Regression for nonnegative skewed dependent variables

Austin Nichols

July 15, 2010

Introduction

Nonnegative skewed outcomes y, e.g.

- labor earnings
- medical expenditures
- trade volume

often modeled using a regression of $\ln(y)$ on X. What about $y = 0$?

Model of the conditional mean

Linear regression of $\ln(y)$ on X assumes

$$E[\ln(y)|X] = Xb$$

but the Poisson quasi-MLE (Gourieroux et al. 1984) or GLM with a log link assumes

$$\ln(E[y|X]) = Xb$$

Only one of these makes sense when y can be zero.

Note that the conditional mean must always be positive, but the actual realized outcome can be zero. GLM with a log link can even accommodate negative outcomes (but `poisson` exits with an error).

When does OLS make sense?

If we write

$$y_i = \exp(X_i b + e_i) = \exp(X_i b) v_i$$

and if we happen to have data where $y_i > 0$ for all i, then we can take logs for

$$\ln(y_i) = X_i b + e_i$$

which motivates the OLS specification. With $y > 0$ always, Manning and Mullahy (2001) provide guidance on when to prefer OLS or GLM (if e is symmetric and homoskedastic, prefer OLS).

Tobit typically not a good alternative

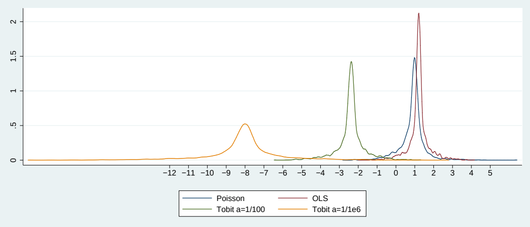
Other common approaches include `tobit` and "two-part" or "hurdle" models. One `tobit` approach puts a small number a for every zero (smaller than the smallest observed positive y), takes logs, and then specifies $\ln(a)$ as the lower limit. See Cameron and Trivedi (2009, p.532), §16.4.2 "Setting the censoring point for data in logs," for one example of this advice.

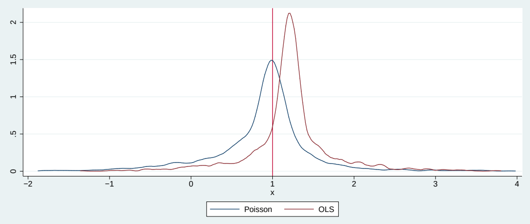
But this approach makes no sense. The choice of a is arbitrary, and affects the estimation. Choosing $a = .01$ results in $\widetilde{lny} = -4.6$ and choosing $a = .000001$ results in $\widetilde{lny} = -13.8$ and there is no obvious reason to prefer one over the other, for example when the smallest positive y is 1.

The only time replacing zero with a small positive number a, taking logs, and running a tobit makes sense is when zero represents the result of a known lower detection limit, or rounding, and y is known to actually be positive in these cases. This is not the case in practice, typically.

Comparison of OLS and Tobit

Graph comparing OLS, Poisson, and Tobit (with *a* equal to one hundredth or one millionth)





The simulation model

We specify a data generating process given by

$$y_i = \exp(X_i b) v_i$$

with v distributed gamma with moderate or no heteroskedasticity. Choose $x = exp(u)$ with u uniform on $(0, 1)$ for moderate skewness in the predictor.

Also tried mixture of gamma, exponential, pareto, mixture of lognormals.

Poisson tended to dominate in every case.

Objects of interest

We are usually interested not in estimating b, but in the marginal effect

$$\frac{\partial E(y|X)}{\partial X}$$

which is straightforward in the Poisson case, and not in the others. Or we might be interested in predictions, or out of sample predictions. Poisson tends to dominate in these cases as well, and sidesteps the pernicious retransformation problem of OLS (Duan 1983, Manning 1988, Mullahy 1998, Ai and Norton 2000, Santos Silva and Tenreyro 2006).

Whatever we are interested in estimating, we are presumably looking to minimize the MSE of that—so looking for a consistent estimator of $\widehat{y}$ (as in Duan 1983) when we are interested in individual predictions (not the mean of predictions in a large sample) makes no sense—we want good small sample performance.

Marginal Effects

Table: MSE of marginal effect estimates (in percentage terms: $\frac{\partial E(y|X)}{\partial X}\frac{1}{E(y|X)}$)

			No Het.			Low Het.	
Variance		N=100	N=1000	N=10000	N=100	N=1000	N=10000
Low	% nonzero	0.005	0.005	0.005	0.314	0.313	0.312
	OLS	0.062	0.006	0.001	0.352	0.029	0.007
	Poisson	0.050	0.005	0.000	0.405	0.055	0.005
	Tobit	0.799	0.604	0.588	148.919	152.241	148.315
	Hurdle (2PM)	0.765	0.588	0.572	13.252	11.259	10.812
Med.	% nonzero	0.111	0.111	0.111	0.601	0.596	0.597
	OLS	0.148	0.014	0.001	1.003	0.120	0.048
	Poisson	0.139	0.013	0.001	1.342	0.142	0.015
	Tobit	8.810	6.893	6.655	153.898	235.285	229.831
	Hurdle (2PM)	7.317	5.961	5.786	52.625	36.228	33.169
High	% nonzero	0.397	0.397	0.397	0.805	0.802	0.802
	OLS	0.312	0.031	0.003	1.791	0.357	0.156
	Poisson	0.377	0.037	0.004	2.136	0.362	0.037
	Tobit	22.270	8.411	6.797	161.239	92.491	90.506
	Hurdle (2PM)	28.004	20.213	19.243	61.426	40.132	39.633

Predictions

Table: MSE of predictions

			No Het.			Low Het.	
Variance		N=100	N=1000	N=10000	N=100	N=1000	N=10000
Low	% nonzero	0.006	0.005	0.005	0.314	0.313	0.312
	OLS	7.785	8.177	8.098	48.063	75.440	68.899
	Poisson	6.472	6.936	6.875	44.839	71.849	65.649
	Tobit	6.604	6.948	6.877	50.427	77.735	71.049
	Hurdle (2PM)	6.580	6.948	6.876	45.952	72.798	66.345
Med.	% nonzero	0.112	0.112	0.111	0.601	0.596	0.597
	OLS	20.244	21.357	21.634	126.013	162.236	179.508
	Poisson	17.327	18.507	18.776	118.111	159.339	176.848
	Tobit	18.390	19.267	19.519	131.283	166.499	183.631
	Hurdle (2PM)	17.682	18.531	18.780	122.786	160.258	177.462
High	% nonzero	0.403	0.397	0.397	0.805	0.802	0.802
	OLS	45.523	58.396	53.134	481.218	444.892	488.549
	Poisson	41.744	54.808	49.852	335.368	442.150	486.921
	Tobit	48.053	61.223	55.865	351.362	451.000	494.182
	Hurdle (2PM)	42.736	54.926	49.861	372.344	443.862	487.583

Hurdle Models

"Hurdle" or "two-part" models (2PM), described by Mullahy (1986) among others, appear in the prior comparison. Why are they popular? Due to the RAND Health Insurance Experiment (Duan et al. 1983, Manning et al. 1987, Newhouse et al. 1993), primarily.

Idea is: a person decides whether to go to the doctor, and then the doctor decides expenditure conditional on $y > 0$. Also easy to run—likelihood is separable, so just run a `probit` (or `logit` or `cloglog` or what have you) using $\mathbf{1}(y > 0)$ as a dummy outcome, then run OLS regression of $\ln(y)$ on X or a truncated regression (`ztp` or `ztnb` or `truncreg`) of y on X. See McDowell (2003) but replace commands with those appropriate in newer Stata.

Two-part assumption

Not all that realistic in reality-you may find yourself getting medical care without any decision on your part; you can also end your medical care if you decide to (in most cases).

Now we need several pieces of the model to be correctly specified, or all estimates are inconsistent.

Also hard to include endogenous explanatory variables in a hurdle model without some unpleasantly strong ML assumptions. Not so with Poisson/GLM: simply adopt a GMM framework.

GMM framework easily accommodates instruments

GMM version of Poisson assumes:

$$\frac{y_i}{\exp(X_i b)} - 1$$

is orthogonal to X_i (uncorrelated in the population, or dgp). If X is endogenous, we can instead assume it is orthogonal to Z where Z is a set of instruments:

$$E\left[\left(\frac{y_i}{\exp(X_i b)} - 1\right)' Z\right] = 0$$

`ivpois` for Stata 10, on SSC, `gmm` in Stata 11.
Manual entry on `gmm` has many examples.

The RAND HIE

Suppose we want to measure the effect of a one percent reduction in the price of health care on health expenditures. In health plans, prices fall as expenditures increase, so regressing spending on price is a bad idea.

In the RAND Health Insurance Experiment (HIE), participants were randomly assigned first-dollar prices; not prices more generally, because every case had a stoploss.

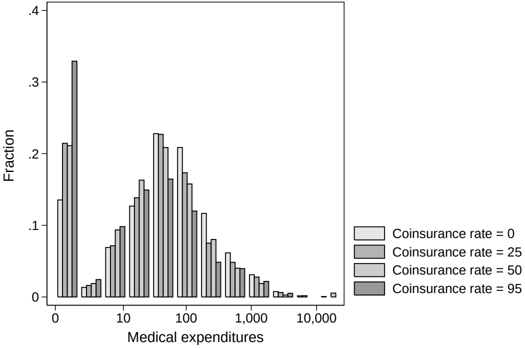
HIE price structure

Expected prices

The price changes during the course of the year; in fact, in the RAND HIE the price is the first dollar price up until the stoploss and then drops to zero; but the shadow price of a bit more health care also has to take into account the chance that you want a lot more later in the year, and spending now lowers the effective price of care later in the year.

Ellis (1986) shows that using expected end-of-year price as a proxy for the actual marginal price (at each point during the plan year) performs very well. But the expected end-of-year price is endogenously determined by spending behavior. I compute expected price over all other individuals in an individual's randomly assigned group and use first dollar price as an instrument for the expected price.

Graph comparing expenditures by first-dollar price



Results

Table: Regressions of medical spending on prices

	(1) Poisson	(2) Poisson using Ep	(3) Poisson using lnEp	(4) GMM-IV using Ep	(5) GMM-IV using lnEp
FDP 25	-0.181 (-1.48)				
FDP 50	0.164 (0.42)				
FDP 95	-0.492 (-3.71)				
Expected price		-0.426 (-2.29)		-0.515 (-3.23)	
ln(Expected price)			-0.153 (-1.37)		-0.167 (-1.65)
Good health	0.366 (1.98)	0.365 (1.98)	0.352 (1.18)	0.318 (2.29)	0.439 (2.44)
Fair health	0.675 (3.93)	0.674 (3.95)	0.854 (3.20)	0.580 (3.03)	0.739 (2.76)
Poor health	1.330 (4.92)	1.345 (4.97)	0.723 (2.33)	1.055 (4.62)	0.626 (2.49)
Child	-0.0799 (-0.29)	-0.0769 (-0.28)	-0.257 (-0.82)	-0.147 (-0.67)	-0.0148 (-0.05)
Female child	-0.365 (-0.86)	-0.366 (-0.87)	0.184 (0.34)	-0.608 (-2.57)	-0.441 (-1.57)
Female	0.425 (3.27)	0.424 (3.27)	0.439 (1.96)	0.448 (3.94)	0.505 (2.94)
Black	-0.671 (-3.82)	-0.690 (-3.80)	-0.615 (-2.16)	-0.519 (-3.23)	-0.503 (-2.26)
Age	0.0105 (2.14)	0.0106 (2.16)	0.0141 (1.68)	0.0134 (2.88)	0.0192 (2.57)
Constant	4.572 (19.66)	4.572 (20.03)	4.071 (9.85)	4.505 (23.18)	3.743 (11.33)
Observations	4146	4146	2277	4146	2277

t statistics in parentheses

Conclusions

"Use a model that could possibly fit your data" seems like simple and obvious advice, and has been offered many times before, sometimes forcefully (e.g. Mullahy 1988, Santos Silva and Tenreyro 2006), but still has not permeated the awareness of many researchers. See e.g.

- Rutledge (2009) regresses ln spending on X, dropping zeros! GLM or GMM is the better alternative.
- Kowalski (2009) compares her method to ivtobit instead of a more reasonable GMM.

These are both common errors, and easily avoided.

There are many other models, zero-inflated or not, for nonnegative outcomes, but few have the robustness of Poisson. Note in particular we need no assumption about conditional variance for consistency, contrary to occasional claims about Poisson.

Practical Guidance

For a specific application, you should run your own simulation. You can run several candidate models on half the data, and see the MSE of the quantity of interest (the other half of the data serves for out of sample predictions), or resample errors to simulate new data in which to estimate (with known coefficients and marginal effects). If you choose half-sample cross-validation, it is easy to run 100 times or so, and get very reliable estimates of MSE for half-samples.

GLM or the equivalent `poisson`, both with a log link, will often "win" this contest.

Note: If you decide on a log link, you may want to call your model "GLM with a log link," rather than a "Poisson" QMLE—some older reviewers believe Poisson regression is only for counts.

Ai, Chunrong and Edward C. Norton. 2000. Standard errors for the retransformation problem with heteroscedasticity. *Journal of Health Economics,* 19(5):697–718

Cameron, A. Colin and Pravin K. Trivedi. 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge.

Cameron, A. Colin and Pravin K. Trivedi. 1991. The role of income and health risk in the choice of health insurance: Evidence from Australia. *Journal of Public Economics,* 45(1): 1–28.

Cameron, A. Colin and Pravin K. Trivedi. 2009. Microeconometrics Using Stata. Stata Press, College Station TX.

Duan, Naihua. 1983. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association,* 78, 605-610.

Duan, Naihua, Willard G. Manning, Carl N. Morris, and Joseph P. Newhouse. 1983. A comparison of alternative models for the demand for medical care. *Journal of Business and Economics Statistics,* 1(2):115-126.

Ellis, Randall P. 1986 "Rational Behavior in the Presence of Coverage Ceilings and Deductibles." *The RAND Journal of Economics,* 17(2): 158–175.

Gourieroux, C., Montfort, A., Trognon, A., 1984. Pseudo-maximum likelihood methods: applications to Poisson models. *Econometrica,* 52, 701-720.

Kowalski, Amanda E. 2009. "Censored Quantile Instrumental Variable Estimates of the Price Elasticity of Expenditure on Medical Care." NBER Working Paper 15085. `http://www.nber.org/papers/w15085`

Manning, Willard G. and John Mullahy. 2001. "Estimating Log Models: To Transform Or Not To Transform?" *Journal of Health Economics,* 20(4): 461–494.

Manning, Willard G. 1998. "The logged dependent variable, heteroscedasticity, and the retransformation problem." *Journal of Health Economics,* 17, 283-295.

Manning, Willard G., Joseph P. Newhouse, Naihua Duan, Emmett B. Keeler, Arleen Liebowitz, and M. Susan Marquis. 1987. "Health insurance and the demand for medical care: evidence from a randomized experiment." *American Economic Review,* 77(3): 251-277.

McDowell, Allen. 2003. "From the help desk: hurdle models." *Stata Journal,* 3(2): 178–184.
http://www.stata-journal.com/sjpdf.html?articlenum=st0040

Mullahy, J., 1986. "Specification and testing of some modified count data models." *Journal of Econometrics,* 33(3):341–365.

Mullahy, J., 1998. "Much ado about two: reconsidering retransformation and the two-part model in health econometrics." *Journal of Health Economics,* 17, 247–281.

Newhouse, Joseph P. and the Insurance Experiment Group. 1993. *Free for all? Lessons from the RAND Health Insurance Experiment.* Harvard University Press, Cambridge.

Rutledge, Matthew S. 2009. "Asymmetric Information and the Generosity of Employer-Sponsored Health Insurance." University of Michigan Working [Job Market] Paper.

Santos Silva, João M. C. and Silvana Tenreyro. 2006. "'The Log of Gravity." *Review of Economics and Statistics,* 88(4): 641–658.

Wooldridge, J.M., 1991. "On the application of robust, regression-based diagnostics to models of conditional means and conditional variances." *Journal of Econometrics,* 47, 5-46.

Wooldridge, J.M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press. Available from Stata.com.