Ordered Data and Count Data Models

- Ordered data
- Count Data Models
  - The Poisson Regression Model
  - Overdispersion
  - Heterogeneity and the Negative Binomial Regression Model
  - Hurdle and Zero-Inflated Poisson Models.

Ordered data

- In some problems, the variate of interest assumes more than two discrete outcomes, but these are inherently *ordered*.
- Examples that have appeared in the literature include the following: Bond ratings; Results of taste tests; Surveys on the degree of satisfaction with some service; The level of insurance coverage taken by a consumer: none, part, or full; Employment: unemployed, part time, or full time

Ordered data

- Zavoina and McElvey (1975) modelled ordered data using the following latent variable framework:

$$Y_i^* = \mathbf{X}_i' \boldsymbol{\beta}_0 + u_i, \qquad Y_i = \begin{cases} 0 & Y_i^* \leq \mu_0 \\ 1 & \mu_0 < Y_i^* \leq \mu_1 \\ 2 & \mu_1 < Y_i^* \leq \mu_2 \\ \vdots & \vdots \\ J-1 & \mu_{J-1} < Y_i^* \leq \mu_{J-1} \\ J & \mu_{J-1} < Y_i^* \end{cases}$$

  where the *threshold parameters* are such that $0 = \mu_0 < \mu_1 < \cdots < \mu_{J-1}$ and Y_i^* is a latent variable.
- If the distribution of u_i is specified, the *unknown parameters* $\boldsymbol{\beta}$ and $\mu_2, \ldots, \mu_{J-1}$ can be estimated by maximum likelihood.
- Assume that u_i has distribution function $F(\cdot)$ and is independent of $\mathbf{X}_i$.

Ordered data

- Notice that

$$
\begin{aligned}
p_0(\mathbf{X}_i, \boldsymbol{\beta}_0) &= \mathcal{P}\left(Y_i = 0|\mathbf{X}_i\right) = \mathcal{P}\left(\mathbf{X}_i'\boldsymbol{\beta}_0 + u_i \leq 0|\mathbf{X}_i\right) \\
&= \mathcal{P}\left(u_i \leq -\mathbf{X}_i'\boldsymbol{\beta}_0|\mathbf{X}_i\right) \\
&= F\left(-\mathbf{X}_i'\boldsymbol{\beta}_0\right)
\end{aligned}
$$

$$
\begin{aligned}
p_1(\mathbf{X}_i, \boldsymbol{\beta}_0) &= \mathcal{P}\left(Y_i = 1|\mathbf{X}_i\right) = \mathcal{P}\left(0 < \mathbf{X}_i'\boldsymbol{\beta}_0 + u_i \leq \mu_1|\mathbf{X}_i\right) \\
&= \mathcal{P}\left(u_i \leq \mu_1 - \mathbf{X}_i'\boldsymbol{\beta}_0|\mathbf{X}_i\right) - \mathcal{P}\left(u_i < -\mathbf{X}_i'\boldsymbol{\beta}_0|\mathbf{X}_i\right) \\
&= F\left(\mu_1 - \mathbf{X}_i'\boldsymbol{\beta}_0\right) - F\left(-\mathbf{X}_i'\boldsymbol{\beta}_0\right) \\
&\ \ \vdots
\end{aligned}
$$

$$
\begin{aligned}
&\vdots \\
p_j(\mathbf{X}_i, \boldsymbol{\beta}_0) &= \mathcal{P}\left(Y_i = j|\mathbf{X}_i\right) = \mathcal{P}\left(\mu_{j-1} < \mathbf{X}_i'\boldsymbol{\beta}_0 + u_i \leq \mu_j|\mathbf{X}_i\right) \\
&= \mathcal{P}\left(u_i \leq \mu_j - \mathbf{X}_i'\boldsymbol{\beta}_0|\mathbf{X}_i\right) - \mathcal{P}\left(u_i < \mu_{j-1} - \mathbf{X}_i'\boldsymbol{\beta}_0|\mathbf{X}_i\right) \\
&= F\left(\mu_j - \mathbf{X}_i'\boldsymbol{\beta}_0\right) - F\left(\mu_{j-1} - \mathbf{X}_i'\boldsymbol{\beta}_0\right) \\
&\vdots \\
p_J(\mathbf{X}_i, \boldsymbol{\beta}_0) &= \mathcal{P}\left(Y_i = J|\mathbf{X}_i\right) = \mathcal{P}\left(\mu_{J-1} < u_i + \mathbf{X}_i'\boldsymbol{\beta}_0|\mathbf{X}_i\right) \\
&= \mathcal{P}\left(u_i > \mu_{J-1} - \mathbf{X}_i'\boldsymbol{\beta}_0|\mathbf{X}_i\right) \\
&= 1 - \mathcal{P}\left(u_i \leq \mu_{J-1} - \mathbf{X}_i'\boldsymbol{\beta}_0|\mathbf{X}_i\right) \\
&= 1 - F\left(\mu_{J-1} - \mathbf{X}_i'\boldsymbol{\beta}_0\right).
\end{aligned}
$$

Ordered data

- Therefore, the log-likelihood function is simply

$$\log L(\theta) = \sum_{i=1}^{n} \sum_{j=0}^{J} \mathbf{1}\left(Y_i = j\right) \log\left[p_j(\mathbf{X}_i, \boldsymbol{\beta})\right]$$

- As in all discrete choice models, the variance of u_i is *not identified*.
- The *ordered-probit* and *ordered-logit* are the most used special cases of this model.

Ordered data

- For the *ordered-probit*

$$F\left(\mu_j - \mathbf{X}_i'\boldsymbol{\beta}_0 | \mathbf{X}_i\right) = \Phi\left(\mu_j - \mathbf{X}_i'\boldsymbol{\beta}_0\right)$$

- For the *ordered-logit*

$$F\left(\mu_j - \mathbf{X}_i'\boldsymbol{\beta}_0 | \mathbf{X}_i\right) = \frac{\exp\left(\mu_j - \mathbf{X}_i'\boldsymbol{\beta}_0\right)}{1 + \exp\left(\mu_j - \mathbf{X}_i'\boldsymbol{\beta}_0\right)}$$

- Interpreting coefficients requires some care. For instance in the *ordered probit* model we have

$$\begin{aligned}
\frac{\partial p_0(\mathbf{X}_i, \boldsymbol{\beta}_0)}{\partial x_k} &= -\beta_{0k}\phi(-\mathbf{X}_i'\boldsymbol{\beta}_0), \frac{\partial p_J(\mathbf{X}_i, \boldsymbol{\beta}_0)}{\partial x_k} = \beta_{0k}\phi(\mu_{J-1} - \mathbf{X}_i'\boldsymbol{\beta}_0) \\
\frac{\partial p_j(\mathbf{X}_i, \boldsymbol{\beta}_0)}{\partial x_k} &= \beta_{0k}[\phi(\mu_{j-1} - \mathbf{X}_i'\boldsymbol{\beta}_0) - \phi(\mu_j - \mathbf{X}_i'\boldsymbol{\beta}_0)], j = 1, ..., J-1
\end{aligned}$$

- For $1 < j < J$, the sign of $\partial p_j(\mathbf{X}_i, \boldsymbol{\beta}_0) \partial x_k$ is ambiguous. It depends on $|\mu_{j-1} - \mathbf{X}_i'\boldsymbol{\beta}_0|$ versus $|\mu_j - \mathbf{X}_i'\boldsymbol{\beta}_0|$ (remember, $\phi(\cdot)$ is symmetric about zero).

Ordered data

- The OP and OL models allow us to obtain the *sign of the partial effects* on $\mathcal{P}(Y > j|\mathbf{X}_i)$: for a continuous variable x_h. For the OP model
$$\frac{\partial \mathcal{P}(Y_i > j|\mathbf{X}_i)}{\partial x_h} = \beta_h \phi(\mu_j - \mathbf{X}_i'\boldsymbol{\beta}),$$
If $\beta_h > 0$, an increase in x_h increases the probability that Y_i is greater than any value j.
- Of course the we can interpret the sign of the parameters in the *latent variable model*.

- A closely related model can be used for *grouped data*.
- **Example:** Income reported in non-overlapping intervals
- In this case, the threshold parameters are the limits of the intervals.
- The main difference is that, for $J > 0$, the variance of u_i is *identified* because the thresholds give information on the scale of u_i.

The Poisson Regression Model

- In many relevant applications, the variate of interest is *the count of the number of occurrences of some event in a given period of time* (rare events).
- Examples include: number of accidents, number of patents, number of takeovers, number of purchases, number of doctor visits, number of jobs and number of trips.
- These data have some very specific characteristics:
  - **Discreteness;**
  - **non-negative;**
  - **Many zeros and a long right-hand tail**.
- In this context, standard linear models are *not appealing* because:
  - **The conditional expectation is necessarily non-negative;**
  - **The data is intrinsically heteroskedastic;**
  - **Do not allow the computation of the probability of events of interest.**

The Poisson Regression Model

- The basic model for count data is the *Poisson regression*, defined by

$$\mathcal{P}\left(Y_i = j|\mathbf{X}_i\right) = \frac{\exp\left(-\lambda(\mathbf{X}_i, \boldsymbol{\beta}_0)\right)\lambda(\mathbf{X}_i, \boldsymbol{\beta}_0)^j}{j!}, \qquad j = 0, 1, 2, \ldots$$

$$\mathrm{E}(Y_i|\mathbf{X}_i) = \mathrm{Var}(Y_i|\mathbf{X}_i) = \lambda(\mathbf{X}_i, \boldsymbol{\beta}_0)$$

- Notice, however, that

$$\mathrm{Var}(Y_i) = \mathrm{E}_x\left[\lambda(\mathbf{X}_i, \boldsymbol{\beta}_0)\right] + \mathrm{Var}_x\left[\lambda(\mathbf{X}_i, \boldsymbol{\beta}_0)\right] \geq \mathrm{E}_x\left[\lambda(\mathbf{X}_i, \boldsymbol{\beta}_0)\right] = \mathrm{E}(Y_i).$$

  where in general, the following specification is adopted: $\lambda(\mathbf{X}_i, \boldsymbol{\beta}_0) = \exp\left(\mathbf{X}_i'\boldsymbol{\beta}_0\right)$.
- Therefore,

$$\frac{\partial \mathrm{E}(Y_i|\mathbf{X}_i)}{\partial \mathbf{X}_i} = \exp\left(\mathbf{X}_i'\boldsymbol{\beta}_0\right)\boldsymbol{\beta}_0$$

- ML estimation of $\boldsymbol{\beta}_0$ is straightforward.

The Poisson Regression Model

- The log-likelihood function, likelihood equations and the Hessian are given by

$$
\begin{aligned}
\log L(\beta) &= \sum_{i=1}^{n} \left[-\exp\left(\mathbf{X}_i'\boldsymbol{\beta}\right) + \left(\mathbf{X}_i'\boldsymbol{\beta}\right) Y_i - \log\left(Y_i!\right)\right] \\
\frac{\partial \log L\left(\widehat{\boldsymbol{\beta}}\right)}{\partial \beta} &= \sum_{i=1}^{n} \left[Y_i - \exp(\mathbf{X}_i'\widehat{\boldsymbol{\beta}})\right] \mathbf{X}_i = 0 \\
\frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta'} &= -\sum_{i=1}^{n} \exp(\mathbf{X}_i'\boldsymbol{\beta})\mathbf{X}_i\mathbf{X}_i'
\end{aligned}
$$

- Notice that the Hessian is *negative definite* for all $\mathbf{X}$ and $\boldsymbol{\beta}$, which facilitates the estimation and ensures the uniqueness of the maximum, **if it exists**.
- The MLE has the usual properties. In particular

$$
\sqrt{n}\left(\hat{\beta}_{ML} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{E}\left(\exp(\mathbf{X}_i'\boldsymbol{\beta}_0)\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\right)
$$

- As usual, inference can be performed using the LR, W and LM tests.

Overdispersion

- The Poisson model imposes (conditional) *equidispersion*, which is very restrictive.
- There are many possible causes for overdispersion:
  - **Measurement error;**
  - **Misspecification of the conditional mean;**
  - **Neglected heterogeneity (random parameter variation).**
- Applied economists tend to focus on the neglected heterogeneity issue, assuming

$$E(Y_i|\mathbf{X}_i, \varepsilon_i) = \exp(\mathbf{X}_i' \boldsymbol{\beta}_0 + \varepsilon_i)$$

$$E\left(\exp\left(\varepsilon_i\right)|\mathbf{X}_i\right) = 1, \qquad \mathrm{Var}\left(\exp\left(\varepsilon_i\right)|\mathbf{X}_i\right) = \sigma^2$$

- In this particular case

$$E(Y_i|\mathbf{X}_i) = E(\lambda(\mathbf{X}_i, \boldsymbol{\beta}_0)|\mathbf{X}_i) = E_\varepsilon\left[\exp(\mathbf{X}_i'\boldsymbol{\beta}_0 + \varepsilon_i)|\mathbf{X}_i\right] = \exp(\mathbf{X}_i'\boldsymbol{\beta}_0)$$

- Therefore, this sort of neglected heterogeneity does not change the form of the conditional expectation of Y_i.
- Gourieroux, Monfort and Trognon (1984) proved the following *powerful result*: If $E(Y_i|\mathbf{X}_i) = \lambda(\mathbf{X}_i, \boldsymbol{\beta}_0)$ is correctly specified and the Likelihood function is constructed using a probability distribution which does not necessarily correspond to the true distribution of the data, but belongs to the *family of linear exponential distributions*, then the *Quasi-Maximum Likelihood* estimator is consistent for $\boldsymbol{\beta}_0$.

Overdispersion

- The family of linear exponential distributions includes the *Poisson Distribution*, the *Normal Distribution* (with fixed variance). the *binomial* (with fixed number of trials), the *gamma distribution* (with fixed shape parameter)
- In this particular context the *Quasi-Maximum Likelihood* estimator is sometimes called *Pseudo-Maximum Likelihood Estimator* by some authors.
- Inference is done using the results presented previously for the Quasi-Maximum Likelihood estimator. In particular since the Poisson pseudo-MLE is consistent in presence of this sort of misspecification, valid inference can be based on

$$\sqrt{n}\left(\widehat{\beta}-\beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, A^{-1}BA^{-1}\right)$$

$$A = \mathrm{E}\left[\exp(\mathbf{X}_i'\beta_0)\mathbf{X}_i\mathbf{X}_i'\right] \qquad B = \mathrm{E}\left[\left(y_i - \exp(\mathbf{X}_i'\beta_0)\right)^2 \mathbf{X}_i\mathbf{X}_i'\right]$$

Overdispersion

Note that

$$\begin{aligned} \mathrm{Var}(Y_i|\mathbf{X}_i) &= \mathrm{E}_\varepsilon\left[\exp(\mathbf{X}_i'\boldsymbol{\beta}_0+\varepsilon_i)\right]+\mathrm{Var}_\varepsilon\left[\exp(\mathbf{X}_i'\boldsymbol{\beta}_0+\varepsilon_i)\right] \\ &= \exp(\mathbf{X}_i'\boldsymbol{\beta}_0)+\sigma^2\exp(2\mathbf{X}_i'\boldsymbol{\beta}_0). \end{aligned}$$

- The presence of *overdispersion* can be tested by testing $H_0:\sigma^2=0$.
- This can be done using the following LM (IM) test statistic (Cox, 1983, and Chesher, 1984)

$$T=\sum_{i=1}^{n}\frac{\left(Y_i-\exp(\mathbf{X}_i'\widehat{\boldsymbol{\beta}})\right)^2-Y_i}{\sqrt{2\sum_{i=1}^{n}\exp(2\mathbf{X}_i'\widehat{\boldsymbol{\beta}})}}\xrightarrow{d}\mathcal{N}(0,1)$$

- Alternatively, we can regress $\left[\left(Y_i - \exp(\mathbf{X}_i'\widehat{\boldsymbol{\beta}})\right)^2 - Y_i\right] \exp(-\mathbf{X}_i'\widehat{\boldsymbol{\beta}})$ on $\exp(\mathbf{X}_i'\widehat{\boldsymbol{\beta}})$ (or on a constant or other functions of $\exp(\mathbf{X}_i'\widehat{\boldsymbol{\beta}})$) and test the significance of the regressor (Cameron & Trivedi, 1986).
- All these tests can also detect *underdispersion*.
- Overdispersion tests are overplayed in the literature:
  1. in practice, the null is almost always rejected;
  2. if this is the only source of misspecification, the **Poisson pseudo-MLE is still consistent**.
- Other specification tests are available, like the *RESET* test which tests if the specification $\mathrm{E}(Y_i|\mathbf{X}_i) = \exp(\mathbf{X}_i'\boldsymbol{\beta}_0)$ is correct. This test can be performed by checking the significance of the additional regressor $\left(\mathbf{X}_i'\widehat{\boldsymbol{\beta}}\right)^2$.

Heterogeneity and the Negative Binomial Regression Model

- The assumption that Y_i has a Poisson distribution conditional of $\mathbf{X}_i$ and ε_i with mean $\lambda_i = \exp(\mathbf{X}_i'\boldsymbol{\beta}_0 + \varepsilon_i)$, leads to the compound Poisson regression model

$$\mathcal{P}(Y_i = j|\mathbf{X}_i, \varepsilon_i) = \frac{\exp[-\exp(\mathbf{X}_i'\boldsymbol{\beta}_0 + \varepsilon_i)]\exp(\mathbf{X}_i'\boldsymbol{\beta}_0 + \varepsilon_i)^j}{j!}$$

$$\mathcal{P}(Y_i = j|\mathbf{X}_i) = \int_{-\infty}^{+\infty} \frac{\exp[-\exp(\mathbf{X}_i'\boldsymbol{\beta}_0 + \varepsilon_i)]\exp(\mathbf{X}_i'\boldsymbol{\beta}_0 + \varepsilon_i)^j}{j!} g(\varepsilon_i)d\varepsilon_i$$

  where $g(\varepsilon_i)$ is the density function of ε_i and we assumed that $\mathbf{X}_i$ and ε_i are independent.

- This model can be made operational in different ways:
  1. *Pseudo maximum likelihood* estimation (discussed previously);
  2. *Parametric estimation* for specified $g(\varepsilon_i)$;
  3. *Semiparametric estimation* of $\boldsymbol{\beta}_0$ and $g(\varepsilon_i)$.

Heterogeneity and the Negative Binomial Regression Model

- If $g(\varepsilon_i)$ is specified, the MLE can be obtained, but the estimator *may not be robust* to departures from the additional distributional assumptions.
- Assuming that $\exp(\varepsilon_i) \sim \Gamma\left(\sigma^{-2}, \sigma^2\right)$, $\mathcal{P}(Y_i = j|\mathbf{X}_i)$ is given by the **negative-binomial** (Cameron and Trivedi (1986).denote it as NegBin II) model:

$$\mathcal{P}(Y_i = j|\mathbf{X}_i) = \frac{\Gamma\left(j + \sigma^{-2}\right)\left[1 + \sigma^{-2}\exp(-\mathbf{X}_i'\boldsymbol{\beta}_0)\right]^{-j}}{\Gamma\left(\sigma^{-2}\right)\Gamma\left(j+1\right)\left(1 + \sigma^2\exp(\mathbf{X}_i'\boldsymbol{\beta}_0)\right)^{\sigma^{-2}}}. \qquad (1)$$

- The Poisson model is obtained as a limiting case when $\sigma^2 \to 0$, but $H_0 : \sigma^2 = 0$ **cannot** be tested with a standard LR or W test.
- If the model (1) is misspecified but $\mathrm{E}(Y_i|\mathbf{X}_i) = \exp(\mathbf{X}_i'\boldsymbol{\beta}_0)$ is correct and σ^{-2} is fixed, the **negative-binomial Psedo-MLE** estimator is consistent for $\boldsymbol{\beta}_0$ This follows from the results of Gourieroux, Monfort and Trognon (1984) and the fact that the *negative-binomial distribution* with σ^{-2} fixed is a member of the family of linear exponential distributions

Heterogeneity and the Negative Binomial Regression Model

- The score test for $H_0 : \sigma^2 = 0$ is the overdispersion test studied before.
- Other parametric alternatives to the Poisson regression are available.
- A *semiparametric alternative* is to assume that ε has a discrete distribution with Q support points $\alpha_1, \ldots, \alpha_Q$ and corresponding probabilities $\pi_1, \ldots, \pi_Q$, leading to

$$\mathcal{P}(Y_i = j|\mathbf{X}_i) = \sum_{q=1}^{Q} \frac{\exp[-\exp(\mathbf{X}_i'\beta + \alpha_q)]\exp(\mathbf{X}_i'\boldsymbol{\beta}_0 + \alpha_q)^j}{j!}\pi_q,$$

Heterogeneity and the Negative Binomial Regression Model

- **For a given** Q, estimation of $\beta, \alpha_1, \ldots, \alpha_Q$ and $\pi_1, \ldots, \pi_{Q-1}$ can be performed by ML.
- This model can be interpreted as *semiparametric approximation* to a compound Poisson model with unspecified distribution.
- This leads to a consistent estimator if Q is *allowed to increase* at an appropriate rate;
- In practice, the value of Q has to be chosen (for example using an information criterion);
- Inference is complicated by the fact that the number of parameters is not fixed;

Hurdle and Zero-Inflated Poisson Models

- In some cases, the population may be contaminated by individuals for which $Y_i \equiv 0$.
- There are two ways to model this type of data. The *Zero-Inflated Poisson Model* and the *Hurdle Model*
- The *Zero-Inflated Poisson Model*: The zero outcome can arise from one of two regimes. In one regime, the outcome is always zero. In the other, the usual Poisson process is at work
- Let Z_i be a bernoulli random variable such that

$$Z_i = \begin{cases} 0 & \text{with } P\left(Z_i = 0|\mathbf{X}_i\right) = p_i \\ 1 & \text{with } P\left(Z_i = 1|\mathbf{X}_i\right) = 1 - p_i \end{cases}$$

where p_i can be a function of the regressors.

- Let $\mathcal{P}\left(Y_i = j|\mathbf{X}_i, Z_i = 1\right) = \pi_i\left(j; \beta_0\right), j = 0, 1, ...$ be the Poisson probability function.
- Let $\mathcal{P}\left(Y_i = 0|\mathbf{X}_i, Z_i = 0\right) = 1$.
- Note that

$$\begin{aligned}\mathcal{P}(Y_i = 0|\mathbf{X}_i) &= \mathcal{P}(Z_i = 0|\mathbf{X}_i)\mathcal{P}\left(Y_i = 0|\mathbf{X}_i, Z_i = 0\right) \\ &\quad + \mathcal{P}(Z_i = 1|\mathbf{X}_i)\mathcal{P}\left(Y_i = 0|\mathbf{X}_i, Z_i = 1\right) \\ &= \mathcal{P}(Z_i = 0|\mathbf{X}_i) + \mathcal{P}(Z_i = 1|\mathbf{X}_i)\mathcal{P}\left(Y_i = 0|\mathbf{X}_i, Z_i = 1\right) \\ &= p_i + (1 - p_i)\pi_i\left(0; \beta_0\right)\end{aligned}$$

- Additionally for $j > 0$:

$$\begin{aligned}\mathcal{P}(Y_i &= j|\mathbf{X}_i) = \mathcal{P}(Z_i = 1|\mathbf{X}_i)\mathcal{P}\left(Y_i = j|\mathbf{X}_i, Z_i = 1\right) \\ &= (1 - p_i)\pi_i\left(j; \beta_0\right)\end{aligned}$$

- Notice that

$$\begin{aligned}\mathrm{E}\left(Y_i|\mathbf{X}_i\right) &= \sum_{j=0}^{\infty} j\mathcal{P}(Y_i = j|\mathbf{X}_i) = \sum_{j=1}^{\infty} j\mathcal{P}(Y_i = j|\mathbf{X}_i) \\ &= (1 - p_i)E\left(Y_i|\mathbf{X}_i, Z_i = 1\right)\end{aligned}$$

- Therefore the standard pseudo maximum likelihood result does not hold here if p_i depends on $\mathbf{X}_i$.
- Then, the log-likelihood function for this *zero-inflated* (Mullahy, 1986) model can bewritten as

$$\begin{aligned}\log L(\beta) &= \sum_{i=1}^{n} \log\{[p_i + (1-p_i)\,\pi_i(0;\beta)]^{\mathbf{1}(Y_i=0)} \\ &\quad \times [(1-p_i)\,\pi_i(Y_i;\beta)]^{\mathbf{1}(Y_i>0)}\}\end{aligned}$$

- The *Hurdle Model* (Mullahy, 1986): A different extension of the basic count data model is obtained by letting the zero and positive observations be generated by different mechanisms.
- In his formulation, a binary probability model determines whether a zero or a nonzero outcome occurs, then, in the latter case we observe always a positive integer $1, 2, 3, ...$
- Consider the Bernoulli random variable

$$W_i = \begin{cases} 1 & \text{with } \mathcal{P}\left(W_i = 1|\mathbf{X}_i\right) = 1 - q_i \\ 0 & \text{with } \mathcal{P}\left(W_i = 0|\mathbf{X}_i\right) = q_i \end{cases}$$

where q_i may depend on $\mathbf{X}_i$.

Hurdle and Zero-Inflated Poisson Model

- $\mathcal{P}(Y_i = 0|\mathbf{X}_i, W_i = 0) = 1.$
- $\mathcal{P}(Y_i = 0|\mathbf{X}_i, W_i = 1) = 0$, and $\mathcal{P}\left(Y_i = j|\mathbf{X}_i, W_i = 1\right) = \pi_i^\star\left(j;\beta_0\right)$, $j = 1, 2, 3, ...$
- In this case

$$
\begin{aligned}
\mathcal{P}(Y_i &= 0|\mathbf{X}_i) = \mathcal{P}(W_i = 0|\mathbf{X}_i)P(Y_i = 0|\mathbf{X}_i, W_i = 0) \\
+\mathcal{P}(W_i &= 1|\mathbf{X}_i)P(Y_i = 0|\mathbf{X}_i, W_i = 1) = q_i
\end{aligned}
$$

- Additionally for $j = 1, 2, ...$

$$
\begin{aligned}
\mathcal{P}(Y_i &= j|\mathbf{X}_i) = \mathcal{P}(W_i = 0|\mathbf{X}_i)P(Y_i = j|\mathbf{X}_i, W_i = 0) \\
+\mathcal{P}(W_i &= 1|\mathbf{X}_i)P(Y_i = j|\mathbf{X}_i, W_i = 1) \\
&= \mathcal{P}(W_i = 1|\mathbf{X}_i)P(Y_i = j|\mathbf{X}_i, W_i = 0) \\
&= (1 - q_i)\,\pi_i^\star\left(j;\beta_0\right)
\end{aligned}
$$

Hurdle and Zero-Inflated Poisson Model

- In this case we have

$$
\begin{aligned}
\mathrm{E}\left(Y_i|\mathbf{X}_i\right) &= \sum_{j=0}^{\infty} j\mathcal{P}(Y_i = j|\mathbf{X}_i) = \sum_{j=1}^{\infty} j\mathcal{P}(Y_i = j|\mathbf{X}_i) \\
&= (1-q_i)\sum_{j=1}^{\infty} j\pi_i^{\star}\left(j;\beta_0\right) \\
&= (1-q_i)E\left[Y_i|\mathbf{X}_i, W_i = 1\right]
\end{aligned}
$$

- Again the standard pseudo maximum likelihood result does not hold here.

Hurdle and Zero-Inflated Poisson Model

- Then, the likelihood function has the form

$$\begin{aligned}\log L(\beta) &= \sum_{i=1}^{n}\{\mathbf{1}(Y_i = 0)(\log q_i) + \mathbf{1}(Y_i > 0)\log(1 - q_i) \\ &\quad +\mathbf{1}(Y_i > 0)\log[\pi_i^{\star}(Y_i;\beta)]\}\end{aligned}$$

- Notice that this function is separable.
- Correlated unobserved heterogeneity can be allowed for and integrated-out numerically.

Hurdle and Zero-Inflated Poisson Model

- Usually, $\pi_i^\star(j;\beta_0)$ is specified as a truncated Poisson of the form

$$\pi_i^\star(j;\beta_0) = \frac{\exp(-\lambda_i)\lambda_i^j}{(1-\exp(-\lambda_i))j!}, \qquad j > 0,$$

  with $\lambda_i = \exp(\mathbf{X}_i'\beta_0)$.
- However, in this model **there is no real truncation** and therefore an equally valid specification would be

$$\pi_i^\star(j;\beta_0) = \frac{\exp(-\lambda_i)\lambda_i^{j-1}}{(j-1)!}, \qquad j > 0.$$

- When the truncated Poisson specification is used and q_i is specified as

$$q_i = \exp(-\exp(\mathbf{X}_i'\gamma_0)),$$

  the null of no hurdle can be tested by testing $H_0 : \beta_0 = \gamma_0$.
- In any case, consistency depends on the distributional assumptions.