Robust-to-endogenous-selection estimators for two-part models, hurdle models, and zero-inflated models

David M. Drukker

Executive Director of Econometrics
Stata

Italian Stata User Group Meeting
15 November 2018

What's this talk about?

- Two-part models, hurdle models, and zero-inflated models are frequently used in applied research

What's this talk about?

- Two-part models, hurdle models, and zero-inflated models are frequently used in applied research
- This talk shows that they all have a surprising robustness property
  - The are robust to endogeneity

What's this talk about?

- Two-part models, hurdle models, and zero-inflated models are frequently used in applied research
- This talk shows that they all have a surprising robustness property
    - The are robust to endogeneity
- Robustness makes estimation much easier
    - No instrument needed

- Many outcomes of interest have mass points on a boundary and are smoothly distributed over a large interior set
  - Hours worked has a mass point at zero and is smoothly distributed over strictly positive values
  - Expenditures on health care, Deb and Norton (2018)

- Many outcomes of interest have mass points on a boundary and are smoothly distributed over a large interior set
  - Hours worked has a mass point at zero and is smoothly distributed over strictly positive values
  - Expenditures on health care, Deb and Norton (2018)
- Three models (or approaches) arose to account for the apparent difference between the distribution of the outcome at the boundary and over the interior
  - Two-part models: Duan, Manning, Morris, and Newhouse (1983), Duan, Manning, Morris, and Newhouse (1984)
  - Hurdle models: Cragg (1971) and Mullahy (1986)
  - Zero-inflated (With-Zeros) models:Mullahy (1986) and Lambert (1992)
  - Standard tools: see Cameron and Trivedi (2005), Winkelmann (2008), and Wooldridge (2010)

Zero-lower-limit models

- The cannonical case is the zero-lower-limit model, $y \geq 0$

$$y = s(\mathbf{x}, \epsilon) G(\mathbf{x}, \eta)$$

  where

  - $\mathbf{x}$ are observed covariates
  - ϵ and η are random disturbances
  - $s(\mathbf{x}, \epsilon) \in \{0, 1\}$ is the selection process,
  - $G(\mathbf{x}, \eta)$ is the the main process
- When $G(\mathbf{x}, \eta) > 0$ we have two-part model or a hurdle model
- When $G(\mathbf{x}, \eta) \geq 0$ we have zero-inflated (or with zeros) model

Two-part models and Hurdle models

$$y = s(\mathbf{x}, \epsilon)G(\mathbf{x}, \eta)$$

- The two-part model was motivated as a flexible model for $\mathbf{E}[y|\mathbf{x}]$
  - It allowed the zeros to come from a different process than the one that generates the outcome over the interior values
- Hurdle models were motivated by the idea of observing a zero until a hurdle was crossed

Zero-inflated/With-zeros models

$$y = s(\mathbf{x}, \epsilon) G(\mathbf{x}, \eta)$$

- Zero-inflated and with-zeros models were motivated by a mixture process
  - $G(\mathbf{x}, \eta) \geq 0$ contributes some of the zeros
  - But there are too many zeros in the data to be explained by the distribution assumed for $G(\mathbf{x}, \eta)$
  - So we observe either a zero or $G(\mathbf{x}, \eta) \geq 0$ with probability determined by $s(\mathbf{x}, \epsilon)$

Value table

Table: $y = s(\mathbf{x}, \epsilon)G(\mathbf{x}, \eta)$ value table

	$G(\mathbf{x}, \eta) = 0$	$G(\mathbf{x}, \eta) > 0$
$s(\mathbf{x}, \epsilon) = 0$	0	0
$s(\mathbf{x}, \epsilon) = 1$	0	$G(\mathbf{x}, \eta)$

- TPMs and HMs only include the right-hand column in which $G(\mathbf{x}, \eta) > 0$
- ZIMs include both columns, because $G(\mathbf{x}, \eta) \geq 0$

Endogeneity?

$$y = s(\mathbf{x}, \epsilon) G(\mathbf{x}, \eta)$$

- If ϵ and η are correlated, there is an endogeneity problem

Endogeneity?

$$y = s(\mathbf{x}, \epsilon) G(\mathbf{x}, \eta)$$

- If ϵ and η are correlated, there is an endogeneity problem
- The original proposers of the TPM claimed that the TPM was robust to endogeneity, but this was rejected by most econometricians
  - The claim of robustness led to the cake debates (Hay and Olsen (1984), Duan et al. (1984))
    This debate went nowhere, because the debate was over whether one log-likelihood was a special case of another
    Wrong way to settle an identification debate

Endogeneity?

$$y = s(\mathbf{x}, \epsilon)G(\mathbf{x}, \eta)$$

- If ϵ and η are correlated, there is an endogeneity problem
- The original proposers of the TPM claimed that the TPM was robust to endogeneity, but this was rejected by most econometricians
  - The claim of robustness led to the cake debates (Hay and Olsen (1984), Duan et al. (1984))
    This debate went nowhere, because the debate was over whether one log-likelihood was a special case of another
    Wrong way to settle an identification debate
  - Section 17.6 of Wooldridge (2010) is representative of the modern position
    He assumes that exogeneity is required and derives an estimator for the case of endogeneity

TPMs and HMs are robust

- Both TPMs and HMs restrict $G(\mathbf{x}, \eta) > 0$, so only the right-hand column of values for y is possible.

TPMs and HMs are robust

- Both TPMs and HMs restrict $G(\mathbf{x}, \eta) > 0$, so only the right-hand column of values for y is possible.
- Drukker (2017) used iterated expectations to show that $\mathbf{E}[y|\mathbf{x}]$ is identified when $s()$ and $G()$ are not mean independent, after conditioning on $\mathbf{x}$.

TPMs and HMs are robust

- Both TPMs and HMs restrict $G(\mathbf{x},\eta) > 0$, so only the right-hand column of values for y is possible.
- Drukker (2017) used iterated expectations to show that $\mathbf{E}[y|\mathbf{x}]$ is identified when $s()$ and $G()$ are not mean independent, after conditioning on $\mathbf{x}$.

$$\begin{aligned}
\mathbf{E}[y|\mathbf{x}] &= \mathbf{E}[s(\mathbf{x},\epsilon)G(\mathbf{x},\eta)|\mathbf{x}] \\
&= \mathbf{E}[s(\mathbf{x},\epsilon)G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=0]\mathbf{Pr}[s(\mathbf{x},\epsilon)=0|\mathbf{x}] \\
&\qquad + \mathbf{E}[s(\mathbf{x},\epsilon)G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]\mathbf{Pr}[s(\mathbf{x},\epsilon)=1|\mathbf{x}] \\
&= \mathbf{E}[0\ G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=0]\mathbf{Pr}[s(\mathbf{x},\epsilon)=0|\mathbf{x}] \\
&\qquad + \mathbf{E}[1\ G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]\mathbf{Pr}[s(\mathbf{x},\epsilon)=1|\mathbf{x}] \\
&= \mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]\mathbf{Pr}[s(\mathbf{x},\epsilon)=1|\mathbf{x}] \qquad (1)
\end{aligned}$$

Estimable robust TPMs and HMs

$$\mathbf{E}[y|\mathbf{x}] = \mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]\ \ \mathbf{Pr}[s(\mathbf{x},\epsilon)=1|\mathbf{x}]$$

Estimable robust TPMs and HMs

$$\mathbf{E}[y|\mathbf{x}] = \mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]\ \ \mathbf{Pr}[s(\mathbf{x},\epsilon)=1|\mathbf{x}]$$

- The data on y nonparametrically identify $\mathbf{Pr}[s(\mathbf{x},\epsilon)=1]$ and $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]$
  - $\mathbf{Pr}[s(\mathbf{x},\epsilon)=1]$:
    When $y=0$, $s(\mathbf{x},\epsilon)=0$
    When $y>0$, $s(\mathbf{x},\epsilon)=1$
  - $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]$:
    When $y>0$, $s(\mathbf{x},\epsilon)=1$ and $y=G(\mathbf{x},\eta)$,
    $\mathbf{E}[y|\mathbf{x},s=1]=\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]$

Estimable robust TPMs and HMs

$$\mathbf{E}[y|\mathbf{x}] = \mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ \ s(\mathbf{x},\epsilon) = 1]\mathbf{Pr}[s(\mathbf{x},\epsilon) = 1|\mathbf{x}]$$

- No exclusion restriction is required to identify $\mathbf{E}[y|\mathbf{x}]$.

Estimable robust TPMs and HMs

$$\mathbf{E}[y|\mathbf{x}] = \mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]\mathbf{Pr}[s(\mathbf{x},\epsilon)=1|\mathbf{x}]$$

- No exclusion restriction is required to identify $\mathbf{E}[y|\mathbf{x}]$.
- Can recover DGP parameters in $s(\mathbf{x},\epsilon)$

Estimable robust TPMs and HMs

$$\mathbf{E}[y|\mathbf{x}] = \mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]\mathbf{Pr}[s(\mathbf{x},\epsilon)=1|\mathbf{x}]$$

- No exclusion restriction is required to identify $\mathbf{E}[y|\mathbf{x}]$.
- Can recover DGP parameters in $s(\mathbf{x},\epsilon)$
- Cannot recover DGP parameters in $G(\mathbf{x},\eta)$, estimate parameters of misspecified model
    - Trade off:
      Estimate $\mathbf{E}[y|\mathbf{x}]$ without an exclusion restriction in exchange for not estimating DGP parameters in $G(\mathbf{x},\eta)$

Estimable robust TPMs and HMs

$$\mathbf{E}[y|\mathbf{x}] = \mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=1]\mathbf{Pr}[s(\mathbf{x},\epsilon)=1|\mathbf{x}]$$

- No exclusion restriction is required to identify $\mathbf{E}[y|\mathbf{x}]$.
- Can recover DGP parameters in $s(\mathbf{x},\epsilon)$
- Cannot recover DGP parameters in $G(\mathbf{x},\eta)$, estimate parameters of misspecified model
    - Trade off:
      Estimate $\mathbf{E}[y|\mathbf{x}]$ without an exclusion restriction in exchange for not estimating DGP parameters in $G(\mathbf{x},\eta)$
- Inference about $\mathbf{E}[y|\mathbf{x}]$ is causal

Why is it robust?

- The feature of the derivation that is essential to this robustness result is that $\mathbf{E}[G(\mathbf{x}, \eta)|\mathbf{x},\ s(\mathbf{x}, \epsilon) = 0]$ is not needed to compute $\mathbf{E}[y|\mathbf{x}]$

Why is it robust?

- The feature of the derivation that is essential to this robustness result is that $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=0]$ is not needed to compute $\mathbf{E}[y|\mathbf{x}]$
- This result is analogous to the robustness result for estimating the averge treatment effect conditional on the treated
  - $\mathbf{E}[y_{1i}|t_i=1]-\mathbf{E}[y_{0i}|t_i=1]$
    Only need conditional mean independence for $\mathbf{E}[y_{0i}|t_i=1]$

Why is it robust?

- The feature of the derivation that is essential to this robustness result is that $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=0]$ is not needed to compute $\mathbf{E}[y|\mathbf{x}]$
- This result is analogous to the robustness result for estimating the averge treatment effect conditional on the treated
    - $\mathbf{E}[y_{1i}|t_i=1]-\mathbf{E}[y_{0i}|t_i=1]$
      Only need conditional mean independence for $\mathbf{E}[y_{0i}|t_i=1]$
- The data on y do not nonparametrically identify $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=0]$
    - If $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=0]$ was required, we would need to impose functional form assumptions to identify it

Why is it robust? (Continued)

- $\mathbf{E}[G(\mathbf{x}, \eta) | \mathbf{x},\ s(\mathbf{x}, \epsilon) = 0]$ is not needed because the boundary values are actual outcome values and not just indicators for censoring

Why is it robust? (Continued)

- $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=0]$ is not needed because the boundary values are actual outcome values and not just indicators for censoring
- If the observations indicated censoring instead of being actual outcome values, we could not model y as the product of $s(\mathbf{x},\epsilon)$ and $G(\mathbf{x},\eta)$ as

$$y = s(\mathbf{x},\epsilon)G(\mathbf{x},\eta)$$

Why is it robust? (Continued)

- $\mathbf{E}[G(\mathbf{x}, \eta)|\mathbf{x},\ s(\mathbf{x}, \epsilon) = 0]$ is not needed because the boundary values are actual outcome values and not just indicators for censoring
- If the observations indicated censoring instead of being actual outcome values, we could not model y as the product of $s(\mathbf{x}, \epsilon)$ and $G(\mathbf{x}, \eta)$ as

$$y = s(\mathbf{x}, \epsilon)G(\mathbf{x}, \eta)$$

- This discussion formally justifies the assertation of Duan, Manning, Morris, and Newhouse (1983) and Duan, Manning, Morris, and Newhouse (1984) that the TPM is robust because it models the observed data

Why is it robust? (Continued)

- $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x},\ s(\mathbf{x},\epsilon)=0]$ is not needed because the boundary values are actual outcome values and not just indicators for censoring
- If the observations indicated censoring instead of being actual outcome values, we could not model y as the product of $s(\mathbf{x},\epsilon)$ and $G(\mathbf{x},\eta)$ as
$$y = s(\mathbf{x},\epsilon)G(\mathbf{x},\eta)$$
- This discussion formally justifies the assertation of Duan, Manning, Morris, and Newhouse (1983) and Duan, Manning, Morris, and Newhouse (1984) that the TPM is robust because it models the observed data
- Essentialy, Drukker (2017) ended the "cake debate" by showing that the TPM is robust.

More identification results

- I have formal identification results for
  - Zero-lower-limit ZIMs under endogneity
  - Two-limit TPMs/HMs under endogneity
  - Two-limit ZIMs under endogneity
- For time, concentrate on cake-debate version of zero-lower-limit TPM.

Cake-debate model

The cake-debate model disccussed in Duan et al. (1984), Hay and Olsen (1984), and section 17.6.3 of Wooldridge (2010) is

$$s(\mathbf{x}, \epsilon) = \begin{cases} 1 & \text{if } \mathbf{x}\boldsymbol{\gamma} + \epsilon > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$G(\mathbf{x}, \eta) = \exp(\mathbf{x}\boldsymbol{\beta} + \eta) \tag{3}$$

$$y = s(\mathbf{x}, \epsilon)G(\mathbf{x}, \eta) \tag{4}$$

$$\begin{pmatrix} \epsilon \\ \eta \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_\eta \\ \rho\sigma_\eta & \sigma_\eta^2 \end{pmatrix}\right) \tag{5}$$

A robust TPM estimator for cake-debate model

A TPM estimator for the parameters of the cake-debate model proceeds by

1. Estimating γ from a probit model of s on $\mathbf{x}$
   - This functional form is justified by the normality of ϵ

A robust TPM estimator for cake-debate model

A TPM estimator for the parameters of the cake-debate model proceeds by

1. Estimating γ from a probit model of s on $\mathbf{x}$
   - This functional form is justified by the normality of ϵ
2. Estimating $\tilde{\beta}$ by a quasi maximum likelihood estimator of a poisson model of y on $\mathbf{x}$ conditional on $s = 1$
   - This functional form takes more work, but I justify it below
   - Note that $\tilde{\beta}$ differs from β
     The endogeneity causes the estimable parameters to differ from the data-generating process parameters
     The estimable parameters are exactly the parameters that we need to estimate $\mathbf{E}[y|\mathbf{x}]$

A robust TPM estimator for cake-debate model

A TPM estimator for the parameters of the cake-debate model proceeds by

1. Estimating γ from a probit model of s on $\mathbf{x}$
   - This functional form is justified by the normality of ϵ
2. Estimating $\tilde{\beta}$ by a quasi maximum likelihood estimator of a poisson model of y on $\mathbf{x}$ conditional on $s = 1$
   - This functional form takes more work, but I justify it below
   - Note that $\tilde{\beta}$ differs from β
     The endogeneity causes the estimable parameters to differ from the data-generating process parameters
     The estimable parameters are exactly the parameters that we need to estimate $\mathbf{E}[y|\mathbf{x}]$
3. Estimating $\mathbf{E}[y|\mathbf{x}]$ by $\Phi(\mathbf{x}\widehat{\gamma})\exp(\mathbf{x}\widehat{\tilde{\beta}} + (\mathbf{x}\widehat{\gamma})^2\widehat{\alpha}_1 + (\mathbf{x}\widehat{\gamma})^3\widehat{\alpha}_2)$

Justifying the cake-debate functional form

- Recall that we need to estimate $\mathbf{E}[G(\mathbf{x}, \eta)|\mathbf{x}\ s(\mathbf{x}, \epsilon) = 1]$ which is the same as $\mathbf{E}[y|\mathbf{x}\ s(\mathbf{x}, \epsilon) = 1]$, because $y = G()$ when $s() = 1$

Justifying the cake-debate functional form

- Recall that we need to estimate $\mathbf{E}[G(\mathbf{x}, \eta)|\mathbf{x}\ s(\mathbf{x}, \epsilon) = 1]$ which is the same as $\mathbf{E}[y|\mathbf{x}\ s(\mathbf{x}, \epsilon) = 1]$, because $y = G()$ when $s() = 1$
- Given the exponential mean model for $G()$ in the cake-debate model, the TPM is going to use an exponential mean for $G()$ conditional on $s() = 1$

Justifying the cake-debate functional form

- Recall that we need to estimate $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x}\ s(\mathbf{x},\epsilon)=1]$ which is the same as $\mathbf{E}[y|\mathbf{x}\ s(\mathbf{x},\epsilon)=1]$, because $y=G()$ when $s()=1$
- Given the exponential mean model for $G()$ in the cake-debate model, the TPM is going to use an exponential mean for $G()$ conditional on $s()=1$
- Given the structure of the model, do there exist $\tilde{\beta}$ for which $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x}\ s(\mathbf{x},\epsilon)=1]=\exp(\mathbf{x}\tilde{\beta})$?
  Yes, sort of

Justifying the cake-debate functional form

- Recall that we need to estimate $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x}\ s(\mathbf{x},\epsilon)=1]$ which is the same as $\mathbf{E}[y|\mathbf{x}\ s(\mathbf{x},\epsilon)=1]$, because $y=G()$ when $s()=1$
- Given the exponential mean model for $G()$ in the cake-debate model, the TPM is going to use an exponential mean for $G()$ conditional on $s()=1$
- Given the structure of the model, do there exist $\tilde{\boldsymbol{\beta}}$ for which $\mathbf{E}[G(\mathbf{x},\eta)|\mathbf{x}\ s(\mathbf{x},\epsilon)=1]=\exp(\mathbf{x}\tilde{\boldsymbol{\beta}})$?
  Yes, sort of
- In an appendix, I show that

$$\mathbf{E}[\exp(\mathbf{x}\boldsymbol{\beta}+\eta)|\mathbf{x},\ \epsilon>-\mathbf{x}\gamma]=\exp(\mathbf{x}\boldsymbol{\beta}+\tilde{q})$$

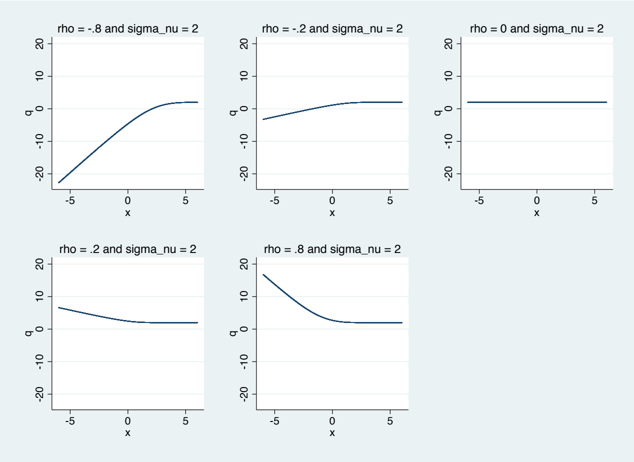
where

$$\tilde{q}=\sigma_\nu^2/2+\ln\left\{\frac{\Phi[(\rho\sigma_\nu+\mathbf{x}\gamma)]}{[1-\Phi(-\mathbf{x}\gamma)]}\right\}$$

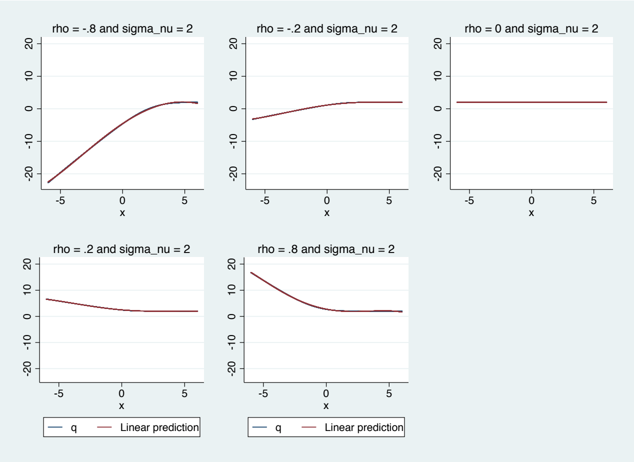
- Plots of

$$\ln\left\{\frac{\Phi[(\rho\sigma_\nu + x)]}{[1-\Phi(-x)]}\right\}$$

for values of ρ and σ_ν



- Plots of correction terms and predicted values from third-order polynomial in x



Example : cakep

```
. cakep expend ages phealth
Iteration 0:   GMM criterion Q(b) =  2.381e-21
Iteration 1:   GMM criterion Q(b) =  1.290e-32
Cake model                                     Number of obs      =      2,000
Selection model: Probit                          Equal to zero      =        946
 Interior model: Poisson                         Greater than zero =      1,054
```

expend	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
select						
ages	.4843445	.0616662	7.85	0.000	.363481	.6052081
phealth	-.32653	.0483122	-6.76	0.000	-.4212202	-.2318399
_cons	.0537728	.035187	1.53	0.126	-.0151923	.122738
interior						
ages	.5183393	.1932158	2.68	0.007	.1396432	.8970354
phealth	.7858247	.1460173	5.38	0.000	.4996361	1.072013
_cons	.4459145	.0919501	4.85	0.000	.2656957	.6261333
poly2						
_cons	1.071851	.7394328	1.45	0.147	-.3774107	2.521113
poly3						
_cons	-1.413192	1.905859	-0.74	0.458	-5.148607	2.322222

Example : cakep

```
. cakep expend ages phealth, polyorder(2)
Iteration 0:   GMM criterion Q(b) =  2.228e-21
Iteration 1:   GMM criterion Q(b) =  3.444e-33
Cake model                                        Number of obs       =      2,000
Selection model: Probit                             Equal to zero     =        946
 Interior model: Poisson                            Greater than zero =      1,054
```

expend	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
select						
ages	.4843445	.0616662	7.85	0.000	.363481	.6052081
phealth	-.32653	.0483122	-6.76	0.000	-.4212202	-.2318399
_cons	.0537728	.035187	1.53	0.126	-.0151923	.122738
interior						
ages	.3901893	.1167197	3.34	0.001	.1614229	.6189557
phealth	.8792678	.1028623	8.55	0.000	.6776613	1.080874
_cons	.4476793	.0915416	4.89	0.000	.2682611	.6270974
poly2						
_cons	.8301923	.5688684	1.46	0.144	-.2847693	1.945154

Monte Carlo with discrete covariates

- A Monte Carlo simulation evaluates the estimation and inference properties of an estimator in finite samples

Monte Carlo with discrete covariates

- A Monte Carlo simulation evaluates the estimation and inference properties of an estimator in finite samples
    - In parametric models, this usually involves comparing point estimates against DGP parameters

Monte Carlo with discrete covariates

- A Monte Carlo simulation evaluates the estimation and inference properties of an estimator in finite samples
  - In parametric models, this usually involves comparing point estimates against DGP parameters
  - The object of interest in a TPM is $\mathbf{E}[y|\mathbf{x}]$, or counter-factual changes in $\mathbf{E}[y|\mathbf{x}]$

Monte Carlo with discrete covariates

- A Monte Carlo simulation evaluates the estimation and inference properties of an estimator in finite samples
  - In parametric models, this usually involves comparing point estimates against DGP parameters
  - The object of interest in a TPM is $\mathbf{E}[y|\mathbf{x}]$, or counter-factual changes in $\mathbf{E}[y|\mathbf{x}]$
  - So the place to start evaluating a TPM estimator is its performance for $\mathbf{E}[y|\mathbf{x}]$

Monte Carlo with discrete covariates

- A Monte Carlo simulation evaluates the estimation and inference properties of an estimator in finite samples
  - In parametric models, this usually involves comparing point estimates against DGP parameters
  - The object of interest in a TPM is $\mathbf{E}[y|\mathbf{x}]$, or counter-factual changes in $\mathbf{E}[y|\mathbf{x}]$
  - So the place to start evaluating a TPM estimator is its performance for $\mathbf{E}[y|\mathbf{x}]$
  - The trick to doing this evaluation is to generate the data using discrete covariates and compare the TPM estimator's estimates of $\mathbf{E}[y|\mathbf{x}]$ with the nonparametric cell-mean estimates (NP estimates)

MC for cakep with discrete

```
. use cake_simd_v2

. summarize cm_1* cm_2* cm_3*, sep(4)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
cm_1_t	2,000	.6099153	0	.6099153	.6099153
cm_1_b	2,000	.6070482	.0726858	.3848774	.8592353
cm_1_se	2,000	.0709065	.0128669	.0428758	.2142889
cm_1_r	2,000	.0645	.2457029	0	1
cm_2_t	2,000	.8341332	0	.8341332	.8341332
cm_2_b	2,000	.8331135	.0825487	.5642096	1.168678
cm_2_se	2,000	.0794897	.0129875	.0498566	.1683961
cm_2_r	2,000	.0635	.2439211	0	1
cm_3_t	2,000	1.119697	0	1.119697	1.119697
cm_3_b	2,000	1.116043	.1235469	.7047904	1.58126
cm_3_se	2,000	.1219146	.0236314	.0673789	.2991421
cm_3_r	2,000	.067	.2500845	0	1

MC for cakep with discrete

```
. summarize cm_4* cm_5* cm_6*, sep(4)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
cm_4_t	2,000	.977028	0	.977028	.977028
cm_4_b	2,000	.9748809	.084304	.6899576	1.343455
cm_4_se	2,000	.084854	.012212	.0552322	.1796997
cm_4_r	2,000	.0505	.2190291	0	1
cm_5_t	2,000	1.382903	0	1.382903	1.382903
cm_5_b	2,000	1.385858	.0805017	1.170033	1.704497
cm_5_se	2,000	.0792886	.0096752	.0607276	.1607804
cm_5_r	2,000	.062	.2412159	0	1
cm_6_t	2,000	1.923175	0	1.923175	1.923175
cm_6_b	2,000	1.939031	.1599157	1.449924	2.505669
cm_6_se	2,000	.1559157	.0278615	.0955437	.3776995
cm_6_r	2,000	.055	.2280373	0	1

MC for cakep with discrete

```
. summarize cm_7* cm_8* cm_9*, sep(4)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
cm_7_t	2,000	1.257671	0	1.257671	1.257671
cm_7_b	2,000	1.255832	.1074974	.9523147	1.693319
cm_7_se	2,000	.1073283	.0195462	.0692952	.2948076
cm_7_r	2,000	.057	.2319006	0	1
cm_8_t	2,000	1.810889	0	1.810889	1.810889
cm_8_b	2,000	1.810946	.124859	1.447053	2.279219
cm_8_se	2,000	.1228508	.0170666	.0881146	.2383234
cm_8_r	2,000	.0555	.2290109	0	1
cm_9_t	2,000	2.568117	0	2.568117	2.568117
cm_9_b	2,000	2.577586	.195092	1.954408	3.511249
cm_9_se	2,000	.1890343	.0292786	.1280508	.4601372
cm_9_r	2,000	.0565	.2309425	0	1

DGP details

1. The two discrete covariates were generated from two correlated normal random variables
2. The selection process is generated from

$$s = x_1\gamma_1 + x_2\gamma_2 + \epsilon > 0$$

where ϵ is a standard normal.

DGP details

1. The main process G is generated as a Gamma random variable with parameters

$$a = exp(x_1\beta_{a1} + x_2\beta_{a2} + \beta_{a0} + .5\eta)$$

$$b = exp(x_1\beta_{b1} + x_2\beta_{b2} + \beta_{b0} + .5\eta)$$

η is a normal random variable that is correlated with ϵ
The mean of G conditional on x_1, x_2, and η is

$$\exp[x_1(\beta_{a1} + \beta_{b1}) + x_2(\beta_{a2} + \beta_{b2}) + (\beta_{a0} + \beta_{b0}) + \eta]$$

The mean of $G()$ has a functional form covered the cake-debate TPM, but it is not Poisson

What coming up?

- Extend `cakep` to handle other TPMs and HMs
    - Rename it when it does more that cake models
- Extend command that currently does TPM version of fractional models
- Extend command that currently does zero-inflated poisson models to other ZIMs
- Write command that for fractional ZIMs

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Cragg, J. G. 1971. Some Statistical Models for Limited Dependent Variables with Applications to the Demand for Durable Goods. *Econometrica* 39(5): 829–844.

Deb, P., and E. C. Norton. 2018. Modeling Health Care Expenditures and Use. *Annual Review of Public Health* 39], pages = 489505.

Drukker, D. M. 2017. Two-part models are robust to endogenous selection. *Economics Letters* 152: 71–72.

Duan, N., W. Manning, C. N. Morris, and J. P. Newhouse. 1983. A Comparison of Alternative Models for the Demand for Medical Care. *Journal of Business and Economic Statistics* 1(2): 115–126.

Duan, N., W. G. Manning, C. N. Morris, and J. P. Newhouse. 1984. Choosing between the Sample-Selection Model and the Multi-part Model. *Journal of Business and Economic Statistics* 2: 283–289.

Hay, J. W., and R. J. Olsen. 1984. Let Them Eat Cake: A Note on Comparing Alternative models for the Demand of Medical Care. *Journal of Business and Economic Statistics* 2(3): 279–282.

Lambert, D. 1992. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics* 34(1): 1–14.

Mullahy, J. 1986. Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics* 33: 341365.

Winkelmann, R. 2008. *Econometric Analysis of Count Data*. 5th ed. Springer.

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