Causal Inference with Observational Data

6. Nonstandard Dependent Variables

Marc F. Bellemare

May 2018

Introduction

With this chapter, we get to the crux of what used to be call "microeconometrics" in the years before the Credibility Revolution.

That is, we get to those models that allow properly modeling the data-generating process (DGP) for the dependent variables we choose to study.

Introduction

Indeed, in applied microeconomics, it is often the case that the dependent variables we are dealing with are not the nice, continuous variables most basic econometrics is made of.

Rather, we often deal with discrete-choice, limited-dependent, count, or duration data as our dependent variables.

Binary Choice Models

Using a linear regression on those variables will usually lead to there being some issues with what is being estimated.

Take the simplest case, for example, and consider a binary dependent variable, i.e., one that only takes on values equal to zero or one.

For instance, it could be that you are interested in estimating the effect of some treatment variable D on whether an individual votes Democratic ($y = 1$) or Republican ($y = 0$), or it could be that you are interested in the effect of some other treatment variable D on whether a household adopts improved rice seeds ($y = 1$) or not ($y = 0$).

Binary Choice Models

Even in this simple case, the choice of estimator can lead to important issues. Estimating the equation

$$y_i = \alpha + \beta x_i + \gamma D + \epsilon_i \tag{1}$$

by ordinary least squares (OLS) leads to estimating what is called a linear probability model (LPM).

The LPM has many advantages, which we will get to later, but let us begin with the disadvantages, since this is a chapter dedicated to estimators that aim to properly model the DGP for the dependent variable at the expense of everything else.

Binary Choice Models

The critiques commonly leveled at the LPM are as follows:

1. The error term ϵ_i has a Bernoulli structure. That is, $Var(\epsilon_i) = p_i(1 - p_i)$, which means that the variance of the error term is nonconstant, and there is heteroskedasticity.
2. The LPM can yield predicted values $\widehat{y}$ of the dependent variable y that lie outside of the unit interval $[0, 1]$.
3. The LPM imposes linearity on the relationship between y and the variables on the RHS of equation 1.

Binary Choice Models

In order to deal with those problems, two common estimators are used: the logit (the name comes from the fact that a logistic distribution is imposed on the error term), and the probit (which imposes a normal distribution on the error term). Let's begin with the former, since it was developed by an economist.

The key to understanding binary choice models is to start from the idea that the binary variable y we observe is a proxy for a latent variable y^*, for example utility. So we observe $y = 1$ if y^* exceeds a certain threshold, and we observe $y = 0$ if y^* falls below that threshold.

Binary Choice Models

Further, suppose that the latent variable y^* is such that

$$y_i^* = \beta x_i + \epsilon_i. \tag{2}$$

Since we do not observe y_i^*, we can set the aforementioned threshold at zero (recall that utility functions are ordinal, not cardinal) and say that we observe $y_i = 1$ if $y_i^* > 0$ and $y_i = 0$ if $y_i^* \leq 0$.

Binary Choice Models

Since $y_i^* = \beta x_i + \epsilon_i$, this means that if $y_i = 1$, then it's because $\beta x_i + \epsilon_i > 0$, or $\epsilon_i > -\beta x_i$. Alternatively, if $y_i = 0$, then it's because $\beta x_i + \epsilon_i \leq 0$, or $\epsilon_i \leq -\beta x_i$. In the case of the probit, ϵ_i follows a standard normal distribution, i.e., $\epsilon_i \tilde{} N(0, 1)$.

Alternatively, knowing that ϵ_i follows a standard normal, we can write that $\Pr(y_i = 1|x_i) = \Phi(\epsilon_i = 1 - \beta x_i)$, where $\Phi(\cdot)$ denotes the standard normal cdf, and that $\Pr(y_i = 0|x_i) = \Phi(\epsilon_i = -\beta x_i)$. By symmetry of the standard normal cdf, it is trivial to show that $\Pr(y_i = 0|x_i) + \Pr(y_i = 1|x_i) = 1$.

Binary Choice Models

A probit is estimated by maximum likelihood (ML), and its likelihood function is such that

$$\ln L(\beta) = \sum_{i=1}^{N} \left\{ y_i \ln \Phi(1 - \beta x_i) + (1 - y_i) \ln \Phi(-\beta x_i) \right\}, \qquad (3)$$

and equation 3 is maximized by choosing parameter vector β using one of the usual algorithms (e.g., Newton-Raphson, Broyden-Fletcher-Goldfarb-Shanno, Berndt-Hall-Hall-Hausman, Davidson-Fletcher-Powell, etc.)

Binary Choice Models

The logit model is conceptually similar to the probit model, save that it replaces the assumption that the error term follows a standard normal distribution by the assumption that it follows a logistic distribution.

Intuitively, a logistic distribution looks almost exactly like a normal distribution, except that it has slightly fatter tails.

Binary Choice Models

A logit is also estimated by maximum likelihood (ML), and its likelihood function is such that

$$\ln L(\beta) = \sum_{i=1}^{N} \left\{ \begin{array}{c} y_i \ln \left(\frac{\exp[\beta x_i]}{1+\exp[\beta x_i]} \right) \\ +(1-y_i) \ln \left(1 - \frac{\exp[\beta x_i]}{1+\exp[\beta x_i]} \right) \end{array} \right\}, \tag{4}$$

and equation 4 is maximized by choosing parameter vector β using one of the usual algorithms.

Binary Choice Models

Should you estimate an LPM, a probit, or a logit?

Though the choice between probit and logit ultimately depends on your tastes (economists tend to prefer the probit because it was developed by one of our own, other disciplines tend to prefer the logit, but they both give very similar answers once marginal effects are computed), the choice between LPM on the one hand and probit or logit on the other hand hinges upon your research design.

Binary Choice Models

Indeed, there are a few problems with estimating either a probit or a logit:

1. Both can lead to identification by functional form. What this means is that you might have a coefficient turn out significant simply because you have imposed a normal or logistic distribution on the error term when, in fact, that coefficient is not really significant.
2. Both the probit and logit are not good choices with dealing with large numbers of fixed effects, due to something called the incidental parameter problem.
3. The LPM with robust standard errors can properly account for the heteroskedasticity problem that arises when dealing with a binary dependent variable.

Binary Choice Models

4. Unless you are interested in forecasting the value of the dependent variable (e.g., you are forecasting elections), the fact that an LPM will yield predicted probabilities outside of the $[0, 1]$ interval is not an issue. Most of us are interested in estimating a coefficient rather than forecasting the dependent variable.

Binary Choice Models

5. It is not clear that imposing a nonlinear relationship is better than imposing a linear relationship on $E(y|x)$. Angrist and Pischke: "If the conditional expectation function (CEF) is linear, as it is for a saturated model, regression gives the CEF—even for LPM. If the CEF is non-linear, regression approximates the CEF. Usually it does it pretty well. Obviously, the LPM won't give the true marginal effects from the right nonlinear model. But then, the same is true for the 'wrong' nonlinear model! The fact that we have a probit, a logit, and the LPM is just a statement to the fact that we don't know what the "right" model is. Hence, there is a lot to be said for sticking to a linear regression function as compared to a fairly arbitrary choice of a non-linear one! Nonlinearity per se is a red herring."

Binary Choice Models

6. Unless you know the precise form of heteroskedasticity, the ML estimator of the parameter vector β is biased and inconsistent if the errors are heteroskedastic in the case of probit and logit.

Points 1, 2, 3, and 4 all militate in favor of estimating LPMs when you are dealing with a binary dependent variable and observational data. If you have experimental data, or if you have a defensible selection-on-observables design, a probit or logit is more easily justified, but even then, points 3 and 4 still matter, and unless you know there is no heteroskedasticity in your application, it's best to stick with the LPM.

Bellemare, Novak, and Steinmetz (2015)

In this paper, we were interested in studying the persistence of female genital cutting (FGC) in West Africa.

Specifically, we knew whether a woman reported having undergone FGC, and we knew whether she supported the practice, either for society at large or for her own daughter, so we were interested in how much the former explained the latter.

More importantly, we were interested in how much the various levels of variation in the data (individual, household, community, and beyond) explained the variation in support for FGC.

Bellemare, Novak, and Steinmetz (2015)

Both our outcome variable (whether a woman supports FGC) and our variable of interest (whether she has undergone FGC) are binary, and so we estimate linear probability models. We have two primary reasons for doing so:

1. We incorporate lots and lots of fixed effects, and
2. We are interested in the contribution of each level of FE to the total variation in support for FGC.

Bellemare, Novak, and Steinmetz (2015)

Table 3
LPM estimation results for whether respondents think FGC should continue—Benin 2011.

	(1)	(2)	(3)	(4)
Dependent variable: = 1 if respondent thinks FGC should continue, = 0 otherwise.				
Underwent FGC	0.056***	0.058***	0.054***	0.006
	(0.012)	(0.012)	(0.011)	(0.016)
Age	0.001	0.001	0.001	0.000
	(0.001)	(0.001)	(0.001)	(0.004)
Age squared	−0.000	−0.000	−0.000	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Primary education	−0.006	−0.006*	−0.006	−0.014
	(0.004)	(0.004)	(0.004)	(0.021)
Secondary education	0.000	0.000	−0.000	−0.001
	(0.005)	(0.005)	(0.005)	(0.016)
Higher education	0.010	0.007	0.010	0.010
	(0.014)	(0.013)	(0.014)	(0.031)
Married	−0.002	−0.002	−0.000	0.004
	(0.005)	(0.005)	(0.005)	(0.016)
Cohabitating	0.017*	0.016*	0.016	0.007
	(0.009)	(0.009)	(0.010)	(0.032)
Widowed	0.014	0.013	0.021	0.052
	(0.014)	(0.014)	(0.014)	(0.059)
Divorced	0.019	0.018	0.012	0.005
	(0.018)	(0.018)	(0.018)	(0.017)
Separated	−0.005	−0.006	−0.002	−0.025
	(0.006)	(0.006)	(0.007)	(0.039)
Television	−0.001	−0.001	−0.002	
	(0.004)	(0.004)	(0.004)	
Radio	−0.004	−0.004	0.001	
	(0.004)	(0.004)	(0.005)	
Electricity	0.004	0.004	0.008*	
	(0.004)	(0.004)	(0.004)	
Urban household	−0.001	−0.001		
	(0.004)	(0.004)		
Constant	−0.028	−0.037*	−0.018	0.040
	(0.026)	(0.022)	(0.018)	(0.072)
Observations	10,477	10,477	10,477	10,477
R-squared	0.402	0.405	0.477	0.915
Interviewer fixed effects	Yes	Yes	Yes	Yes
Ethnicity fixed effects	Yes	Yes	Yes	Yes
Religion fixed effects	Yes	Yes	Yes	Yes
Region fixed effects	No	Yes	Yes	Yes
Village fixed effects	No	No	Yes	Yes
Household fixed effects	No	No	No	Yes

Robust standard errors in parentheses, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Bellemare, Novak, and Steinmetz (2015)

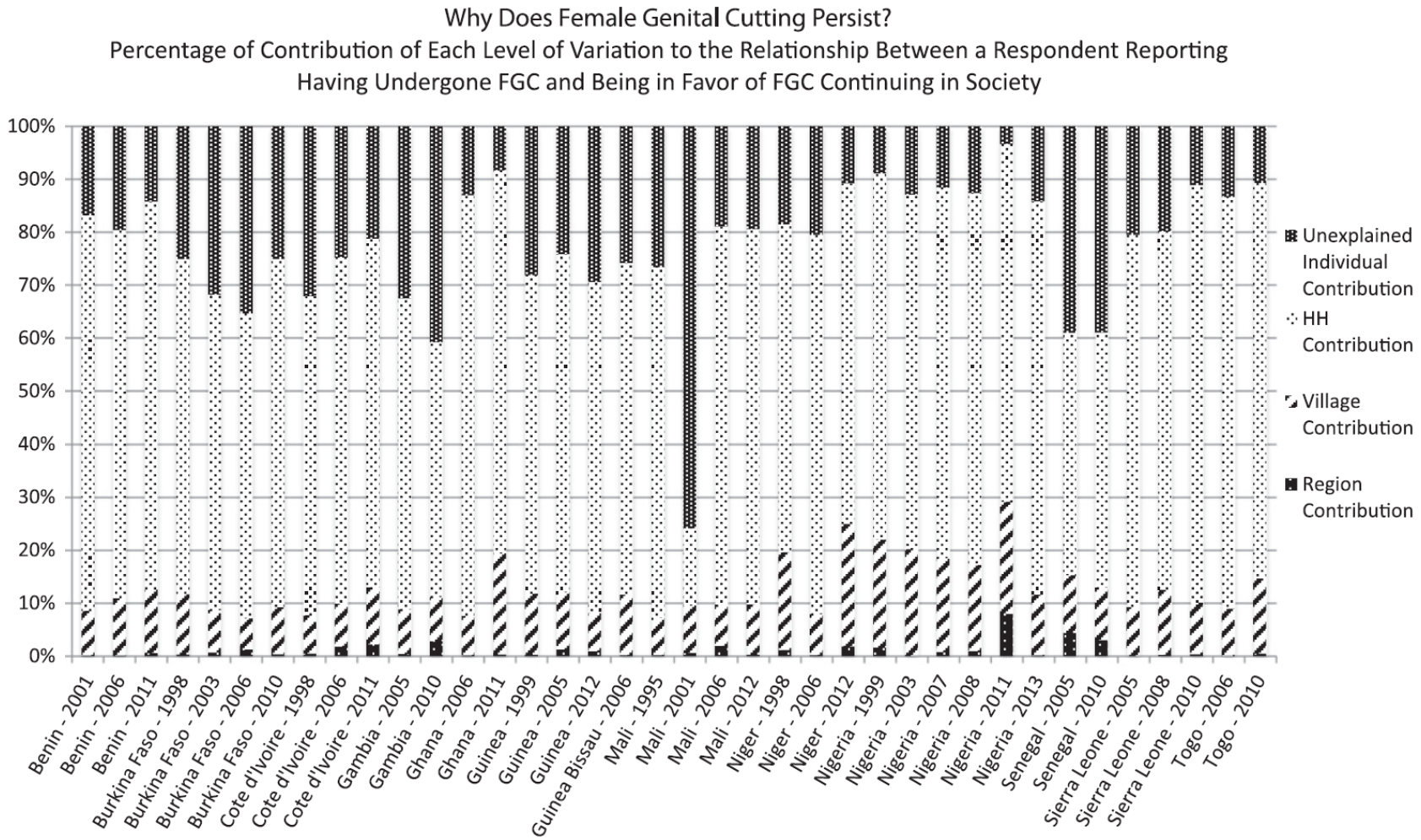

Fig. 1. Percentage contribution of each level of variation to the persistence of FGC in West Africa.

Binary Choice Models

The are several variants to the probit and logit. Here are some, discussed in intuitive terms.

Bivariate probit, with or without selection. A bivariate probit is a system of equations, one for each of two binary dependent variables y_1 and y_2. A bivariate probit with selection is such that y_2 is observed if and only if $y_1 = 1$. The two equations are estimated simultaneously by full-information maximum likelihood (FIML; see Rivers and Vuong, 1988), which imposes that their error terms ϵ_1 and ϵ_2 follow a bivariate normal, such that $\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \tilde{} NB(0, 0, 1, 1, \rho)$, where ρ is the correlation between the two error terms. This suggests that a test of the null hypothesis of error independence $H_0 : \rho = 0$ can be used to determine whether one should estimate a bivariate probit or two separate probits.

Binary Choice Models

Semiparametric binary choice models. Manski (1975) developed a semiparametric estimator for binary choice variables called the maximum score estimator—semiparametric because although it involves estimating some parameters, it makes no distributional assumption. This is only very rarely used anymore, as it does not add much to our understanding, and it takes a while to converge.

Binary Choice Models

Ordered probit. When the dependent variable only takes a finite number of ordered, categorical values (e.g., when people are asked if they agree strongly, agree, neither agree nor disagree, disagree, or disagree strongly, or when they have to rate something on a one- to five-star basis), we can adapt the apparatus of the probit and estimate an ordered probit.

Binary Choice Models

Multinomial logit. When the dependent variable takes a finite number of categorical but non-ordered values (e.g., whether an individual walks, bikes, drives, or takes the bus to work), we can study their choice by using multinomial models. Here, the logit is almost always used because a multinomial probit integrates over a multidimensional normal distribution, which has no closed-form solution unlike the equivalent logit. One drawback is that the logit imposes the assumption that irrelevant alternatives are independent. For example, that an individual's choice to walk, bike, drive, or take the bus to work is unaffected by the addition of an extra choice, e.g., skiing to work.

Limited Dependent Variables

A variable y is said to be truncated if we don't observe the values above or below a certain threshold, or both; a variable y is said to be censored if its value is only partially known.

In both cases, we talk of a limited dependent variable—limited in the sense that it only takes a certain range of values.

Limited Dependent Variables

Again, assume we do not observe a latent variable y_i^*. Rather, we observe $y_i = y_i^*$ if $y_i^* > 0$, and we observe $y_i = 0$ if $y_i^* \leq 0$. Again, we are interested in estimating

$$y_i^* = \beta x_i + \epsilon_i. \tag{5}$$

To properly account for the truncated or censored nature of the DGP, here, too, we have to rely on nonlinear procedures.

Limited Dependent Variables

A popular such procedure is the tobit model, named after the article in which James Tobin (1958) explored the properties of the estimator. The likelihood function for a tobit is such that

$$\ln L(\beta) = \sum_{i=1}^{N} \left\{ \begin{array}{c} I(y_i^* > 0) \ln \left(\frac{1}{\sigma} \phi \left(\frac{y_i - \beta x_i}{\sigma} \right) \right) \\ + [1 - I(y_i^* > 0)] \ln \left(1 - \Phi \left(\frac{\beta x_i}{\sigma} \right) \right) \end{array} \right\}, \quad (6)$$

where ϕ denotes a standard normal pdf, Φ denotes a standard normal cdf, and since $\epsilon_i \tilde{} N(0, \sigma^2)$, σ is the standard deviation of the error term in equation 5.

Limited Dependent Variables

Type I tobit. Note that censoring need not occur at zero, as we laid out above. It could be the case that there is censoring for values of y_i^* below y_L, for values of y_i^* above y_U, or both, such that we only observe values of y_i^* in the $[y_L, y_U]$ interval. Most statistical packages will allow you to specify your own threshold(s).

Limited Dependent Variables

Type II tobit. This type of tobit allows for there to be selection. That is, it allows for different sets of covariates to affect first whether we observe y_i (i.e., whether $y_i^* > 0$ and $y_i = y_i^*$, or whether $y_i^* \leq 0$ and $y_i = 0$) and second the extent of $y_i = y_i^*$ conditional on $y_i^* > 0$ and y_i being observed.

In Bellemare and Barrett (2006), for example, we estimate two such models—one net purchases, and one for net sales—which we append to the end choices of an ordered probit model for whether a household is a net buyer, autarkic, or a net seller.

Limited Dependent Variables

Type III tobit. This is simply a bivariate tobit.

Limited Dependent Variables

As with the probit and logit, the tobit does not lend itself well to fixed effects, nor does it lend itself to the use of robust standard errors. In addition, estimating a tobit can also lead to identification via functional form.

As such, unless you have experimental data or a defensible selection-on-observables design, it is perhaps best to live with a small amount of bias from not properly taking into account the censoring or truncation, and to stick to the linear model.

Count Data

Count data models are useful when the dependent variable you are dealing with takes only non-negative integer values.

The two most common models in such cases is the Poisson model and the negative binomial model, which is a version of the Poisson model that accounts for overdispersion in the dependent variable.

But note that entire books have been written about count data models; see Cameron and Trivedi (1998) for a complete treatment.

Count Data

Indeed, one of the (extremely) restrictive assumptions which the Poisson model makes is that the mean of the distribution of y and its variance are equal.

In cases where that is not true (almost always, this is because $\sigma^2 > \mu$ rather than the other way around, and there is overdispersion), then the estimation procedure has to substitute a distribution that can account for this departure from the Poisson assumption.

Count Data

Starting from the assumption that the mean λ of a predicted Poisson distribution is such that $\lambda = \exp\{\beta x_i\}$, the log-likelihood function for a Poisson regression is such that

$$\ln L(\beta) = \sum_{i=1}^{N} \{y_i \beta x_i - \exp\{\beta x_i\} - \ln(y_i!)\} . \tag{7}$$

This log-likelihood function can be adapted to the case where y is distributed negative binomial instead of Poisson, and usually, statistical packages will test the hypothesis that there is neither over- nor under-dispersion when estimating a Poisson model, so that you can determine whether a negative binomial would be better.

Count Data

A separate class of Poisson and negative binomial models are zero-inflated Poisson and zero-inflated negative binomial models.

As their name indicates, those are to be used when you are interested in modeling the selection going into strictly positive values of y.

That is, the first stage accounts for whether y is positive or zero, and the second stage accounts for how much of y there is conditional on $y > 0$. As with all selection models, this often requires heroic assumptions.

Duration Data

Duration models refer to models used to study time durations.

Those models are often referred to by the name survival analysis—a name which comes from the biomedical sciences, since their interest is to study patient survival in response to certain shocks or procedures. In this case, too, there have been entire books dedicated to those models; see Lancaster (1992) for a complete if dated treatment.

Duration Data

Duration data present their own set of challenges because they are conceptually similar to censored data. Sometimes, we do not observe when a certain episode of something has begun; at other times, we do not observe when it has ended; and at yet other times, we observe neither its beginning nor its end.

In this class, I will focus on three estimators—one nonparametric, and two parametric.

Duration Data

The nonparametric way to deal with duration data is by using a Kaplan-Meier estimator, which graphs the frequency of the sample which "survives" at any given time.

Here is an example of a Kaplan-Meier estimator of the duration of the hungry season experienced by a sample of households in rural Madagascar, from Bellemare and Novak (2017).

Bellemare and Novak (2017)

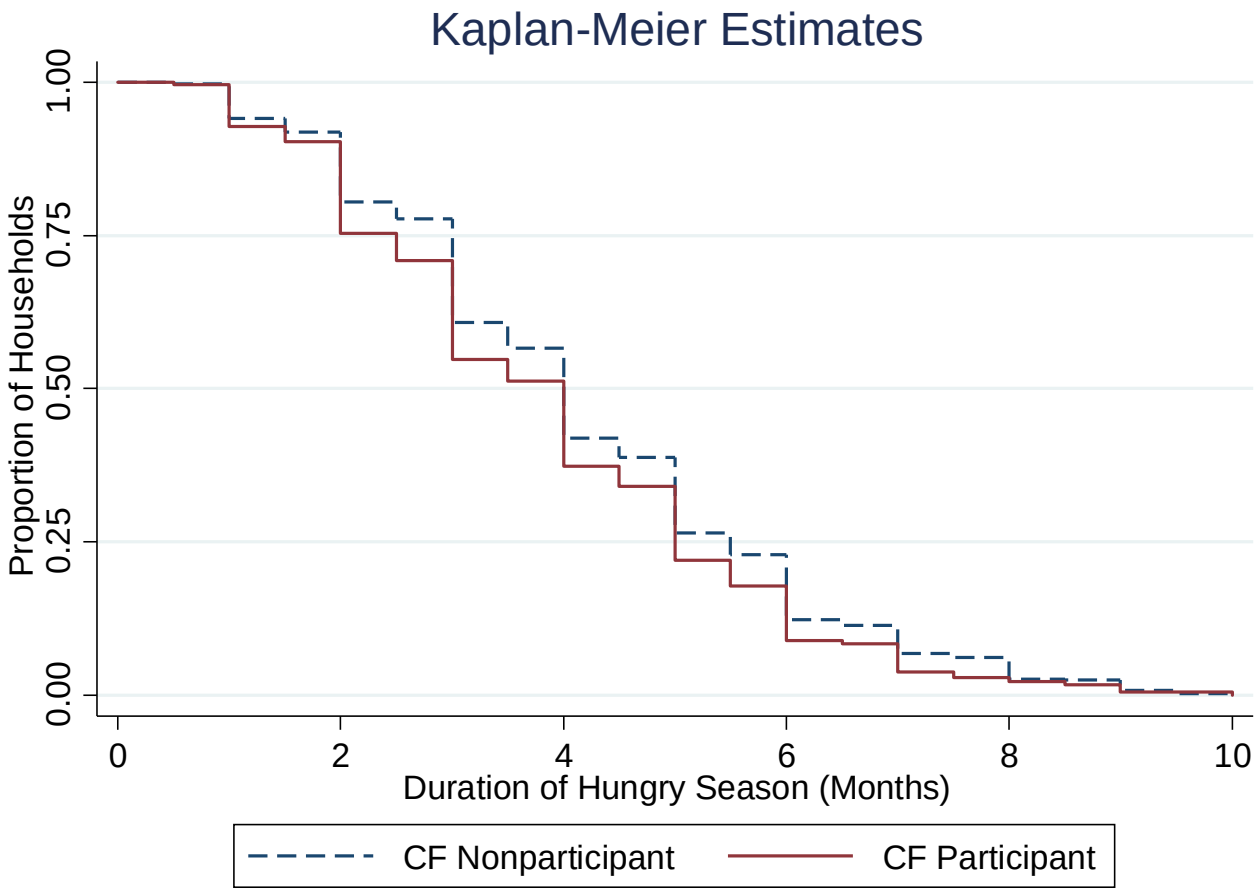

Figure: An example of a Kaplan-Meier estimator in practice. (Source: Bellemare and Novak, 2016.)

Duration Data

Note that in the figure, because the hungry season is measured in months and everyone experiences it, everyone's start date is set equal to zero, and at the beginning, everyone is in the sample.

Progressively, however, people exit the hungry season, up until the point where no one experiences it anymore, around 10 months.

From the figure, it also looks as though those households that participate in contract farming exit the hungry season at a faster rate than those that do not—a finding that was confirmed by our parametric analysis.

Duration Data

The Kaplan-Meier estimator does not control for any confounding factor and, as a such, it is a purely descriptive tool.

To control for confounding factors, these notes present two choices: (i) survival time regressions and (ii) Cox proportional hazards models.

Both models have their advantages and disadvantages, but note that their coefficients should be interpreted in the opposite way one would interpret an OLS coefficient.

Duration Data

That is, whereas regressing a duration on some covariates would tell you by how much each covariate reduces that duration (if its effect is well identified), the same coefficient in a survival time or Cox proportional hazards model would tell you by how much the likelihood of exiting some condition (in the Bellemare and Novak, 2017 example, the condition is the hungry season) changes.

As such, we can learn useful things from each of OLS and duration or survival models.

Bellemare and Novak (2017)

Table 2a. Estimation Results for OLS, Cox Proportional Hazard, and Survival-time Regressions Omitting WTP Variables

Variables Dependent Variable: Duration of Hungry Season	OLS	Cox	Survival Time
Contract farming participant	−0.294** (0.142)	0.150** (0.062)	0.171** (0.070)

Bellemare and Novak (2017)

Table 2b. Estimation Results for OLS, Cox Proportional Hazard, and Survival-time Regressions Including WTP Variables

Variables Dependent Variable: Duration of Hungry Season	OLS	Cox	Survival Time
Contract farming participant	−0.277* (0.145)	0.166*** (0.063)	0.188*** (0.071)

Duration Data

As always, the issue is that you should only really estimate survival time or Cox proportional hazards models if you have good identification or a selection-on-observables design.

In Bellemare and Novak (2017), we had the latter, and so we supplemented our linear regressions with duration analyses, but it is rare that you can do so.

The only other occasion I would encourage you to do it is if you have experimental data.

Summary

- There exist estimators to account for the discrete, limited-dependent, count, or duration nature of various dependent variables.
- But like the various moves a karateka is taught throughout her training, you are taught those estimators so that you do not have to use them.

Summary

- The reason why you should not use those estimators is that in the absence of experimental data, it is hard to know where the identification is coming from.
- In many cases, identification is entirely possible because of the functional form you are imposing on the error term or the dependent variable, and this is not the kind of identification we want.

Summary

- In other cases, specific estimators do not lend themselves to using fixed effects, to incorporating instrumental variables, to dealing with heteroskedasticity, or all of those at once.
- Almost always, the estimators just discussed are used to avoid some amount of bias due to the nonstandard nature of the dependent variable, but the assumptions they rely on leave the door open for more bias. It is better to take care of identification issues before you properly account for the DGP behind your dependent variable.

Summary

- That said, you will commonly be asked by people who should know better why you haven't used a probit or a logit instead of the LPM you judiciously used. You should always have the fancier, likelihood-based model in an appendix, if only to satisfy the curiosity of your reviewers.

Summary

- Still, you can sometimes combine bits and pieces of likelihood functions to account for increasingly complex decisions, and some journals will be interested. This is especially so if your new estimator can study phenomena that are different than the one you are studying. In Bellemare and Barrett (2006), for example, we developed an ordered tobit estimator to study the marketing behavior of pastoralist households in Ethiopia and Kenya, and thus we made a dual contribution to the literature.