COUNT DATA AND OTHER NONNEGATIVE RESPONSES

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction
2. Poisson Regression
3. Negative Binomial Models
4. Hurdle Models
5. Binomial Regression
6. Endogenous Explanatory Variables
7. Panel Data

1. INTRODUCTION

- A count variable is one that takes on nonnegative integer values. In the leading case, there is no natural upper bound, so the support of y is $\{0, 1, 2, \ldots\}$. In other cases, there is an upper bound, and it can even change by individual: (n_i, y_i) is drawn with n_i a positive integer and then $y_i \in \{0, 1, \ldots, n_i\}$. (For example, n_i is number of employees and y_i is the number who participate in an optional pension plan.)

- Focus here is on count data, but the quasi-likelihood methods can be applied to any nonnegative response.

- For the most part, count data are analyzed from one of two perspectives:

(1) We are mainly interested in $E(y|\mathbf{x})$ – and so we want consistent estimators of the mean parameters without additional assumptions – but we would like our estimators to at least recognize the count nature of y and be efficient in some situations.

(2) We are interested in other features of $D(y|\mathbf{x})$, and so we use various models for $D(y|\mathbf{x})$ and apply MLE.

- For later: If the data are censored or truncated, MLE is independsable, which means we should choose a flexible model for $D(y|\mathbf{x})$.

- If we our interest is in $E(y|\mathbf{x})$ and we observe y over its entire range, then, as usual, a linear model might provide a good approximation to the average partial effects. At a minimum, the linear model results can be compared with APEs from other methods. Goodness-of-fit can also be compared across different models of conditional means.
- The drawbacks of a linear model are that it will not ensure $\hat{E}(y|\mathbf{x}) \geq 0$ for all relevant vectors $\mathbf{x}$ and it may not give sensible partial effects for extreme values of $\mathbf{x}$.

- We cannot use $\log(y)$ in interesting applications because y_i is typically zero for a nontrivial fraction of the observations. Sometimes $\log(1 + y)$ is used as the dependent variable in linear regression. But this transformation has several shortcomings.
- First, $\log(1) = 0$, so this transformation does not help with the pile-up-at-zero problem. $w \equiv \log(1 + y)$ is still a discrete, nonnegative variable.

- Second, even if we assume

$$E[\log(1+y)|\mathbf{x}] = \mathbf{x}\boldsymbol{\beta} \tag{1}$$

– not a great assumption because $\log(1+y) \geq 0$ – how do we interpret the β_j? We cannot "undo" the log:

$$E(y|\mathbf{x}) \neq \exp(\mathbf{x}\boldsymbol{\beta}) - 1 \tag{2}$$

(Note that the RHS could be negative for some values of **x**.) Generally, we cannot find $E(y|\mathbf{x})$, so we cannot find partial effects on $E(y|\mathbf{x})$.

- As an exercise, suppose you are willing to assume $\log(1+y) = \mathbf{x}\boldsymbol{\beta} + v$ where $D(v|\mathbf{x}) = D(v)$ and $E(v) = 0$. (Independence between v and $\mathbf{x}$ is questionable because $v \geq -\mathbf{x}\boldsymbol{\beta}$ is required.) Show that $E(y|\mathbf{x}) = \eta \exp(\mathbf{x}\boldsymbol{\beta}) - 1$ for some $\eta > 1$.
- Because y changes discretely, we cannot use the approximation that the change in logs is roughly the proportionate change in y, especially starting at $y = 0$. In other words, the β_j need not be good approximations to semi-elasticities or elasticities.

- One reason to use $\log(1 + y)$ rather than y in a linear regression is to guard against outliers (large counts). But we still face the problem of uncovering $E(y|\mathbf{x})$.
- A better solution is to model $E(y|\mathbf{x})$ directly. We will mainly use an exponential function: $E(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta})$ but other functional forms are possible.

2. POISSON REGRESSION

Setup and Estimation

• If we were to start with a distribution for $D(y|\mathbf{x})$, when y is an unbounded count variable, the Poisson is natural. Recall that the distribution is completely characterized by its mean.

• Let $\mu(\mathbf{x}) = E(y|\mathbf{x})$ where $y \in \{0, 1, 2, \ldots\}$ and $\mu(\cdot) > 0$. Then the conditional distribution is Poisson if the density is

$$f(y|\mathbf{x}) = \exp[-\mu(\mathbf{x})][\mu(\mathbf{x})]^y / y! \qquad (3)$$

where $y! = 1 \cdot 2 \cdot \cdot \cdot (y-1) \cdot y$ and $0! = 1$.

• If $m(\mathbf{x}, \boldsymbol{\beta})$ is the parametric model of the mean, then the model of the density is

$$f(y|\mathbf{x}; \boldsymbol{\beta}) = \exp[-m(\mathbf{x}, \boldsymbol{\beta})][m(\mathbf{x}, \boldsymbol{\beta})]^{y}/y! \tag{4}$$

or, with $m(\mathbf{x}, \boldsymbol{\beta}) = \exp(\mathbf{x}\boldsymbol{\beta})$,

$$f(y|\mathbf{x}; \boldsymbol{\beta}) = \exp[-\exp(\mathbf{x}\boldsymbol{\beta})]\exp(y\mathbf{x}\boldsymbol{\beta})/y! \tag{5}$$

• The exponential is by far the most popular, and can be made flexible by including nonlinear functions (such as logs, squares, and interactions) in **x**.

• The Poisson distribution is a member of the linear exponential family. So, for *any* $y \geq 0$ with $E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}_o)$, where $m(\mathbf{x}, \boldsymbol{\beta}) > 0$ for all $\mathbf{x}$ and $\boldsymbol{\beta}$, the Poisson QMLE is consistent for $\boldsymbol{\beta}_o$ regardless of arbitrary misspecification of other distributional features.

• It is easily seen that the score of the quasi-log likelihood has zero conditional mean when evaluated at $\boldsymbol{\beta}_o$:

$$\ell_i(\boldsymbol{\beta}) = -m(\mathbf{x}_i, \boldsymbol{\beta}) + y_i \log[m(\mathbf{x}_i, \boldsymbol{\beta})] - \log(y_i!) \tag{6}$$

$$\begin{aligned} \mathbf{s}_i(\boldsymbol{\beta}) &= -\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta})' + y_i \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta})'/m(\mathbf{x}_i, \boldsymbol{\beta}) \\ &= \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta})'[y_i - m(\mathbf{x}_i, \boldsymbol{\beta})]/m(\mathbf{x}_i, \boldsymbol{\beta}) \end{aligned} \tag{7}$$

- Therefore, when $E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}_o)$,

$$E[\mathbf{s}_i(\boldsymbol{\beta}_o)|\mathbf{x}_i] = \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o)'[E(y_i|\mathbf{x}_i) - m(\mathbf{x}_i, \boldsymbol{\beta}_o)]/m(\mathbf{x}_i, \boldsymbol{\beta}_o) = \mathbf{0}. \tag{8}$$

- In the exponential case, the score is particularly simple:

$$\mathbf{s}_i(\boldsymbol{\beta}) = \mathbf{x}_i'[y_i - \exp(\mathbf{x}_i\boldsymbol{\beta})] \tag{9}$$

an so the FOC is

$$\sum_{i=1}^{N} \mathbf{x}_i'[y_i - \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})] = \mathbf{0}.$$

- It follows that when $x_{i1} \equiv 1$ – by far the leading case – the residuals $\hat{u}_i \equiv y_i - \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})$ sum to zero.
- The fitted values are $\hat{y}_i = \exp(\mathbf{x}_i\hat{\boldsymbol{\beta}})$, and so in the exponential case with a constant, the average of the fitted values equals $\bar{y}$. (Recall this also holds for a linear model estimated by OLS and a logit model estimated by Bernoulli MLE.)

• The Hessian in the exponential case is also simple and negative semi-definite for any value of $\boldsymbol{\beta}$:

$$\mathbf{H}_i(\boldsymbol{\beta}) = -\exp(\mathbf{x}_i\boldsymbol{\beta})\mathbf{x}_i'\mathbf{x}_i \tag{10}$$

• For other conditional mean choices, the Hessian depends on y_i and is not always negative semi-definite.

• From our MLE notation, in general,

$$\mathbf{A}(\mathbf{x}_i,\boldsymbol{\beta}_o) = -E[\mathbf{H}_i(\boldsymbol{\beta}_o)|\mathbf{x}_i] = \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i,\boldsymbol{\beta}_o)'\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i,\boldsymbol{\beta}_o)/m(\mathbf{x}_i,\boldsymbol{\beta}_o) \tag{11}$$

$$\mathbf{A}_o = E[\mathbf{A}(\mathbf{x}_i,\boldsymbol{\beta}_o)] = -E[\mathbf{H}_i(\boldsymbol{\beta}_o)]. \tag{12}$$

- The inner part of the sandwich is

$$\mathbf{B}_o = E[\mathbf{s}_i(\boldsymbol{\beta}_o)\mathbf{s}_i(\boldsymbol{\beta}_o)'] = E[u_i^2 \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o)' \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o) / m(\mathbf{x}_i, \boldsymbol{\beta}_o)^2] \tag{13}$$

$$= E[\tau_o^2(\mathbf{x}_i) \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o)' \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i, \boldsymbol{\beta}_o) / m(\mathbf{x}_i, \boldsymbol{\beta}_o)^2] \tag{14}$$

where

$$\tau_o^2(\mathbf{x}_i) \equiv E(u_i^2 | \mathbf{x}_i) = Var(y_i | \mathbf{x}_i) \tag{15}$$

is the *true* variance of y_i given $\mathbf{x}_i$.

- Robust variance matrix if $m(\mathbf{x}_i, \boldsymbol{\beta})$ is correctly specified:

$$\left(\sum_{i=1}^{N} \nabla_{\boldsymbol{\beta}}\hat{m}_i' \nabla_{\boldsymbol{\beta}}\hat{m}_i / \hat{m}_i\right)^{-1} \left(\sum_{i=1}^{N} \hat{u}_i^2 \nabla_{\boldsymbol{\beta}}\hat{m}_i' \nabla_{\boldsymbol{\beta}}\hat{m}_i / \hat{m}_i^2\right) \cdot \left(\sum_{i=1}^{N} \nabla_{\boldsymbol{\beta}}\hat{m}_i' \nabla_{\boldsymbol{\beta}}\hat{m}_i / \hat{m}_i\right)^{-1} \tag{16}$$

or with multiplication by a degrees-of-freedom adjustment, $N/(N - P)$.

- The usual Poisson MLE variance matrix estimator,

$$\left(\sum_{i=1}^{N} \nabla_{\boldsymbol{\beta}} \hat{m}_i' \nabla_{\boldsymbol{\beta}} \hat{m}_i / \hat{m}_i\right)^{-1} \tag{17}$$

is valid only under the Poisson variance assumption:

$$Var(y|\mathbf{x}) = E(y|\mathbf{x}). \tag{18}$$

(However, other features of the Poisson distribution may be misspecified.)

- The GLM variance assumption for the Poisson QMLE is

$$Var(y|\mathbf{x}) = \sigma_o^2 E(y|\mathbf{x}) \tag{19}$$

for some $\sigma_o^2 > 0$. The case of overdispersion (relative to the Poisson assumption), $\sigma_o^2 > 1$, is common in applications. But underdispersion, $\sigma_o^2 < 1$, can occur.

- The Pearson residuals are $\hat{u}_i/\sqrt{\hat{m}_i}$ and the usual estimate of σ_o^2 is

$$\hat{\sigma}^2 = (N - P)^{-1} \sum_{i=1}^{N} \hat{u}_i^2/\hat{m}_i. \tag{20}$$

• The GLM asymptotic variance estimate is

$$\hat{\sigma}^2\left(\sum_{i=1}^{N}\nabla_{\boldsymbol{\beta}}\hat{m}_i'\nabla_{\boldsymbol{\beta}}\hat{m}_i/\hat{m}_i\right)^{-1}. \tag{21}$$

The GLM standard errors are the MLE standard errors multiplied by $\hat{\sigma}$ (and often $\hat{\sigma} > 1$).

• For multiple restrictions, Wald test is easy to compute using any of the variance matrix estimates [preferably the fully robust form in (16).]

- Under the GLM variance assumption a quasi-LR statistic is justified:

$$QLR = 2(\mathcal{L}_{ur} - \mathcal{L}_r)/\hat{\sigma}^2, \tag{22}$$

where $\hat{\sigma}^2$ is from the unrestricted model, has an asymptotic χ^2_Q distribution under H_0.

- With overdispersion it is clear that the usual LR statistic, $2(\mathcal{L}_{ur} - \mathcal{L}_r)$, will be too large on average.

Estimation and Intepretation with an Exponential Mean

- The Stata command for Poisson regression with an exponential function is one of the following (in order of decreasing robustness for inference):

```
glm y x1 ... xK, fam(poisson) link(log) robust
glm y x1 ... xK, fam(poisson) link(log)
scale(x2)
glm y x1 ... xK, fam(poisson) link(log)
```

- The option "`scale(x2)`" means to use the Pearson estimate of σ^2 and to report those standard errors.
- The first command is equivalent to

```
poisson y x1 ... xK, robust
```

and the last `glm` command is equivalent to

```
poisson y x1 ... xK
```

which means that the Poisson distribution (actually, the variance assumption) is taken to be correct in inference.

• If $E(y|\mathbf{x}) = \exp(\beta_1 + \beta_2 x_2 + \ldots + \beta_K x_K)$ then

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \beta_j \exp(\mathbf{x}\boldsymbol{\beta}). \tag{23}$$

• The APE is consistently estimated as

$$\hat{\beta}_j \left[N^{-1} \sum_{i=1}^{N} \exp(\mathbf{x}_i \hat{\boldsymbol{\beta}}) \right] = \hat{\beta}_j \bar{y} \tag{24}$$

because as shown earlier (from the first-order condition), $\bar{\hat{y}} = \bar{y}$.

• Consequently for (roughly) continuous x_j, the Poisson coefficients multiplied by the sample average is comparable to the OLS estimates $\hat{\gamma}_j$ from y_i on $1, x_{i2}, \ldots, x_{iK}$.

• For most purposes, the $\hat{\beta}_j$ are more interesting than the scaled coefficients because

$$\beta_j = \frac{\partial \log E(y|\mathbf{x})}{\partial x_j}, \tag{25}$$

so $100\beta_j$ is roughly the ceteris paribus percentage change in $E(y|\mathbf{x})$ when $\Delta x_j = 1$.

• If $x_j = \log(z_j)$, then $\hat{\beta}_j$ is the estimated elasticity of $E(y|\mathbf{x})$ with respect to z_j.

• Should compute the fully robust standard errors. Also use the GLM version of the standard errors to get an estimate of σ.

- Other functional forms are easily accomodated. For example, if

$$E(y|\mathbf{z}) = \exp(\beta_1 + \beta_2 z_1 + \beta_3 z_1^2 + \mathbf{z}_2\boldsymbol{\beta}_4), \tag{26}$$

then

$$\frac{\partial \log E(y|\mathbf{z})}{\partial z_1} = \beta_2 + 2\beta_3 z_1 \tag{27}$$

and so $100(\beta_2 + 2\beta_3 z_1)$ is roughly the percentage change in $E(y|\mathbf{z})$ when $\Delta z_1 = 1$. If β_2 and β_3 have opposite signs, the turning point in the quadratic is $z_1^* = |\beta_2/(2\beta_3)|$.

- Goodness-of-Fit: To measure how well the mean predicts y_i, can use the squared correlation between y_i and $\hat{y}_i$ as an R-squared. (Basing it on Poisson log likelihood is too restrictive.) Or, use $1 - SSR/SST$.

- There are ways to test the Poisson variance assumption and the GLM variance assumption; see Chapter 18. Why test? First, to see whether a more efficient estimator might be available (see next section). Second, to see whether, if the full distribution is of interest, whether the Poisson is obviously deficient.

- If we want to estimate, say, $P(y > j|\mathbf{x})$, then we should use a flexible distribution, not the Poisson.

Efficiency of Poisson QMLE

• A loose motivation for using Poisson QMLE for count responses is that the Poisson distribution is traditionally used for modeling count data. But the assumption that the variance equals the mean is too restrictive across all applications.

• Nevertheless, it turns out the Poisson QMLE is the efficient estimator in a certain class of estimators under the GLM assumption,

$$Var(y|\mathbf{x}) = \sigma_o^2 E(y|\mathbf{x}) \tag{28}$$

• Precisely, consider any $\sqrt{N}$-asymptotically normal estimator that is $\sqrt{N}$-consistent under only the conditional mean assumption, assumption

$$E(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\beta}_o) \text{ for some } \boldsymbol{\beta}_o \in \mathcal{B}. \tag{29}$$

Then the Poisson QMLE has a smaller asymptotic variance if the Poisson GLM variance assumption holds.

• It is important to understand that there are more efficient estimators than the Poisson QMLE if we use the mean specification *along with*

$$Var(y|\mathbf{x}) = \sigma_o^2 m(\mathbf{x}, \boldsymbol{\beta}_o) \tag{30}$$

because this adds additional information useful for estimating $\boldsymbol{\beta}_o$. But then such estimators generally would be inconsistent if the GLM variance assumption does not hold.

• Generally one needs to be specific about the comparison class of estimators when discussing efficiency.

3. NEGATIVE BINOMIAL MODELS

- A useful alternative to Poisson regression is the class of Negative Binomial regression models.
- Two useful approaches are what have been dubbed the NegBing I and NegBin II models (Cameron and Trivedi (1986)).
- It is important to distinguish between cases where the full distribution is assumed correctly specified, with parameters estimated by MLE, and a QMLE two-step framework. The latter only requires correct specification of the conditional mean.

MLE Estimation of NegBin II

• One way to derive NegBin II is to add an unobservable to the standard Poisson model:

$$y_i|(\mathbf{x}_i, c_i) \sim Poisson[c_i m(\mathbf{x}_i, \boldsymbol{\beta})] \tag{31}$$

$$c_i|\mathbf{x}_i \sim Gamma(\eta^{-2}, \eta^{-2}) \tag{32}$$

$$E(c_i) = 1,\ Var(c_i) = \eta^2 \tag{33}$$

• Adding c_i is one device to obtain a more flexible distribution for $D(y_i|\mathbf{x}_i)$. We need the density of y_i conditional on $\mathbf{x}_i$.

- Can show the log-likelihood function for each i is

$$\ell_i(\boldsymbol{\beta}, \eta^2) = \eta^{-2} \log\left[\frac{\eta^{-2}}{\eta^{-2} + m(\mathbf{x}_i, \boldsymbol{\beta})} \right] + y_i \log\left[\frac{m(\mathbf{x}_i, \boldsymbol{\beta})}{\eta^{-2} + m(\mathbf{x}_i, \boldsymbol{\beta})} \right] \qquad (34)$$
$$+ \log[\Gamma(y_i + \eta^{-2})/\Gamma(\eta^{-2})]$$

where $\Gamma(\bullet)$ is the gamma function: for $r > 0$,

$$\Gamma(r) = \int_0^\infty u^{r-1} \exp(-u) du \qquad (35)$$

• MLE is relatively straightfoward. Technically, it has no known robustness properties for estimating $\boldsymbol{\beta}$ if the density is misspecified, but it often seems to give similar estimates to the Poisson QMLE.

• In Stata with an exponential mean:

```
nbreg y x1 ... xK, disp(mean)
```

• Stata labels $\alpha = \eta^2$.

• NegBin II implies that there *must* be overdispersion, and it must be a function of the mean.

• To see this, we can find the mean and variance of y_i conditional only on $\mathbf{x}_i$ (and we do not even need to assume a Gamma distribution for c_i):

$$E(y_i|\mathbf{x}_i) = E[E(y_i|\mathbf{x}_i, c_i)|\mathbf{x}_i] = E(c_i|\mathbf{x}_i)m(\mathbf{x}_i, \boldsymbol{\beta}) = m(\mathbf{x}_i, \boldsymbol{\beta}) \tag{36}$$

and

$$\begin{aligned} Var(y_i|\mathbf{x}_i) &= E[Var(y_i|\mathbf{x}_i, c_i)|\mathbf{x}_i] + Var[E(y_i|\mathbf{x}_i, c_i)|\mathbf{x}_i] && (37) \\ &= E[c_i m(\mathbf{x}_i, \boldsymbol{\beta})|\mathbf{x}_i] + Var[c_i m(\mathbf{x}_i, \boldsymbol{\beta})|\mathbf{x}_i] \\ &= E(c_i|\mathbf{x}_i)m(\mathbf{x}_i, \boldsymbol{\beta}) + Var(c_i|\mathbf{x}_i)[m(\mathbf{x}_i, \boldsymbol{\beta})]^2 \\ &= m(\mathbf{x}_i, \boldsymbol{\beta}) + \eta^2[m(\mathbf{x}_i, \boldsymbol{\beta})]^2 && (38) \end{aligned}$$

• So, the variance for NegBin II does not allow underdisperion because $Var(y_i|\mathbf{x}_i) \geq E(y_i|\mathbf{x}_i)$, with strict inequality if $\eta^2 > 0$. Plus, it rules out the GLM variance assumption, $Var(y|\mathbf{x}) = \sigma^2 m(\mathbf{x}, \boldsymbol{\beta})$, unless $\eta^2 = 0$ and $\sigma^2 = 1$.

• Generally, the dispersion of $D(y_i|\mathbf{x}_i)$ is defined as

$$Disperson(\mathbf{x}_i) = \frac{Var(y_i|\mathbf{x}_i)}{E(y_i|\mathbf{x}_i)} \tag{39}$$

and is a function of $\mathbf{x}_i$.

• For the GLM variance assumption, the dispersion is constant and equal to σ^2. For NegBin II, the dispersion is linear in the mean:

$$Disperson(\mathbf{x}_i) = 1 + \eta^2 m(\mathbf{x}_i, \boldsymbol{\beta}) \tag{40}$$

Two-Step QMLE for NegBin II

• The variance expression in (38) suggests a two-step approach. Estimate $\boldsymbol{\beta}$ by, say, Poisson regression. Letting $u_i = y_i - E(y_i|\mathbf{x}_i)$, note that $E(u_i^2|\mathbf{x}_i) = m(\mathbf{x}_i,\boldsymbol{\beta}) + \eta^2[m(\mathbf{x}_i,\boldsymbol{\beta})]^2$. Therefore,

$$\eta^2 = E\left[\frac{u_i^2 - m(\mathbf{x}_i,\boldsymbol{\beta})}{[m(\mathbf{x}_i,\boldsymbol{\beta})]^2}\right], \tag{41}$$

so, letting $\breve{u}_i$ be the Poisson residuals, define

$$\hat{\eta}^2 = N^{-1}\sum_{i=1}^{N}\frac{[\breve{u}_i^2 - m(\mathbf{x}_i,\breve{\boldsymbol{\beta}})]}{[m(\mathbf{x}_i,\breve{\boldsymbol{\beta}})]^2} \tag{42}$$

• Importantly, for fixed η^2, say $\bar{\eta}^2$, the log likelihood

$$l_i(\boldsymbol{\beta}) = \bar{\eta}^{-2} \log\left[\frac{\bar{\eta}^{-2}}{\bar{\eta}^{-2} + m(\mathbf{x}_i, \boldsymbol{\beta})} \right] + y_i \log\left[\frac{m(\mathbf{x}_i, \boldsymbol{\beta})}{\bar{\eta}^{-2} + m(\mathbf{x}_i, \boldsymbol{\beta})} \right] \tag{43}$$

is in the LEF. (Term depending on gamma function no longer needed.)

• Exercise: Show that the score evaluated at "true" $\boldsymbol{\beta}$ (call it $\boldsymbol{\beta}_o$) has zero conditional mean whenever the mean is correctly specified (with arbitrary misspecification of other aspects of the distribution).

• Whether or not the variance function is correct,

$\hat{\eta}^2 \xrightarrow{p} E\{[u_i^2 - m(\mathbf{x}_i, \boldsymbol{\beta})]/[m(\mathbf{x}_i, \boldsymbol{\beta})]^2\}$. The two-step QMLE is like using a fixed value of η^2. In effect, we just take

$$\bar{\eta}^2 \equiv plim_{N\to\infty}(\hat{\eta}^2)$$

and apply the LEF results.

• Can also show that plugging in $\hat{\eta}^2$ for η^2 does not affect the $\sqrt{N}$-asymptotic distribution of the 2SQMLE. (Verify that key result holds for two-step M-estimation.)

- The two-step NeBin II QMLE is asymptotically equivalent to weighted NLS using estimated variance function $m(\mathbf{x}_i, \breve{\boldsymbol{\beta}}) + \hat{\eta}^2[m(\mathbf{x}_i, \breve{\boldsymbol{\beta}})]^2$.
- Can also fix η^2 at given value, such as $\eta^2 = 1$ for Geometric distribution.
- Using two or more different QMLEs in the LEF with the same mean function (Poisson, Geometric, two-step NegBin, NLS) can provide evidence of conditional mean misspecification.

MLE Estimation of NegBin I

- A different parameterization of the NegBin distribution, called NegBin I, has the same mean function $m(\mathbf{x}_i, \boldsymbol{\beta})$ but variance

$$(1 + \eta^2)m(\mathbf{x}_i, \boldsymbol{\beta}) \tag{44}$$

so the dispersion is contant, $1 + \eta^2$.

- Like NegBin II, NegBin I allows only overdispersion.

• If, for example, the mean is correctly specified and $Var(y_i|\mathbf{x}_i) = \sigma^2 m(\mathbf{x}_i,\boldsymbol{\beta})$ with $\sigma^2 < 1$, neither NegBin I nor NegBin II estimated by MLE is consistent (although the two-step NegBin II estimator would be). A practical issue is how much the estimates differ from Poisson regression.

- In Stata with an exponential mean, NegBin I is estimated as

```
nbreg y x1 x2 ... xK, disp(const)
```

- The reported estimates are interpreted just as with Poisson regression, since the mean is exponential (just like NegBin II). Stata uses $\alpha = \eta^2$.
- For NegBin I and II, need to average $\exp(\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_K x_{iK})$ to get APEs. The average of fitted values is not $\bar{y}$.

- Goodness-of-Fit just like in Poisson case. Use squared correlation between actual and fitted values or an SSR version.
- Remember, only OLS (or NLS, if we used a nonlinear mean function) chooses the parameters to maximize the R-squared. The other estimators maximize the (quasi-) log likelihood function.

- Example: GROGGER.DTA. Data on men in California born in 1960 or 1961. All had been arrested at least once previously. Data from 1986.
- *narr*86 is the number of times the man was arrested during 1986. Are there deterrent effects to prior convictions or sentence lengths? What is the effect of incarceration? What about labor market opportunities?

```
. des

Contains data from crime1.dta
  obs:         2,725
 vars:            19                          6 Nov 1996 10:54
 size:       155,325 (98.5% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
narr86          byte   %9.0g                  # times arrested, 1986
nfarr86         byte   %9.0g                  # felony arrests, 1986
nparr86         byte   %9.0g                  # property crme arr., 1986
pcnv            float  %9.0g                  proportion of prior convictions
avgsen          float  %9.0g                  avg sentence length, mos.
tottime         float  %9.0g                  time in prison since 18 (mos.)
ptime86         byte   %9.0g                  mos. in prison during 1986
qemp86          float  %9.0g                  # quarters employed, 1986
inc86           float  %9.0g                  legal income, 1986, $100s
durat           float  %9.0g                  recent unemp duration
black           byte   %9.0g                  =1 if black
hispan          byte   %9.0g                  =1 if Hispanic
born60          byte   %9.0g                  =1 if born in 1960
pcnvsq          float  %9.0g                  pcnv^2
pt86sq          int    %9.0g                  ptime86^2
inc86sq         float  %9.0g                  inc86^2
```

```
. tab narr86

   # times |
  arrested, |
       1986 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,970       72.29       72.29
          1 |        559       20.51       92.81
          2 |        121        4.44       97.25
          3 |         42        1.54       98.79
          4 |         12        0.44       99.23
          5 |         13        0.48       99.71
          6 |          4        0.15       99.85
          7 |          1        0.04       99.89
          9 |          1        0.04       99.93
         10 |          1        0.04       99.96
         12 |          1        0.04      100.00
------------+-----------------------------------
      Total |      2,725      100.00

. sum pcnv avgsen ptime86 inc86

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        pcnv |      2725    .3577872     .395192          0          1
      avgsen |      2725    .6322936    3.508031          0       59.2
     ptime86 |      2725     .387156    1.950051          0         12
       inc86 |      2725    54.96705    66.62721          0        541
```

```
. reg narr86 pcnv avgsen tottime ptime86 inc86 qemp86 black hispan born60, robust

Linear regression                                      Number of obs =    2725
                                                       F(  9,  2715) =   25.93
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0725
                                                       Root MSE      =  .82873

------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |   -.131886   .0335876    -3.93   0.000    -.1977458   -.0660262
      avgsen |  -.0113316   .0141409    -0.80   0.423    -.0390595    .0163963
     tottime |   .0120693   .0131776     0.92   0.360    -.0137699    .0379084
     ptime86 |  -.0408735   .0067985    -6.01   0.000    -.0542043   -.0275426
       inc86 |  -.0014617   .0002289    -6.38   0.000    -.0019106   -.0010128
      qemp86 |  -.0513099    .014205    -3.61   0.000    -.0791636   -.0234562
       black |   .3270097   .0584381     5.60   0.000     .2124221    .4415973
      hispan |   .1938094   .0401625     4.83   0.000     .1150572    .2725616
      born60 |   -.022465    .032094    -0.70   0.484    -.0853961    .0404661
       _cons |    .576566   .0426021    13.53   0.000     .4930302    .6601019
------------------------------------------------------------------------------
```

```
. glm narr86 pcnv avgsen tottime ptime86 inc86 qemp86 black hispan born60,
      fam(poisson)

Generalized linear models                          No. of obs      =      2725
Optimization     : ML                              Residual df     =      2715
                                                   Scale parameter =         1
Deviance         =  2822.184873                    (1/df) Deviance =  1.039479
Pearson          =  4118.079859                    (1/df) Pearson  =  1.516788

Variance function: V(u) = u                        [Poisson]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  1.657806
Log likelihood   = -2248.761092                    BIC             = -18654.07

------------------------------------------------------------------------------
             |                 OIM
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.4015713   .0849712    -4.73   0.000    -.5681117   -.2350308
      avgsen |  -.0237723    .019946    -1.19   0.233    -.0628658    .0153212
     tottime |   .0244904   .0147504     1.66   0.097    -.0044199    .0534006
     ptime86 |  -.0985584   .0206946    -4.76   0.000    -.1391192   -.0579977
       inc86 |  -.0080807    .001041    -7.76   0.000     -.010121   -.0060404
      qemp86 |  -.0380187   .0290242    -1.31   0.190    -.0949051    .0188677
       black |   .6608376   .0738342     8.95   0.000     .5161252      .80555
      hispan |   .4998133   .0739267     6.76   0.000     .3549196     .644707
      born60 |  -.0510286   .0640518    -0.80   0.426    -.1765678    .0745106
       _cons |  -.5995888   .0672501    -8.92   0.000    -.7313966    -.467781
------------------------------------------------------------------------------
```

```
. glm narr86 pcnv avgsen tottime ptime86 inc86 qemp86 black hispan born60,
      fam(poisson) robust

Generalized linear models                          No. of obs      =       2725
Optimization     : ML                              Residual df     =       2715
                                                   Scale parameter =          1
Deviance         =  2822.184873                    (1/df) Deviance =   1.039479
Pearson          =  4118.079859                    (1/df) Pearson  =   1.516788

                                                   AIC             =   1.657806
Log pseudolikelihood = -2248.761092                BIC             =  -18654.07

------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.4015713   .1011619    -3.97   0.000    -.5998449   -.2032976
      avgsen |  -.0237723   .0236078    -1.01   0.314    -.0700427    .0224981
     tottime |   .0244904   .0205023     1.19   0.232    -.0156934    .0646741
     ptime86 |  -.0985584   .0223035    -4.42   0.000    -.1422724   -.0548445
       inc86 |  -.0080807   .0012276    -6.58   0.000    -.0104867   -.0056747
      qemp86 |  -.0380187   .0341509    -1.11   0.266    -.1049532    .0289158
       black |   .6608376   .0994572     6.64   0.000     .4659051      .85577
      hispan |   .4998133   .0923874     5.41   0.000     .3187374    .6808892
      born60 |  -.0510286   .0811403    -0.63   0.529    -.2100606    .1080034
       _cons |  -.5995888   .0893463    -6.71   0.000    -.7747044   -.4244732
------------------------------------------------------------------------------
```

```
. glm narr86 pcnv avgsen tottime ptime86 inc86 qemp86 black hispan born60,
      fam(poisson) sca(x2)

Generalized linear models                          No. of obs      =       2725
Optimization     : ML                              Residual df     =       2715
                                                   Scale parameter =          1
Deviance         =  2822.184873                    (1/df) Deviance =   1.039479
Pearson          =  4118.079859                    (1/df) Pearson  =   1.516788

                                                   AIC             =   1.657806
Log likelihood   = -2248.761092                    BIC             =  -18654.07

------------------------------------------------------------------------------
             |                 OIM
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.4015713   .1046488    -3.84   0.000    -.6066791   -.1964634
      avgsen |  -.0237723   .0245651    -0.97   0.333    -.0719191    .0243745
     tottime |   .0244904   .0181663     1.35   0.178    -.0111149    .0600957
     ptime86 |  -.0985584   .0254871    -3.87   0.000    -.1485122   -.0486047
       inc86 |  -.0080807   .0012821    -6.30   0.000    -.0105935   -.0055679
      qemp86 |  -.0380187   .0357456    -1.06   0.288    -.1080788    .0320414
       black |   .6608376   .0909327     7.27   0.000     .4826127    .8390624
      hispan |   .4998133   .0910466     5.49   0.000     .3213652    .6782614
      born60 |  -.0510286   .0788849    -0.65   0.518    -.2056401     .103583
       _cons |  -.5995888   .0828238    -7.24   0.000    -.7619206    -.437257
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion)
```

```
. di sqrt(1.5168)
1.2315843

. * The estimate of sigma is about 1.232.

. predict narr86h_p
(option mu assumed; predicted mean narr86)

. corr narr86 narr86h_p
(obs=2725)

             |   narr86 narr86~p
-------------+------------------
      narr86 |   1.0000
   narr86h_p |   0.2775   1.0000

. di .2775^2
.07700625

. * Somewhat better fit than linear model.
```

```
. * Average marginal (or partial) effects:

. margeff

Average partial effects after glm
      y  = log(narr86)

------------------------------------------------------------------------------
    variable |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.1623973   .0427474    -3.80   0.000    -.2461807    -.078614
      avgsen |  -.0096136   .0099407    -0.97   0.333    -.0290969    .0098697
     tottime |    .009904   .0073557     1.35   0.178    -.0045129     .024321
     ptime86 |   -.039922   .0104625    -3.82   0.000    -.0604282   -.0194158
       inc86 |  -.0032679   .0005325    -6.14   0.000    -.0043115   -.0022243
      qemp86 |  -.0153749    .014467    -1.06   0.288    -.0437298    .0129799
       black |   .3278389   .0601313     5.45   0.000     .2099838    .4456941
      hispan |   .2349835   .0535485     4.39   0.000     .1300302    .3399367
      born60 |   -.020474    .030841    -0.66   0.507    -.0809213    .0399732
------------------------------------------------------------------------------
```

```
. * Marginal effects at averages:

. mfx

Marginal effects after glm
      y  = predicted mean narr86 (predict)
         =  .32918187
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
    pcnv |    -.13219      .03412   -3.87   0.000   -.19907  -.06531   .357787
  avgsen |  -.0078254      .00808   -0.97   0.333   -.02367  .008019   .632294
 tottime |   .0080618      .00598    1.35   0.177  -.003655  .019779   .838752
 ptime86 |  -.0324437      .00835   -3.89   0.000    -.0488 -.016087   .387156
   inc86 |    -.00266      .00039   -6.81   0.000  -.003426 -.001894    54.967
  qemp86 |  -.0125151      .01182   -1.06   0.290  -.035691  .010661   2.30903
   black*|     .27712      .04734    5.85   0.000   .184332  .369908   .161101
  hispan*|   .1914481      .03993    4.79   0.000   .113184  .269713   .217615
  born60*|  -.0166821      .02561   -0.65   0.515   -.06688  .033516   .362569
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

```
. * Now estimate NegBin II model:

. nbreg narr86 pcnv avgsen tottime ptime86 inc86 qemp86 black hispan born60,
        disp(mean)

Negative binomial regression                      Number of obs   =       2725
                                                  LR chi2(9)      =     266.12
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -2157.628                        Pseudo R2       =     0.0581

------------------------------------------------------------------------------
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.4770963   .1033295    -4.62   0.000    -.6796183   -.2745743
      avgsen |  -.0173385   .0261171    -0.66   0.507    -.0685272    .0338501
     tottime |   .0197394   .0192325     1.03   0.305    -.0179557    .0574344
     ptime86 |  -.1073997    .025074    -4.28   0.000    -.1565439   -.0582555
       inc86 |  -.0077126   .0011465    -6.73   0.000    -.0099596   -.0054656
      qemp86 |  -.0504884   .0351857    -1.43   0.151    -.1194511    .0184743
       black |   .6560406   .0923594     7.10   0.000     .4750195    .8370617
      hispan |   .5048465   .0895663     5.64   0.000     .3292998    .6803932
      born60 |   -.046412   .0776384    -0.60   0.550    -.1985804    .1057564
       _cons |  -.5637368   .0827121    -6.82   0.000    -.7258495   -.4016242
-------------+----------------------------------------------------------------
    /lnalpha |  -.0738912   .1177617                     -.3046999    .1569175
-------------+----------------------------------------------------------------
       alpha |   .9287728   .1093739                      .7373446    1.169899
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) =  182.27 Prob>=chibar2 = 0.000
```

```
. margeff

Average partial effects after nbreg
      y  = E(narr86) (expected number of counts)

------------------------------------------------------------------------------
    variable |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.1933489   .0429863    -4.50   0.000    -.2776004   -.1090974
      avgsen |  -.0070266    .010587    -0.66   0.507    -.0277767    .0137235
     tottime |   .0079996   .0078007     1.03   0.305    -.0072894    .0232886
     ptime86 |  -.0436086    .010438    -4.18   0.000    -.0640668   -.0231504
       inc86 |  -.0031256   .0004821    -6.48   0.000    -.0040705   -.0021807
      qemp86 |   -.020461   .0143138    -1.43   0.153    -.0485155    .0075935
       black |   .3256315    .061895     5.26   0.000     .2043195    .4469436
      hispan |   .2382802   .0535439     4.45   0.000     .1333361    .3432244
      born60 |  -.0186745   .0305162    -0.61   0.541     -.078485    .0411361
------------------------------------------------------------------------------
```

```
. predict narr86h_nb2
(option n assumed; predicted number of events)

. corr narr86 narr86h_nb2
(obs=2725)

             |   narr86 narr86~2
-------------+------------------
      narr86 |   1.0000
 narr86h_nb2 |   0.2735   1.0000

. di .2735^2
.07480225

. corr narr86h_p narr86h_nb2
(obs=2725)

             | narr86~p narr86~2
-------------+------------------
   narr86h_p |   1.0000
 narr86h_nb2 |   0.9982   1.0000

.* Compute proportionate change in mean if pcnv increases by .1:

. di .1*-.4015713
-.04015713
```

```
. * NLS with exponential mean:

. glm narr86 pcnv avgsen tottime ptime86 inc86 qemp86 black hispan born60,
      fam(normal) link(log) robust

Generalized linear models                          No. of obs      =       2725
Optimization     : ML                              Residual df     =       2715
                                                   Scale parameter =   .6812465
Deviance         =  1849.584158                    (1/df) Deviance =   .6812465
Pearson          =  1849.584158                    (1/df) Pearson  =   .6812465

                                                   AIC             =   2.457709
Log pseudolikelihood = -3338.628406                BIC             =  -19626.67

------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.1869304   .1079617    -1.73   0.083    -.3985315    .0246707
      avgsen |  -.0350896   .0218319    -1.61   0.108    -.0778793    .0077001
     tottime |     .03202   .0209271     1.53   0.126    -.0089963    .0730363
     ptime86 |  -.0783047   .0219865    -3.56   0.000    -.1213974   -.035212
       inc86 |  -.0107211   .0018603    -5.76   0.000    -.0143672   -.0070749
      qemp86 |   .0051333   .0375761     0.14   0.891    -.0685146    .0787811
       black |   .6837919   .1104365     6.19   0.000     .4673403    .9002434
      hispan |   .5239022   .1075846     4.87   0.000     .3130403    .7347641
      born60 |  -.0533576   .0973706    -0.55   0.584    -.2442006    .1374853
       _cons |  -.6944945   .0945734    -7.34   0.000     -.879855   -.509134
------------------------------------------------------------------------------
```

```
. margeff

Average partial effects after glm
      y  = log(narr86)

------------------------------------------------------------------------------
    variable |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |  -.0750057   .0367633    -2.04   0.041    -.1470604    -.002951
      avgsen |  -.0140797   .0071365    -1.97   0.049     -.028067   -.0000924
    tottime |    .012848   .0049705     2.58   0.010      .003106      .02259
    ptime86 |  -.0314518    .008387    -3.75   0.000      -.04789   -.0150136
       inc86 |  -.0043018   .0008093    -5.32   0.000    -.0058879   -.0027157
      qemp86 |   .0020597   .0141107     0.15   0.884    -.0255967    .0297162
       black |   .3389282   .0523359     6.48   0.000     .2363517    .4415046
      hispan |   .2459323   .0501628     4.90   0.000      .147615    .3442495
      born60 |  -.0212337   .0278328    -0.76   0.446     -.075785    .0333176
------------------------------------------------------------------------------

. * The big differences in Poisson and NLS suggests functional form
. * misspecification of the conditional mean (not variance!)
. * Adding quadratics in pcinv, ptime86, and inc86 shows they are jointly
. * significant but give some odd turning points.
```

```
. * NLS minimizes the SSR, which is not the same as maximizing the correlation
. * between y and yhat. Still, NLS might actually fit the mean better (and it
. * does):

. predict narr86h_nls
(option mu assumed; predicted mean narr86)

. corr narr86 narr86h_nls
(obs=2725)

             |   narr86 narr86~s
-------------+------------------
      narr86 |   1.0000
 narr86h_nls |   0.2829   1.0000


. di .2829^2
.08003241
```

```
. * Now use Poisson QMLE to test for joint significance of quadratics in some
. * variables:

. glm narr86 pcnv avgsen tottime ptime86 inc86 qemp86 black hispan born60
      pcnvsq pt86sq inc86sq, fam(poisson) robust


------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |   1.153133   .3721125     3.10   0.002     .4238059     1.88246
      avgsen |   -.025572   .0280475    -0.91   0.362     -.080544       .0294
     tottime |   .0121518   .0231558     0.52   0.600    -.0332328    .0575364
     ptime86 |   .6836811   .0975803     7.01   0.000     .4924273    .8749349
       inc86 |  -.0120712   .0017862    -6.76   0.000    -.0155721   -.0085703
      qemp86 |   .0230132   .0362222     0.64   0.525     -.047981    .0940073
       black |   .5913914   .0993506     5.95   0.000     .3966677    .7861151
      hispan |   .4220377   .0924178     4.57   0.000     .2409021    .6031732
      born60 |  -.0929425   .0799599    -1.16   0.245    -.2496609     .063776
      pcnvsq |  -1.795063   .4297431    -4.18   0.000    -2.637344   -.9527822
      pt86sq |  -.1034404   .0160526    -6.44   0.000    -.1349029   -.0719779
     inc86sq |   .0000207   5.03e-06     4.11   0.000     .0000108    .0000305
       _cons |   -.709866    .088544    -8.02   0.000    -.8834092   -.5363229
------------------------------------------------------------------------------
```

```
. test pcnvsq pt86sq inc86sq

 ( 1)  [narr86]pcnvsq = 0
 ( 2)  [narr86]pt86sq = 0
 ( 3)  [narr86]inc86sq = 0

           chi2(  3) =   78.31
         Prob > chi2 =    0.0000

. * Compute the turning points for the variables in quadratics:

. di .6837/(2*.1034)
3.3060928

. * So the turning point for the ptime86 variable is at just over three months,
. * which is somewhat puzzling because the effect of prison time is positive
. * up until that point. It does imply the effect of ptime86 gets stronger
. * as it heads to 12.
```

```
. di .0121/(2*.000021)
288.09524

. sum inc86

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       inc86 |      2725    54.96705    66.62721          0        541

. count if inc86 > 288
   14

. * The turning point for inc86 is acceptable because only 14 observations
. * are to the right of the minimum value, 288.

. di 1.153/(2*1.795)
.32116992

. sum pcnv

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        pcnv |      2725    .3577872     .395192          0          1

. count if pcnv > .32
 1316

. * pcnv has a puzzling turning point.
```

```
. * Now NLS on expanded model:

. glm narr86 pcnv avgsen tottime ptime86 inc86 qemp86 black hispan born60
     pcnvsq pt86sq inc86sq, fam(normal) link(log) robust

------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pcnv |    1.20703    .424864     2.84   0.004     .3743119    2.039748
      avgsen |  -.0569116   .0438451    -1.30   0.194    -.1428464    .0290233
     tottime |   .0249159     .02692     0.93   0.355    -.0278463    .0776781
     ptime86 |   .6230321   .1264642     4.93   0.000     .3751668    .8708975
       inc86 |   -.012419   .0022894    -5.42   0.000    -.0169061   -.007932
      qemp86 |   .0170401   .0406964     0.42   0.675    -.0627234    .0968036
       black |   .5397976   .1196815     4.51   0.000     .3052261    .7743691
      hispan |   .4398227   .1086155     4.05   0.000     .2269403    .6527051
      born60 |  -.0994788   .0988655    -1.01   0.314    -.2932516     .094294
      pcnvsq |  -1.659975   .5038294    -3.29   0.001    -2.647463   -.6724878
      pt86sq |  -.0923192   .0220002    -4.20   0.000    -.1354389   -.0491996
     inc86sq |   .0000196   5.14e-06     3.82   0.000     9.56e-06    .0000297
       _cons |  -.7030874   .0950016    -7.40   0.000    -.8892871   -.5168877
------------------------------------------------------------------------------

. * NLS and Poisson estimates now seem closer. Could compute APEs.
```

4. HURDLE MODELS

- As with a corner solution that is continuous over positive values, we can specify hurdle models for count data.
- The idea again is that the mechanisms determining $y_i = 0$ versus $y_i > 0$ may be different (but related to some common factors).
- If $h(\cdot|\mathbf{x}, \boldsymbol{\delta})$ denotes a count density conditional on $\mathbf{x}$, and $G(\mathbf{x}, \boldsymbol{\gamma})$ is a model for $P(y = 0|\mathbf{x})$, then a general density for a hurdle model is

$$f(0|\mathbf{x},\boldsymbol{\theta}) = G(\mathbf{x},\boldsymbol{\gamma}) \tag{45}$$

$$f(y|\mathbf{x},\boldsymbol{\theta}) = [1 - G(\mathbf{x},\boldsymbol{\gamma})]\frac{h(y|\mathbf{x},\boldsymbol{\delta})}{[1 - h(0|\mathbf{x},\boldsymbol{\delta})]},\ y = 1,2,\ldots. \tag{46}$$

- To nest common models (Poisson, NegBin I & II), choose

$$G(\mathbf{x},\boldsymbol{\gamma}) = h(0|\mathbf{x},\boldsymbol{\gamma}), \tag{47}$$

so that when $\boldsymbol{\gamma} = \boldsymbol{\delta}$, $f(y|\mathbf{x},\boldsymbol{\theta}) = h(y|\mathbf{x},\boldsymbol{\delta})$, $y = 0,1,2,\ldots.$

- Suppose $h(y|\mathbf{x},\boldsymbol{\delta})$ is the Poisson distribution with mean $\exp(\mathbf{x}\boldsymbol{\beta})$. Then

$$h(0|\mathbf{x},\boldsymbol{\beta}) = \exp[-\exp(\mathbf{x}\boldsymbol{\beta})] \tag{48}$$

and so choose

$$G(\mathbf{x},\boldsymbol{\gamma}) = \exp[-\exp(\mathbf{x}\boldsymbol{\gamma})] \tag{49}$$

The density is then

$$f(0|\mathbf{x},\boldsymbol{\theta}) = \exp[-\exp(\mathbf{x}\boldsymbol{\gamma})] \tag{50}$$

$$f(y|\mathbf{x},\boldsymbol{\theta}) = \{1-\exp[-\exp(\mathbf{x}\boldsymbol{\gamma})]\}\frac{\exp[-\exp(\mathbf{x}\boldsymbol{\beta})]\exp(\mathbf{x}\boldsymbol{\beta})^{y}/y!}{\{1-\exp[-\exp(\mathbf{x}\boldsymbol{\beta})]\}}, \tag{52}$$

$$y = 1,2,\ldots.$$

- The MLE of $\boldsymbol{\gamma}$ is easily seen to be the MLE for a binary response, defining $w_i = 1[y_i > 0]$, so that

$$P(w_i = 1|\mathbf{x}_i) = 1 - \exp[-\exp(\mathbf{x}_i\boldsymbol{\gamma})] \tag{53}$$

- Then, $\boldsymbol{\beta}$ can be estimated by MLE using the truncated Poisson distribution (that is, conditional on $y_i \geq 1$).
- For more flexibility, use, say, NegBin II for $h(\cdot)$ and $G(\cdot)$.

5. BINOMIAL REGRESSION

- Now suppose y_i is a count variable taking values in $\{0, 1, \ldots, n_i\}$ for an integer $n_i > 0$. A random draw consists of $(y_i, n_i, \mathbf{x}_i)$ and, as usual, the sample size is N.
- For example, $n_i = 30$ for all i and y_i is the number of days in the last 30 that a person has smoked marijuana. Or, n_i is number of adult children in a family and y_i is the number of who attended college.

- A natural starting point is to view y_i as the number of “successes” out of n_i independent Bernoulli (zero-one) trials, with chance of success $0 < p(\mathbf{x}_i, \boldsymbol{\beta}) < 1$. Typically, $p(\mathbf{x}_i, \boldsymbol{\beta}) = \Phi(\mathbf{x}_i\boldsymbol{\beta})$ or $p(\mathbf{x}_i, \boldsymbol{\beta}) = \Lambda(\mathbf{x}_i\boldsymbol{\beta})$.

- Under the previous assumptions, y_i given $(n_i, \mathbf{x}_i)$ has a $Binomial[n_i, p(\mathbf{x}_i, \boldsymbol{\beta})]$ distribution.
- The mean and variance are

$$E(y_i|n_i, \mathbf{x}_i) = n_i p(\mathbf{x}_i, \boldsymbol{\beta}) \tag{54}$$

$$Var(y_i|n_i, \mathbf{x}_i) = n_i p(\mathbf{x}_i, \boldsymbol{\beta})[1 - p(\mathbf{x}_i, \boldsymbol{\beta})]. \tag{55}$$

- Given standard functional forms for $p(\mathbf{x}_i, \boldsymbol{\beta})$, it is easy to obtain partial effects on the mean.

• The Binomial log likelihood is

$$l_i(\boldsymbol{\beta}) = y_i \log[p(\mathbf{x}_i, \boldsymbol{\beta})] + (n_i - y_i) \log[1 - p(\mathbf{x}_i, \boldsymbol{\beta})] + \log\{n_i!/[y_i!(n_i - y_i)!]\}$$

and we drop the last term.

• MLE estimation is straightforward.

• Importantly, the Binomial density is in the linear exponential family, so only $E(y_i|n_i, \mathbf{x}_i)$ needs to be correctly specified to consistently estimate $\boldsymbol{\beta}$.

- It is easy to devise cases – particularly when the underlying Bernoulli trials for each i are correlated (so $y_i = w_{i1} + w_{i2} + \ldots + w_{i,n_i}$) – where the binomial variance function is incorrect.
- Later we will discuss so-called “cluster sampling,” which is more appropriate if we actually observe the individual w_{ir} along with covariates $\mathbf{x}_{ir}$.
- Fully robust inference is straightforward.
- The GLM variance assumption is

$$Var(y_i|n_i, \mathbf{x}_i) = \sigma^2 n_i p(\mathbf{x}_i, \boldsymbol{\beta})[1 - p(\mathbf{x}_i, \boldsymbol{\beta})] \tag{57}$$

for $\sigma^2 > 0$.

• As before, a consistent estimator of σ^2 is based on the sum of squared weighted residuals,

$$\hat{\sigma}^2 = (N-P)^{-1} \sum_{i=1}^{N} \hat{u}_i^2 / \hat{v}_i \tag{58}$$

$$\hat{u}_i = y_i - n_i p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \tag{59}$$

$$\hat{v}_i = n_i p(\mathbf{x}_i, \hat{\boldsymbol{\beta}})[1 - p(\mathbf{x}_i, \hat{\boldsymbol{\beta}})] \tag{60}$$

• There can be overdispersion ($\sigma^2 > 1$) or underdispersion ($\sigma^2 < 1$).

- In Stata:

```
glm y x1 ... xK, fam(binomial n) link(logit)
robust
glm y x1 ... xK, fam(binomial n) link(probit)
sca(x2)
glm y x1 ... xK, fam(binomial n) link(logit)
```

The last command produces the MLE standard errors and inference.

The variable *n* must be defined by you as the number of “trials,” such as *nkids* for the number of children in a family.

- To estimate the APE for a continuous x_j,

$$\widehat{APE}_j = N^{-1} \sum_{i=1}^{N} n_i \frac{\partial p(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial x_j}$$

If $p(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) = G(\mathbf{x}_i \hat{\boldsymbol{\beta}})$ (by far the most common case, where usually $G(\cdot) = \Phi(\cdot)$ or $\Lambda(\cdot)$),

$$\widehat{APE}_j = \hat{\beta}_j \left[N^{-1} \sum_{i=1}^{N} n_i g(\mathbf{x}_i \hat{\boldsymbol{\beta}}) \right]$$

where $g(\cdot)$ is the derivative of $G(\cdot)$.

• For a discrete change in one or more elements of $\mathbf{x}$,

$$\widehat{APE} = N^{-1} \sum_{i=1}^{N} n_i [G(\mathbf{x}_i^{(1)}\hat{\boldsymbol{\beta}}) - G(\mathbf{x}_i^{(0)}\hat{\boldsymbol{\beta}})]$$

• When `margeff` is used in Stata after GLM estmation, it computes the APEs on $G(\mathbf{x}_i\boldsymbol{\beta})$, the response probability for the underlying binary outcomes. It does not compute the APEs on $E(y|n, \mathbf{x})$.

6. ENDOGENOUS EXPLANATORY VARIABLES

- An exponential regression function is very convenient for nonnegative responses. IV methods and control function methods have been worked out to handle endogeneity. (In the CF cases, we will cover the continuous and binary cases.)

- With a single EEV, write

$$E(y_1|\mathbf{z}, y_2, c_1) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + c_1), \tag{61}$$

where c_1 is the omitted variable. (Extensions to general nonlinear functions of $(\mathbf{z}_1, y_2)$ are immediate; we just add those functions with linear coefficients. Leading cases are polynomials and interactions.)

• Suppose first that y_2 has a standard linear reduced form with an additive, independent error:

$$y_2 = \mathbf{z}\boldsymbol{\pi}_2 + v_2 = \mathbf{z}_1\boldsymbol{\pi}_{21} + \mathbf{z}_2\boldsymbol{\pi}_{22} + v_2 \tag{62}$$

$$D(c_1, v_2|\mathbf{z}) = D(c_1, v_2), \tag{63}$$

so that (c_1, v_2) is independent of $\mathbf{z}$. As in linear and probit models, for identification we need $\boldsymbol{\pi}_{22} \neq \mathbf{0}$.

• The independence of v_2 and $\mathbf{z}$ effectively rules out discrete y_2.

• We can write

$$E(y_1|\mathbf{z}, y_2) = E(y_1|\mathbf{z}, v_2) = E[\exp(c_1)|v_2]\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2). \tag{64}$$

• Suppose we can write $c_1 = \rho_1 v_2 + e_1$ where e_1 is independent of v_2; always holds if (c_1, v_2) are jointly normal. Then $E[\exp(c_1)|v_2] = \exp(\eta_1 + \rho_1 v_2)$ where $\exp(\eta_1) = E[\exp(e_1)]$. Then

$$E(y_1|\mathbf{z}, y_2) = E(y_1|\mathbf{z}, v_2) = \exp(\eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + \rho_1 v_2). \tag{65}$$

• Because $\mathbf{z}_1$ should always contain an intercept, a two-step procedure based on this mean identifies only $\eta_1 + \delta_{11}$. But this is fine because the average partial effects depend on $\eta_1 + \delta_{11}$. To see this, the average structural function is

$$
\begin{aligned}
ASF(\mathbf{z},y_2) &= E_{c_1}[\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + c_1)] &(66)\\
&= E_{c_1}[\exp(c_1)]\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + c_1)\\
&= E_{(v_2,e_1)}[\exp(\rho_1 v_2 + e_1)]\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2)\\
&= E_{v_2}[\exp(\rho_1 v_2)]\exp(\eta_1 + \mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2) &(67)
\end{aligned}
$$

where $E_{(v_2,e_1)}[\exp(\rho_1 v_2 + e_1)] = E_{v_2}[\exp(\rho_1 v_2)\exp(\eta_1)]$ follows from iterated expectations and independence of e_1 and v_2.

• We will be able to estimate the scale term out front via sample averages. So, the APEs depend on the intercept $\eta_1 + \delta_{11}$. In what follows, absorb η_1 into δ_{11}.

- Two-step CF estimation procedure. (1) Estimate the reduced form for y_2 and obtain the residuals. (2) Include $\hat{v}_2$, along with $\mathbf{z}_1$ and y_2, in a QMLE in the LEF. Especially if y_1 is a count variable, Poisson QMLE is attractive, or the two-step NegBin II. For a continuous y_1, might use the Exponential distribution (special case of Gamma).

• A (fully robust) t test of $H_0 : \rho_1 = 0$ is valid as a test that y_2 is exogenous. Average partial effects on the mean are obtained from

$$\left[N^{-1} \sum_{i=1}^{N} \exp(\hat{\rho}_1 \hat{v}_{i2}) \right] \exp(\mathbf{z}_1 \hat{\boldsymbol{\delta}}_1 + \hat{\alpha}_1 y_2). \tag{68}$$

• Can use bootstrap for standard errors.

• Proportionate effects on the expected value, that is elasticities and semi-elasticities, do not depend on the scale factor out front in $[\cdot]$.

EXAMPLE: Data in FERTIL2.DTA. Effects of schooling on fertility in Botswana. Treat education as a continuous variable. As an IV for education, use a dummy variable for being born in the first half of the year. (Of course, must first establish partial correlation with education.)

```
. tab children

  number of |
     living |
   children |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,132       25.96       25.96
          1 |        907       20.80       46.76
          2 |        696       15.96       62.71
          3 |        528       12.11       74.82
          4 |        392        8.99       83.81
          5 |        255        5.85       89.66
          6 |        197        4.52       94.18
          7 |        134        3.07       97.25
          8 |         68        1.56       98.81
          9 |         32        0.73       99.54
         10 |         13        0.30       99.84
         11 |          3        0.07       99.91
         12 |          3        0.07       99.98
         13 |          1        0.02      100.00
------------+-----------------------------------
      Total |      4,361      100.00
```

```
. tab educ

   years of |
  education |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        906       20.78       20.78
          1 |         60        1.38       22.15
          2 |        104        2.38       24.54
          3 |        142        3.26       27.79
          4 |        194        4.45       32.24
          5 |        234        5.37       37.61
          6 |        298        6.83       44.44
          7 |      1,162       26.65       71.08
          8 |        184        4.22       75.30
          9 |        232        5.32       80.62
         10 |        527       12.08       92.71
         11 |         33        0.76       93.46
         12 |        165        3.78       97.25
         13 |         19        0.44       97.68
         14 |         36        0.83       98.51
         15 |         25        0.57       99.08
         16 |         17        0.39       99.47
         17 |         15        0.34       99.82
         18 |          3        0.07       99.89
         19 |          4        0.09       99.98
         20 |          1        0.02      100.00
------------+-----------------------------------
      Total |      4,361      100.00
```

```
. * First use OLS and Poisson regression assuming educ exogenous.

. reg children educ age agesq tv electric spirit protest catholic, robust

Linear regression                                      Number of obs =    4358
                                                       F(  8,  4349) =  725.54
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5727
                                                       Root MSE      =  1.4538

------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   -.076893   .0064872   -11.85   0.000    -.0896111   -.0641749
         age |   .3382634   .0192102    17.61   0.000     .3006016    .3759252
       agesq |   -.002683   .0003516    -7.63   0.000    -.0033722   -.0019937
          tv |  -.2056831   .0825702    -2.49   0.013    -.3675628   -.0438035
    electric |  -.2929425   .0740342    -3.96   0.000    -.4380873   -.1477976
      spirit |   .1297104    .056653     2.29   0.022     .0186417    .2407791
     protest |   .0727998    .066177     1.10   0.271    -.0569409    .2025404
    catholic |    .094514   .0787555     1.20   0.230     -.059887     .248915
       _cons |  -4.355587   .2484493   -17.53   0.000    -4.842674     -3.8685
------------------------------------------------------------------------------
```

```
. glm children educ age agesq tv electric spirit protest catholic,
      fam(poisson) robust

Generalized linear models                          No. of obs      =       4358
Optimization     : ML                              Residual df     =       4349
                                                   Scale parameter =          1
Deviance         =   4090.58415                    (1/df) Deviance =   .9405804
Pearson          =  3419.999356                    (1/df) Pearson  =   .7863875

                                                   AIC             =   3.027522
Log pseudolikelihood = -6587.970483                BIC             = -32353.03

------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0254056   .0026802    -9.48   0.000    -.0306587   -.0201525
         age |   .3663358   .0092647    39.54   0.000     .3481773    .3844944
       agesq |  -.0044562   .0001433   -31.11   0.000     -.004737   -.0041754
          tv |  -.1136942   .0439245    -2.59   0.010    -.1997847   -.0276037
    electric |  -.1333479   .0375466    -3.55   0.000    -.2069379   -.0597579
      spirit |   .0310247   .0251135     1.24   0.217    -.0181969    .0802463
     protest |   .0060793   .0297304     0.20   0.838    -.0521912    .0643499
    catholic |   .0029979   .0366356     0.08   0.935    -.0688065    .0748023
       _cons |  -5.765003   .1476478   -39.05   0.000    -6.054387   -5.475619
------------------------------------------------------------------------------
```

```
. * The estimated variance-mean ratio is about .786, so there is underdisperion
. * in this application.

. margeff

Average partial effects after glm
      y  = log(children)

------------------------------------------------------------------------------
    variable |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0576149   .0061909    -9.31   0.000    -.0697487    -.045481
         age |   .8493913   .0231569    36.68   0.000     .8040046     .894778
       agesq |  -.0101047    .000328   -30.81   0.000    -.0107475   -.0094619
          tv |  -.2459013   .0897211    -2.74   0.006    -.4217514   -.0700513
    electric |  -.2878202   .0756018    -3.81   0.000     -.435997   -.1396435
      spirit |   .0705026     .05793     1.22   0.224    -.0430382    .1840433
     protest |   .0138084   .0677342     0.20   0.838    -.1189481     .146565
    catholic |   .0068061    .083297     0.08   0.935     -.156453    .1700652
------------------------------------------------------------------------------
```

```
. glm children educ age agesq tv electric  spirit protest catholic,
      fam(poisson) sca(x2)

Generalized linear models                          No. of obs      =       4358
Optimization     : ML                              Residual df     =       4349
                                                   Scale parameter =          1
Deviance         =   4090.58415                    (1/df) Deviance =   .9405804
Pearson          =  3419.999356                    (1/df) Pearson  =   .7863875

                                                   AIC             =   3.027522
Log likelihood   = -6587.970483                    BIC             =  -32353.03

------------------------------------------------------------------------------
             |                 OIM
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0254056   .0026197    -9.70   0.000    -.0305401   -.0202712
         age |   .3663358     .00854    42.90   0.000     .3495977     .383074
       agesq |  -.0044562   .0001261   -35.33   0.000    -.0047034    -.004209
          tv |  -.1136942   .0416736    -2.73   0.006     -.195373   -.0320154
    electric |  -.1333479   .0337158    -3.96   0.000    -.1994297   -.0672661
      spirit |   .0310247   .0226257     1.37   0.170    -.0133209    .0753703
     protest |   .0060793   .0266835     0.23   0.820    -.0462194     .058378
    catholic |   .0029979   .0346797     0.09   0.931    -.0649731    .0709689
       _cons |  -5.765003   .1423751   -40.49   0.000    -6.044053   -5.485953
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion)

. * The GLM standard errors are, generally, slightly less than the
. * fully robust ones.
```

```
. * Reduced form for educ. Omitted IV from fertility equation is frsthalf

. reg educ frsthalf age agesq tv electric  spirit protest catholic

      Source |       SS       df       MS              Number of obs =    4358
-------------+------------------------------           F(  8,  4349) =  203.40
       Model |  18293.8049     8  2286.72561           Prob > F      =  0.0000
    Residual |  48892.9866  4349  11.2423515           R-squared     =  0.2723
-------------+------------------------------           Adj R-squared =  0.2709
       Total |  67186.7914  4357  15.4204249           Root MSE      =   3.353

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    frsthalf |  -.6881822   .1021737    -6.74   0.000    -.8884947   -.4878696
         age |  -.1093716   .0380656    -2.87   0.004    -.1839997   -.0347436
       agesq |  -.0006491   .0006275    -1.03   0.301    -.0018792     .000581
          tv |   2.623495   .2077932    12.63   0.000     2.216115    3.030876
    electric |   2.103403   .1733896    12.13   0.000     1.763471    2.443335
      spirit |   .6109302   .1287208     4.75   0.000     .3585718    .8632886
     protest |   1.839693   .1480621    12.43   0.000     1.549416     2.12997
    catholic |   2.188532   .1894826    11.55   0.000      1.81705    2.560015
       _cons |   8.321511   .5515747    15.09   0.000     7.240143    9.402878
------------------------------------------------------------------------------

. * So frsthalf is strongly correlated with educ.

. predict v2h, resid
(3 missing values generated)
```

```
. ivreg children age agesq tv electric  spirit protest catholic
        (educ = frsthalf), robust

Instrumental variables (2SLS) regression               Number of obs =    4358
                                                       F(  8,  4349) =  695.91
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5527
                                                       Root MSE      =  1.4874

------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.1700919   .0646806    -2.63   0.009    -.2968988   -.0432849
         age |   .3279273   .0207317    15.82   0.000     .2872826    .3685721
       agesq |  -.0027435   .0003533    -7.77   0.000    -.0034361   -.0020509
          tv |   .0419927   .1926421     0.22   0.827     -.335684    .4196695
    electric |  -.0931133   .1524638    -0.61   0.541    -.3920202    .2057935
      spirit |   .1865364   .0700722     2.66   0.008     .0491592    .3239136
     protest |   .2442842   .1367228     1.79   0.074    -.0237621    .5123305
    catholic |   .2980737   .1624449     1.83   0.067    -.0204011    .6165485
       _cons |  -3.611507   .5758888    -6.27   0.000    -4.740543   -2.482472
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   age agesq tv electric spirit protest catholic frsthalf
------------------------------------------------------------------------------
```

```
. glm children educ v2h age agesq tv electric  spirit protest catholic,
    fam(poisson) robust

Generalized linear models                          No. of obs      =      4358
Optimization     : ML                              Residual df     =      4348
                                                   Scale parameter =         1
Deviance         =  4088.317584                    (1/df) Deviance =  .9402754
Pearson          =  3416.432176                    (1/df) Pearson  =   .785748

                                                   AIC             =  3.027461
Log pseudolikelihood = -6586.837199                BIC             = -32346.92

------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0697537   .0281717    -2.48   0.013    -.1249692   -.0145381
         v2h |   .0447756   .0282491     1.59   0.113    -.0105917    .1001429
         age |   .3614873   .0098567    36.67   0.000     .3421685    .3808061
       agesq |  -.0044861   .0001433   -31.31   0.000    -.0047669   -.0042053
          tv |   .0038535   .0881183     0.04   0.965    -.1688551    .1765621
    electric |  -.0376484   .0692252    -0.54   0.587    -.1733274    .0980305
      spirit |   .0578503   .0304184     1.90   0.057    -.0017687    .1174693
     protest |   .0875434   .0594771     1.47   0.141    -.0290296    .2041163
    catholic |   .0999684   .0719366     1.39   0.165    -.0410248    .2409616
       _cons |   -5.41201   .2710455   -19.97   0.000     -5.94325   -4.880771
------------------------------------------------------------------------------
```

```
. margeff

Average partial effects after glm
      y  = log(children)

------------------------------------------------------------------------------
    variable |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.1582985   .0641851    -2.47   0.014    -.2840989    -.032498
         v2h |   .1015305   .0641333     1.58   0.113    -.0241685    .2272295
         age |   .8376609   .0243225    34.44   0.000     .7899896    .8853322
       agesq |  -.0101725   .0003292   -30.90   0.000    -.0108178   -.0095272
          tv |   .0087523   .2005306     0.04   0.965    -.3842804     .401785
    electric |  -.0841601   .1517847    -0.55   0.579    -.3816526    .2133324
      spirit |   .1317427   .0712942     1.85   0.065    -.0079913    .2714768
     protest |   .2035704   .1445524     1.41   0.159    -.0797471    .4868879
    catholic |    .236228     .17872     1.32   0.186    -.1140568    .5865128
------------------------------------------------------------------------------

. * Only marginal evidence of endogeneity, but estimated effect differs by a lot.
```

• In the case just treated, under similar assumptions can justify a plug-in method: insert $\hat{y}_2$ for y_2 and then use QMLE in the second stage. But it has little to offer over the CF method (and does not yield a very easy test).

• Now suppose y_2 is binary,

$$y_2 = 1[\mathbf{z}\boldsymbol{\pi}_2 + v_2 \geq 0],\ v_2|\mathbf{z} \sim Normal(0,1). \tag{69}$$

• With $E(y_1|\mathbf{z}, y_2, c_1) = \exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + c_1)$, it is tempting to try the following. First, estimate a probit model for y_2 and obtain the fitted probabilities, $\Phi(\mathbf{z}\hat{\boldsymbol{\pi}}_2)$. In a second stage, plug $\Phi(\mathbf{z}\hat{\boldsymbol{\pi}}_2)$ in for y_2 and use, say, Poisson regression of y_1 on $\mathbf{z}_1, \Phi(\mathbf{z}\hat{\boldsymbol{\pi}}_2)$.

• This plug-in method does not consistently estimate the parameters or average partial effects. It acts as if we can pass the expected value through the exponential function. The (incorrect!) argument goes like this:

$$
\begin{aligned}
E(y_1|\mathbf{z}) &= E[\exp(\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 y_2 + c_1)|\mathbf{z}] \\
&= \exp[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1 E(y_2|\mathbf{z}) + E(c_1|\mathbf{z})] \\
&= \exp[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1\Phi(\mathbf{z}\boldsymbol{\pi}_2) + 0] \quad = \exp[\mathbf{z}_1\boldsymbol{\delta}_1 + \alpha_1\Phi(\mathbf{z}\boldsymbol{\pi}_2)].
\end{aligned}
\tag{70}
$$

Unfortunately, the second equality is wrong.

- As shown by Terza (1998), a control function method can be applied when (c_1, v_2) has a joint normal distribution and is independent of $\mathbf{z}$.
- In order to implement a CF approach, we need to find $E(y_1|\mathbf{z}, y_2) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1)E[\exp(c_1)|\mathbf{z}, y_2]$, where $\mathbf{x}_1$ is a function of $(\mathbf{z}_1, y_2)$ which would almost certainly include y_2 linearly and possibly interacted with elements of $\mathbf{z}_1$.

• Let $\tau_1^2 = Var(c_1)$ and $\rho_1 = Cov(v_2, c_1)$, so that $c_1 = \rho_1 v_2 + e_1$ where $e_1|\mathbf{z}, v_2 \sim$Normal$(0, \tau_1^2 - \rho_1^2)$. Then

$$\begin{aligned} E(y_1|\mathbf{z}, v_2) &= E[\exp(e_1)]\exp(\mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1 v_2) \\ &= \exp((\tau_1^2 - \rho_1^2)/2)\exp(\mathbf{x}_1\boldsymbol{\beta}_1 + \rho_1 v_2) \\ &= \exp((\tau_1^2 - \rho_1^2)/2 + \mathbf{x}_1\boldsymbol{\beta}_1)\exp(\rho_1 v_2). \end{aligned} \tag{71}$$

To find $E(y_1|\mathbf{z}, y_2)$, we have

$$E(y_1|\mathbf{z}, y_2) = \exp((\tau_1^2 - \rho_1^2)/2)\exp(\mathbf{x}_1\boldsymbol{\beta}_1)E[\exp(\rho_1 v_2)|\mathbf{z}, y_2]. \tag{72}$$

- Can show

$$E[\exp(\rho_1 v_2)|\mathbf{z}, y_2 = 1] = E[\exp(\rho_1 v_2)|\mathbf{z}, v_2 > -\mathbf{z}\boldsymbol{\pi}_2] \\ = \exp(\rho_1^2/2)\Phi(\rho_1 + \mathbf{z}\boldsymbol{\pi}_2)/\Phi(\mathbf{z}\boldsymbol{\pi}_2). \quad (73)$$

Similarly,

$$E[\exp(\rho_1 v_2)|\mathbf{z}, y_2 = 0] = \exp(\rho_1^2/2)[1 - \Phi(\rho_1 + \mathbf{z}\boldsymbol{\pi}_2)]/[1 - \Phi(\mathbf{z}\boldsymbol{\pi}_2)]$$

and so

$$E(y_1|\mathbf{z}, y_2) = \exp(\tau_1^2/2 + \mathbf{x}_1\boldsymbol{\beta}_1)\{\Phi(\rho_1 + \mathbf{z}\boldsymbol{\pi}_2)/\Phi(\mathbf{z}\boldsymbol{\pi}_2)\}^{y_2} \\ \cdot \{[1 - \Phi(\rho_1 + \mathbf{z}\boldsymbol{\pi}_2)]/[1 - \Phi(\mathbf{z}\boldsymbol{\pi}_2)]\}^{(1-y_2)}. \quad (74)$$

• If $\mathbf{x}_1$ contains unity, as it should, then only $\tau_1^2/2 + \beta_{11}$ is identified, along with the other elements of $\boldsymbol{\beta}_1$, ρ_1, and $\boldsymbol{\pi}_2$. This is just fine because the average structural function is $ASF(\mathbf{z}_1, y_2) = E_{c_1}[\exp(\mathbf{x}_1\boldsymbol{\beta}_1 + c_1)] = \exp(\tau_1^2/2 + \mathbf{x}_1\boldsymbol{\beta}_1)]$, and so the intercept that is identified is exactly what we want for computing APEs.

• So just absorb $\tau_1^2/2$ into the intercept.

• Two-step CF method: (1) Estimate the probit model of y_2 on $\mathbf{z}$ to obtain the MLE, $\hat{\boldsymbol{\pi}}_2$. (2) Estimate the above mean function, with $\hat{\boldsymbol{\pi}}_2$ in place of $\boldsymbol{\pi}_2$. We can use nonlinear least squares or a quasi-MLE, such as the Poisson.

- Inference should account for the two-step estimation, either using the delta method or bootstrap.
- A simple test of $H_0 : \rho_1 = 0$ is available. The derivative of the mean function with respect to ρ_1, evaluated at $\rho_1 = 0$, is $\exp(\mathbf{x}_1\boldsymbol{\beta}_1)[\lambda(\mathbf{z}\boldsymbol{\pi}_2)]^{y_2}[-\lambda(-\mathbf{z}\boldsymbol{\pi}_2)]^{(1-y_2)}$, where $\lambda(\cdot)$ is the IMR.
- Simple variable addition test of $\rho_1 = 0$: add the variable $y_2 \log[\lambda(\mathbf{z}\hat{\boldsymbol{\pi}}_2)] - (1 - y_2)\log[\lambda(-\mathbf{z}\hat{\boldsymbol{\pi}}_2)]$ to the exponential model $\exp(\mathbf{x}_1\boldsymbol{\beta}_1)$. For each i define $\hat{r}_{i2} = y_{i2}\log[\lambda(\mathbf{z}_i\hat{\boldsymbol{\pi}}_2)] - (1 - y_{i2})\log[\lambda(-\mathbf{z}_i\hat{\boldsymbol{\pi}}_2)]$ (called a *generalized residual*) and then use a QMLE to estimate the artificial mean function $\exp(\mathbf{x}_{i1}\boldsymbol{\beta}_1 + \rho_1\hat{r}_{i2})$, and use a robust t statistic for $\hat{\rho}_1$.

- Unfortunately, adding $\hat{r}_{i2}$ to an exponential regression does not solve the endogeneity problem; it is only justified as a test. For "small" amounts of endogeneity, that is, ρ_1 "close" to zero, it might be approximately valid. But how "small" it needs to be is unclear, and then maybe ignoring endogeneity is sufficient, anyway.

• An alternative to CF approaches is an IV approach. It is nice because, as in the linear case, it can be applied regardless of the nature of y_2.

• Write $\mathbf{x}_1 = \mathbf{g}_1(\mathbf{z}_1, \mathbf{y}_2)$ as any function of exogenous and endogenous variables. If we start with

$$E(y_1|\mathbf{z}, \mathbf{y}_2, c_1) = \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + c_1) \tag{75}$$

then we can use a transformation due to Mullahy (1997) to consistently estimate $\boldsymbol{\beta}_1$ by method of moments.

- Can write

$$\begin{aligned} y_1 &= \exp(\mathbf{x}_1\boldsymbol{\beta}_1 + c_1)a_1 = \exp(\mathbf{x}_1\boldsymbol{\beta}_1)\exp(c_1)a_1 \\ E(a_1|\mathbf{z},\mathbf{y}_2,c_1) &= 1. \end{aligned} \tag{76}$$

- We can write

$$\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1 = \exp(c_1)a_1$$

- If c_1 is independent of $\mathbf{z}$ then

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1|\mathbf{z}] = E[\exp(c_1)|\mathbf{z}] = E[\exp(c_1)] = 1, \tag{77}$$

where the last equality is just a normalization that defines the intercept in $\boldsymbol{\beta}_1$.

• Therefore, we have conditional moment conditions

$$E[\exp(-\mathbf{x}_1\boldsymbol{\beta}_1)y_1 - 1|\mathbf{z}] = 0, \tag{78}$$

which depends on the unknown parameters $\boldsymbol{\beta}_1$ and observable data. Any function of $\mathbf{z}$ can be used as instruments in a nonlinear GMM procedure. An important issue in implementing the procedure is choosing instruments.

• The CF methods are convenient for testing, but the IV method can work for any kind of y_2 (continuous, binary, corner, count, fraction, and so on).

7. PANEL DATA

- Let $\{(\mathbf{x}_{it}, y_{it}) : t = 1, \ldots, T\}$ be a random draw for cross section i, where $y_{it} \geq 0$. We are thinking of cases where y_{it} is a count variable, but several methods can be applied for any nonnegative response.
- Can start with a standard linear unobserved effects model estimated by FE!
- The most common model for the conditional mean allows multiplicative in the heterogeneity:

$$E(y_{it}|\mathbf{x}_{it}, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{79}$$

where $c_i \geq 0$ is the unobserved effect and $\mathbf{x}_{it}$ would incude a full set of year dummies in most cases.

- There is no difficulty in replacing $\exp(\mathbf{x}_{it}\boldsymbol{\beta})$ with a general function $m_t(\mathbf{x}_{it}, \boldsymbol{\beta}) > 0$ but the exponential model is by far the most popular.
- Notice that if we start with

$$E(y_{it}|\mathbf{x}_{it}, r_i) = \exp(\mathbf{x}_{it}\boldsymbol{\beta} + g_i) \tag{80}$$

then we can take $c_i = \exp(g_i)$.

- As in the linear case, many estimation methods assume strict exogeneity of the covariates conditional on c_i:

$$E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},c_i) = E(y_{it}|\mathbf{x}_{it},c_i). \tag{81}$$

- Adding independence between c_i and $\mathbf{x}_i$ – a random effects approach – and using $E(c_i) = 1$ as a normalization,

$$E(y_{it}|\mathbf{x}_i) = \exp(\mathbf{x}_{it}\boldsymbol{\beta}). \tag{82}$$

- Various estimation methods can be used to account for the serial dependence in $\{y_{it}\}$ conditional on $\mathbf{x}_i$.
- For example, a simple approach is the pooled Poisson quasi-MLE, which only requires

$$E(y_{it}|\mathbf{x}_{it}) = \exp(\mathbf{x}_{it}\boldsymbol{\beta}), \tag{83}$$

and so we do not even need to impose strict exogeneity (because we have effectively dropped the heterogeneity).
- Pooled Poisson regression is likely to be inefficient. So, can apply GEE with the Poisson family. Can specify the working correlation to be exchangeable or unstructured.

- The error term that we nominally apply the constant conditional correlation assumption to is

$$e_{it} \equiv \frac{[y_{it} - \exp(\mathbf{x}_{it}\boldsymbol{\beta})]}{\exp(\mathbf{x}_{it}\boldsymbol{\beta}/2)}$$

- Stata commands:

```
xtgee y x1 ... xK, fam(poisson) corr(exch)
robust
xtgee y x1 ... xK, fam(poisson) corr(uns) robust
```

• If one believes the first two moments of the Poisson distribution conditional on c_i,

$$E(y_{it}|\mathbf{x}_i, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{84}$$

$$Var(y_{it}|\mathbf{x}_i, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \tag{85}$$

along with $D(c_i|\mathbf{x}_i) = D(c_i)$, and conditional uncorrelatedness:

$$Cov(y_{it}, y_{is}|\mathbf{x}_i, c_i) = 0, \; t \neq s,$$

then the GEE Poisson variance-covariance matrix is wrong.

- The derivation of $Var(y_{it}|\mathbf{x}_i)$ follows the same argument as in the derivation in the cross section case:

$$Var(y_{it}|\mathbf{x}_i) = \exp(\mathbf{x}_{it}\boldsymbol{\beta}) + \eta^2 \exp(2\mathbf{x}_{it}\boldsymbol{\beta}) \quad (86)$$

- For the covariances conditional on $\mathbf{x}_i$:

$$\begin{aligned} Cov(y_{it}, y_{is}|\mathbf{x}_i) &= E[Cov(y_{it}, y_{is}|\mathbf{x}_i, c_i)|\mathbf{x}_i] + Cov[E(y_{it}|\mathbf{x}_i, c_i), E(y_{is}|\mathbf{x}_i, c_i)|\mathbf{x}_i] \\ &= 0 + Var(c_i|\mathbf{x}_i) \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \exp(\mathbf{x}_{is}\boldsymbol{\beta}) \\ &= \eta^2 \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \exp(\mathbf{x}_{is}\boldsymbol{\beta}) \quad (8 \end{aligned}$$

• Could use multivariate WNLS using this variance-covariance structure. Use simple moment estimators for η^2.

• The conditional correlations are not constant:

$$Corr(y_{it}, y_{is}|\mathbf{x}_i) = \frac{\eta^2 \exp(\mathbf{x}_{it}\boldsymbol{\beta}) \exp(\mathbf{x}_{is}\boldsymbol{\beta})}{\sqrt{[\exp(\mathbf{x}_{it}\boldsymbol{\beta}) + \eta^2 \exp(2\mathbf{x}_{it}\boldsymbol{\beta})][\exp(\mathbf{x}_{is}\boldsymbol{\beta}) + \eta^2 \exp(2\mathbf{x}_{is}\boldsymbol{\beta})]}}.$$

• Rather than just first and second moment assumptions, suppose we maintain

$$D(y_{it}|\mathbf{x}_i, c_i) \sim Poisson[c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})] \tag{88}$$

$$D(c_i|\mathbf{x}_i) \sim Gamma(\delta, \delta), \tag{89}$$

where $E(c_i) = 1$ and $Var(c_i) = 1/\delta$, and conditional independence:

$$D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_i, c_i) \tag{90}$$

then we have the **random effects Poisson model**.

- Can show that the log likelihood for observation i is (up to addive factors)

$$\ell_i(\boldsymbol{\theta}) = \sum_{t=1}^{T} y_{it}\mathbf{x}_{it} + \delta\log(\delta) - \log[\Gamma(\delta)]$$
$$+ \log\left[\sum_{t=1}^{T}\exp(\mathbf{x}_{it}\boldsymbol{\beta}) + n_i\right] - (n_i + \delta)\log\left[\sum_{t=1}^{T}\exp(\mathbf{x}_{it}\boldsymbol{\beta}) + \delta\right]$$

where $\Gamma(\cdot)$ is the gamma function and $n_i = y_{i1} + \ldots + y_{iT}$.

- Maximizing the sample log likelihood is relatively straightforward.

Available in Stata as

```
xtpoisson y x1 ... xK, re
```

• This estimator has no known robustness properties if any of the assumptions are violated. In particular, it, like RE probit, requires the conditional independence assumption in (90). GEE is more robust but less efficient if all of the RE assumptions hold.

• In the pooled Poisson, GEE Poisson, and Poisson RE approaches, can implement a Chamberlain-Mundlak correlated random effects (CRE) device by assuming

$$c_i = \exp(\psi + \bar{\mathbf{x}}_i \boldsymbol{\xi}) a_i, \tag{91}$$

where a_i is independent of $\mathbf{x}_i$ with unit mean. Then

$$E(y_{it}|\mathbf{x}_i) = \exp(\psi + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\xi}). \tag{92}$$

• So, use any of the previous methods by adding $\bar{\mathbf{x}}_i$ as a set of covariates. Can include time-constant covariates, say $\mathbf{z}_i$, if desired.

- Stata commands (assuming time averages have been generated using, say, `egen`)

```
glm y x1 ... xK x1bar ... xKbar, fam(poisson)
cluster(id)
xtgee y x1 ... xK x1bar ... xKbar, fam(poisson)
corr(exch) robust
xtpoisson y x1 ... xK x1bar ... xKbar, re
```

- Pooled Poisson and GEE only use $E(y_{it}|\mathbf{x}_i) = \exp(\psi + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi})$. The Poisson RE method requires that $D(y_{it}|\mathbf{x}_i, c_i)$ is Poisson, a_i in (91) has a $Gamma(\delta, \delta)$ distribution, and conditional independence over time.

• An important estimator that can be used under just

$$E(y_{it}|\mathbf{x}_i, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})$$

is the conditional MLE derived under a Poisson distributional assumption and the conditional independence assumption.

• It is often called the **fixed effects Poisson estimator**. It is best characterized as a conditional MLE (like fixed effects logit). But, in this case, $\hat{\boldsymbol{\beta}}$ turns out to be identical to using pooled Poisson QMLE and treating the c_i as parameters to estimate (one for each i). (This is a rare case, like the linear model, where "estimating" the unobserved effects does not result in an incidental parameters problem for estimating $\boldsymbol{\beta}$.)

- For FE Poisson, we *nominally* start with strict exogeneity and the Poisson distributional assumptions,

$$D(y_{it}|\mathbf{x}_i, c_i) \sim Poisson[c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})], \tag{93}$$

and conditional independence,

$$D(y_{i1}, \ldots, y_{iT}|\mathbf{x}_i, c_i) = \prod_{t=1}^{T} D(y_{it}|\mathbf{x}_i, c_i), \tag{94}$$

but put no restrictions on $D(c_i|\mathbf{x}_i)$.

- Let $n_i = y_{i1} + \ldots + y_{iT}$ be the total number of counts.

• Can show that the joint distribution of $(y_{i1}, \ldots, y_{iT})$ conditional on $(n_i, \mathbf{x}_i, c_i)$ is multinomial with probabilities

$$p_t(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})}{\sum_{r=1}^{T} c_i \exp(\mathbf{x}_{ir}\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}_{it}\boldsymbol{\beta})}{\sum_{r=1}^{T} \exp(\mathbf{x}_{ir}\boldsymbol{\beta})} \tag{95}$$

so that the heterogeneity c_i has disappeared.

• Time-constant variables drop out, as in linear case. For example, γz_i would come out in front as $\exp(\gamma z_i)$ and cancel in the numerator and demoninator.

- The FE Poisson estimator maximimizes the resulting log-likelihood function. For each i,

$$l_i(\boldsymbol{\beta}) = \sum_{t=1}^{T} y_{it} \log[p_t(\mathbf{x}_i, \boldsymbol{\beta})]. \tag{96}$$

- In Stata:

```
xtpoisson y x1 ... xK, fe
```

- Important result: The Poisson distribution can be arbitrarily misspecified, and any kind of serial correlation can be present, and the the FEP estimator is consistent provided

$$E(y_{it}|\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT},c_i) = E(y_{it}|\mathbf{x}_{it},c_i) = c_i\exp(\mathbf{x}_{it}\boldsymbol{\beta}).$$

- In particular, y_{it} need not even be a count variable. It could be continuous, or a corner. However, the mean $c_i\exp(\mathbf{x}_{it}\boldsymbol{\beta})$ should make logical sense.
- We do require strict exogeneity.
- See Wooldridge (1999, Journal of Econometrics) for a general proof.

• Whether or not y_{it} is a count, should make inference fully robust to serial correlation and violation of the Poisson distribution.

• The score can be written as

$$\mathbf{s}_i(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}\mathbf{p}(\mathbf{x}_i,\boldsymbol{\beta})'\mathbf{W}(\mathbf{x}_i,\boldsymbol{\beta})[\mathbf{y}_i - n_i\mathbf{p}(\mathbf{x}_i,\boldsymbol{\beta})]$$

where $\mathbf{p}(\mathbf{x}_i,\boldsymbol{\beta})$ is the $T \times 1$ vector with elements $p_t(\mathbf{x}_i,\boldsymbol{\beta})$ and $\mathbf{W}(\mathbf{x}_i,\boldsymbol{\beta})$ is the $T \times T$ diagonal matrix with elements $1/p_t(\mathbf{x}_i,\boldsymbol{\beta})$.

• See text, Section 18.7.4, for verification that $E[\mathbf{s}_i(\boldsymbol{\beta}_o)|\mathbf{x}_i] = \mathbf{0}$ (when one is careful to indicate the true value).

• As usual, the robust variance matrix estimator of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ has the sandwich form, $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$ with

$$\hat{\mathbf{A}} = N^{-1}\sum_{i=1}^{N} n_i \nabla_{\boldsymbol{\beta}}\mathbf{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}})'\mathbf{W}(\mathbf{x}_i, \hat{\boldsymbol{\beta}})\nabla_{\boldsymbol{\beta}}\mathbf{p}(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) \tag{99}$$

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N} \mathbf{s}_i(\hat{\boldsymbol{\beta}})\mathbf{s}_i(\hat{\boldsymbol{\beta}})' \tag{100}$$

• If the Poisson distribution is correct and independence holds, both conditional on $(\mathbf{x}_i, c_i)$, then $\hat{\mathbf{A}}^{-1}$ can be used.

- Fully robust form in Stata (as an "ado" file):

```
xtpqml y x1 ... xK, fe
```

- In effect, `xtpqml` has superceded `xtpoisson`.
- Can cluster at a higher level of aggregation (we will discuss later).

For example, if have a few firms per industry, and lots of industries, might allow within-industry correlation:

```
xtset firmid year
xtpqml y x1 ... xK, fe cluster(industid)
```

EXAMPLE: The patents-R&D relationship. 226 firms over 10 years.

Data compiled by NBER (update?). Need to allow for substantial lag.

```
. use patent

. des cusip year patents rnd lrnd

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
cusip           float  %9.0g                  firm identifier
year            byte   %9.0g                  72 through 81
patents         int    %9.0g                  patents applied for
rnd             float  %9.0g                  R&D expend, current mill $
lrnd            float  %9.0g                  log(1+rnd)

. tab patents if year == 81

    patents |
applied for |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        125       55.31       55.31
          1 |         37       16.37       71.68
          2 |          8        3.54       75.22
          3 |          7        3.10       78.32
          4 |          4        1.77       80.09
          5 |          4        1.77       81.86
          6 |          5        2.21       84.07
          7 |          5        2.21       86.28
          8 |          1        0.44       86.73
          9 |          2        0.88       87.61
```

10	1	0.44	88.05
11	2	0.88	88.94
12	1	0.44	89.38
13	1	0.44	89.82
14	3	1.33	91.15
15	1	0.44	91.59
16	1	0.44	92.04
17	1	0.44	92.48
18	2	0.88	93.36
20	1	0.44	93.81
24	2	0.88	94.69
27	1	0.44	95.13
28	1	0.44	95.58
33	1	0.44	96.02
35	1	0.44	96.46
39	1	0.44	96.90
40	1	0.44	97.35
41	1	0.44	97.79
48	1	0.44	98.23
51	1	0.44	98.67
58	1	0.44	99.12
97	1	0.44	99.56
168	1	0.44	100.00
Total	226	100.00	

```
. xtreg patents lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4 lrnd_5 lrnd_6 y79-y81,
        fe cluster(cusip)

Fixed-effects (within) regression               Number of obs      =       904
Group variable: cusip                           Number of groups   =       226

R-sq:  within  = 0.1117                         Obs per group: min =         4
       between = 0.3831                                        avg =       4.0
       overall = 0.2870                                        max =         4

                                (Std. Err. adjusted for 226 clusters in cusip)
------------------------------------------------------------------------------
             |               Robust
     patents |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lrnd |  -4.891047   4.487274    -1.09   0.277    -13.73351    3.951412
      lrnd_1 |  -8.770371   5.825649    -1.51   0.134    -20.25018     2.70944
      lrnd_2 |  -1.399383   3.024928    -0.46   0.644    -7.360195    4.561428
      lrnd_3 |  -3.218844   3.173328    -1.01   0.312    -9.472087    3.034399
      lrnd_4 |   -8.89406   4.729909    -1.88   0.061    -18.21464    .4265244
      lrnd_5 |  -4.574966   5.090455    -0.90   0.370    -14.60603    5.456098
      lrnd_6 |   -13.7178   6.755444    -2.03   0.043    -27.02983   -.4057713
         y79 |   2.282507   1.051129     2.17   0.031     .2111897    4.353824
         y80 |   .6192547   1.261851     0.49   0.624    -1.867302    3.105811
         y81 |  -9.543918   2.827286    -3.38   0.001    -15.11526   -3.972572
       _cons |    93.2246   18.58406     5.02   0.000     56.60353    129.8457
-------------+----------------------------------------------------------------
         rho |  .93762215   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. gen lpatents = log(1 + patents)

. xtreg lpatents lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4 lrnd_5 lrnd_6 y79-y81,
        fe cluster(cusip)

Fixed-effects (within) regression               Number of obs      =       904
Group variable: cusip                           Number of groups   =       226

R-sq:  within  = 0.4905                         Obs per group: min =         4
       between = 0.7607                                        avg =       4.0
       overall = 0.5018                                        max =         4

                                (Std. Err. adjusted for 226 clusters in cusip)
------------------------------------------------------------------------------
             |               Robust
    lpatents |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lrnd |  -.2326016   .1544488    -1.51   0.133    -.5369528    .0717496
      lrnd_1 |  -.2410225   .1543456    -1.56   0.120    -.5451702    .0631252
      lrnd_2 |  -.1720482    .135934    -1.27   0.207    -.4399149    .0958184
      lrnd_3 |  -.0878612   .1543296    -0.57   0.570    -.3919775    .2162552
      lrnd_4 |  -.1987401   .1446014    -1.37   0.171    -.4836863    .0862061
      lrnd_5 |   .0657994   .1778924     0.37   0.712    -.2847489    .4163477
      lrnd_6 |  -.3073906   .1660589    -1.85   0.065    -.6346203    .0198391
         y79 |    .042687   .0475658     0.90   0.370    -.0510444    .1364183
         y80 |  -.0160116   .0593244    -0.27   0.787     -.132914    .1008909
         y81 |   -.727999   .0781183    -9.32   0.000    -.8819359    -.574062
       _cons |   3.638258   .3846061     9.46   0.000     2.880368    4.396149
-------------+----------------------------------------------------------------
     sigma_u |   3.034257
     sigma_e |  .50399388
         rho |  .97315113   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. poisson patents lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4 lrnd_5 lrnd_6 y79-y81,
          cluster(cusip)

Poisson regression                                Number of obs   =        904
                                                  Wald chi2(10)   =     878.50
Log pseudolikelihood = -7209.5811                 Prob > chi2     =     0.0000

                                (Std. Err. adjusted for 226 clusters in cusip)
------------------------------------------------------------------------------
             |               Robust
     patents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lrnd |   .7310869   .3421723     2.14   0.033     .0604416    1.401732
      lrnd_1 |  -.3529964   .3324251    -1.06   0.288    -1.004538    .2985449
      lrnd_2 |  -.2556217   .4482509    -0.57   0.568    -1.134177    .6229339
      lrnd_3 |   .5953963   .5041567     1.18   0.238    -.3927328    1.583525
      lrnd_4 |   .3925039   .2426363     1.62   0.106    -.0830545    .8680624
      lrnd_5 |  -.0209466   .3541238    -0.06   0.953    -.7150165    .6731232
      lrnd_6 |  -.3143883    .436545    -0.72   0.471    -1.170001    .5412242
         y79 |  -.1004726   .0618751    -1.62   0.104    -.2217455    .0208004
         y80 |  -.4389727   .0639321    -6.87   0.000    -.5642773   -.3136681
         y81 |  -1.871174   .0994861   -18.81   0.000    -2.066163   -1.676185
       _cons |   .8720311   .2072225     4.21   0.000     .4658825     1.27818
------------------------------------------------------------------------------
```

```
. test lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4 lrnd_5 lrnd_6

 ( 1)  [patents]lrnd = 0
 ( 2)  [patents]lrnd_1 = 0
 ( 3)  [patents]lrnd_2 = 0
 ( 4)  [patents]lrnd_3 = 0
 ( 5)  [patents]lrnd_4 = 0
 ( 6)  [patents]lrnd_5 = 0
 ( 7)  [patents]lrnd_6 = 0

           chi2(  7) =  224.78
         Prob > chi2 =    0.0000

. lincom lrnd + lrnd_1 + lrnd_2 + lrnd_3 + lrnd_4 + lrnd_5 + lrnd_6

 ( 1)  [patents]lrnd + [patents]lrnd_1 + [patents]lrnd_2 + [patents]lrnd_3
        + [patents]lrnd_4 + [patents]lrnd_5 + [patents]lrnd_6 = 0

------------------------------------------------------------------------------
     patents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .775034   .0710549    10.91   0.000     .6357689     .914299
------------------------------------------------------------------------------
```

```
. xtpoisson patents lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4 lrnd_5 lrnd_6 y79-y81, fe
note: 19 groups (76 obs) dropped because of all zero outcomes

Conditional fixed-effects Poisson regression    Number of obs      =       828
Group variable: cusip                           Number of groups   =       207

                                                Obs per group: min =         4
                                                               avg =       4.0
                                                               max =         4

                                                Wald chi2(10)      =   2588.77
Log likelihood  = -1288.6346                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     patents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lrnd |   .0171501   .0969203     0.18   0.860    -.1728101    .2071103
      lrnd_1 |   .0147816   .1171385     0.13   0.900    -.2148056    .2443688
      lrnd_2 |   .1145972   .0772455     1.48   0.138    -.0368013    .2659956
      lrnd_3 |  -.0886588   .0812003    -1.09   0.275    -.2478084    .0704909
      lrnd_4 |  -.0889191   .1093972    -0.81   0.416    -.3033336    .1254955
      lrnd_5 |   .4899219   .1348606     3.63   0.000        .2256    .7542438
      lrnd_6 |   .2129892   .1270982     1.68   0.094    -.0361188    .4620972
         y79 |  -.1165952   .0283495    -4.11   0.000    -.1721592   -.0610312
         y80 |  -.4133889   .0463611    -8.92   0.000     -.504255   -.3225228
         y81 |  -1.785541   .0709727   -25.16   0.000    -1.924645   -1.646437
------------------------------------------------------------------------------
```

```
. lincom lrnd + lrnd_1 + lrnd_2 + lrnd_3 + lrnd_4 + lrnd_5 + lrnd_6

 ( 1) [patents]lrnd + [patents]lrnd_1 + [patents]lrnd_2 + [patents]lrnd_3
      + [patents]lrnd_4 + [patents]lrnd_5 + [patents]lrnd_6 = 0

------------------------------------------------------------------------------
     patents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .6718622   .1905634     3.53   0.000     .2983647     1.04536
------------------------------------------------------------------------------

. * But the above standard errors assume the Poisson variance assumption and
. * conditional independence.
```

```
. xtpqml patents lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4 lrnd_5 lrnd_6 y79-y81, fe
note: 19 groups (76 obs) dropped because of all zero outcomes

Conditional fixed-effects Poisson regression    Number of obs      =       828
Group variable: cusip                           Number of groups   =       207

                                                Obs per group: min =         4
                                                               avg =       4.0
                                                               max =         4

------------------------------------------------------------------------------
     patents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lrnd |   .0171501   .0969203     0.18   0.860    -.1728101    .2071103
      lrnd_1 |   .0147816   .1171385     0.13   0.900    -.2148056    .2443688
      lrnd_2 |   .1145972   .0772455     1.48   0.138    -.0368013    .2659956
      lrnd_3 |  -.0886588   .0812003    -1.09   0.275    -.2478084    .0704909
      lrnd_4 |  -.0889191   .1093972    -0.81   0.416    -.3033336    .1254955
      lrnd_5 |   .4899219   .1348606     3.63   0.000        .2256    .7542438
      lrnd_6 |   .2129892   .1270982     1.68   0.094    -.0361188    .4620972
         y79 |  -.1165952   .0283495    -4.11   0.000    -.1721592   -.0610312
         y80 |  -.4133889   .0463611    -8.92   0.000     -.504255   -.3225228
         y81 |  -1.785541   .0709727   -25.16   0.000    -1.924645   -1.646437
------------------------------------------------------------------------------
```

Calculating Robust Standard Errors...

patents	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
patents						
lrnd	.0171501	.1362715	0.13	0.900	-.2499372	.2842374
lrnd_1	.0147816	.149009	0.10	0.921	-.2772706	.3068339
lrnd_2	.1145972	.0554412	2.07	0.039	.0059344	.2232599
lrnd_3	-.0886588	.0889173	-1.00	0.319	-.2629335	.085616
lrnd_4	-.0889191	.1358352	-0.65	0.513	-.3551512	.1773131
lrnd_5	.4899219	.1846058	2.65	0.008	.1281011	.8517427
lrnd_6	.2129892	.2252369	0.95	0.344	-.2284671	.6544455
y79	-.1165952	.0386929	-3.01	0.003	-.1924318	-.0407585
y80	-.4133889	.0679516	-6.08	0.000	-.5465717	-.2802061
y81	-1.785541	.1304135	-13.69	0.000	-2.041147	-1.529936

Wald chi2(10) = 472.12 Prob > chi2 = 0.0000

```
. lincom lrnd + lrnd_1 + lrnd_2 + lrnd_3 + lrnd_4 + lrnd_5 + lrnd_6

 ( 1) [patents]lrnd + [patents]lrnd_1 + [patents]lrnd_2 + [patents]lrnd_3
      + [patents]lrnd_4 + [patents]lrnd_5 + [patents]lrnd_6 = 0

------------------------------------------------------------------------------
     patents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .6718622   .3317594     2.03   0.043     .0216257    1.322099
------------------------------------------------------------------------------

. * The robust 95% CI for the long run elasticity is much wider than the CI that
. * maintains the Poisson distribution and serial independence. The LR elasticity
. * is (barely) statistically different from zero at the 5% leve, but not
. * statistically different from unity.
```

- A simple test to detect violation of strict exogeneity is to add $\mathbf{w}_{i,t+1}$ to the FE Poisson estimation and test its joint significance, where $\mathbf{w}_{i,t+1}$ is a subset of $\mathbf{x}_{i,t+1}$ that varies (at least somewhat) across i and t and which is suspected of violating strict exogeneity. As usual, a fully robust statistic should be used.

```
. sort cusip year

. gen lrndp1 = lrnd[_n+1] if year < 81
(226 missing values generated)

. xtpqml patents lrnd lrnd_1 lrnd_2 lrnd_3 lrnd_4 lrnd_5 lrnd_6 y79-y80 lrndp1, fe
note: 20 groups (60 obs) dropped because of all zero outcomes

Conditional fixed-effects Poisson regression    Number of obs      =       618
Group variable: cusip                           Number of groups   =       206

------------------------------------------------------------------------------
     patents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lrnd |   .3451131   .1164738     2.96   0.003     .1168286    .5733977
      lrnd_1 |   .0177909   .1310983     0.14   0.892     -.239157    .2747388
      lrnd_2 |   .0946511   .0791481     1.20   0.232    -.0604763    .2497785
      lrnd_3 |  -.0679123   .0828895    -0.82   0.413    -.2303727    .0945482
      lrnd_4 |   .0443248   .1372362     0.32   0.747    -.2246532    .3133029
      lrnd_5 |   .4613278     .14832     3.11   0.002      .170626    .7520296
      lrnd_6 |   .0658392   .1415702     0.47   0.642    -.2116333    .3433118
         y79 |  -.0857226   .0325261    -2.64   0.008    -.1494727   -.0219726
         y80 |  -.3748064   .0565806    -6.62   0.000    -.4857023   -.2639105
      lrndp1 |  -.4417111   .1032046    -4.28   0.000    -.6439883   -.2394338
------------------------------------------------------------------------------
```

```
Calculating Robust Standard Errors...
------------------------------------------------------------------------------
     patents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
patents      |
        lrnd |   .3451131   .0998228     3.46   0.001      .149464    .5407623
      lrnd_1 |   .0177909   .1486163     0.12   0.905    -.2734917    .3090735
      lrnd_2 |   .0946511    .050323     1.88   0.060    -.0039802    .1932823
      lrnd_3 |  -.0679123   .0822041    -0.83   0.409    -.2290293    .0932048
      lrnd_4 |   .0443248   .1940994     0.23   0.819    -.3361031    .4247527
      lrnd_5 |   .4613278   .1864198     2.47   0.013     .0959517    .8267039
      lrnd_6 |   .0658392   .1976914     0.33   0.739    -.3216287    .4533072
         y79 |  -.0857226   .0461068    -1.86   0.063    -.1760902     .004645
         y80 |  -.3748064   .0716907    -5.23   0.000    -.5153177   -.2342951
      lrndp1 |  -.4417111   .1257841    -3.51   0.000    -.6882434   -.1951787
------------------------------------------------------------------------------
Wald chi2(10) =   198.11                            Prob > chi2 =   0.0000
```

- The rejection of strict exogeneity of the R&D variable is pretty strong, although the sign is a bit hard to interpret.

Estimation Under Sequential Exogeneity

• RE Poisson, FE Poisson, and GEE all assume strict exogeneity of $\{\mathbf{x}_{it} : t = 1, 2, \ldots, T\}$ conditional on c_i. Pooled Poisson QMLE (or other pooled methods) do not require strict exogeneity but they effectively rule out correlation between c_i and $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ (as do RE and GEE methods unless we included time averages)

• Now we assume only sequential exogeneity of $\{\mathbf{x}_{it} : t = 1, 2, \ldots, T\}$ conditional on c_i with an exponential regression function:

$$E(y_{it}|\mathbf{x}_{it}, \ldots, \mathbf{x}_{i1}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}).$$

• We are silent on whether

$$E(y_{it}|\mathbf{x}_{iT}, \ldots, \mathbf{x}_{it}, \ldots \mathbf{x}_{i1}, c_i) = E(y_{it}|\mathbf{x}_{it}, \ldots, \mathbf{x}_{i1}, c_i)$$

but we think it might not be true.

- As in the linear case, this setup applies to models with lagged dependent variables – so, say, $y_{i,t-1}$ is in $\mathbf{x}_{it}$, or functions of $y_{i,t-1}$, such as $1[y_{i,t-1} = 0]$ and $1[y_{i,t-1} > 0]\log(y_{i,t-1})$ – and also finite distributed lag (FDL) models, where $\mathbf{x}_{it} = (\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \ldots, \mathbf{z}_{i,t-Q})$.
- We need to choose $\mathbf{x}_{it}$ appropriately so that no further lags of elements in $\mathbf{x}_{it}$ matter.

- By definition we can write

$$y_{it} = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})r_{it} \tag{101}$$

$$E(r_{it}|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1},c_i) = 1. \tag{102}$$

- Viewing $\{r_{it} : t = 1,\ldots,T\}$ as multiplicative "shocks," this setup allows $\mathbf{x}_{i,t+1}$ to be correlated with r_{it}, which is necessarily true if $\mathbf{x}_{it}$ contains functions of $y_{i,t-1}$. It can also can also be true when there is feedback in static or FDL models.

- Generally, $\{r_{it}\}$ is serially correlated, although when $\mathbf{x}_{it}$ contains lags of y_{it}, the intention is probably that

$$E(r_{it}|\mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}, c_i) = E(r_{it}|\mathbf{x}_{it}, \ldots, \mathbf{x}_{i1}, c_i) = 1$$

in which case $\{r_{it}\}$ would not be serially correlated. In finite distributed lag models, with $\mathbf{x}_{it} = (\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \ldots, \mathbf{z}_{i,t-Q})$, would expect serial correlation.

• How do we obtain moments that can be used to estimate $\boldsymbol{\beta}$? We can write, for $t = 1, \ldots, T-1$,

$$y_{it} - y_{i,t+1}\left[\frac{\exp(\mathbf{x}_{it}\boldsymbol{\beta})}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})}\right] = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})r_{it} - c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})r_{i,t+1} \tag{103}$$

$$= c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})(r_{it} - r_{i,t+1}) \tag{104}$$

• Using only the condition $E(r_{it}|\mathbf{x}_{it}, \ldots, \mathbf{x}_{i1}, c_i) = 1$ we can show that the RHS has zero mean conditional on $(\mathbf{x}_{it}, \ldots, \mathbf{x}_{i1}, c_i)$:

$$E[c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})(r_{it} - r_{i,t+1})|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1},c_i] \quad (105)$$
$$= c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})E(r_{it} - r_{i,t+1}|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1},c_i) \quad (106)$$
$$= c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})(1-1) = 0; \quad (107)$$

note that $E(r_{i,t+1}|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1},c_i) = 1$ by iterated expectations.

• Therefore,

$$E\{[y_{it} - y_{i,t+1}\exp((\mathbf{x}_{it} - \mathbf{x}_{i,t+1})\boldsymbol{\beta})|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1}] = 0. \quad (108)$$

• Because these moment conditions depend only on observed data and the parameter vector $\boldsymbol{\beta}$, GMM can be used to estimate $\boldsymbol{\beta}$, and fully robust inference is straightforward.

• Because the moment conditions depend on the change in the explanatory variables, GMM approach might suffer from a weak instruments problem. [That is, $\mathbf{x}_{it} - \mathbf{x}_{i,t+1}$ is only weakly correlated with functions of $(\mathbf{x}_{it}, \ldots, \mathbf{x}_{i1})$.]

• Choice of instruments is not obvious. What might be some good approximations to the optimal instruments?

• If violation of strict exogeneity is due only to a lagged dependent variable, can use a conditional MLE approach. For example, suppose

$$D(y_{it}|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i1}, y_{i0}, c_i) = Poisson[c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta})]$$

where, say, $\mathbf{x}_{it}$ can be any function of $(\mathbf{z}_{it}, y_{i,t-1})$. (Adding lags of $\mathbf{z}_{it}$, or further lags of y_{it}, is relatively straightforward with several time periods.). This assumption implies correct dynamics as well as strict exogeneity of $\{\mathbf{z}_{it} : t = 1, \ldots, T\}$.

• As usual, the presence of dynamics and heterogeneity in nonlinear models raises an "initial conditions" problem. A simple solution is to model the dependence between c_i and $(\mathbf{z}_i, y_{i0})$:

$$c_i = \exp(\psi + \mathbf{z}_i\boldsymbol{\gamma} + \xi y_{i0})a_i \tag{110}$$

$$D(a_i|\mathbf{z}_i, y_{i0}) = Gamma(\delta, \delta) \tag{111}$$

where $E(a_i) = 1$ and $\delta = 1/\eta^2 = 1/Var(a_i)$.

• As shown in Wooldridge (2005, Journal of Applied Econometrics), the resulting likelihood function is identical to the Poisson RE likelihood with explanatory variables

$$(\mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_i, y_{i0}) \tag{112}$$

in the case $\mathbf{x}_{it} = (\mathbf{z}_{it}, y_{i,t-1})$.

• So, to implement the method, generate $\mathbf{z}_i$ and y_{i0} so that they appear on every line (time period) of data for each i.

Contemporaneous Endogeneity

• How can we handle heterogeneity and contemporaneously endogenous explanatory variables? There are control function and GMM approaches, with the former being more convenient but imposing more restrictions.

- Papke and Wooldridge (2008, Journal of Econometrics) propose a control function approach that allows contemporaneous endogeneity and for heteroegeneity to be correlated with the instruments.
- We can start with an omitted variables formulation:

$$E(y_{it1}|\mathbf{z}_i, y_{it2}, c_{i1}, r_{it1}) = \exp(\mathbf{z}_{it1}\boldsymbol{\delta}_1 + \alpha_1 y_{it2} + c_{i1} + r_{it1}),$$

where c_{i1} is unobserved heterogeneity and r_{it1} is a time-varying omitted variable.

- The $\{\mathbf{z}_{it}\}$ – including the excluded instruments – are assumed to be strictly exogenous here. We must have at least one time-varying IV.

- If y_{it2} is (roughly) continuous we might specify

$$y_{it2} = \psi_2 + \mathbf{z}_{it}\boldsymbol{\pi}_2 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_2 + v_{it2},$$

where we have imposed the Chamberlain-Mundlak device to allow heterogeneity affecting y_{it2} to be correlated with $\mathbf{z}_i$ through the time average, $\bar{\mathbf{z}}_i$.

- If we also specify

$$c_{i1} = \psi_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + a_{i1}$$

then we can write

$$E(y_{it1}|\mathbf{z}_i, y_{it2}, v_{it1}) = \exp(\psi_1 + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \alpha_1 y_{it2} + v_{it1}),$$

where $v_{it1} = a_{i1} + r_{it1}$.

- It is reasonable (but not completely general) to assume (v_{it1}, v_{i2}) is independent of $\mathbf{z}_i$.

- If we specify $E[\exp(v_{it1})|v_{it2}] = \exp(\eta_1 + \rho_1 v_{it2})$ (as would be true under joint normality), we obtain the estimating equation

$$E(y_{it1}|\mathbf{z}_i, y_{it2}, v_{it2}) = \exp(\kappa_1 + \mathbf{z}_{it1}\boldsymbol{\delta}_1 + \alpha_1 y_{it2} + \bar{\mathbf{z}}_i\boldsymbol{\xi}_1 + \rho_1 v_{it2}).$$

- Now we can apply a simple two-step method. (1) Obtain the residuals $\hat{v}_{it2}$ from the pooled OLS estimation y_{it2} on 1, $\mathbf{z}_{it}$, $\bar{\mathbf{z}}_i$ across t and i. (2) Use a pooled NLS or QMLE (perhaps the Poisson or NegBin II) to estimate the exponential function, where $(\bar{\mathbf{z}}_i, \hat{v}_{it2})$ are explanatory variables along with $(\mathbf{z}_{it1}, y_{it2})$. (As usual, a fully set of time period dummies is a good idea in the first and second steps).
- Note that y_{it2} is not strictly exogenous in the estimating equation. and so GEE should not be used. GMM with carefully constructed moments could be.

- Estimating the ASF is straightforward:

$$\widehat{ASF}_t(\mathbf{z}_{t1}, y_{t2}) = N^{-1} \sum_{i=1}^{N} \exp(\hat{\kappa}_1 + \mathbf{z}_{t1}\hat{\boldsymbol{\delta}}_1 + \hat{\alpha}_1 y_{t2} + \bar{\mathbf{z}}_i \hat{\boldsymbol{\xi}}_1 + \hat{\rho}_1 \hat{v}_{it2});$$

that is, we average out $(\bar{\mathbf{z}}_i, \hat{v}_{it2})$.

- Test the null of contemporaneous exogeneity of y_{it2} by using a fully robust t statistic on $\hat{v}_{it2}$.

- A GMM approach can be applied if the instruments satisfy a sequential exogeneity assumption; we do not need strict exogeneity:

$$y_{it} = c_i \exp(\mathbf{x}_{it}\boldsymbol{\beta}) r_{it} \tag{113}$$

$$E(r_{it}|\mathbf{z}_{it}, \ldots, \mathbf{z}_{i1}, c_i) = 1, \tag{114}$$

which contains the the case with sequentially exogenous $\{\mathbf{x}_{it}\}$ as a special case ($\mathbf{z}_{it} = \mathbf{x}_{it}$).

- Now start with the transformation

$$\frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})} = c_i(r_{it} - r_{i,t+1}). \tag{115}$$

- In the sequential exogeneity case,

$E(r_{it}|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1},c_i) = E(r_{i,t+1}|\mathbf{x}_{it},\ldots,\mathbf{x}_{i1},c_i) = 1$, and so multiplying the moment conditions by any function of $\mathbf{x}_{it}$ is allowed. We get the previous moment conditions by multiplying through by $\exp(\mathbf{x}_{it}\boldsymbol{\beta})$.

- Cannot multiply through by $\exp(\mathbf{x}_{it}\boldsymbol{\beta})$ if $\mathbf{x}_{it}$ is contemporaneously endogenous (correlated with r_{it}).

• Can easily show that $E[c_i(r_{it} - r_{i,t+1})|c_i, \mathbf{z}_{it}, \ldots, \mathbf{z}_{i1}) = 0$, which leads to the moment conditions

$$E\left[\frac{y_{it}}{\exp(\mathbf{x}_{it}\boldsymbol{\beta})} - \frac{y_{i,t+1}}{\exp(\mathbf{x}_{i,t+1}\boldsymbol{\beta})} \middle| \mathbf{z}_{it}, \ldots, \mathbf{z}_{i1} \right] = 0, \ t = 1, \ldots, T-1. \tag{116}$$

• Using these directly generally causes computational problems. For example, if $x_{itj} \geq 0$ for some j and all i and t, with strict inequality in some cases – for example, if x_{itj} is a time dummy – then the moment conditions can be made arbitarily close to zero by choosing β_j larger and larger.

• Windmeijer (2002, Economics Letters) suggested (effectively) multiplying through by $\exp(\boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\beta})$ where

$$\boldsymbol{\mu}_{\mathbf{x}} \equiv T^{-1}\sum_{r=1}^{T} E(\mathbf{x}_{ir}). \tag{117}$$

In other words, $\boldsymbol{\mu}_{\mathbf{x}}$ is the average of the $E(\mathbf{x}_{it})$ across t. Notice that $\exp(\boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\beta})$ is a constant and so the orthogonality conditions are not changed.

• The modified moment conditions are

$$E\left[\frac{y_{it}}{\exp[(\mathbf{x}_{it}-\boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta}]}-\frac{y_{i,t+1}}{\exp[(\mathbf{x}_{i,t+1}-\boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\beta}]}\,\middle|\,\mathbf{z}_{it},\ldots,\mathbf{z}_{i1}\right]=0. \tag{118}$$

• As a practical matter, replace $\boldsymbol{\mu}_{\mathbf{x}}$ with the overall sample average,

$$\bar{\mathbf{x}}=(NT)^{-1}\sum_{i=1}^{N}\sum_{r=1}^{T}\mathbf{x}_{ir}. \tag{119}$$

• The deviated variables, $\mathbf{x}_{it}-\bar{\mathbf{x}}$, will always take on positive and negative values, and this seems to solve the GMM computational problem. (But more work could be done on this, especially in models with time dummies.)

- The sample moments look like

$$\sum_{i=1}^{N}\sum_{t=1}^{T-1}\mathbf{g}_{it}'\left[\frac{y_{it}}{\exp[(\mathbf{x}_{it}-\bar{\mathbf{x}})\boldsymbol{\beta}]}-\frac{y_{i,t+1}}{\exp[(\mathbf{x}_{i,t+1}-\bar{\mathbf{x}})\boldsymbol{\beta}]}\right]$$

where $\mathbf{g}_{it} \equiv \mathbf{g}_t(\mathbf{z}_{it},\ldots,\mathbf{z}_{i1})$ is a function of the instruments up through time t. Or, stack these over the time periods for more efficiency.

- As usual, we use GMM with an optimal weighting matrix to set the sample moments as close to zero as possible.

- For computing standard errors and conducting statistical inference, we can probably ignore the sampling variation in $\bar{\mathbf{x}}$ (it is an estimator of $\boldsymbol{\mu}_{\mathbf{x}}$) in computing standard errors. The sampling variation in $\hat{\boldsymbol{\beta}}$, given that $\hat{\boldsymbol{\beta}}$ is based on a kind of differencing, likely swamps that in $\bar{\mathbf{x}}$.
- The earlier moment conditions under sequential exogeneity replace $\bar{\mathbf{x}}$ with $\mathbf{x}_{it}$.

- An alternative approach is to multiply through by $\exp(\boldsymbol{\mu}_{\mathbf{x}_t} + \boldsymbol{\mu}_{\mathbf{x}_{t+1}})$ to get

$$E\left[\frac{y_{it}\exp(\boldsymbol{\mu}_{\mathbf{x}_{t+1}}\boldsymbol{\beta})}{\exp[(\mathbf{x}_{it} - \boldsymbol{\mu}_{\mathbf{x}_t})\boldsymbol{\beta}]} - \frac{y_{i,t+1}\exp(\boldsymbol{\mu}_{\mathbf{x}_t}\boldsymbol{\beta})}{\exp[(\mathbf{x}_{i,t+1} - \boldsymbol{\mu}_{\mathbf{x}_{t+1}})\boldsymbol{\beta}]} \middle| \mathbf{z}_{it},\ldots,\mathbf{z}_{i1} \right] = 0,$$

and then replace $\boldsymbol{\mu}_{\mathbf{x}_r}$, $r = t,\ t+1$, with its sample analog,

$$\bar{\mathbf{x}}_r = N^{-1}\sum_{i=1}^{N}\mathbf{x}_{ir}.$$

- Results in demeaning the covariates within each time period.