

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.
 - Models for Panel Data

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:

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

where the *threshold parameters* are such that $0 = \mu_0 < \mu_1 < \dots < \mu_{J-1}$ and Y_i^* is a latent variable.

- If the distribution of u_i is specified, the *unknown parameters* $\boldsymbol{\beta}$ and μ_2, \dots, μ_{J-1} can be estimated by maximum likelihood.
- Assume that u_i has distribution function $F(\cdot)$ and is independent of \mathbf{X}_i .

- Notice that

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

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

⋮

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

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

- Therefore, the log-likelihood function is simply

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

- 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*

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

- For the *ordered-logit*

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

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

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

- 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).

- 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*.

Ordered data

- 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

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

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

- Notice, however, that

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

where in general, the following specification is adopted:

$$\lambda(\mathbf{X}_i, \boldsymbol{\beta}_0) = \exp(\mathbf{X}_i' \boldsymbol{\beta}_0).$$

- Therefore,

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

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

The Poisson Regression Model

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

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

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

$$\sqrt{n} (\hat{\beta}_{ML} - \beta_0) \xrightarrow{d} \mathcal{N} \left(0, E(\exp(\mathbf{X}'_i\beta_0) \mathbf{X}_i \mathbf{X}'_i)^{-1} \right)$$

- 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

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

$$E(\exp(\varepsilon_i) | \mathbf{X}_i) = 1, \quad \text{Var}(\exp(\varepsilon_i) | \mathbf{X}_i) = \sigma^2$$

- In this particular case

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

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

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

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

Note that

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

- 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)

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

Overdispersion

- Alternatively, we can regress $\left[\left(Y_i - \exp(\mathbf{X}'_i \hat{\boldsymbol{\beta}}) \right)^2 - Y_i \right] \exp(-\mathbf{X}'_i \hat{\boldsymbol{\beta}})$ on $\exp(\mathbf{X}'_i \hat{\boldsymbol{\beta}})$ (or on a constant or other functions of $\exp(\mathbf{X}'_i \hat{\boldsymbol{\beta}})$) and test the significance of the regressor (Cameron & Trivedi, 1986).
- All these tests can also detect *underdispersion*.
- Overdispersion tests are overlapped 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 that checks the moment condition

$$E \left[(Y_i - \exp(\mathbf{X}'_i \boldsymbol{\beta}_0)) (\mathbf{X}'_i \boldsymbol{\beta}_0)^2 \right] = 0$$

- In practice, the test can be performed by checking the significance of the additional regressor $(\mathbf{X}'_i \hat{\boldsymbol{\beta}})^2$.

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 $\exp(\varepsilon_i) \sim \Gamma(\sigma^{-2}, \sigma^2)$, $\mathcal{P}(Y_i = j | \mathbf{X}_i)$ is given by the **negative-binomial** (Cameron and Trivedi (1986). denote it as NegBin II) model:

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

- The Poisson model is obtained as a limiting case when $\sigma^2 \rightarrow 0$, but $H_0 : \sigma^2 = 0$ **cannot** be tested with a standard LR or W test.
- If the model (1) is misspecified but $E(Y_i | \mathbf{X}_i) = \exp(\mathbf{X}'_i \boldsymbol{\beta}_0)$ is correct and σ^{-2} is fixed, the **negative-binomial Pseudo-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 $\alpha_1, \dots, \alpha_Q$ and corresponding probabilities π_1, \dots, π_Q , leading to

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

Heterogeneity and the Negative Binomial Regression Model

- For a given Q , estimation of $\beta, \alpha_1, \dots, \alpha_Q$ and π_1, \dots, π_{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

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

where p_i can be a function of the regressors.

Hurdle and Zero-Inflated Poisson Models

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

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

- Additionally for $j > 0$:

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

- Notice that

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

Hurdle and Zero-Inflated Poisson Models

- 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 be written as

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

Hurdle and Zero-Inflated Poisson Model

- 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 $1, 2, 3, \dots$
- Consider the Bernoulli random variable

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

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 $\mathcal{P}(Y_i = j | \mathbf{X}_i, W_i = 1) = \pi_i^*(j; \beta_0)$, $j = 1, 2, 3, \dots$
- In this case

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

- Additionally for $j = 1, 2, \dots$

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

Hurdle and Zero-Inflated Poisson Model

- In this case we have

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

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

Hurdle and Zero-Inflated Poisson Model

- Then, the likelihood function has the form

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

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

Hurdle and Zero-Inflated Poisson Model

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

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

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

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

- 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.

Define $Y_i = (Y_{i1}, \dots, Y_{iT})'$ and $j_i = (j_{i1}, \dots, j_{iT})'$, and let

$$\begin{aligned} P(Y_{it} = j_{it} | \mathbf{X}_{it}, \varepsilon_i) &= \frac{\exp(-\lambda_{it}) \lambda_{it}^{j_{it}}}{j_{it}!} \\ \lambda_{it} &= \exp(\mathbf{X}'_{it} \beta_0 + \varepsilon_i) \\ &= \exp(\mathbf{X}'_{it} \beta_0) \alpha_i, i = 1, \dots, n, t = 1, \dots, T \end{aligned}$$

where ε_i is a random variable and $\alpha_i = \exp(\varepsilon_i)$.

Models for Panel Data

The Pooled Poisson regression model

- We must assume that $E(\alpha_i | \mathbf{X}_{it})$ is a constant (normalized to 1)
- Based on this assumption we have $E(y_{it} | \mathbf{X}_{it}) = \exp(\mathbf{X}'_{it}\beta_0)$
- The *Poisson Quasi-logLikelihood* is given by (up to additive constants):

$$\log L(\beta) = \sum_{i=1}^n \sum_{t=1}^T [Y_{it}\mathbf{X}'_{it}\beta - \exp(\mathbf{X}'_{it}\beta)]$$

- The Poisson Quasi-logLikelihood estimator is consistent under mild assumptions.
- Inference must be based on a robust (to heteroskedasticity and dependence) covariance estimator.
- Inclusion of time dummies in the model is generally recommended.

Models for Panel Data

A more efficient estimator

- We require additional assumptions:
 - ① *independence* of the elements of $Y_i = (Y_{i1}, \dots, Y_{iT})$, conditional on ε_i and $\mathbf{X}_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{iT})'$;
 - ② the following distributional assumptions:
 - ① $\mathcal{P}(Y_{it} = j_{it} | \mathbf{X}_i, \varepsilon_i) = \exp(-\lambda_{it}) \lambda_{it}^{j_{it}} / j_{it}!$, where $\lambda_{it} = \exp(\mathbf{X}'_{it} \beta_0 + \varepsilon_i)$ (the *Poisson model*);
 - ② distribution of ε_i is *known* and *independent* of \mathbf{X}_i .

Remark: Assumption 2(a) implies that

$\mathcal{P}(Y_{it} = j_{it} | \mathbf{X}_i, \varepsilon_i) = \mathcal{P}(Y_{it} = j_{it} | \mathbf{X}_{it}, \varepsilon_i)$ and, consequently *strict-exogeneity* of the regressors: $E(Y_{it} | \mathbf{X}_{i1}, \dots, \mathbf{X}_{iT}, \varepsilon_i) = E(Y_{it} | \mathbf{X}_{it}, \varepsilon_i)$

Models for Panel Data

A more efficient estimator

- In this case, $L(\beta) = \prod_{i=1}^n L_i(\beta)$, where

$$L_i(\beta) = \int_{-\infty}^{+\infty} \left[\prod_{t=1}^T \frac{\exp(-\exp(\mathbf{X}'_{it}\beta)\alpha_i) (\exp(\mathbf{X}'_{it}\beta)\alpha_i)^{j_{it}}}{j_{it}!} \right] g(\alpha_i) d\alpha$$

- If $\alpha_i = \exp(\varepsilon_i)$ is assumed to have a *gamma distribution*, the model has a closed form based on the negative-binomial distribution.
- Often, it is assumed that α_i has a *log-normal distribution* (no closed form).
- Consistency depends, of course, on the validity of the distributional assumptions.

Models for Panel Data

The fixed effects estimator

- There is a consistent *fixed-effects* estimator for the Poisson model, that does not require independence between α_i and the regressors.
- Now it is assumed that for $i = 1, \dots, n, t = 1, \dots, T$ we have:

$$P(Y_{it} = j_{it} | \mathbf{X}_i, \varepsilon_i) = \frac{\exp(-\lambda_{it}) \lambda_{it}^{j_{it}}}{j_{it}!}$$
$$\lambda_{it} = \exp(\mathbf{X}'_{it} \beta_0 + \varepsilon_i) = \exp(\mathbf{X}'_{it} \beta_0) \alpha_i,$$

where ε_i is a random variable and $\alpha_i = \exp(\varepsilon_i)$. Recall that this assumption implies that $\mathcal{P}(Y_{it} = j_{it} | \mathbf{X}_i, \varepsilon_i) = \mathcal{P}(Y_{it} = j_{it} | \mathbf{X}_{it}, \varepsilon_i)$ and, consequently *strict-exogeneity* of the regressors:

$$E(Y_{it} | \mathbf{X}_{i1}, \dots, \mathbf{X}_{iT}, \varepsilon_i) = E(Y_{it} | \mathbf{X}_{it}, \varepsilon_i) = \lambda_{it}$$

- This estimator requires independence of the elements of $Y_i = (Y_{i1}, \dots, Y_{iT})'$, conditional on ε_i and \mathbf{X}_i ;
- By the additivity property of the Poisson distribution, we have that

$$\sum_{t=1}^T Y_{it} | \varepsilon_i, \mathbf{X}_i \sim \text{Poisson} \left(\sum_{t=1}^T \lambda_{it} \right).$$

Models for Panel Data

The fixed effects estimator

- It turns out that the distribution of Y_i conditional on \mathbf{X}_i , α_i and $\sum_{t=1}^T Y_{it}$ does not depend on α_i .
- Indeed, we have (for $j_i = (j_{i1}, \dots, j_{iT})'$):

$$\begin{aligned} & \mathcal{P} \left(Y_i = j_i \mid \mathbf{X}_i, \varepsilon_i, \sum_{t=1}^T Y_{it} = \sum_{t=1}^T j_{it} \right) \\ &= \frac{\left(\sum_{t=1}^T j_{it} \right)!}{\prod_{t=1}^T j_{it}!} \prod_{t=1}^T \left(\frac{\exp(\mathbf{X}'_{it} \beta_0)}{\sum_{t=1}^T \exp(\mathbf{X}'_{it} \beta_0)} \right)^{j_{it}}. \end{aligned}$$

- Write

$$p_t(\mathbf{X}_i, \beta_0) = \frac{\exp(\mathbf{X}'_{it} \beta_0)}{\sum_{t=1}^T \exp(\mathbf{X}'_{it} \beta_0)}.$$

Models for Panel Data

The fixed effects estimator

- Estimation is simple due to the the fact that the log-likelihood function (up to additive constants) is similar to that of the Conditional Logit model:

$$\log(L(\beta)) = \sum_{i=1}^n \sum_{t=1}^T Y_{it} \log(p_t(\mathbf{X}_i, \beta))$$

- Wooldridge (1999) shows that the estimator **is consistent** provided that $E(Y_{it} | \mathbf{X}_{i1}, \dots, \mathbf{X}_{iT}, \varepsilon_i)$ is correctly specified even if:
 - 1 Y_{it} is *not Poisson*.
 - 2 the elements of $Y_i = (Y_{i1}, \dots, Y_{iT})$ are *not independent*, conditional on α_i and \mathbf{X}_i .
- Naturally, if these assumptions do not hold, inference must be based on a *robust* (to heteroskedasticity and dependence) covariance matrix.