp\left(x_i'b\right)\right)$ and, in this sense, deliver the parameters of the best LIFE approximation to the conditional expectation.

[4] Although (4) resembles the objective function of the quasi-generalized NLS estimator discussed by Gourieroux, Monfort and Trognon (1984, pp. 688/9), $\hat{\beta}_w$ is a standard PML estimator because $w\left(x_i, \hat{\beta}_w\right)$ does not have to be proportional to the conditional variance of y_i.

When $w(x_i, \beta_w^*) = 1$, this result matches the seminal finding of White (1981), but the results for other weighting functions appear to be new.[5]

Interpretation of misspecified models

Since all PML estimators are consistent for β^0 when (2) holds, the choice between different PML estimators is often based on attempts to maximize the efficiency of the estimator by choosing weights that better approximate the conditional variance of y (see, e.g., Manning and Mullahy 2001, and Santos Silva and Tenreyro, 2006). However, in the likely scenario where (2) does not hold, different PML estimators will identify different sets of parameters, and therefore the precision of the estimators becomes a second-order question.[6] That is, when different PML estimators lead to significantly different results, it is reasonable to choose the estimator based on the asymptotic properties of the approximation to the conditional expectation it delivers, rather than on the specification of the conditional variance. For that, it is important to understand what different estimators do.

The objective function of the gamma PML estimator depends on the size of the residual $(y_i - \exp(x_i'\beta_w^*))$ relative to the fitted value $\exp(x_i'\beta_w^*)$, but not on the size of the residual itself. This implies that when the fitted values are close to zero, even small residuals will have a sizable contribution to the objective function, with the reverse happening when the fitted values are large. Hence, the gamma PML estimator leads to an approximation to $\mu(x_i)$ that emphasizes the fit in areas where the fitted values are close to zero, as is made clear by its first-order conditions.

In contrast, the NLS estimator only takes into account the size of the residual, completely ignoring its relative size. Therefore, as made clear by its first-order conditions, the NLS estimator leads to an approximation to $\mu(x_i)$ that emphasizes the fit in areas with

[5] We note that, although we focus on estimators of LIFE models, related results can be obtained for all other generalized linear models (Nelder and Wedderburn, 1972), including the logit and probit.

[6] Our result suggests that (2) can be tested by estimating β simultaneously using moment conditions of different PML estimators and performing a J-test (Hansen, 1982). Such test is closely related to the reweighting test suggested by White (1981).

large fitted values, which are typically associated with larger errors. In between these two extremes, PPML has an objective function that takes into account both the absolute and relative sizes of the residuals.[7] Therefore, this estimator does not put particular weight on any observations, as is confirmed by its first-order conditions in which all observations have the same weight.

Finally, we note that, in general, β_w^* depends on two sets of weights. Indeed, besides the weight $w(x_i, \beta_w^*)$ that defines the estimator, (6) shows that, as in White (1981), β_w^* generally also depends on $f(x)$, the probability density of the regressors, which is implicitly used as a weighting function when taking expectations over x.[8] The exception to this is the case where (2) holds, because then we have that $\beta_w^* = \beta^0$, whatever the distribution of x.[9] Consequently, under misspecification, all estimators will provide a better approximation to $\mu(x_i)$ in regions where the regressors have higher density. In contrast, the approximation may be poor in regions where the regressors have low density. We also note that, because β_w^* depends on $f(x)$ when (2) does not hold, the interpretation of $\exp(x_i'\beta_w^*)$ as an approximation will change when data are obtained by some form of exogenously stratified sampling. Indeed, in this case, $\exp(x_i'\beta_w^*)$ is the best LIFE approximation to the conditional expectation in the artificial population induced by the sampling scheme, but it may be possible to estimate the parameters of the best LIFE approximation to $\mu(x_i)$ in the population if suitable sampling weights are available (see, e.g., Wooldridge, 1999).

These results are illustrated in Figure 1 for the case where there is a single regressor. Figure 1 displays a conditional expectation (solid black line) of the form $E[y_i|x_i] = 0.01 \times x_{1i}^{1.5}$, and the corresponding approximations obtained by estimating a LIFE model using gamma PML (dashed line), PPML (long-dashed line) and NLS (dotted line), as well as the density of the regressor (solid grey line at the bottom). In the top panel

[7]Notice that, in the Poisson case, the contributions to the objective function in (4) can be seen as the product between the absolute and relative residuals.

[8]For simplicity, we do not distinguish between continuous and discrete variables.

[9]The fact that β_w^* depends on the distribution of x introduces additional sampling variability that is unrelated to the one induced by the randomness of y given x. As noted by the Buja et al. (2019), this implies that valid inference about β_w^* has to be based on a misspecification-robust covariance matrix.

$x_1 \sim \mathcal{N}(200, 40^2)$ while in the bottom one $x_1 \sim U(75, 325)$. As expected, when x_1 is normal, all approximations are good in the region where the regressor is more dense, with the NLS approximation being better for large values of the regressor and the gamma PML approximation being better for lower values; the Poisson approximation is between the other two for extreme values of the regressor. When x_1 has a uniform distribution, the approximations are better in the tails, but less good in the centre. The two panels in this figure clearly illustrate the joint roles $w(x_i, \beta_w^*)$ and $f(x)$ play in obtaining an approximation to $\mu(x_i)$.

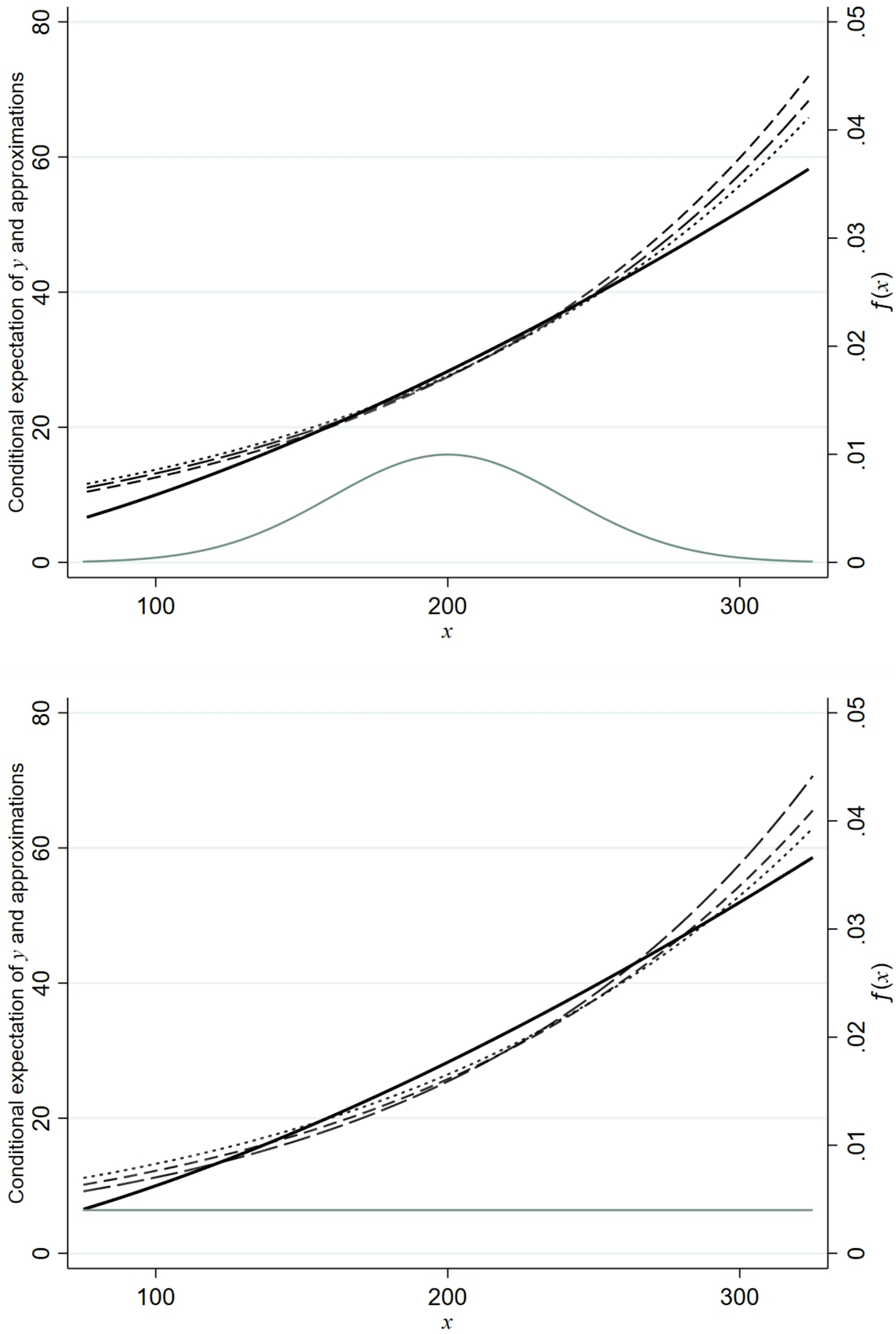
IDENTIFICATION OF AVERAGE MARGINAL EFFECTS

While it is reassuring that estimates of misspecified LIFE models provide best LIFE approximations to the conditional expectation, in a minimum weighted MSE sense, researchers are often interested in parameters and marginal effects. The estimated parameters depend on the distribution of regressors, and specific results are available only for restricted classes of distributions.

An earlier literature has considered the case where the true conditional expectation is of linear index form, the link function is misspecified, and regressors have a multivariate normal or an elliptical distribution. Ruud (1983) shows, for such regressors, that using a logit model rather than a probit model estimates the slope parameters up to scale. Li and Duan (1989) extend this result to any generalized linear model with a misspecified link, i.e., including the ordinary least squares estimator of the linear model (OLS), as well as the LIFE models considered here. With estimation up to scale, hypothesis tests for $H_0 : \beta = 0$ remain valid and, for models with $k > 1$, ratios of coefficients are identified so that statements about substitutability between regressors can be made.

The PPML estimator of a LIFE model with normally distributed regressors has an additional useful property that seems to have gone unnoticed until now: it identifies the average marginal effect $E[\partial\mu(x_i)/\partial x_i]$, a population parameter that researchers often care about. Importantly, PPML is the only estimator of a LIFE model with this property and it does not require the conditional expectation function $\mu(x_i)$ to be of single-index form.

Figure 1: Approximations to the conditional expectation



Note: The figure displays conditional expectations (solid black lines) and the corresponding optimal exponential approximations obtained with gamma PML (dashed lines), PPML (long-dashed lines), and NLS (dotted lines), as well as the densities of the regressor (solid grey lines at the bottom), which are normal in the top panel and uniform in the bottom one.

To establish the general result, recall (5) showing that the PPML estimand is the solution to the population first-order condition

$$E_x[(\mu(x_i) - \exp(x_i'^{*}))\, x_i] = 0, \tag{7}$$

where we use β^* to denote the probability limit of the PPML estimator.

We start by considering the model with constant and a single regressor. It follows from (7) that the first-order condition for the slope can be re-written as

$$Cov(x_{1i}, \mu(x_{1i})) = Cov(x_{1i}, \exp(\beta_0^* + \beta_1^* x_{1i})), \tag{8}$$

where β_0^* and β_1^* are the elements of β^*.

Stein's Lemma (Stein, 1981) states that if x is normally distributed and the relevant moments exist, then for any differentiable function $h(x)$, $Cov(x, h(x)) = Var(x)E(\partial h(x)/\partial x)$. Assuming that its conditions are satisfied, we can apply the Lemma on both sides of (8) to obtain the equality $Var(x_{1i})E\left[\partial \mu(x_{1i})/\partial x_{1i}\right] = Var(x_{1i})E[(\exp(\beta_0^* + \beta_1^* x_{1i})\beta_1^*]$, and therefore

$$E\left[\partial \mu(x_{1i})/\partial x_{1i}\right] = E[\exp(\beta_0^* + \beta_1^* x_{1i})]\beta_1^*. \tag{9}$$

Thus, as long as PPML is used for estimation and the regressor is normally distributed, the estimated AME of x_1 obtained with a LIFE model (the right-hand side of equation 9) is equal to the AME of the conditional expectation $\mu(x_{1i})$ even if the model is misspecified. A similar result exists for the OLS estimator of a linear working model (e.g., Wooldridge 2010, p. 579, who draws on Stoker, 1986), which, under normality of x, also identifies the AME for arbitrary $\mu(x)$.[10]

Often, researchers are interested in relative marginal effects, or semi-elasticities. When the exponential model is misspecified, the relative marginal effects $\left(\partial \mu(x_{1i})/\partial x_{1i}\right)/\mu(x_{1i})$ are not constant. However, under normality of x_1, the PPML estimand β_1^* identifies the AME relative to the mean of y, since from the first-order condition it follows that

$$E(y_i) = E(\mu(x_i)) = E[\exp(\beta_0^* + \beta_1^* x_{1i})],$$

[10] The result applies to all estimators that solve an unweighted zero-correlation moment condition between the working model's residuals and the regressors. Other examples include the logit maximum likelihood estimator, and a probit estimated by solving the sample analog of $E[(y - \Phi(x'\beta))x] = 0$, where $\Phi(\cdot)$ denotes the normal CDF.

and hence $\beta_1^* = E[\partial\mu(x_i)/\partial x_{1i}]/E[\mu(x_i)]$.

This result generalizes to multiple regressors. Let x be multivariate normal with non-singular covariance matrix Σ_x, and let $h(x)$ be a differentiable function of x with partial derivatives $\nabla h(x) = \left(\frac{\partial h(x)}{\partial x_1}, \ldots, \frac{\partial h(x)}{\partial x_k}\right)'$. Then, if the relevant moments exist, $Cov(x, h(x)) = \Sigma_x E(\nabla h(x))$. The PPML first-order conditions for the k-variate regression model imply that $Cov(x_i, \mu(x_i)) = Cov(x_i, \exp(x_i'^*))$. Substituting the terms from Stein's equality, writing $\nabla\mu(x) = \left(\frac{\partial\mu(x)}{\partial x_1}, \ldots, \frac{\partial\mu(x)}{\partial x_k}\right)'$, and pre-multiplying both sides by Σ_x^{-1}, we obtain

$$E(\nabla\mu(x_i)) = E[\exp(x_i'^*)]\beta^*,$$

which again implies that, for normally distributed regressors, the PPML estimator of the AMEs of the regressors in a LIFE model is equal to the AMEs of the conditional expectation even under misspecification.

The main limitation of these results remains the assumption of a normally distributed design matrix, which is not realistic in many applications and is a necessary condition for the results to hold: under misspecification, PPML identifies the AMEs of the regressors *iff* the regressors are normally distributed (see Ross, 2011, p. 215).[11] The numerical experiments in Section 5 document how the bias of PPML estimates of the AMEs increases with departures from normality. In these experiments, the relative bias remains moderate, and for departures that preserve symmetry, PPML generally performs better than gamma PML or NLS.

In cases where x is not normally distributed, one could be tempted to use a weighted version of (7) such that the marginal distribution of x becomes approximately normal. Indeed, Ruud (1986) suggests such a procedure and shows that it can identify the parameters of a single-index model up to scale. In the present context, where the objective is to estimate AMEs, the value of this approach is limited by the fact that it would identify the AME for the artificial population induced by the weighting scheme, which may not be interesting.

Another set of robustness results exists for binary regressors. Consider a model where the single regressor, x_{1i}, is binary. It follows from the PPML first-order conditions that

[11] We are grateful to a referee for bringing this result to our attention.

$E(y_i) = E[\exp(\beta_0^* + \beta_1^* x_{1i})]$ and $E(y_i x_{1i}) = E\left[\exp(\beta_0^* + \beta_1^* x_{1i}) x_{1i}\right]$. By applying iterated expectations to both sides of these equalities, and letting $p = \Pr(x_{1i} = 1)$, we can write

$$pE(y_i|x_{1i} = 1) + (1-p)E(y_i|x_{1i} = 0) = p\exp(\beta_0^* + \beta_1^*) + (1-p)\exp(\beta_0^*),$$

$$pE(y_i|x_{1i} = 1) = p\exp(\beta_0^* + \beta_1^*),$$

from where it follows that $E(y_i|x_{1i} = 1) = \exp(\beta_0^* + \beta_1^*)$ and $E(y_i|x_{1i} = 0) = \exp(\beta_0^*)$. Therefore, the average treatment effect is given by $\exp(\beta_0^* + \beta_1^*) - \exp(\beta_0^*)$, and the relative treatment effect by $\exp(\beta_1^*) - 1$.

Since the model with a single binary regressor is saturated, there is no approximation error and the model is trivially "correct". However, Negi and Wooldridge (2022) show that the argument extends to an exponential regression with additional, potentially continuous, covariates, as long as those are distributed independently of x_{1i}. The leading application is that of a randomized treatment x_{1i} where a covariate adjustment is considered for efficiency reasons and the exponential regression is used because the outcome is non-negative. In this case, for a model with $k = 2$, we have that

$$E[\exp(\beta_0^* + \beta_1^* + \beta_2^* x_{2i})] - E[\exp(\beta_0^* + \beta_2^* x_{2i})]$$

identifies the average treatment effect even if the link is misspecified, as long as PPML is used for the estimation of a LIFE model.

NUMERICAL ILLUSTRATION

In this section we present numerical evidence to illustrate how misspecification affects the ability of a LIFE model to identify AMEs. Specifically, we consider a population where

$$E\left[y_i|x_{1i}\right] = \mu\left(x_{1i}\right) = \frac{10\exp\left(x_{1i}\right)}{\left(1+\exp\left(x_{1i}\right)\right)^{\theta}}, \qquad (10)$$

and compare the AME in the population with their estimates obtained with different PML estimators of a LIFE model like (1), with $k = 1$. For the sake of comparison, we also include OLS estimates for the linear model. It is unclear ex-ante, whether the linear or the exponential approximation to (10) provides better estimates of the AME. Only for normally distributed regressors (see Case 1 below) do we know that both PPML and OLS are unbiased.

The degree of misspecification is indexed by $\theta \in \{0.25, 0.50, 0.75\}$. For these values of θ, $\mu(x_{1i})$ has no upper bound but increases more slowly for larger θ; that is, the LIFE model will be a better approximation for small θ, whereas the fit of the linear model improves as θ increases.[12]

We consider several distributions of x_{1i}, which is always standardized to have 0 mean and variance $\sigma^2 \in \{0.25, 1\}$. In Case 1, x_{1i} is a mixture of a normal and a uniform distribution, being drawn from the normal with probability π, and from the uniform with probability $1 - \pi$. Therefore, x_{1i} has a symmetrical distribution that is platykurtic for $\pi < 1$, and both the PPML and OLS estimators identify the population AMEs for $\pi = 1$. In Case 2, x_{1i} follows a $\chi^2_{(\kappa)}$ distribution with $\kappa \in \{6, 12, 24\}$. In this case the distribution is leptokurtic but it approaches the normal as $\kappa \to \infty$. Moreover, because the distribution is asymmetric with zero mean, the bulk of values of x_{1i} is negative, but the left tail is short. Case 3 is the mirror image of Case 2, with $-x_{1i} \sim \chi^2_{(\kappa)}$, where again $\kappa \in \{6, 12, 24\}$. Finally, in Case 4, we have that $x_{1i} = (-1)^{s_i} x_i^*$, where $x_i^* \sim \chi^2_{(\kappa)}$ and s_i, is a Bernoulli variable with $\Pr(s_i = 1) = 0.5$. As before, we consider cases with $\kappa \in \{6, 12, 24\}$. For a given κ, x_{1i} has the same degree of kurtosis as the variables considered in Cases 2 and 3, but now it has a symmetric distribution.

The range of cases we consider allows us to see how the results in Section 3 and Section 4 interact. From Section 4, we know that PPML and OLS will identify the AMEs in Case 1, with $\pi = 1$, and we will see how different departures from normality affect their ability to identify the AMEs. Other PML estimators of the LIFE model generally do not identify the AMEs under misspecification, but our results provide information on how sensitive they are to misspecification under different scenarios. From Section 3, we know that different PML estimators of misspecified LIFE models will provide approximations to $\mu(x_{1i})$ whose quality varies over the range of x. Therefore, we expect that the gamma PML estimator will perform well in Case 2 because in this case the bulk of observations will have low values of $\mu(x_{1i})$; conversely, we expect NLS to have an advantage in Case

[12] Note that for $\theta = 0$ the LIFE model is correctly specified.

3. Because PPML and OLS give the same weight to all observations, we expect them to perform better in Cases 1 and 4, where x_{1i} has a symmetrical distribution.

For each of the four cases described above, we generate $\mu(x_{1i})$ for $1,000$ draws of the relevant distribution of x_{1i}, evaluate $\partial\mu(x_{1i})/\partial x_{1i}$, obtain the pseudo-true values of the parameters by solving the relevant moment conditions, and compute the estimated AMEs. We then repeat the process $10,000$ times and average the results. That is, we evaluate $E[\partial\mu(x_{1i})/\partial x_{1i}]$ and its approximations using Monte Carlo integration over the distribution of x_{1i}. Note that we never actually generate y_i, and therefore our results abstract from sampling noise and are only informative about the results that should be expected in very large samples, but are valid for any distribution of y_i.

Tables 1 and 2 report the results for each of the four cases. Each table displays, for the different combinations of parameters we consider, the population AME (denoted AME) and the differences between this and the estimated AMEs obtained with PPML (labelled PPML), gamma PML (labelled GPML), and non-linear least squares (labelled NLS) estimators of the LIFE model, as well as with the OLS estimator of the linear model (labelled OLS).

The results in Table 1 show that, as expected, PPML and OLS identify the AMEs when x_{1i} is normally distributed ($\pi = 1$), while the same is not true for the other estimators.[13] Naturally, the performance of the PPML and OLS estimators deteriorates as π goes to 0. Interestingly, the performance of the NLS estimator improves as the distribution of x_{1i} approaches the uniform, dominating the other estimators of the LIFE model for $\pi = 0$. In turn, for all values of π and θ, the gamma PML estimator is by far the worst performer among the estimators of the LIFE model. Finally, for $\pi < 1$, the OLS results very much depend on the value of θ, being very poor when $\mu(x_{1i})$ grows quickly, but being excellent when $\mu(x_{1i})$ grows more slowly.

Turning to Table 2, the results for Case 2 show that the gamma PML estimator has an excellent performance when $\sigma^2 = 0.25$, but is often outperformed by PPML when $\sigma^2 = 1$. In contrast, NLS has a poor performance with $\sigma^2 = 0.25$, and is dominated by PPML,

[13] The situation is very similar to the top panel of Figure 1: We know that PPML gives the AME. The LIFE model fitted by NLS is flatter, and hence there is a downward bias in the AMEs, whereas the GPML curve is steeper than PPML, leading to an upward bias.

but not by the gamma PML estimator, when $\sigma^2 = 1$. In Case 3, it is NLS that has the best performance, while the gamma PML estimator has the worst performance among the estimators of the LIFE model. PPML is in between these two estimators, and its performance naturally improves with κ. For Case 4, PPML leads to the best results of all estimators of the LIFE model, with the gamma PML estimator dominating NLS for $\sigma^2 = 0.25$, but being almost dominated by it for $\sigma^2 = 1$. As in Case 1, in Cases 2 to 4, the OLS performance varies from excellent to very poor with the value of θ.

Overall, with the chosen distributions and parameters, it is noteworthy that the relative biases of the PML estimators of the LIFE model tend to be rather small. For example, in almost all cases it is below 10 percent of the true AME when $\sigma^2 = 0.25$. Moreover the bias of these estimators varies little with the degree of misspecification, changing little as θ goes from 0.25 to 0.75. By contrast, the OLS bias is very sensitive to the value of θ, and can be very large, both in absolute and relative terms, especially when $\theta = 0.25$ and the distribution is leptokurtic. Hence, the use of the linear model to estimate AMEs cannot be recommended unless the regressors are close to being normally distributed, or $\mu(x)$ has little curvature.

Table 1: Average marginal effects and biases for Case 1

		$\sigma^2 = 0.25$					$\sigma^2 = 1.0$				
π	θ	AME	PPML	GPML	NLS	OLS	AME	PPML	GPML	NLS	OLS
1.0	0.25	7.91	0.00	0.13	−0.10	−0.00	9.82	0.01	0.69	−0.23	−0.01
	0.50	5.43	0.00	0.17	−0.14	−0.00	5.89	0.01	0.77	−0.34	−0.00
	0.75	3.65	0.00	0.17	−0.14	0.00	3.52	0.00	0.66	−0.34	0.00
0.5	0.25	7.91	0.03	0.12	−0.05	−0.11	9.74	0.10	0.62	−0.17	−0.48
	0.50	5.43	0.05	0.17	−0.07	−0.03	5.87	0.14	0.73	−0.21	−0.12
	0.75	3.65	0.05	0.17	−0.06	0.01	3.50	0.15	0.65	−0.17	0.04
0.0	0.25	7.90	0.07	0.12	0.02	−0.22	9.67	0.24	0.55	0.07	−0.96
	0.50	5.43	0.09	0.17	0.03	−0.05	5.85	0.32	0.69	0.10	−0.23
	0.75	3.64	0.10	0.17	0.03	0.03	3.49	0.31	0.63	0.10	0.08

Note: AME is the true average marginal effect. The remaining columns show the difference between AME and its estimates obtained by different methods. The regressor has a normal distribution with probability π, and a uniform distribution with probability $1 - \pi$, and is standardized to have zero mean and variance σ^2.

Table 2: Average marginal effects and biases for Cases 2 to 4

		$\sigma^2 = 0.25$					$\sigma^2 = 1.0$				
κ	θ	AME	PPML	GPML	NLS	OLS	AME	PPML	GPML	NLS	OLS
Case 2											
6	0.25	8.00	−0.17	0.01	−0.30	2.30	11.06	−0.27	0.99	−0.33	8.45
	0.50	5.44	−0.27	−0.04	−0.46	1.00	6.02	−0.43	0.61	−0.70	2.53
	0.75	3.64	−0.32	−0.11	−0.50	0.30	3.50	−0.49	0.22	−0.83	0.57
12	0.25	7.97	−0.12	0.04	−0.24	1.50	10.56	−0.20	0.81	−0.33	4.71
	0.50	5.44	−0.19	0.02	−0.36	0.68	5.97	−0.31	0.63	−0.62	1.61
	0.75	3.64	−0.22	−0.03	−0.39	0.22	3.50	−0.35	0.34	−0.70	0.41
24	0.25	7.95	−0.08	0.07	−0.20	1.01	10.29	−0.14	0.74	−0.32	2.90
	0.50	5.44	−0.13	0.06	−0.29	0.47	5.94	−0.22	0.67	−0.55	1.07
	0.75	3.64	−0.15	0.03	−0.31	0.16	3.51	−0.24	0.43	−0.60	0.30
Case 3											
6	0.25	7.84	0.12	0.25	0.05	−1.48	9.24	0.23	0.74	0.03	−2.91
	0.50	5.43	0.20	0.38	0.09	−0.80	5.82	0.35	1.02	0.07	−1.48
	0.75	3.66	0.26	0.44	0.13	−0.35	3.57	0.42	1.06	0.13	−0.62
12	0.25	7.86	0.09	0.22	0.01	−1.11	9.37	0.18	0.71	−0.04	−2.89
	0.50	5.43	0.15	0.32	0.04	−0.59	5.84	0.27	0.95	−0.03	−1.11
	0.75	3.66	0.19	0.37	0.06	−0.25	3.55	0.31	0.95	0.01	−0.44
24	0.25	7.87	0.07	0.19	−0.01	−0.81	9.48	0.13	0.70	−0.09	−1.75
	0.50	5.43	0.11	0.28	−0.01	−0.42	5.85	0.20	0.90	−0.11	−0.82
	0.75	3.65	0.14	0.31	0.01	−0.17	3.54	0.23	0.87	−0.08	−0.31
Case 4											
6	0.25	7.92	−0.10	0.14	−0.27	0.41	10.15	−0.26	1.09	−0.41	2.73
	0.50	5.44	−0.13	0.18	−0.37	0.10	5.92	−0.34	0.97	−0.71	0.52
	0.75	3.65	−0.12	0.17	−0.36	−0.03	3.54	−0.29	0.71	−0.74	−0.02
12	0.25	7.91	−0.05	0.13	−0.20	0.19	9.96	−0.16	0.86	−0.37	1.19
	0.50	5.43	−0.07	0.18	−0.27	0.05	5.90	−0.19	0.86	−0.59	0.24
	0.75	3.65	−0.06	0.17	−0.26	−0.02	3.53	−0.16	0.68	−0.59	−0.02
24	0.25	7.91	−0.03	0.13	−0.16	0.10	9.89	−0.09	0.77	−0.33	0.58
	0.50	5.44	−0.03	0.17	−0.21	0.02	5.89	−0.10	0.82	−0.50	0.12
	0.75	3.65	−0.03	0.17	−0.21	−0.01	3.52	−0.08	0.67	−0.48	−0.01

Note: AME is the true average marginal effect. The remaining columns show the difference between AME and its estimates obtained by different methods. In case 2 the regressor has a χ^2 distribution with κ degrees of freedom. In Case 3 the regressor has a χ^2 distribution with κ degrees of freedom multiplied by -1. In case 4 the regressor has a χ^2 distribution with κ degrees of freedom that is multiplied by -1 with probability 0.5. In all cases, the regressor is standardized to have zero mean and variance σ^2.

Our findings confirm that the quality of the results obtained with the different estimators of a LIFE model is heavily dependent on the distribution of the regressors. In our setting, there are cases in which the gamma PML and the NLS estimators perform very well, but both estimators can also perform poorly. In contrast, the PPML estimator was never the worst performer and it dominates the other estimators when x_{1i} has a symmetric distribution that is not platykurtic. Finally, as could be expected, we note that the results are generally better for $\sigma^2 = 0.25$ than for $\sigma^2 = 1$.

It is naturally difficult to generalize these results to more complex settings but, together with our earlier findings, it is reasonable to conclude that PPML is always a relatively safe choice, in the sense that it may not be the best, but it is also unlikely to be the worst performer. Of course, having additional information on the nature of the problem, researchers may be justified to choose a different approach.

We conclude by reiterating that our results are only informative about the asymptotic performance of the estimators. In finite samples, it is also important to take into consideration how robust the estimators are to departures from assumptions about the heteroskedasticity pattern. The results in Manning and Mullahy (2001) and Santos Silva and Tenreyro (2006) suggest that the NLS can be particularly inefficient and have a significant finite-sample bias, and therefore its use is difficult to recommend.

CONCLUDING REMARKS

We show that, asymptotically, pseudo maximum likelihood estimators of misspecified exponential models provide optimal approximations to the conditional expectation, in a weighted mean squared error sense, and we give conditions under which their Poisson pseudo maximum likelihood estimator identifies the average marginal effects of the regressors. These results provide an additional motivation for the use of exponential models, and suggest that researchers should carefully consider the interpretation of the estimates under misspecification when choosing between estimators for these models. Moreover, our results establish a striking similarity between the properties of a misspecified linear model estimated by OLS and those of a misspecified exponential model estimated by Poisson pseudo maximum likelihood.

Which of the estimators of an exponential model is preferable will, of course, depend on the purpose of the model and on the nature of the problem, but our results show that the Poisson pseudo maximum likelihood estimator is generally a reasonable choice under misspecification, and has properties that are not shared by the other estimators. It therefore has a range of characteristics that make it a suitable workhorse for the estimation of exponential regression models.

REFERENCES

Breinlich, H., Novy, D., and Santos Silva, J.M.C. (2024). "Trade, Gravity and Aggregation," *The Review of Economics and Statistics*, forthcoming.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. (2019). "Models as Approximations I: Consequences Illustrated with Linear Regression," *Statistical Science*, 34, 523-544.

Cox, D.R. (1961). "Tests of Separate Families of Hypotheses," *in* Neyman, J. (ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4.1, 105-123, Berkeley (CA): University of California Press.

Drake, C. and McQuarrie, A. (1995). "A Note on the Bias due to Omitted Confounders," *Biometrika,* 82, 633-638.

Efron, B. (1967). "The Two-Sample Problem with Censored Data," *in* Le Cam, L.M. and Neyman, J. (eds.) *Proceedings of the Fifth Berkeley Symposium*, vol 5.4, 831-853, Berkeley (CA): University of California Press.

Flury, B.D. and Tarpey, T. (2006). "Self-Consistency–II," *in* Kotz, S., Read, C.B., Balakrishnan, N., Vidakovic, B., and Johnson, N.L. (eds.) *Encyclopedia of Statistical Sciences: Re-Se.*, vol 11, 7542-7545. New York (NY): Wiley-Interscience.

Gail, M.H., Wieand, S., and Piantadosi, S. (1984). "Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates," *Biometrika*, 71, 431-444.

Goldberger, A. (1991). *A Course in Econometrics*, Cambridge (MA): Harvard University Press.

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

Hansen, L.P. (1982). "Large Sample Properties of Generalized Methods of Moments Estimators," *Econometrica*, 50, 1029-1054.

Heckman, J.J. and Willis, R.J. (1977). “A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women,” *Journal of Political Economy*, 85, 27-58.

Kukush, A., Schneeweis, H., and Wolf, R. (2004). “Three Estimators for the Poisson Regression Model with Measurement Errors,” *Statistical Papers*, 45, 351-368.

Kullback, S. and Leibler, R.A. (1951). “On Information and Sufficiency,” *Annals of Mathematical Statistics*, 22, 79-86.

Li, K.-C. and Duan, N. (1989). “Regression Analysis Under Link Violation,” *The Annals of Statistics*, 17, 1009-1052.

Manning, W.G. and Mullahy, J. (2001). “Estimating Log Models: To Transform or Not to Transform?,” *Journal of Health Economics*, 20, 461-494.

Mullahy, J. (1986). “Specification and Testing in some Modified Count Data Models,” *Journal of Econometrics*, 33, 341-365.

Negi, A. and Wooldridge, J.M. (2021). “Revisiting Regression Adjustment in Experiments with Heterogeneous Treatment Effects, ” *Econometric Reviews*, 40, 504-534.

Nelder, J.A. and Wedderburn, R.W.M. (1972). “Generalized Linear Models,” *Journal of the Royal Statistical Society A*, 135, 370-384.

Peng, L. (2012). “Self-Consistent Estimation of Censored Quantile Regression,” *Journal of Multivariate Analysis*, 105, 368-379.

Powell, J.L. (1986). “Symmetrically Trimmed Least Squares Estimation for Tobit Models,” *Econometrica*, 54, 1235-1460.

Ross, N. (2011). “Fundamentals of Stein’s Method,” *Probability Surveys*, 8, 210-293.

Ruud, P. (1983). “Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecifications of Distribution in Multinomial Discrete Choice Models,” *Econometrica*, 51, 225-228.

Ruud, P. (1986). " Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution," *Journal of Econometrics*, 32, 157-187.

Santos Silva, J.M.C. and Tenreyro, S. (2006). "The Log of Gravity," *The Review of Economics and Statistics*, 88, 641-658.

Santos Silva, J.M.C. and Tenreyro, S. (2022). "The Log of Gravity at 15," *Portuguese Economic Journal*, 21, 423-437.

Stein, C.S. (1981). "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135-1151.

Stoker, T.M. (1986). "Consistent Estimation of Scaled Coefficients, "*Econometrica*, 54, 1461-1481.

Tobin, J. (1958). "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24-36.

White, H. (1981). "Consequences and Detection of Misspecified Nonlinear Regression Models," *Journal of the American Statistical Association*, 76, 419-433.

White, H. (1982). "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1-25.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*, 5th ed., Berlin: Springer-Verlag.

Wooldridge, J.M. (1999). "Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples," *Econometrica*, 67, 1385-1406.

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed., Cambridge (MA): MIT Press.