MLE: Categorical and Limited Dependent Variables

Unit 2: Ordered, Multinomial, Count, and Limited Dependent Variables

3b. Negative Binomial and Zero-Inflated Models

PS2730-2021

Weeks 8-9

Professor Steven Finkel

Limitations of the Poisson Model for Count Data

- Limitations of Poisson regression: variance of count distributions often does not equal the mean, and Poisson models often do not account sufficiently for the number of 0s in the data
- Problem of *overdispersion* in count distributions limits Poisson regression; occurs because of contagion effects (where the outcomes are not independent), or due to unobserved heterogeneity across units, where factors that are not included in the model are producing higher variances for a given conditional mean based on the Xs
- Overdispersion problem is addressed (potentially) with the **Negative Binomial Model**, which allows for a heterogeneity term in the distribution (similar to a "random effect" in panel analysis)
- Problem of "too many" 0s is complicated; solutions depend on assumptions about the data generating process that produces 0s versus positive counts. Possible solutions: **Zero Inflated Models** (which we will cover) or **Hurdle Models** (which we won't except in passing)

Negative Binomial Model

- Poisson model for the latent rate of occurrence of an outcome in a given time interval

$$\mu_i = E(y \mid X) = \exp^{XB}$$

- Negative Binomial Model: adds an additional parameter to allow for unit-level unobserved heterogeneity in the expected count or latent rate. In Poisson, all units with same Xs have the same expected count or rate; here we allow there to be some error due to unobservables that produces higher or lower counts for given units than expected from the observed X.
- Estimate

$$\tilde{\mu}_i = E(y \mid X) = \exp^{XB+\varepsilon_i}$$

with ε_i representing the unit level error term in the latent rate

- Relationship to our original rate equation:

$$\tilde{\mu}_i = \exp^{XB+\varepsilon_i} = \exp^{XB} \exp^{\varepsilon_i} = \mu_i \delta_i$$

$$\text{where } \delta_i = \exp^{\varepsilon_i}$$

- NBR proceeds by identifying this model with the assumption that the mean of δ=1
- This means that we have the same **average rate** as in Poisson

$$E(\tilde{\mu}_i) = E(\mu_i \delta_i) = \mu_i E(\delta_i) = \mu_i$$

- And the predicted counts still follow a Poisson distribution:

$$\Pr(y_i \mid x, \delta) = \frac{\exp^{-\tilde{\mu}_i} (\tilde{\mu}_i^{\,y_i})}{y_i!} = \frac{\exp^{-\mu_i \delta_i} (\mu_i \delta_i^{\,y_i})}{y_i!}$$

- But: not straightforward to solve since δ is unknown
- We assume a distribution for δ to make some headway. NBM assumes that δ follows a "gamma distribution". Gamma is a distribution for positive outcomes governed by one parameter **alpha** which affects its shape. When alpha (α) is very small, gamma looks like a normal distribution, when alpha is very big, gamma looks extremely skewed, like in Long 232, where *v i*s 1/(α).
- So the negative binomial distribution results from a "mixture" of the poisson and gamma distributions
- So large *v* = small α = more normal; small *v*=large α = more skewed
- Then:

$$\Pr(y_i \mid x) = \frac{\Gamma(y_i + v_i)}{y_i!\Gamma(v_i)} \left(\frac{\Gamma(v_i)}{(v_i + \mu_i)} \right)^{v_i} \left(\frac{\mu_i}{(v_i + \mu_i)} \right)^{y_i}$$

- Which is also the P_i for the likelihood function of the NBM

- This leads to the same expected mean of the rate for Poisson and Negative Binomial, but a difference variance of the counts in NB:

$$\text{Var}(y \mid X) = \mu_i(1 + \frac{\mu_i}{v_i}) = \exp^{x_i B}(1 + \frac{\exp^{x_i B}}{v_i})$$

- This allows the variance of the counts in the NB model to be greater than the mean, since both μ and v are positive
- As $v \rightarrow \infty$ (and $\alpha \rightarrow 0$), then Var(y)= μ and we're back at Poisson!
- In terms of alpha (or $1/v_i$), assuming common v_i:

$$\text{Var}(y \mid X) = \mu_i(1 + \alpha\mu_i) = \mu_i + \alpha\mu_i^2$$

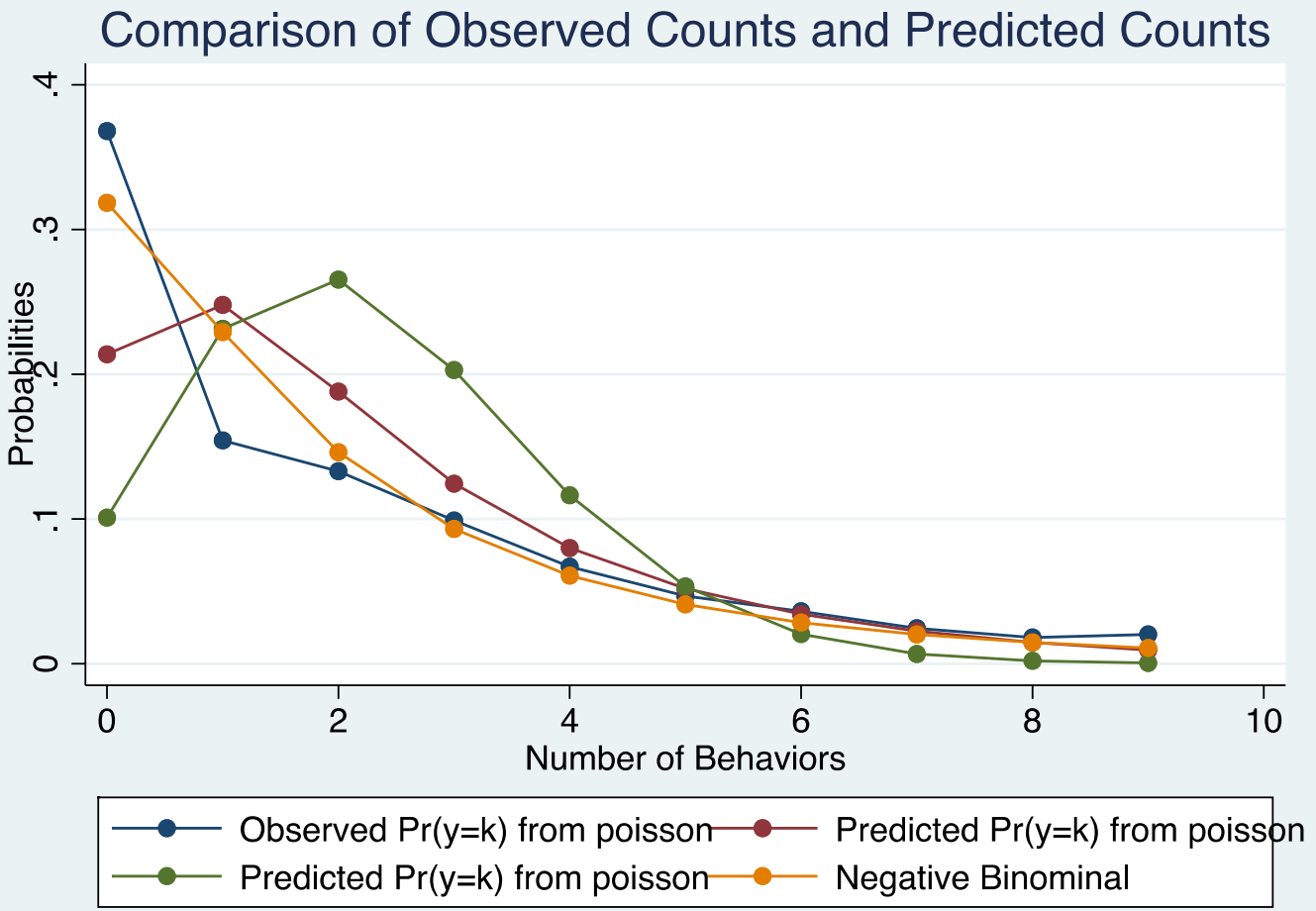
- So we have a model with same mean structure as Poisson, but increased variance governed by a gamma-distributed parameter v (α) to be estimated from the data. The bigger α, the more the additional parameter adds variance to the counts, which will account for more 0s and fewer high counts than Poisson

- Suggests a test for whether the additional parameter α is needed, since inclusion will result in inefficient estimation should it not account for the distribution of counts better than Poisson (or, alternatively, increase the log-likelihood sufficiently).

- One test: a one-tailed "Z test" of H_0: α=0, since α must be positive
- Alternative test: LR comparison of log-likelihoods for NB and Poisson as

$$G^2 = 2(\ln L_{NBRM} - \ln L_{PRM})$$

- which also tests whether α=0. Also requires adjustment of p value since alpha can only be positive
- Other interpretations (predicted probabilities, effects of independent variables on expected rates, etc.) straightforward and similar to Poisson (with caveat that predicted P formula differs as on slide 5)



- Observed, univariate (no independent variables) and multivariate Poisson all underpredict 0 counts considerably; negative binomial model substantially better and less overprediction of high counts as well. Looks like a better choice!

Zero Inflated Models

- "Zero-inflated" models available if 0s are not simply prevalent in the data to a larger degree than Poisson or NB would predict, but rather if they are *structurally* different than positive counts
- That is, in some instances there may be groups of individuals/units that *cannot* be on 0, e.g. Long's scientific productivity research where some scholars are not in jobs where research output is possible; or battle deaths from conflicts, where most countries or dyads are not (could not be?") in a state of conflict (so-called "peace zeros"–Bagozzi 2015)
- The 0s are "inflated" as a result, beyond what Possion or NB would predict
- Also: in these cases, we don't know what a 0 actually is – a "structural" or "inflated" 0, or a 0 that is a normal prediction from a Possion or NB model
- So "Zero Inflated" Models developed to take this into account

- Assume the population is made up of two groups
- People/units in group 1 are those who will *always* be 0 – what are called the **"structural 0s"**, and
- People/units in group 2 whose rates are generated by normal count processes, and which *might* include 0 counts in the given time interval
- Example: number of fish caught as the DV; two groups of 0s – people who *never* fish, and people who are just bad fishermen
- We model the first group as a logit or a probit; the second as a Poisson or Negative Binomial
- The first equation is called the "inflation" equation; the second the "count" equation
- Put them together and you arrive at the **Zero Inflated Poisson** (ZIP) or **Zero-Inflated Negative Binomial** (ZINB), depending on the nature of the count equation

Zero-Inflated Poisson (ZIP)

- Probability of being in Group 1, the "Always 0" group, is ψ_i, with the probability predicted by individual/unit-level independent variables

$$\Pr(y_i = 0) = \psi_i = \frac{\exp(z\gamma)}{1+\exp(z\gamma)} \quad \text{for logit}$$

$$\Pr(y_i = 0) = \psi_i = 1 - \Phi(z\gamma) \quad \text{for probit}$$

- Probability of a given count, for those in Group 2, assuming Poisson:

$$\Pr(y \mid \mu, \psi_i = 0) = \frac{\exp^{-\mu}(\mu^y)}{y!}$$

- With $\mu_i = E(y \mid X) = \exp^{XB}$

IMPORTANT NOTE: I'm using z and γ for the variables in the inflation equation and x and β for the variables in the count equation – they *can* be the same or they can be different, though

Zero-Inflated Poisson (ZIP)

- Putting the two together gives the ZIP model

$$\Pr(y = 0 \mid \mu, \psi_i) = \psi_i + (1 - \psi_i)\exp^{-\mu_i})$$

$$\Pr(y = y_i > 0 \mid \mu, \psi_i) = (1 - \psi_i)\frac{\exp^{-\mu_i}(\mu^{y_i})}{y_i!}$$

- With the equation predicting 0s consisting of probability of being a structural 0 **plus** the 0s from the Poisson-distributed count among the non-structural 0s, and the equation predicting positive counts consisting of the Poisson distributed counts for the non-structural 0 group
- In words: ψ is the probability of an "excess" or "inflated" 0.
- If : ψ=0, the inflation equation is just the Poisson predicted 0 counts and the whole model is just a Poisson regression

- The conditional mean and variance in the ZIP differs from regular Poisson

$$E(y_i \mid x,z) = (0 * \psi_i) + (1-\psi_i) * \mu_i = \mu_i - \psi_i \mu_i$$
$$\text{Var}(y_i \mid x,z) = \mu_i (1-\psi_i)(1+\psi_i \mu)$$

- Which says that the expected rate, given the Xs, is lowered by a factor $\psi\mu$, or the fraction of inflated 0s in the population
- The expected rate among those "eligible" for a non-zero count (i.e., not in the structural 0 group) is μ
- The variance is the Poisson variance weighted by the size of Group 1: the bigger the structural 0 group, the bigger the variance, relative to normal Poisson
- Again: If $\psi = 0$, everything reduces to regular Poisson!!!

Zero-Inflated Negative Binomial (ZINB)

- Similar process if we think the count portion of the model has additional 0s as in the Negative Binomial model considered earlier, independent of the structural 0 issue
- Probability of being in Group 1, the "Always 0" group, is ψ_i, with the probability predicted by individual/unit-level independent variables

$$\Pr(y_i = 0) = \psi_i = \frac{\exp(z\gamma)}{1+\exp(z\gamma)} \quad \text{for logit}$$

$$\Pr(y_i = 0) = \psi_i = \Phi(z\gamma) \quad \text{for probit}$$

- Probability of a given count, for those in Group 2, assuming NB:

$$\Pr(y_i > 0 \mid \mu_i, \psi, \delta) = \frac{\exp^{-\mu_i\delta_i}(\mu_i\delta_i^{\,y_i})}{y_i!}$$

- With δ following the gamma distribution as in earlier NB slides

- So ZINB

$$\Pr(y=0\,|\,\mu,\psi_i)=\psi_i+(1-\psi_i)\left(\frac{(v_i)}{(v_i+\mu_i)}\right)^{v_i}$$

$$\Pr(y=y_i>0\,|\,\mu,\psi_i)=(1-\psi_i)\frac{\Gamma(y_i+v_i)}{y_i!\Gamma(v_i)}\left(\frac{\Gamma(v_i)}{(v_i+\mu_i)}\right)^{v_i}\left(\frac{\mu_i}{(v_i+\mu_i)}\right)^{y_i}$$

- Mean and Variance of ZINB

$$E(y_i\,|\,x,z)=(0*\psi_i)+(1-\psi_i)*\mu_i=\mu_i-\psi_i\mu_i$$

$$\mathrm{Var}(y_i\,|\,x,z)=\mu_i(1-\psi_i)(1+\mu(\psi_i+\alpha))$$

- with α being the additional variance component from NB versus Poisson

Interpretations in Zero-Inflated Models

1. Probability of Being in the Structural Zero Group, from coefficients in the ”inflation” equation

$$\Pr(y_i = 0) = \psi_i = \frac{\exp(z\gamma)}{1+\exp(z\gamma)}$$

2. Probability of a Count, given NOT being in the Structural Zero Group, with rate predicted from the coefficients in the “count” equation

$$\Pr(y \mid \mu, \psi_i = 0) = \frac{\exp^{-\mu}(\mu^y)}{y!}$$

Poisson reported here; adjust accordingly for NB Model

$$\mu_i = E(y \mid X) = \exp^{XB}$$

- Probability of being either an "inflated" or "true count" 0:

$$\Pr(y=0\,|\,x,z)=\frac{\exp^{\gamma z}}{1+\exp^{\gamma z}}+\frac{1}{1+\exp^{\gamma z}}\exp^{-x\beta_i}$$

which means you can separate the "true count" and "inflated" 0s once you've estimated the model, and see how X/Z relates to each

- The unconditional rate, given the presence of structural 0s:

$$E(y_i\,|\,x,z)=\exp^{x_i\beta}-(\frac{\exp^{z\gamma}}{1+\exp^{z\gamma}})\exp^{x_i\beta}$$

- The rate for a case NOT in structural 0 group:

$$E(y_i\,|\,\psi=0,x)=\exp^{x_i\beta}$$

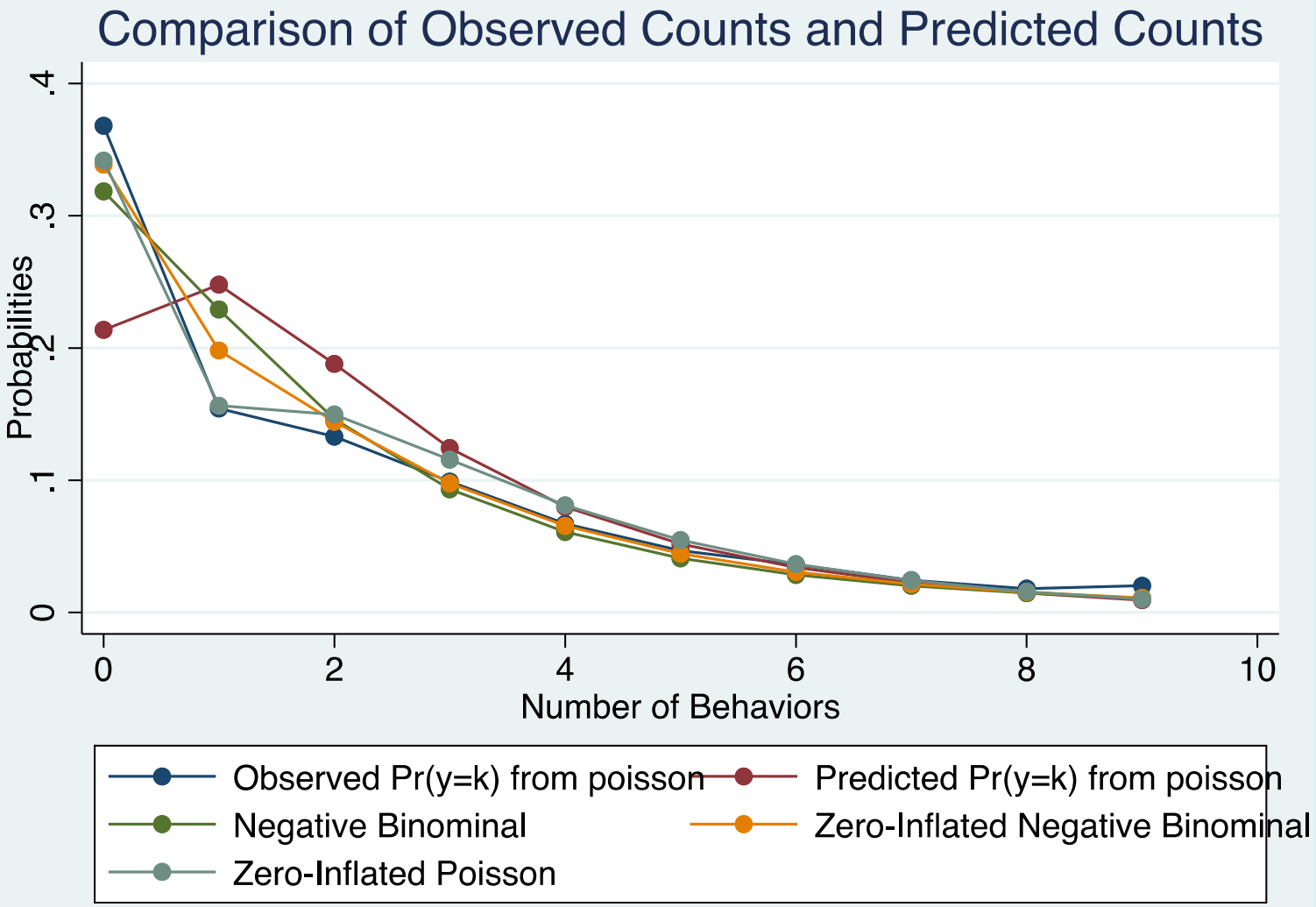
Comparing Count Models

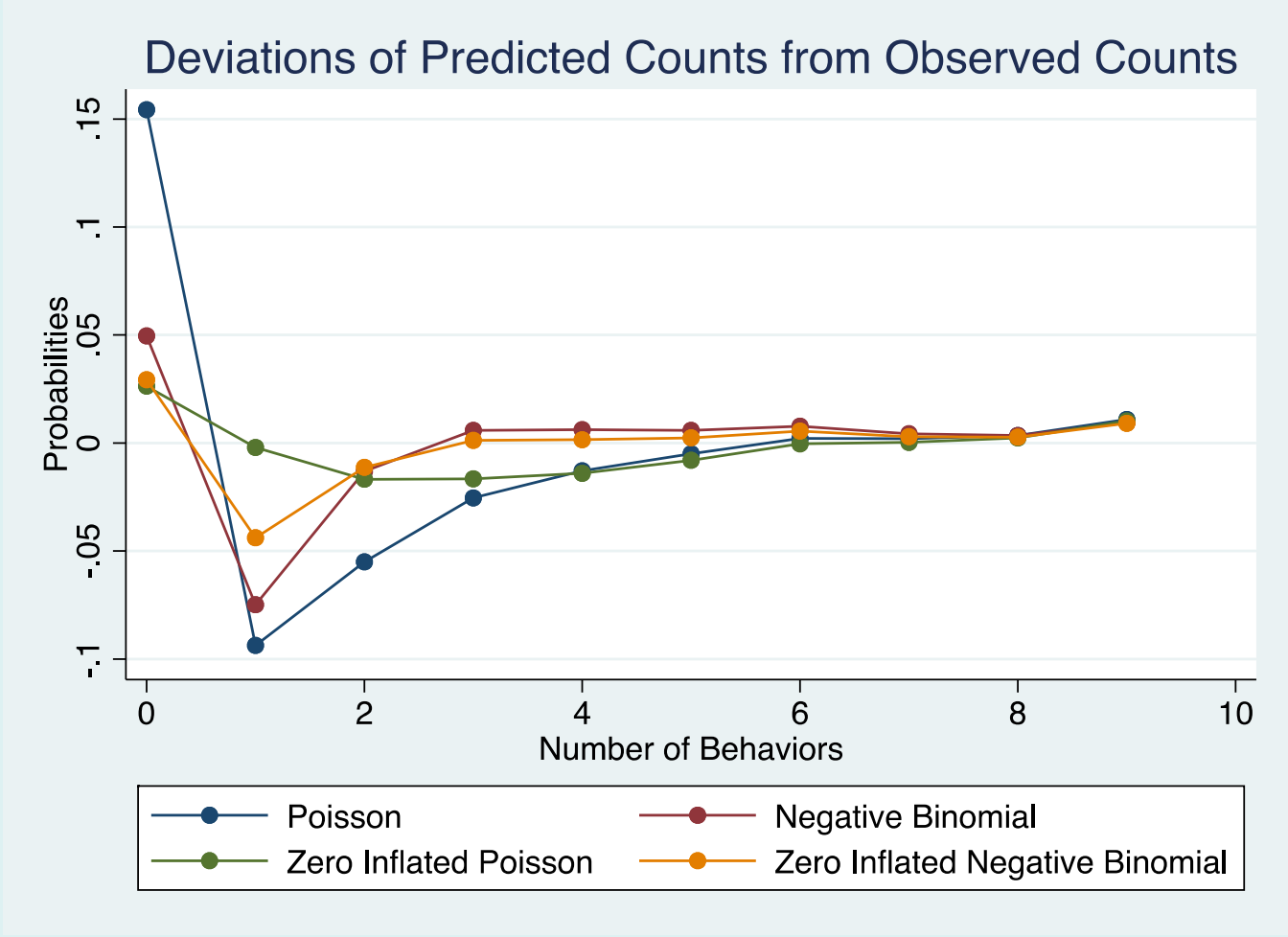
- 4 different models we've considered: Poisson, Negative Binomial, Zero-Inflated Poisson and Zero-Inflated Negative Binomial. How to choose between them all?
- One way is to eyeball the counts that are produced by each model and see which one fits the observed data (see below)
- But what about more formal statistical tests?
- If models are nested, then can test via LR test
  - Poisson and Negative Binomial are nested; test H_o: $\alpha = 0$ via LR test comparing the LL between models with and without the alpha parameter
  - ZINF Poisson and ZINF Binomial are also nested; test H_o: $\alpha = 0$ via similar LR test (need to "force" Stata to do this)

- How to compare other models?

- Zero-inflated Poisson and Poisson are not formally nested; Zero-inflated Negative Binomial and Negative Binomial are not formally nested
- Several possibilities
- 1) Vuong test of non-nested models
  - Logic: Take ratio of the predicted likelihoods for a unit from M1 and M2; if M1 is "better" this ratio should be greater than 1, beyond sampling error
  - Calculate the log-ratio as $m_i = ln \frac{\widehat{P(1)}(y_i|x)}{\widehat{P(2)}(y_i|x)}$
  - Test H_o: $\bar{m} = 0$ This would mean that the ratio is 1 in the population, so ln(1)=0
  - Vuong V=$\frac{\sqrt{N}\bar{m}}{s_m}$ where s_m is the standard deviation of the m_i
  - If V> 1.96 then reject H_0
- 2) Compare AIC and BIC values for the alternative models, following the decision rules discussed earlier in the course

Informal: Compare the observed and predicted counts from all models and see which fits best





Calculated as Observed Count Minus Predicted Count for each count

Long-Freese "countfit" routine in SPOST does all of this, formal and informal tests, including graphs