ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Winkelmann, Rainer

Conference Paper

Underreporting in count data models: some new results and an application to labor mobility

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2026: Behavioral Economics

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Winkelmann, Rainer (2026) : Underreporting in count data models: some new results and an application to labor mobility, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2026: Behavioral Economics, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/343743

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Underreporting in count data models: some new results and an application to labor mobility

Rainer Winkelmann*

University of Zurich

January 2026

Abstract

Winkelmann and Zimmermann (1993), although only published as a working paper, has initiated a small but persistent literature on underreporting in count data models. Their Poisson-logistic (Pogit) model was initially developed to estimate the determinants of the number of job offers during a given time interval, when the researcher only observes the number of accepted offers. Later Pogit applications can be found in epidemiology, engineering, public health, and criminology, among others. I review some of the recent developments in the literature. I also consider an extension of the original model where the reporting process can be estimated separately from the event generating process due to the availability of a validation dataset.

Keywords: binomial thinning, infeasible Poisson, validation data, two-step Pogit

JEL classification: C21, C25, J62

* Department of Economics, University of Zurich, E-mail: rainer.winkelmann@econ.uzh.ch.

1 Introduction

Winkelmann and Zimmermann (1993, henceforth WZ93) discuss underreporting in the context of a structural model of labor mobility: over time, employed workers receive outside-offers with a given frequency. When receiving a offer, the worker decides whether or not to accept. In data available to WZ93, they observe the number of job changes, and hence accepted offers, during a 10 year period for a sample of workers in Germany. The number of total offers is unobserved. However, by imposing functional form assumptions, WZ93 show how features of the distribution of the number of wage offers can be identified even if only the number of accepted offers is observed. Clearly, the number of accepted offers is less or equal to the total number of wage offers, so the inference problem is equal to the one arising for underreported count data in general.[1] WZ93 refer to their proposed approach as the Poisson-logistic regression, or Pogit model.

Although never published in a journal, WZ93 has initiated a small but lasting literature on underreporting in count data models. To give some examples from the applied literature, the Pogit model has been used in labor economics (worker absenteeism in Winkelmann, 1996), epidemiology (tuberculosis incidence in Stoner et al., 2019, and Shaweno et al., 2017; cervical cancer deaths in Dvorzak and Wagner, 2016), public health (number of abortions in Tennekoon, 2017), criminology (assaults in Moreno and Girón, 1998; property crimes in Papadopoulos, 2014), highway engineering (traffic accidents in Wood et al., 2016) and in the study of natural hazards (volcano eruptions in Stoner, 2018).

On the methodological side, a number of papers have clarified identification issues that can arise within the Pogit model, pointing to the need for outside, or prior, information to obtain a model that is formally identified, as well as an estimator that is stable and well-behaved in finite samples (e.g., Santos Silva and Papadopoulos, 2012, Dvorzak and Wagner, 2016, Brennan et al., 2021).

One such source of outside information considered in the literature is the availability of a validation dataset, for which it is "known" that events are properly counted (e.g. Dvorzak and Wagner, 2016). However, a formal incorporation of such validation information in a frequentist set-up, where

[1] In contrast, *mis*-reporting, i.e. the possibility that a reported count under- or overstates the true count, would require a different framework, see e.g. Li et al. (2003).

parameters are estimated by maximum likelihood, is to the best of my knowledge missing so far.

This paper attempts to fill this gap, by suggesting a new two-step Pogit estimator for underreported count data. Simulation experiments show that it works very well, allowing consistent parameter estimation as if the true count was known, although there is some cost to pay in terms of reduced efficiency. An application to determinants of individual labor mobility shows that the two-step Pogit estimator based on direct job changes corrects very well for underreporting, as estimates are very similar to those obtained for the infeasible Poisson regression that uses the total number of job changes as dependent variable.

2 The Pogit model

The model proposed by WZ93 assumed a two-step data generating process known as "binomial thinning". Let Y_i^* represent the (latent) total number of counts for observations $i = 1, 2, \ldots, n$. In a first step, Y_i^* is drawn according to a Poisson distribution:

$$Y_i^* \sim Poisson(\lambda_i) \tag{1}$$

In the second step, events are filtered through a reporting process, where event i has probability p_i of being reported. This filtering process can be written as a convolution operation

$$Y_i = \sum_{j=1}^{Y_i^*} B_j \tag{2}$$

where $Y_i^* \in \mathbb{N}_0$, $B_j \in \{0, 1\}$ are i.i.d. Bernoulli distributed random variables with success probability p_i, and where an empty sum (when Y_i^*=0) is defined to be zero. Equivalently,

$$Y_i|Y_i^* \begin{cases} \sim Binomial(Y_i^*, p_i) & \text{for } Y_i^* = 1, 2, \ldots \\ = 0 & \text{for } Y_i^* = 0 \end{cases} \tag{3}$$

Note that there are two types of zeros in model (3), one arising by chance, as actual events are randomly not reported, and one arising deterministically, when no event occurred.

Under the stated assumptions, it is straightforward to show that the number of reported counts

is Poisson distributed with parameter $p_i\lambda_i$:[2]

$$Y_i \sim Poisson(p_i\lambda_i) \tag{4}$$

So a simple Poisson regression with Y_i as dependent variable identifies the overall mean $\mu_i = p_i\lambda_i$, but not its constituent parts, p_i and λ_i. A high observed average count may be due to either a high underlying rate or a high reporting rate. Without further assumptions, it is impossible to disentangle these two possibilities and learn something about λ_i and its determinants. I introduce such assumptions in the next subsection.

Before that, it is worthwhile reflecting on the various assumption already made by the model, a Poisson distribution for the true count and a completely random reporting process, which also means independence of the count and the reporting: the reporting rate p_i and the occurrence rate λ_i do not correlate in the population. Below, when introducing covariates, the independence requirement will be weakened to that of conditional independence. Also, the assumptions rule out "occurrence dependence" whereby for instance more counts increase or lower the probability of reporting.

2.1 Introducing covariates

With covariates, it becomes possible to disentangle effects operating through the rate intensity to that from reporting. Specifically assume as in WZ93 that the Poisson rate for the total number of counts Y_i^* conditional on a set of covariates x_i is given by

$$\lambda_i = \exp(\beta_0 + x_i'\beta) \tag{5}$$

and that the reporting probability conditional on covariates z_i follows a logit model

$$p_i = \frac{\exp(\alpha_0 + z_i'\alpha)}{1 + \exp(\alpha_0 + z_i'\alpha)} \tag{6}$$

[2]The underreporting consider here is probabilistic. Another distribution would arise if underreporting were deterministic. Rose (2026) considers the case where m "proto-events" are needed to produce one actual event, and the distribution of proto-events is Poisson. For example, for $m = 2$, $Y = 0$ for $Y^* \in \{0, 1\}$, $Y = 1$ for $Y^* \in \{2, 3\}$, and so forth.

It follows that the Poisson parameter for observed counts Y_i is given by[3]

$$\lambda_i^{\text{pog}} \equiv p_i \lambda_i = \frac{\exp(\alpha_0 + z_i'\alpha)}{1 + \exp(\alpha_0 + z_i'\alpha)} \exp(\beta_0 + x_i'\beta) \tag{7}$$

The set of variables in z_i can be identical to x_i, overlap with x_i, or be completely distinct from x_i. For obvious reasons, the resulting model has been named the Poisson-logistic, or Pogit, regression model. The assumed non-linearity of the exponential and logit functions allows to separately identify the determination of the total counts and of reporting, but there are some limits. For example, the model is not globally identified because when $\alpha = 0$, the constant in the Pogit model is given by $\beta_0 + \alpha_0 - \log(1 + \exp(\alpha_0))$ which gives two unknown parameters for one estimable quantity.

Another important observation has been made by Papadopoulos and Santos Silva (2012). Let z_i and x_i contain the same covariates, now including a constant, without any exclusion restriction, then

$$\begin{aligned}\lambda_i^{\text{pog}} &= \frac{\exp(x_i'\alpha)}{1 + \exp(x_i'\alpha)} \exp(x_i'\beta) = \frac{1}{1 + \exp(-x_i'\alpha)} \exp(x_i'\beta) \\ &= \frac{\exp(-x_i'\alpha)}{1 + \exp(-x_i'\alpha)} \exp(x_i'(\beta + \alpha))\end{aligned}$$

so the parameter pairs (α, β) and $(-\alpha, (\beta + \alpha))$ give the same conditional probability $\Pr(Y_i|x_i)$. Identification of the model parameters can be achieved by a sign restriction on α, or by an exclusion restriction: there are covariates that matter for the count but do not affect the reporting probability. Depending on application, it can be easier or harder to justify such exclusions.[4]

2.2 Consequences of underreporting

Given that underreporting poses conceptual and empirical challenges, it is a legitimate question what are the consequences of simply ignoring underreporting? The answer depends both on the purpose of the analysis, prediction or estimating effects, as well as on the nature of underreporting.

[3]This form of a conditional expectation function (CEF) is not specific to the Poisson model with logit reporting function. The same CEF can arise in zero-inflated and hurdle Poisson models (see Staub and Winkelmann, 2013)

[4]WZ93 used a simple economic model of labor mobility, where labor demand conditions determined the wage offer distribution but not the acceptance decision which in turn depended only on human capital variables.

For prediction, underreporting that is unaccounted for directly leads to underprediction, which is a problem.

For estimated effects, say slopes in a standard Poisson model, the consequences are less clear, because they depend on the nature of the reporting process. Three cases can be distinguished. First, assume that there is a constant reporting probability. This affects only estimation of the constant in the Poisson regression, and estimated semi-elasticities are unchanged.

Second, the same conclusion holds if the reporting probability varies with i, but the variation is random and independent of x. In this case, underreporting acts like unobserved heterogeneity. The only difference is that the multiplicative factor p_i is bound between 0 and 1, rather than non-negative as is usually assumed. In either case, the marginal count distribution is no longer Poisson, as reflected by overdispersion, but Poisson regression remains consistent for the semi-elasticities, due to its pseudo-maximum likelihood property (e.g. Winkelmann, 2008). Under particular assumptions, it is possibe to derive the exact distribution model. For example, if p_i has a standard uniform distribution, then the marginal distribution of Y is available in closed form (Winkelmann, 1996):

$$\Pr(Y_i = y_i) = \int_0^1 \frac{e^{-\lambda_i p}(\lambda p)^{y_i}}{y_i!}\mathrm{d}p = \lambda^{-1}\left(1 - e^{-\lambda_i}\sum_{j=0}^{y_i}\frac{\lambda^j}{j!}\right) \qquad y_i = 0, 1, \ldots \qquad ,$$

although this does not represent a known "standard" distribution type.[5]

Third, and perhaps most interestingly, the reporting probability is neither constant nor entirely random but related to x: meaning that it either depends on x (or part of x) itself, or on factors z that are correlated with x. Ignoring underreporting in this case means that the count process slopes β are estimated inconsistently, an instance of omitted variable bias.

2.3 Underreporting versus censoring

In principle, underreporting could also be understood as an instance of censoring. When a count of Y_i is observed, we know that $Y_i^* \geq Y_i$. If Y_i^* has probability function $f(y_i^*)$, then $\Pr(Y_i^* \geq y_i) =$

[5]Fader and Hardie (2000) derive the marginal distribution of Y_i when p has a beta distribution and in addition, multiplicative unobserved heterogeneity in the Poisson rate λ_i is modelled using a gamma distribution.

$\sum_{j=y_i}^{\infty} f(j)$, where $f(j)$ could be, for instance, the Poisson probability function. Hence, rather than focusing on the distribution of Y_i, the censoring approach focuses on the conditional distribution of $Y_i^*|Y_i$. The main shortcoming of this approach is that the ML estimator does not exist if all observations are subject censoring. Intuitively, $\Pr(Y_i^* \geq y_i)$ will tend to one for all units, as the mean parameter of $f(j)$ goes to infinity.

As a consequence, a censored Poisson regression model requires a mix of censored and uncensored observations, i.e., information on which counts are underreported and which ones are not. Such information is usually not available. Oliveira et al. (2017) discuss a model with a random censoring mechanism that is independent of the counts, embedding it in a Bayesian posterior analysis. Plausible external information on consoring status exists in the application of Caudill and Mixon (1995) to fertility data. The outcome variable is completed fertility, but the sample includes women who are still in their childbearing years, defined here as being younger than 40 years. Hence, the observed number of children in censored (or an undercount of completed fertility) for those aged 39 or younger, and it is uncensored and thus complete for women aged 40 or older.

2.4 Estimation

Estimation of the Pogit parameters α and β by maximum likelihood or pseudo-maximum likelihood appears straightforward at first, provided identification is ensured. Given an independent sample of size n, the log likelihood function has the form

$$l(\alpha, \beta; y, z, x) = c + \sum_{i=1}^{n} \left\{ -\frac{\exp(z_i'\alpha + x_i'\beta)}{1 + \exp(z_i'\alpha)} + y_i[z_i'\alpha + x_i'\beta - \ln(1 + \exp(z_i'\alpha))] \right\}.$$

Let $\theta = [\alpha \ \beta]'$ and $\Lambda(z_i'\alpha) = \exp(z_i'\alpha)/(1 + \exp(z_i'\alpha))$. Then

$$\begin{aligned} \frac{\partial l(\theta; y, z, x)}{\partial \theta} &= \sum_{i=1}^{n} \left(\frac{y_i - \lambda_i^{\text{pog}}}{\lambda_i^{\text{pog}}} \right) \frac{\partial \lambda_i^{\text{pog}}}{\partial \theta} \\ &= \sum_{i=1}^{n} (y_i - \lambda_i^{\text{pog}}) \begin{bmatrix} x_i \\ z_i(1 - \Lambda(z_i'\alpha)) \end{bmatrix}. \end{aligned} \tag{8}$$

Since x_i contains an intercept, the first-order conditions

$$\left. \frac{\partial l(\theta; y, z, x)}{\partial \theta} \right|_{\hat{\theta}} = 0$$

imply that the sum of the residuals $\hat{u}_i = y_i - \hat{\lambda}_i^{\text{pog}}$ is equal to zero. WZ93 provide the expression for the observed Hessian, and optimization can be based on second-order approximations, using either the observed Hessian as in Newton-Raphson or the expected Hessian in the BHHH algorithm.

However, there are two problems that can make actual estimation difficult. First, the observed Hessian matrix is not necessarily negative definite (see WZ93), due to diagonal terms where $y_i - \lambda_i^{\text{pog}} < -\exp(x_i'\beta)(1 - \Lambda(z_i'\alpha))$ which can happen when the reporting probability is large and y_i is below its mean. WZ93 therefore advise to repeat the optimization for various starting values addresses the problem of potential multi-modality of the log-likelihood function for a given sample.

Second, Brennan et al. (2021) show that estimation of α in the Pogit necessarily lacks precision. They analyse the inverse of the Fisher information matrix, i.e. the Cramér Rao lower bound, in a two-parameter Pogit model: p and λ are each determined by a single covariate, and covariates are independent. They find that the standard deviation of $\hat{\alpha}$ grows at least proportionally to the value of α. This means that the signal-to-noise ratio does not depend on α, and the probability of incorrectly determining whether p is an increasing or decreasing function of z is at least constant, for a given sample size n. They conclude that additional information must be used to decrease the variance of $\hat{\alpha}$ and thereby obtain meaningful estimates. For example, such outside information could impose bounds on p, or restrict the sign of α.

Given these caveats, I explore in the following section an alternative avenue for addressing underreporting in Poisson regression models, by separating the estimation of α and β when validation data ara available.

3 Using validation data

The basic idea underlying the use of external validation data is separating the estimation of α and β, i.e., to use an auxiliary dataset to estimate the reporting equation, and a second dataset to estimate the count equation. The requirement is that the validation dataset has information on Y^*, Y and z, whereas the main dataset has information on Y, z and x. As before, z and x can be identical covariates, overlapping or disjoint.

Consider the validation data first. If Y_i^* and Y_i are observed, it is possible to compute reporting

ratios for each i as long as $Y_i^* \neq 0$: $w_i = Y_i/Y_i^*$. If $Y_i, Y_i^* \in \mathbb{N}$ for all i (i.e., no zeros) the Poisson first-order condition

$$\sum_{i=1}^{n}(Y_i^* - \exp(x_i'\beta))x_i = 0$$

is equivalent to

$$\sum_{i=1}^{n} 1/w_i(Y_i - w_i \exp(x_i'\beta))x_i = 0$$

so the Poisson regression using the full counts is equivalent to a weighted Poisson regression using the underreported counts where the reporting rate is used as an offset, and where the first-order condition is weighted by the inverse reporting rate.[6]

This result is of limited use, unfortunately, since zero counts are present in most applications, and often constitute a high proportion of all data. Moreover, this equivalence does not help in correcting for underreporting in a separate dataset, where information on Y_i^*, and thus on individual reporting rates, is missing.

Instead, it is a better idea to use the validation dataset in order to build a prediction model for reporting rates, conditional on covariates z, that can then be applied to new data, by adjusting the conditional expectation function to the presence of underreporting. Suppose that a logit model has been estimated. Predicted reporting probabilities $\hat{p}_i$ are obtained as

$$\hat{p}_i = \frac{\exp(\hat{\alpha}_0 + z_i'\hat{\alpha})}{1 + \exp(\hat{\alpha}_0 + z_i'\hat{\alpha})} \tag{9}$$

A modified Pogit model replaces the unknown reporting probability p_i by the estimate $\hat{p}_i$, and accordingly postulates that $Y_i \sim Poisson(\hat{p}_i\lambda_i)$, where $\hat{p}_i\lambda_i = \exp(x_i'\beta + \log\hat{p}_i)$, yielding the first-order condition

$$\sum_{i=1}^{n}(Y_i - \exp(x_i'\beta + \log\hat{p}_i))x_i = 0 \tag{10}$$

This estimator can be easily implemented in two-steps. First, a fractional response model is estimated by pseudo maximum likelihood for the validation data (see Papke and Wooldridge, 1996).

[6] Although w_i is well defined for $Y_i = 0$ and $Y_i^* > 0$ (a zero reporting probability), using weights $1/w_i$ requires that $Y_i > 0$ as well. Also, we cannot scale Y_i up to get Y_i^*, because the product of $Y_i = 0$ with any factor remains zero.

This needs to be done on a subset of observations for which $Y_i^* > 0$, because $Y_i^* = 0$ implies that $Y_i = 0$, and the reporting rate is not defined in this case. Second, a standard Poisson regression with logarithmic offset is estimated. Robust standard errors can be used, while bootstrapping standard errors would in addition account for the first-step estimation error.

3.1 Numerical illustration

In this section, I investigate the performance of the proposed estimation method in a small simulation experiment. I generate a sequence of artificial datasets for Y^* and Y, depending on covariates, and document that the two-step method outlined in the previous section on average recovers the true regression coefficients and thus performs as if the true count, not the underreported one, was used as dependent variable in a Poisson regression. For simplicity, I estimate both components, the reporting function and the count equation, using the same dataset, but results would be the same, if the validation data set is collected independently of the Poisson estimation dataset, as long as the underlying data generating process (DGP) is stable.

The characteristics of the DGP are as follows: the sample size n is either 100 or 1000. Two covariates are included:

$$x_{1i} = \exp(u_i) - 1, \qquad x_{2i} = x_{1i} + z_i$$

where pseudo-random numbers for u_i are drawn from a uniform distribution, whereas pseudo-random numbers for z_i are drawn from a standard normal distribution. It follows that x_{1i} is non-negative with a mean of approximately 0.72. The distribution is asymmetric and skewed to the right. Moreover, x_{1i} and x_{2i} are positively correlated ($\rho \approx 0.44$). The true count process coefficients are $\beta_0 = 0$, $\beta_1 = 1$ and $\beta_2 = 0.1$. Poisson pseudo-random numbers with mean $\lambda_i = \exp(x_{1i} + 0.1x_{2i})$ yield observations for Y_i^*. In this set-up, the proportion of zero counts is about 15%, the mean of Y_i^* is about 2.6, and the largest generated count is about 17.

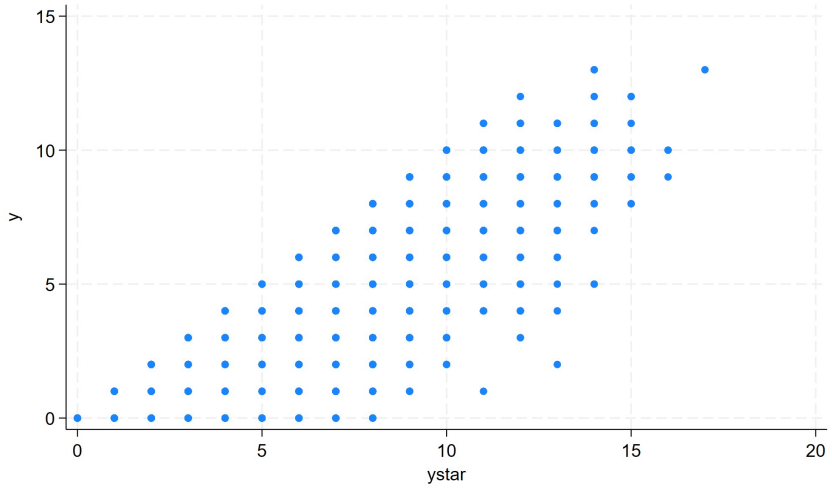
Fo the reporting process, I assume that it depends on x_1 only, as actual counts are reported with probability

$$p_i = \frac{\exp(-1 + x_{1i})}{1 + \exp(-1 + x_{1i})},$$

Underreported counts Y_i are then drawn as pseudo-random numbers from a binomial distribution with parameters Y_i^* and p_i. Since the binomial distribution is not defined for $Y_i^* = 0$, but we know that $Y_i^* = 0$ implies that $Y_i = 0$, Y_i is coded accordingly. As a result, the proportion of zeros increases to about 40% for the underreported counts, the mean is about 1.3, and the largest observed value drops to about 13.

Figure 1 displays a scatter plot of Y_i against Y_i^*, for a typical sample. It is evident how Y_i^* is an upper bound for Y_i. Zero observed counts are observed for true counts up to eight. What the figure does not show is the joint frequency distribution which would attribute much more mass to points in the left part of the plot, and to those about halfway between the 45 degree line and the abcissa, since the average reporting rate is about 46%.

Figure 1: Scatterplot of Y_i against Y_i^*



For the thus obtained sample, I first estimate a fractional response model, with the ratio Y_i^*/Y_i as dependent variable (dropping all observations for which $Y_i = 0$) and x_{1i} as covariate, and use the estimates to obtain predicted reporting probabilities $\hat{p}_i$.[7] I then estimate a Poisson regression with Y_i as dependent variable and $\log \hat{p}_i$ as offset. These estimates can be compared to those obtained from a Poisson regression without offset, and to the infeasible Poisson regression with Y_i^* as dependent variable.

[7]The exclusion restriction for x_{2i}, $\alpha_2 = 0$, is imposed here, but this is not necessary. The results remain similar if the fractional response model includes x_{2i}.

Repeating this process many times, in my case for $r = 1000$ replications, allows me to recover the means and standard deviations of the respective sampling distributions for the two slope parameters associated with x_{1i} and x_{2i}. These are shown in Table 1. The left panel displays results for $n = 100$, the right panel that for $n = 1000$.

The table contains results for three different estimators. Two-step Pogit was introduced in the previous section. The infeasible Poisson estimator uses Y_i^* as dependent variable, and hence does not suffer from underreporting. While the two-step Pogit as implemented here also makes use of Y_i^*, in order to estimate the reporting model, the difference is that this could have been done in a smaller validation dataset. In this case, observations on Y_i^* are not needed in the dataset on which estimation of the count model is based. Finally, the naive Poisson estimator simply uses Y_i as dependent variable.

Table 1: Simulation results

	$r = 100$		$r = 1000$	
	mean	sd	mean	sd
Two-step Pogit				
$\hat{\beta}_1$	0.996	(0.187)	0.997	(0.058)
$\hat{\beta}_2$	0.101	(0.091)	0.100	(0.028)
Infeasible Poisson				
$\hat{\beta}_1$	1.004	(0.137)	0.999	(0.045)
$\hat{\beta}_2$	0.098	(0.062)	0.100	(0.019)
Naive Poisson				
$\hat{\beta}_1$	1.510	(0.211)	1.511	(0.067)
$\hat{\beta}_2$	0.102	(0.091)	0.100	(0.028)

The results show that both two-step Pogit and infeasible Poisson estimates are centered at the true parameter value, regardless of sample size. The only difference is that the infeasible estimator is more efficient, as the standard deviations are between 29 and 48 percent higher for two-step Pogit. An increase in sample size leads to the expected overall reduction in sampling variability.

By contrast, the naive Poisson estimator for β_1 is seriously biased: it on average overestimates the true effect of x_1 by about 50 percent. The reason is clear in this DGP, since x_1 has a positive effect not only on the count rate, but also on the reporting rate, whereas the naive Poisson regression conflates these two effects, a classical instance of omitted variable bias. In addition, the naive Poisson model also leads to increased standard errors, both relative to two-step Pogit and infeasible Poisson, such that the mean squared estimation error is the highest among the three estimators.

4 Application to labor mobility

In this section I reanalyze the data used in WZ93. The WZ93 data set is based on the *German Socio-Economic Panel* (GSOEP) which was provided by the Deutsches Institut für Wirtschaftsforschung (DIW, see Wagner, Burkhauser and Behringer, 1993, for further information). The WZ93 subsample covers 1962 male individuals between the age of 25 and 50 in 1974. Labor mobility is measured by the number of different jobs held during the ten-year period 1974-1984. This information was collected retrospectively in the second (1985) wave of the panel. The total number of job changes is simply the total number of jobs minus one. In the data, 56 percent of men held a single job, so the number of job changes is zero for them. The highest number of job changes was 15, and the mean count 0.9;

As mentioned in the introduction, WZ93 pursued a structural model of labor mobility, where the Pogit model allowed to distinguish between the number of wage offers and the number of accepted offers. To this end, they focused on "voluntary", or "direct" job changes, defined as those without an intervening spell of unemployment. Here, I consider total labor mobility, the sum of direct job changes and job changes involving an unemployment spell, as outcome of interest, and I ask: considering direct job changes as an undercount of total mobility, can we use the two-step Pogit estimator to correct for undercounting and thereby recover the parameters of the infeasible Poisson regression with the total number of changes as dependent variable. Since I again estimate both models on the same dataset, the infeasible Poisson regression results are observable, constituting a ground truth, against which the two-step Pogit estimator can be benchmarked.

Table 2 displays the regression results. It shows the parameter estimates for nine covariates,

as well as robust standard errors in parentheses. The first two columns refer to the two steps of the Pogit model, columns 3 and 4 to Poisson regressions for the total and direct number of job changes, respectively. This is different from WZ93, who jointly estimated the first two columns of the table, with the number of direct job changes as dependent variable. Since they did not have access to validation data, exclusion restrictions needed to be imposed instead. Here, no such exclusion restrictions are necessary to obtain stable parameter estimates, and identical covariates are used in both parts of the model.

Table 2: Determinants of # of Total Job Changes

	Reporting Logit	Pogit Step II	Infeasible Poisson	Naive Poisson
Education$*10^{-1}$	0.496	-0.300	-0.332	-0.137
	(0.304)	(0.168)	(0.148)	(0.167)
Experience$*10^{-1}$	0.072	-0.800	-0.804	-0.769
	(0.225)	(0.159)	(0.140)	(0.159)
Experience$^2 * 10^{-2}$	-0.049	0.139	0.150	0.119
	(0.071)	(0.050)	(0.046)	(0.050)
Union	0.181	-0.356	-0.368	-0.292
	(0.138)	(0.093)	(0.082)	(0.093)
Single	-0.636	0.210	0.188	-0.049
	(0.219)	(0.160)	(0.131)	(0.160)
German	-0.192	-0.299	-0.156	-0.367
	(0.158)	(0.127)	(0.107)	(0.127)
Qualified White Collar	0.386	-0.067	-0.134	0.067
	(0.280)	(0.196)	(0.180)	(0.196)
Ordinary White Collar	0.649	-0.027	-0.026	0.184
	(0.330)	(0.192)	(0.164)	(0.191)
Qualified Blue Collar	0.105	0.104	0.164	0.147
	(0.166)	(0.117)	(0.099)	(0.116)
Constant	0.003	1.145	1.137	0.501
	(0.352)	(0.192)	(0.172)	(0.191)
Observations	860	1962	1962	1962

Source: German Socio-Economic Panel, see WZ93 for details. Own calculations.

The reporting function in columns 1 is estimated using a fractional logit model on the validation data. Since this requires $Y_i^* > 0$, but the total job change count has 56 percent zeros, only 860 observations are available for this part of the model. Once the reporting function is estimated on the validation data, its predictions can be used to generate a logarithmic offset for any dataset, and information on Y_i^* is no longer needed, provided the same covariates x_i are observed. Here, however, I use the reporting model "internally", which allows for a direct comparison with the otherwise infeasible Poisson results for the total counts, displayed in column 3 of Table 2.

Summarizing the evidence, I find that the two-step Pogit estimator provides a good approximation to the infeasible Poisson model, whereas the differences between the naive and the infeasible Poisson estimators are substantial. For example, the two-step Pogit estimates and the infeasible Poisson estimates have always the same sign, differences are minor and certainly not statistically significant, none of which is true for the naive Poisson regression.

For example, being single is associated with an increase job change rate of 21 and 19 percent, respectively, in the two-step Pogit and infeasible models, but with a decline of 5 percent in the naive model that does not account for the fact that the effect of being single in the reporting model is negative and large. That means that the fraction of direct job changes to total job changes is smaller for singles than for married men, i.e. their job transitions are more likely to involve an unemployment spell. Ignoring this selection effect leads to erroneous conclusions when extrapolating the Poisson results for direct job changes to the determinants of total job changes.

5 Conclusions

This paper revisits the problem of underreporting in count data models and contributes to the literature by proposing a simple and practical two-step estimator that exploits validation data to separately identify the event-generating and reporting processes. Building on the Poisson-logistic (Pogit) framework introduced by Winkelmann and Zimmermann (1993), the approach sidesteps some of the well-known identification and precision problems associated with joint maximum likelihood estimation by decoupling the estimation of reporting probabilities from that of the latent

count intensity. When suitable validation data are available, the proposed estimator recovers the parameters of interest consistently and performs comparably to an infeasible Poisson model that directly observes true counts, albeit with some loss in efficiency.

Simulation evidence confirms that the two-step Pogit estimator corrects the potentially severe biases arising from naive Poisson regressions when underreporting is systematically related to covariates. An empirical application to labor mobility using data on German workers further illustrates the relevance of the approach in practice. Treating direct job changes as an underreported measure of total mobility, the two-step estimator closely reproduces the results of the infeasible benchmark model, while the naive specification yields qualitatively different – and in some cases misleading – conclusions. These findings underscore that ignoring underreporting is not innocuous when reporting probabilities vary with observed characteristics.

While the results highlight the value of validation data in empirical work with underreported counts, they also point to a limitation of the approach, as such data need to be available in the first place. Depending on application, it can make sense to obtain reporting rates for a small sample, when collecting the full count is possible but expensive. The correction can then be used in a much larger dataset. In other cases, aggregate data are already available from official statistics, say by region and age-bracket, and these can then be used to estimate rates of underreporting in survey data.

While the analysis here is confined to Poisson models with binomial thinning, the logic extends naturally to richer count data settings, including models with overdispersion or alternative reporting mechanisms. Future research may explore such extensions, as well as the optimal design of validation studies when data collection is costly.

References

Baker, R. (2026) A model for underdispersed count data, *Statistical Papers*, forthcoming.

Brennan, J., M. Bannick, N. Kassebaum, L. Wilner, A. Thomson, and A. Aravkin (2021) Analysis and methods to mitigate effects of under-reporting in count data, arXiv preprint arXiv:2109.12247

Caudill, B.S., and Mixon, F.G. (1995) Modeling household fertility decisions: estimation and testing of censored regression models for count data. *Empirical Economics*, 20, 183-196.

Dvorzak, M., and Wagner, H. (2016) Sparse Bayesian modelling of underreported count data, *Statistical Modelling*, 16, 24-46.

Fader, P.S. and B. G. S. Hardie (2000) A note on modelling underreported Poisson counts, *Journal of Applied Statistics*, 27, 953-964.

Li, T., Trivedi, P. K., and Guo, J. (2003) Modeling response bias in count: A structural approach with an application to the National Crime Victimization Survey data. *Sociological Methods & Research*, 31(4), 514-544.

Moreno, E., and J. Girón (1998) Estimating with incomplete count data a Bayesian approach, *Journal of Statistical Planning and Inference*, 66, 147-159.

Oliveira, G.L., Loschi, R.H., and Assuncao R.M. (2017) A random-censoring Poisson model for underreported data, *Statistics in Medicine*, 36, 4873-4892.

Papadopoulos, G. (2014) Immigration status and property crime: an application of estimators for underreported outcomes, *IZA Journal of Migration*, 3:12.

Papadopoulos, G. and Santos Silva, J.M.C. (2012) Identification issues in some double-index models for non-negative data. *Economics Letters* 117, 365-367.

Papke, L. and J. Wooldridge (1996) Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates, *Journal of Applied Econometrics*, 11, 619-32.

Shaweno, D., Trauer, J. M., Denholm, J. T., and McBryde, E. S. (2017) A novel Bayesian geospatial method for estimating tuberculosis incidence reveals many missed tb cases in Ethiopia, *BMC Infectious Diseases*, 17, 662.

Staub, K.E. and Winkelmann, R. (2013) Consistent estimation of zero-inflated count models. *Health Economics* 22, 673-686.

Stoner, O. (2018) Correcting under-reporting in historical volcano data, *Proceedings of the 33rd International Workshop on Statistical Modelling*, 1, 288-292.

Stoner, O., T. Economou and G. da Silva (2019) A Hierarchical Framework for Correcting Under-Reporting in Count Data, *Journal of the American Statistical Association*, 114, 1481-1492.

Tennekoon, V. S. (2017) Counting unreported abortions: A binomial-thinned zero-inflated Poisson model. *Demographic Research*, 36, 41-72.

Wagner, G.G., R.V. Burkhauser, and F. Behringer (1993) The English Language Public Use File of the German Socio-Economic Panel, *Journal of Human Resources* 28, 429-433.

Winkelmann, R., and Zimmermann, K. F. (1993) Poisson-logistic Regression, Discussion Papers, Department of Economics, University of Munich, 93(18).

Winkelmann, R. (1996) Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics* 21, 575-587.

Winkelmann, R. (1996) Under-reporting of purchases of port wine: a correction, *Journal of Applied Statistics*, 28, 637.

Winkelmann, R. (2008) *Econometric analysis of count data*, 5th ed., Springer Berlin Heidelberg.

Wood, J. S., Donnell, E. T. and Fariss, C. J. (2016) A method to account for and estimate underreporting in crash frequency research, *Accident Analysis & Prevention* 95, 57-66