

Models for Count Outcomes, Part II

These notes borrow heavily (sometimes verbatim) from Long 1997, Regression Models for Categorical and Limited Dependent Variables, and Long & Freese, 2003 Regression Models for Categorical Dependent Variables Using Stata, Revised Edition, and also the 2006 2nd edition of Long & Freese.

Models for Truncated Counts

Sometimes observations with outcomes equal to zero are missing from the sample because of the way the data are collected. For example, we may not have a list of every Sociologist; we only have a list of those who have published at least one article. Or, a survey of how often people visit the shopping mall may be done of people who are currently at the mall. Or, if you have bought a TV, the warranty card may ask you how many other TVs you have. In each case, observations with a value of 0 are not included in the sample. Zero-truncated count models are designed for such situations.

Long & Freese (2006) go through the math on pp. 382-383. A key thing to note is that the adverse effects of over-dispersion are worse with truncated models. Estimates are biased and inefficient if there is overdispersion. You should estimate a zero-truncated negative binomial model to test for overdispersion.

The `ztp` (zero truncated Poisson) and `ztnb` (zero truncated negative binomial) commands can be used. Output is similar to the `poisson` and `nbreg` commands. If zero counts are missing from your data because of the way the data were collected, and zero counts are generated by the same process as positive counts, interpretation is also similar.

Hurdle Models

Sometimes you may believe that zeros are generated by a different process from that of positive counts. Zero is a “hurdle” that you have to get past before reaching positive counts (but everyone has a nonzero probability of doing so). Hurdle regression models combine a binary model (e.g. logit) to predict zeros with a zero-truncated Poisson or zero-truncated negative binomial model to predict nonzero counts.

Long & Freese (2006, pp. 387-393) show how to estimate hurdle models, even though there is no “official” Stata command for doing so. After the book went to press, Hilbe released several user-written commands for estimating different types of hurdle models (`findit hurdle`). These include `hpllogit` (poisson hurdle model) and `hnblogit`. Compare the following results with those reported by Long & Freese (2006) on p. 389:

```
. use "D:\Soc73994\Long2003\couart2.dta", clear
. hnblogit art fem mar kid5 phd ment, nolog
```

```
Negative Binomial-Logit Hurdle Regression      Number of obs   =      915
                                                Wald chi2(5)    =      49.77
Log likelihood = -1552.5966                    Prob > chi2     =      0.0000
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	

logit						
fem	-.2511511	.1591052	-1.58	0.114	-.5629916	.0606894
mar	.3262336	.1808182	1.80	0.071	-.0281637	.6806308
kid5	-.2852487	.1111304	-2.57	0.010	-.5030603	-.0674371
phd	.0222194	.0795571	0.28	0.780	-.1337097	.1781485
ment	.0801214	.0130181	6.15	0.000	.0546064	.1056363
_cons	.236796	.2955189	0.80	0.423	-.3424104	.8160024

negbinomial						
fem	-.2446712	.0972181	-2.52	0.012	-.4352153	-.0541272
mar	.1034172	.1094297	0.95	0.345	-.1110611	.3178955
kid5	-.1532593	.0722291	-2.12	0.034	-.2948257	-.011693
phd	-.0029336	.0480673	-0.06	0.951	-.0971438	.0912766
ment	.0237382	.0042868	5.54	0.000	.0153362	.0321402
_cons	.3551248	.1968307	1.80	0.071	-.0306564	.740906

/lnalpha	-.6034757	.2249916	-2.68	0.007	-1.044451	-.1625003

```
AIC Statistic =      3.407
```

The logit equation tells you what affects the likelihood of clearing the zero “hurdle.” Women, and those with kids under 5, are less likely to clear the hurdle. Married people, those who went to more prestigious PHD institutions, and those whose mentors are more productive are more likely to clear the hurdle. (Not all effects are significant though.)

The negbinomial part of the model, the coefficients indicate whether increases in the variable increase or decrease productivity. A variable can be significant in one part of the model, but not in the other part.

Long & Freese (2006) show how to get the predicted probabilities of different counts, e.g. 0 articles, one article, etc. I am not sure how to get these predictions using Hilbe’s commands, so if you want them you may have to use Long and Freese’s rather lengthy procedure.

Zero-Inflated Count Models

Zero-inflated models assume that there are two latent groups. One group has no chance of going beyond zero, e.g. they might be scientists in fields or companies that do not allow publishing. We call this Group A, the Always Zero Group. Members of the other group may have a zero count, but the probability of having a positive count is nonzero, e.g. a scientist who could publish may or may not do so. We call this Group –A, the Not Always Zero Group. Zero-Inflated models allow for this possibility, thereby increasing the conditional variance and the probability of zero counts.

Estimating such models is a 3-step process. First, you model membership into the latent groups. Then, you model the counts for those in Group –A (Not Always Zero). Finally, you compute observed probabilities as a mixture of the probabilities for the two groups.

The commands are `zip` and `zinb`. They include an `inflate` option. The vars specified in the `inflate` option are used to predict group membership.

Long and Freese give examples and show how to make interpretation of results easier.

Comparisons of Count Models

Long & Freese’s `countfit` command makes it easy to compare the results of PRM, NBRM, ZIP, and ZINB models.

```
. countfit art fem mar kid5 phd ment, inflate(ment fem)
```

Variable	PRM	NBRM	ZIP	ZINB
art				
Gender: 1=female 0=male	0.799	0.805	0.812	0.836
	-4.11	-2.98	-3.31	-2.40
Married: 1=yes 0=no	1.168	1.162	1.142	1.150
	2.53	1.83	2.01	1.72
Number of children < 6	0.831	0.838	0.849	0.845
	-4.61	-3.32	-3.77	-3.22
PhD prestige	1.013	1.015	0.993	1.001
	0.49	0.42	-0.24	0.04
Article by mentor in last 3 yrs	1.026	1.030	1.018	1.025
	12.73	8.38	8.09	7.07
Constant	1.356	1.292	1.874	1.465
	2.96	1.85	5.54	2.69
lnalpha				
Constant		0.442		0.375
		-6.81		-7.06
inflate				
Article by mentor in last 3 yrs			0.876	0.470
			-3.23	-2.55
Gender: 1=female 0=male			1.120	2.868
			0.42	1.40
Constant			0.484	0.275
			-3.15	-2.14
Statistics				
alpha		0.442		
N	915.000	915.000	915.000	915.000
ll	-1651.056	-1560.958	-1605.644	-1552.034
bic	3343.026	3169.649	3272.659	3172.257
aic	3314.113	3135.917	3229.288	3124.068

legend: b/t

Comparison of Mean Observed and Predicted Count

Model	Maximum Difference	At Value	Mean Diff
PRM	0.091	0	0.026
NBRM	-0.015	3	0.006
ZIP	0.052	1	0.014
ZINB	-0.019	3	0.008

PRM: Predicted and actual probabilities

Count	Actual	Predicted	Diff	Pearson
0	0.301	0.209	0.091	36.489
1	0.269	0.310	0.041	4.962
2	0.195	0.242	0.048	8.549
3	0.092	0.135	0.043	12.483
4	0.073	0.061	0.012	2.174
5	0.030	0.025	0.005	0.760
6	0.019	0.010	0.009	6.883
7	0.013	0.004	0.009	17.815
8	0.001	0.002	0.001	0.300
9	0.002	0.001	0.001	1.550
Sum	0.993	0.999	0.259	91.964

NBRM: Predicted and actual probabilities

Count	Actual	Predicted	Diff	Pearson
0	0.301	0.304	0.003	0.028
1	0.269	0.272	0.003	0.039
2	0.195	0.180	0.014	1.066
3	0.092	0.106	0.015	1.818
4	0.073	0.060	0.013	2.753
5	0.030	0.033	0.004	0.348
6	0.019	0.018	0.000	0.004
7	0.013	0.010	0.003	0.719
8	0.001	0.006	0.005	3.593
9	0.002	0.004	0.001	0.456
Sum	0.993	0.993	0.062	10.824

ZIP: Predicted and actual probabilities

Count	Actual	Predicted	Diff	Pearson
0	0.301	0.298	0.003	0.022
1	0.269	0.217	0.052	11.526
2	0.195	0.210	0.016	1.095
3	0.092	0.142	0.050	16.281
4	0.073	0.076	0.002	0.071
5	0.030	0.034	0.005	0.612
6	0.019	0.014	0.005	1.346
7	0.013	0.005	0.008	9.840
8	0.001	0.002	0.001	0.447
9	0.002	0.001	0.001	1.985
Sum	0.993	0.999	0.143	43.225

ZINB: Predicted and actual probabilities

Count	Actual	Predicted	Diff	Pearson
0	0.301	0.312	0.012	0.396
1	0.269	0.256	0.013	0.611
2	0.195	0.181	0.014	0.981
3	0.092	0.110	0.019	2.889
4	0.073	0.063	0.010	1.524
5	0.030	0.035	0.005	0.709
6	0.019	0.019	0.000	0.004
7	0.013	0.010	0.003	0.710
8	0.001	0.006	0.005	3.397
9	0.002	0.003	0.001	0.302
Sum	0.993	0.995	0.081	11.522

Tests and Fit Statistics

PRM	BIC= -2896.289	AIC= 3.622	Prefer	Over	Evidence
vs NBRM	BIC= -3069.666	dif= 173.377	NBRM	PRM	Very strong
	AIC= 3.427	dif= 0.195	NBRM	PRM	
	LRX2= 180.196	prob= 0.000	NBRM	PRM	p=0.000
vs ZIP	BIC= -2966.657	dif= 70.367	ZIP	PRM	Very strong
	AIC= 3.529	dif= 0.093	ZIP	PRM	
	Vuong= 4.133	prob= 0.000	ZIP	PRM	p=0.000
vs ZINB	BIC= -3067.059	dif= 170.769	ZINB	PRM	Very strong
	AIC= 3.414	dif= 0.208	ZINB	PRM	
NBRM	BIC= -3069.666	AIC= 3.427	Prefer	Over	Evidence
vs ZIP	BIC= -2966.657	dif= -103.010	NBRM	ZIP	Very strong
	AIC= 3.529	dif= -0.102	NBRM	ZIP	
vs ZINB	BIC= -3067.059	dif= -2.608	NBRM	ZINB	Positive
	AIC= 3.414	dif= 0.013	ZINB	NBRM	
	Vuong= 2.069	prob= 0.019	ZINB	NBRM	p=0.019
ZIP	BIC= -2966.657	AIC= 3.529	Prefer	Over	Evidence
vs ZINB	BIC= -3067.059	dif= 100.402	ZINB	ZIP	Very strong
	AIC= 3.414	dif= 0.115	ZINB	ZIP	
	LRX2= 107.221	prob= 0.000	ZINB	ZIP	p=0.000

Both the NBRM & ZINB consistently fit better than either the PRM or ZIP. BIC favors NBRM; AIC likes ZINB.

