NBER WORKING PAPER SERIES

HIDDEN FIGURES: UNCOVERING QUANTITIES BEHIND ZEROS WITH ECONOMETRICS

Anirban Basu

Working Paper 31632
http://www.nber.org/papers/w31632

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2023, revised September 2023

I thank Edward Norton and John Mullahy for their helpful comments—all opinions and errors and mine. Forthcoming in the Volume of Contributions in Economic Analysis published by Emerald Publishing in honour of Professor Andrew Jones (ed. Francesco Moscone and Badi Baltagi). No funding was received for writing this chapter. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Hidden Figures: Uncovering Quantities Behind Zeros with Econometrics
Anirban Basu
NBER Working Paper No. 31632
August 2023, revised September 2023
JEL No. C10,D0,I0

ABSTRACT

This chapter reviews the econometric approaches typically used to deal with the spike of zeros when modeling non-negative outcomes such as expenditures, income, or consumption.

Anirban Basu
The CHOICE Institute
Departments of Pharmacy,
Health Services, and Economics
University of Washington, Seattle
1959 NE Pacific St., Box - 357660
Seattle, WA 98195
and NBER
basua@uw.edu

1. INTRODUCTION

Many economic outcomes are nonnegative in nature. These typically include gross income, wealth (in some cases), patents, consumption expenditures, and utilizations of commodities. The empirical distribution of these outcomes is usually skewed to the right (i.e., long tails) and often kurtotic (i.e., fat tails). Several statistical approaches (e.g., transformation models, nonlinear method of moments, generalized linear models, etc.) have been developed to deal with these features (see Jones 2000 for a review of related statistical models). However, special attention is sometimes needed to deal with another common part of such distributions – a spike in zeros. Often, alternate behavioral assumptions may be required to understand the economic rationale for why zeros were generated in the data, how they should be modeled, how the model parameters should be estimated, and how the results should be interpreted.

A key distinction is understanding whether the goal is to model the true zeros or some other hidden figures behind the zeros observed in the data. The data generation process for both the true zeros and the hidden figures can have rich behavioral content, which may trigger the use of some multi-part models and hurdles. However, uncovering the hidden figures and the marginal effects with respect to these hidden figures would require selection models (e.g., Heckman selection model) with appropriate exclusion restrictions. In this chapter, I briefly review some of these questions and solutions.

2. OVERVIEW OF STATISTICAL MODELING OF NON-NEGATIVE OUTCOMES

To simplify concepts, let us focus on understanding the effect of one binary treatment variable, *D*, on a non-negative outcome, *Y*. For now, I assume that *D* was assigned (pseudo-) randomly across individuals, conditional on a set of observed covariates *X*. Let the potential outcomes (Y_j, *j*=0,1) associated with each level of D be given as:

$$Y_j = \mu(X; \alpha_j) + U_j,\ j = 0,1, \qquad \text{Eq. 1}$$

where α_j represents the parameter of the potential outcome models, and *U* represents a stochastic error term ($U \perp \mathrm{D} \mid \mathrm{X},$ where ⊥ signifies statistical independence). These potential outcomes can be considered consumption outcomes generated based on some generic utility maximization process conditional of *D* and *X.* An example would be understanding the effect of genetic therapy on emergency department use expenditures or the impact of insurance coverage of nicotine patches on smoking. In both cases, assume that the choices of treatments (genetic theory or the coverage) were based on selection on a set of observed covariates *X*.

The potential outcomes, being non-negative in nature, could include zeros. These zeros are considered corner solutions for maximization of consumption utility (Jones 1989a; 1989b). Statistically, having zeros in these outcomes is nothing special. They can usually be treated as any other value in the distribution of these outcomes. The observed outcome is given as follows:

$$Y = D \cdot Y_1 + (1 - D) \cdot Y_0 \qquad \textit{Eq. 2}$$

Analysts never know the true functional form of *μ*(). However, it is often assumed that the expectation of the observed outcome is typically nonlinear with respect to D as long as *μ*() is a non-linear function. Therefore, the expectation of the observed Y can be empirically modeled as a function of *D*, allowing *Y* to follow any standard non-negative probability distribution functions such as Poisson, Gamma, or Exponential. Since Y is non-negative, its expectation is strictly positive, and hence, the typical mean function is given as:

$$E(Y|D, X) = h(\beta_0 + \beta_1 \cdot D + \beta_2 \cdot X) \qquad \textit{Eq. 3}$$

and model parameters estimated via full-information maximum likelihood (FIML) or Quasi-likelihood (QL) maximization approaches (McCullagh 1983; McCullagh and Nelder 1989).

To under the difference between FIML and QL, a quick review of distribution theory is needed. The shape of a distribution of random variables is characterized by a series of moments (m) (sometimes expressed as standardized moments by scaling with a power of the variance). For example, the first moment is the expected value, the second central moment is the variance, and the third and fourth standardized moments are the skewness and kurtosis, respectively. There can be an infinite number of moments. A parametric distribution consists of one or a vector of parameters defined at the population level that completely specifies these moments. For example, a gamma distribution has two parameters, $\{a,b\}$, which in turn defines the moments, e.g., m1 = a/b, m2 = a/b^2, m3 = $2/\sqrt{a}$, m4 = $6/a$, and so on. A statistical model expresses functional forms for the first and second sample moments using a set of model parameters (e.g., β in Eq. 3, distinct from the distribution parameters). These model parameters or their functionals are typically the targets for inference. When we estimate these model parameters from the data at hand using a FIML approach, where a parametric distribution is specified, we link these model parameters to the structural parameters of the specified parametric distribution that specifies all the population moments. For example, in a log-linear model, one specifies an exponential mean model for the sample mean. For a FIML approach with, say, a log-normal distribution, this mean-model specification is followed to impose the full structure for all the moments of the parametric distribution onto the data at hand. When the data follows such a structure, we get a maximum likelihood estimator (MLE) for our statistical model parameters, generating the minimum variance estimator. However, if the data deviates considerably from the structure of the parametric distribution, then even when our mean (and variance model) may be correctly specified, MLE could be inconsistent.

In contrast, analysts often use the method of moments estimators (popular in economics) or quasi-likelihood estimators (popular in statistics) that relax the stringent FIML requirements (McCullagh, 1983; McCullagh and Nelder, 1989); Blough et al. (1999). A full parametric distribution is not explicitly specified for the dependent variable in these approaches. Instead, the sample moment models are set up to reflect a finite population moment following a specific distribution. Estimation of the model parameters then follows minimizing some form of weighted least-squares (Wedderburn 1974). The advantage of these approaches is that if the specified sample moments models are correct, one will get consistent estimates for our model parameters. Therefore, these approaches are robust to misspecification of the underlying distribution function of the dependent variable. However, there are no free lunches. QL estimators are often inefficient compared to well-specified MLEs.

The quasi-likelihood approach was developed to estimate parameters linked to the exponential family of distributions, e.g., Exponential, Gamma, Poisson, etc. One characteristic of this distribution family is that the second moment can often be expressed as a function of a dispersion parameter and the first moment. Therefore, if we specify a mean model for the data, the QL approach uses this specification to define the population mean and variance. It then only imposes this mean-variance relationship to the data at hand, but not any structural relationships with the higher order moments that would be defined under a specific parametric distribution.

2.1 Least squares

Sometimes researchers use transformation models to avoid running non-linear specifications for Eq. (3). For example, a common practice in economics is to log-transform a non-negative outcome so that the transformed outcome can then be modeled linearly as a function of *D*, *X*, and possible their interactions.

$$\text{Ln}\,(Y) = \beta_0 + \beta_1 \cdot D + \beta_2 \cdot X + \varepsilon \qquad \textit{Eq. 4}$$

Estimation can proceed following FIML or QL. For example, in FIML, one can assume $\varepsilon \sim Normal(0, \sigma^2)$ and carry out a log-likelihood maximization. Alternatively, one can use minimize Ordinary Least Squares (OLS), which represents a QL approach.

The challenges of interpreting the parameters of log-transformed outcomes have been widely reported (Manning 1998; Manning and Mullahy 2001). Specifically, such a model represents the conditional mean of the Log(Y), i.e., E(Ln(Y) | *D*, *X*). This formulation represents the conditional geometric mean for the distribution of Y, not the arithmetic mean, which is the target for inference in Eq(3). Consequently, the marginal effects of D on the geometric mean of Y are directly obtained from the regression parameters. Still, the marginal effect of D on the arithmetic mean requires complex transformation processes (e.g., Duan's smearing estimate, Duan (1983)) that can fundamentally rely on modeling all dependence of all the higher-order moments of Y on both *D* and *X*. Similar challenges arises with other transformation approaches such as inverse hyperbolic sine.

Transformation models become even more challenging in the presence of zeros in the data for Y. Typically, authors add an arbitrary constant to Y and then take a log transformation. Alternatively, they use an inverse hyperbolic sine transformation that is similar to the natural log transformation but allows for zeros. A recent paper by Mullahy and Norton (2022) shows that these transformation models have an extra parameter generally not determined by theory but whose values have enormous consequences for point estimates. As these parameters go to extreme values, estimated marginal effects on outcomes' natural scales approach those of either an untransformed linear regression or a normed linear probability model. Chen and Roth (2023) explain why marginal effects from such transformation models are not interpretable in the presence of zeros. Intuitively, they show that the individual-level percentage effect is not welldefined for individuals whose outcome changes from zero to non-zero when receiving treatment. These papers suggest that if the goal of the analysis is to obtain consistent estimates for the marginal effects of, say, *D* on the conditional arithmetic mean, as in Eq. (3), then one should not rely on transformations, especially in the presence of zeros.

3. THE WORLD OF EXCESS ZEROS

The term “excess zeros” is usually not clearly defined. In statistical terms, if the proportion of zeros in the data exceeds that implied by a specific distribution from the exponential family, data is presumed to be overdispersed (due to excess zeros). Overdispersion can occur due to other values in the data, and the designation of excess would vary based on the reference distribution (typically a Poisson distribution). Most often, when zeros lead to a bimodal distribution of the empirical data, analysts consider them to be excess. Mullahy (1997) describes them to be a consequence of unobserved heterogeneity.

3.1. True Zeros

An important feature of non-negative outcomes in specific applications is the presence of a significant density spike at zeros. These zero-inflated data arise from many individuals reporting/recording, for some reason, no consumption. Econometric modeling of this feature of the Y distribution often requires understanding the structural data-generating process (DGP) and specifying the target parameter of interest in the context of such a structure. The simplest assumption regarding such a data-generating process is that these zeros are all due to the results of a corner solution of a maximization process, as described above. Under a DGP where the observed zeros are the “true” zeros, the sample data can be modeled using the ML and QL methods described above. In some cases, a two-part model (TPM) may be used that conditionally separates the expectation of the dependent variable as:

$$E(Y|D,X) = \Pr(Y > 0|D,X) \cdot E(Y|D,X,Y > 0) \qquad \textit{Eq. 5}$$

As the name suggests, this model can be estimated using two parts - the first part consists of a binary outcome of whether Y>0 (Mullahy 1998). A logistic or a probit model can be used for the first part, and either can be estimated following FIML or QL techniques. The second part models the expectation of Y given Y>0. This part can be modeled with any of the FIML or QL approaches described in Section 2.

One FIML approach to this estimation involves estimating parameters for both parts of the models jointly. Based on Tobin's (1958) work, a Tobit model was formulated where one explicitly considers the utility function underlying the observed choices. This latent utility function (Y^*) is assumed to follow a Gaussian distribution, Normal (μ^*, σ^2). The corner solutions arise as follows:

$$\begin{aligned} Y &= 0 \quad \text{if } Y^* \leq 0 \\ &= Y^* \quad \text{if } Y^* > 0 \end{aligned} \qquad \textit{Eq. 6}$$

Under this formulation, one can directly maximize a likelihood for Y^*, following observed data *Y*. Specifically, consider a linear mean model for Y^*: $\mu^* = E(Y^*|D,X) = \beta_0 + \beta_1 \cdot D + \beta_2 \cdot X$, and $Z^0 = -\frac{\mu^*}{\sigma}$. Then,

$$Pr(Y > 0|D,X) = P\,r(Y^* > 0) = (1 - \Phi(Z^0)),$$

$E(Y|Y^* > 0, D, X) = [\mu^* + \sigma \cdot \lambda(Z^0)]$, and

$$E(Y|\, D, X) = (1 - \Phi(Z^0)) \cdot [\mu^* + \sigma \cdot \lambda(Z^0)], \qquad \textit{Eq. 7}$$

where $\lambda(Z^0) = -\frac{\emptyset(Z^0)}{(1-\Phi(Z^0))}$ is the inverse Mills Ratio, $\emptyset()$is a standard normal density function, and $\Phi()$ is a standard normal cumulative distribution function. It is helpful to note that the Tobit model mirrors the principle of a TPM (as in (5)). Still, its estimation is based on a Gaussian likelihood function, accounting for the truncation of zeros in the second part.

There are primarily three advantages of a TPM over a Tobit model. One can specify different functional forms for the mean models of the two parts. These mean models can be nonlinear in *D,* and *X.* One does not have to formally deal with the truncation of zeros in the second part while using a QL approach.

3.2. Richer Behavioral Assumptions for Zeros

Zeros can often comprise richer behavioral assumptions than corner solutions for consumption decisions. Given the substantive context of the data, one must conceptualize and defend a richer DGP. Following Deaton and Irish (1984) and Jones's work (1989a; 1989b), I describe data on cigarette smoking behavior where zeros may camouflage other underlying behaviors.

A survey of smoking behavior often collects data using the question “How many cigarettes have you smoked in the past 30 days?” or “On how many of the PAST 30 DAYS did you smoke a cigarette?”. Of course, most national surveys of cigarette consumption will ask other related questions to fully understand the dynamic nature of cigarette behavior. However, suppose one wants to model the above question in silo. In that case, one can easily perceive that some reported zeros may result from alternate data-generating processes rather than a corner solution of a consumption decision. Let’s consider those alternative DGPs.

Two distinct behaviors could lead to zeros in reported data. An individual may be a never smoker, i.e., the reported zero is a manifestation of a distinct decision process on whether to smoke at all. Alternatively, an individual may be a smoker, but their consumption decision reached a corner solution of zero cigarettes last month for various reasons. For example, the individual may take a break and quit smoking for some time. These behaviors can be represented formally as:

Reported/Observed outcomes (*Y*) represents $Y = P \cdot Y^{**}$, where *P* is the smoking participation decision, and Y^{**} is the latent potential consumption for everyone, irrespective of their smoking status.

Participation Decision (driven by latent utility U)

$$U = g(D, Z;\ \alpha) + u = \ \alpha_0 + \alpha_1 \cdot D + \alpha_2 \cdot Z + \ u \qquad \textit{Eq. 8}$$

Such that $P = 1$ *if* $U > 0$, $P = 0$ otherwise. *Z* is a set of covariates affecting participation, which may or may not be different from *X.*

Among those who chose to participate, Consumption Decision (following (6)):

$$Y^{**} = 0 \quad \text{if } Y^{*} \leq 0$$

$$= Y^{*} \quad \text{if } Y^{*} > 0\text{, and}$$

$$Y^{*} = h(D, X;\ \beta) + v = \ \beta_0 + \beta_1 \cdot D + \beta_2 \cdot X \ + \ v. \qquad \textit{Eq. 9}$$

A joint likelihood for the data under these DGPs is given as follows:

For data representing zeros:

$$L^0 = \prod_0[Pr(U \leq 0) \cdot \Pr(Y^* \leq\ 0\ |U > 0)] \qquad = \prod_0[\Pr(u \leq -g(D,Z;\ \alpha)) \cdot \Pr(v \leq -h(D,X;\ \beta)\ |u > -g(D,Z;\ \alpha))] \qquad \textit{Eq. 10}$$

For data representing non-zeros:

$$L^+ = \prod_+[\Pr(P = 1) \cdot \Pr(Y^* >\ 0\ |\mathrm{P} = 1) \cdot \mathrm{f}(Y^*|Y^* > 0, P = 1)] = \prod_+[\Pr(u > -g(D,Z;\ \alpha)) \cdot \Pr(v > -h(D,X;\ \beta)\ |u > -g(D,Z;\ \alpha)) \cdot \mathrm{f}(Y^*|v > -h(D,X;\ \beta), u > -g(D,Z;\ \alpha))] \qquad \textit{Eq. 11}$$

This setup gave rise to the double-hurdle model in consumption economics (Blundell and Meghir, 1987; Jones,1989a; Jones and Yen, 2000), where the first hurdle represented the participation decision while the second hurdle represented the corner solution to the consumption decision. It is immediately clear from (10) and (11) that the joint distribution of errors u and v is unknown, and additional information is required to identify the parameters of this likelihood function. However, certain behavioral assumptions can help restrict the dependence between errors u and v, and simply the likelihood, which can facilitate identification. I describe three such behavioral assumptions below.

3.3. Hidden Figures Behind Zeros

There is an additional layer of complexity when one believes that the observed zeros are masking hidden quantities. This usually arises in the context of the first hurdle. Even though the first hurdle represents the participation decision, a researcher may ask about potential consumption decisions by those who never smoke. The main reason why such questions become relevant and modeling the observed outcomes would produce biased results relates to self-selection bias.

Self-selection bias arises through many channels. One way it appears is when the observed outcomes represent a selected population segment, but we want to make an inference about the entire population. For example, suppose our (young adults) smoking data were collected by interviewing young adults in colleges. However, we would like to make inferences about the smoking levels of all young adults in the population, some of whom may not go to college. In such cases, even though one can infer the participation rate among any observably identical group of individuals in the sample, this quantity may not reflect the participation rate for an observably identical group in the population. One reason is that some of the zeros at the first hurdle in the sample may not be zeros at the population level. Consequently, even if one invokes the richer behavioral models in Section 3.2 and applies certain behavioral restrictions, modeling the observed outcome versus the potential outcome beneath the self-selection would require different empirical strategies.

Another way self-selection bias arises is during the evaluation of the effect of a treatment or policy, even when the treatment or policy was allocated in a random or pseudo-random manner in the sample. This is often known as the generalizability issue of randomized experiments.

In the previous two examples, the self-selection bias can affect not just zeros but every outcome level. In contrast, sometimes self-selection occurs even when we can obtain a representative population sample, mainly affecting zeros. Such self-selection bias can arise if outcomes for certain groups of people are censored due to endogenous selection into those groups. For example, one wants to evaluate the effect of a smoking cessation program assigned randomly to a representative group of individuals. However, some individuals reside in counties with stringent public and workplace smoking bans. Compared to counties without such prohibitions, individuals living in the ban-counties must have higher smoking inertia to overcome the shadow price of smoking and smoke the marginal cigarette. Consequently, we expect a larger spike in zeros in the smoking data for individuals residing in the ban counties. This does not invalidate the treatment effect estimate of the cessation policy, as the policy was randomly assigned. However, this treatment effect reflects the impact on the observed outcomes in the context of the current policies in place. Suppose the analyses aimed to estimate the full impact of the cessation policy without the encumbrance of other complementary policies. In that case, we must acknowledge that some of the observed zeros would represent non-zero quantities without these complementary policies.

3.4. Empirical Strategies with Behavioral Restrictions

(1) First-Hurdle Dominance

The dominance restriction implies that individuals always smoke once the first hurdle is passed. Consequently, the second hurdle is irrelevant, and none of the zeros are generated via a consumption decision. This results in a simplification of the likelihood function in (10) and (11):

For data representing zeros:

$$L^0 = \prod\nolimits_0[\Pr(U \leq 0)] \quad = \prod\nolimits_0[\Pr(u \leq -g(D,Z;\ \alpha))] \qquad \textit{Eq. 12}$$

For data representing non-zeros:

$$L^+ = \prod\nolimits_+[\Pr(U > 0) \cdot \mathrm{f}(Y^* |\, P = 1)] = \prod\nolimits_+[\Pr(u > -g(D,Z;\ \alpha)) \cdot \mathrm{f}(Y^* |\, u > -g(D,Z;\ \alpha))] \qquad \textit{Eq. 13}$$

If the goal is to estimate the effect of a covariate on the observed outcomes, then a two-part QL model or a Tobit model can be employed (See Section 3.1).

If the goal is to estimate potential consumption after accounting for certain self-selection behaviors, i.e., what would have the consumption levels of certain non-smokers if a smoking ban had not been in effect, a Heckman selection model can be employed (also known as the "Heckit" model, Heckman 1976, 1979). This model is identical to the standard Tobit model in (7) under the Gaussian distributional assumptions (except that the original Heckman model was conceptualized as a two-step estimator). The only difference is that the Tobit model is used for modeling the corner solution of a consumption decision. In contrast, the Heckit model is used to model participation and consumption decisions, where the latter does not have a corner solution. More importantly, Heckit recovers parameters for the potential consumption decisions if everyone had participated.

It is important to note that the identification in the Heckit model relies entirely on distributional assumptions. One needs sufficient variability of *D* and *X*, independent of the inverse Mills ratio, to consistently identify the regression parameters. Alternatively, the identification in the Heckit model can be much improved if one has an exclusion restriction (e.g., an instrumental variable) in the first stage model for participation (Manning et al. 1987).

(2) Independence

The Independence assumption implies independence between the participation and consumption equations (technically, $u \perp v$). However, unlike the dominance assumption, corner solutions for consumption are allowed beyond participation. Here the likelihood functions in (10) and (11) simplifies to:

For data representing zeros:

$$L^0 = \prod_0 [\Pr(U \leq 0) \cdot \Pr(Y^* \leq 0)] \qquad = \prod_0 [\Pr(u \leq -g(D, Z;\, \alpha)) \cdot \Pr(v > -h(D, X;\, \beta))] \quad \textit{Eq. 14}$$

For data representing non-zeros:

$$L^{+} = \prod_{+}[\Pr(U > 0) \cdot \Pr(Y^{*} > 0) \cdot \mathrm{f}(Y^{*}|Y^{*} > 0)] = \prod_{+}[\Pr(u > -g(D,Z;\ \alpha)) \cdot \Pr(v > -h(D,X;\ \beta) \cdot \mathrm{f}(Y^{*}|v > -h(D,X;\ \beta))] \qquad \textit{Eq. 15}$$

When inferring observed outcomes, one can also use a two-part QL model or a Tobit model (See Section 3.1). However, these models alone cannot distinguish between the participation and consumption zeros from the corner solutions. One needs additional information to decompose the observed zeros into these two parts. For example, if one obtained data on whether individuals have ever smoked, they can use that as a proxy for the participation equation. In such a case, a three-part QL model, or a two-part model, with the participation model being a binary model and the consumption equation being a Tobit model (also known as the Cragg Model), can be used. To an extent, this is a double-hurdle model, but the independence assumption helps to identify the model using the data at hand.

(3) Complete Dominance (Independence and Dominance)

Finally, if one assumes both independence and dominance, the outcomes can be modeled directly using any single equation model, as in (3), and using a TPM. The zeros are deemed true, and no separate potential consumption decisions exist.

3.4. Marginal Effects

One of the challenges of using non-linear and multi-step estimators is calculating the marginal effects of covariates. These are no longer apparent by looking at the coefficients of any one or multiple regressions. Here, we focus on *incremental effects* (ξ) of a binary variable (*D*), which is the difference in the expected value of the target outcome between two levels of *D, marginalized* over the distribution of all other *X* in the model. Similar discussions for computing the average marginal effects for continuous variables can be found in Dow and Norton (2003).

Let the hat ($\widehat{\ }$) on a parameter or a functional of parameters indicate that those parameters have been estimated from the data at hand. Following (3) for a single equation model, an estimator for the incremental effect of *D* is given as:

$$\hat{\xi} = \hat{h}(\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot X) - \hat{h}(\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot X) \qquad \textit{Eq. 16}$$

This approach is also known as the method of re-cycled predictions, where one turns on and off the *D* indicator for everyone in the sample and predicts the expected outcomes for all, takes the sample average of the predictions and calculates a difference.

For TPM, the method of recycled predictions can be used to estimate incremental effects for each part of the model (Belotti et al. 2015). However, to obtain the incremental effects for the overall consumption, one must first compute the predictions for overall consumption for everyone in the sample under the two levels of *D* and then take the difference:

$$\hat{\xi} = \hat{g}(\hat{\alpha}_0 + \hat{\alpha}_1 \cdot 1 + \hat{\alpha}_2 \cdot X) \cdot \hat{h}(\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot X)$$

$$- \hat{g}(\hat{\alpha}_0 + \hat{\alpha}_1 \cdot 0 + \hat{\alpha}_2 \cdot X) \cdot \hat{h}(\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot X) \qquad \textit{Eq. 17}$$

Interestingly, after accounting for participation, the Tobit and the Heckit models follow the same principle in computing the average incremental effects of D on overall (or potential) consumption.

Inference on these incremental effects is based on computing their variances, which can follow standard Taylor series approximations (e.g., Delta method) or non-parametric bootstrap approaches.

4. SELECTED EMPIRICAL APPLICATIONS

There is a large literature demonstrating the application of these methods. I highlight a few applications here. Blundell and Meghir (1987) used the double-hurdle model to estimate the Engel curve for clothing, the first hurdle arises from the observation that purchases of clothing are infrequent and that many households will not record any during the 2 weeks of the survey. Jones (1989a) applied double-hurdle models to study cigarette consumption data from the General Household Survey in the UK and separated participation from consumption decisions. He later used these models to panel data on cigarette consumption (Jones 1989b). Yen and Jones (1996) expand on these models to incorporate the effect of addiction on participation and consumption. Grootendorst (1995) invoked both the dominance and the independence assumptions in modeling healthcare utilization data using a two-part model. Similar models were used by Street et al. (1999) to model pharmaceutical expenditures in Russia, by Laporte et al. (2008) to model healthcare utilization in Canada, and by Parente and Evans (1998) to model medical care use in the US. Maciejewski et al. (2012) used a correlated two-part model on specialty care expenditure data to relax the independence assumption between participation and consumption. However, they did not explicitly assume dominance. Deb et al. (2014) model healthcare expenditure dynamically, allowing for contemporaneous interdependence between the participation and the consumption equations through a copula model. However, they invoke the dominance option and use a single hurdle model, which they implement using a two-part specification. Green et al. (2018) modeled the demand for illegal drugs using a double-inflated double-hurdle model that differentiated between nonparticipants, participant misreporters, and infrequent consumers. They assume independence across these three equations.

Zweifel et al. (1999) modeled healthcare cost data using a two-step Heckman model and concluded that the main demographic driver of healthcare costs was time to death rather than age. Using similar data, Seshamani and Gray (2004) replicated Zweifel et al.'s results and showed that, when using a two-part model, both time to death and age were significant predictors of healthcare costs. These analyses highlight the importance of proper interpretation of results from these models. The Heckman two-step is used to solve a self-selection issue where only those with longer time to death are likely to have non-zero expenditures. On the other hand, the TPM model observed costs without trying to correct for self-selection. Which model is correct would depend on the relevant question at hand. Another example of the use of the Heckman selection model is by Porterfield and DeRigine (2011), which examined whether medical home use affected out-of-pocket expenditures in a special population. They appear to correct for a self-selection mechanism where some families used medical homes but reported zero expenditures in the self-reported data, and the authors intended to uncover the hidden figures behind these zeros. A similar approach was adopted by Wirtz et al. (2012) to model the effect of health insurance in Mexico on self-reported out-of-pocket expenditures on medicines.

5. CONCLUSIONS

Zeros in consumption data have rich economic content. Structural approaches to understanding the data-generating processes for these zeros, especially those driven by behavioral criteria, dictate the choice of econometric modeling and the role of hidden figures behind the zeros.

Applied researchers should articulate the behavioral assumptions made and the target outcome for inference when selecting their econometric approach to modeling these data.

6. REFERENCES

Belotti, F., P. Deb, W.G. Manning, E.C. Norton (2015) “twopm: Two-part models.” *Stata Journal* 15(1):3-20.

Blough, D. K., Madden, C. W. And Hornbrook, M. C. (1999). Modeling Risk Using Generalized Linear Models. *Journal Of Health Economics* **18**, 153–171.

Blundell, R. W. And C. Meghir (1987), 'Bivariate Alternative To The Univariate Tobit Model', Journal Of Econometrics, 33 179-200.

Box, G. E. P. And Cox, D. R. (1964). An Analysis Of Transformations. *Journal Of The Royal Statistical Society* B **26**, 211–252.

Chen, J. and J. Roth. (2023). Log with zeros? Some problems and solutions. *Quarterly Journal of Economics,* In Press.

Deaton, A., and M. Irish. (1984). Statistical models for zero expenditures in household budgets. *Journal of Public Economics* **23**, 59-80.

Deb P, Trivedi Pk, Zimmer Dm. Cost-Offsets Of Prescription Drug Expenditures: Data Analysis Via A Copula-Based Bivariate Dynamic Hurdle Model. Health Econ. 2014 Oct;23(10):1242-59.

Dow, W.H., Norton, E.C. Choosing Between and Interpreting the Heckit and Two-Part Models for Corner Solutions. *Health Services & Outcomes Research Methodology* **4**, 5–18 (2003).

Duan, N. (1983). Smearing Estimate: A Nonparametric Retransformation Method. *Journal Of The American Statistical Association* **78**, 605–610.

Greene W, Harris Mn, Srivastava P, Zhao X. Misreporting And Econometric Modelling Of Zeros In Survey Data On Social Bads: An Application To Cannabis Consumption. Health Econ. 2018 Feb;27(2):372-389.

Grootendorst Pv. A Comparison Of Alternative Models Of Prescription Drug Utilization. Health Econ. 1995 May-Jun;4(3):183-98.

Heckman, J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5: 475–492.

Heckman J.1 979. Sample selection bias as a specification error. Econometrica 47: 153–161.

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

Manning, W. G. (1998). The Logged Dependent Variable, Heteroscedasticity, And The Retransformation Problem. *Journal Of Health Economics* **17**, 283–295.

Manning, W. G. And Mullahy, J. (2001). Estimating Log Models: To Transform Or Not To Transform? *Journal Of Health Economics* **20**, 461–494.

McCullagh, P. (1983). Quasi-Likelihood Functions. *Annals Of Statistics* **11**, 59–67.

McCullagh, P. And Nelder, J. A. (1989). *Generalized Linear Models*, 2nd Edn. London: Chapman And Hall.

Mullahy, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics **12**, 337-350.*

Mullahy, J. (1998). Much Ado About Two: Reconsidering Retransformation And The Two-Part Model In Health Econometrics. *Journal Of Health Economics* **17**, 247–281.

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. *NBER Working Paper* 30735.

Jones, A. M. (1989). A Double-Hurdle Model Of Cigarette Consumption. *Journal Of Applied Econometrics*, *4*(1), 23-39.

Jones AM. The UK Demand For Cigarettes 1954-1986, A Double-Hurdle Approach. J Health Econ. 1989 Mar;8(1):133-41.

Jones, A. M., & Yen, S. T. (2000). A Box-Cox Double-Hurdle Model. *Manchester School*, *68*(2), 203-221.

Jones, A. M. (2000). Health Econometrics. In *Handbook Of Health Economics* (Vol. 1a). Elsevier.

Laporte A, Nauenberg E, Shen L. Aging, Social Capital, And Health Care Utilization In Canada. Health Econ Policy Law. 2008 Oct;3(Pt 4):393-411.

Maciejewski Ml, Liu Cf, Kavee Al, Olsen Mk. How Price Responsive Is The Demand For Specialty Care? Health Econ. 2012 Aug;21(8):902-12.

Seshamani M, Gray A. Ageing And Health-Care Expenditure: The Red Herring Argument Revisited. Health Econ. 2004 Apr;13(4):303-14.

Street, A., Jones, A. M., & Furuta, A. (1999). Cost-Sharing And Pharmaceutical Utilisation And Expenditure In Russia. *Journal Of Health Economics*, *18*(4), 459-472.

Parente St, Evans Wn. Effect Of Low-Income Elderly Insurance Copayment Subsidies. Health Care Financ Rev. 1998 Winter;20(2):19-37.

Porterfield Sl, Derigne L. Medical Home And Out-Of-Pocket Medical Costs For Children With Special Health Care Needs. Pediatrics. 2011 Nov;128(5):892-900.

Tobin, James (1958). "Estimation Of Relationships For Limited Dependent Variables". Econometrica. 26 (1): 24–36.

Wedderburn, R. W. M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, And The Gauss–Newton Method. *Biometrika* **61**, 439–447.

Wirtz Vj, Santa-Ana-Tellez Y, Servan-Mori E, Avila-Burgos L. Heterogeneous Effects Of Health Insurance On Out-Of-Pocket Expenditure On Medicines In Mexico. Value Health. 2012 Jul-Aug;15(5):593-603.

Yen, S. T., & Jones, A. M. (1996). Individual Cigarette Consumption And Addiction: A Flexible Limited Dependent Variable Approach. *Health Economics*, *5*(2), 105-117.

Zweifel P, Felder S, Meier M. Ageing Of Population And Health Care Expenditure: A Red Herring?Healthecon1999;8: 485–496