1 Overview

basic infor about the paper Count and count-like data in finance.

- 2022, Journal of Financial Economics
  - authors
    - Malcolm I. Wardlaw (University of Georgia)
    - Zack Liu (University of Houston)
    - Jonathan B. Cohn(University of Texas at Austin)

research background

- many applications deals with count and count-like outcome.
  - ex: number of patents, tons of toxic emissions,number of workplace injuries.
- feature of count and count-like outcome
  - non-negative outcome
  - many zeros
- some econometric approaches used
  - linear regression
  - log-linear regression
  - log1plus regression
  - Poisson regression
  - others: negative binomial , zero-inflated poisson

summary of the paper

- do not use linear regression
- it is risky to use log-linear regression
- do not use bullshit log1plus regression
- use Poisson regression

format of the presentation heuristic problems solving to understand the key takeaways of this paper.

2 Some concepts and definitions

Definition 1 (mean and variance of log normal distribution). Suppose X follows log normal distribution with $\ln(X) \sim N(\mu, \sigma^2)$ then

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$Var(X) = \left[\exp\left(\sigma^2\right) - 1\right] \exp\left(2\mu + \sigma^2\right).$$

Definition 2 (IHS transformation). Inverse hyperbolic sine transformation, i.e,

$$sinh^{-1}(y) = \ln\left(x + \sqrt{x^2+1}\right), -\infty < x < \infty.$$

Problem 1 (lognormal and normal mean variance relationship). Suppose X follows log normal distribution and we know $E(X)$ and $Var(X)$ and please derive μ and σ^2.

solution:

From $E(X)$ and $Var(X)$ formulas in Definition 1, we have

$$\sigma^2 = \ln(\frac{Var(X)}{[E(X)]^2} + 1)$$

$$\mu = \ln[E(X)] - \frac{1}{2}\ln(\frac{Var(X)}{[E(X)]^2} + 1).$$

Example 1 (exposure variable in Poisson regression). The exposure variable accounts for the varying amount of time or space each observation represents. It helps to scale the counts appropriately, making the model more accurate when dealing with data collected over different periods or areas. For instance, if you're modeling the number of accidents per day in different cities, the exposure variable would represent the number of days observed for each city.

3 Key takeaways

3.1 linear regression

Exercise 1 (true model is linear and estimate linear regression). Suppose the true model is

$$y = x\beta + \varepsilon$$

with observable data $(x_i, y_i)_{i=1}^N$ and y_i is a count variable and we estimate a linear regression and obtain $\hat{\beta}$.

implication:

1. prediction $\hat{y} = x\hat{\beta}$ may be negative thus use linear model $y = x\beta + \varepsilon$ is not suitable to model count,count-like data.

2. In the paper, they also argue linear regression suffers from efficiency loss.

Exercise 2 (**true model is nonlinear and estimate linear regression**)**.** Suppose the true model is

$$y = e^{x\beta}\eta$$

with observable data $(x_i, y_i)_{i=1}^N$ and $E(\eta|x) = 1$ and we estimate a linear regression

$$y = x\pi + v$$

and obtain $\hat{\pi}$.

Problem 2. Show that $\hat{\pi}$ is inconsistent for β.
solution:
obvious due to OLS formula: $\hat{\pi} = (\sum x_i^2)^{-1}(\sum x_i y_i) = (\sum x_i^2)^{-1}(\sum x_i e^{x_i\beta}\eta_i)$.
implication:

1. when the true model is nonlinear and estimate a linear regression, consistency is difficult to ensure.
2. efficiency loss may exist compared to using nonlinear models like Poisson regression.

3.2 Log-linear regression

Exercise 3 (**log-linear regression**)**.** Suppose the true model is

$$y = e^{x\beta}\eta$$

and $E(\eta|x) = 1$ so that $E(y|x) = e^{x\beta}$.With observable data $(x_i, y_i)_{i=1}^N$, we can estimate a log-linear regression

$$\ln y = x\beta + \ln(\eta)$$

and obtain estimate $\hat{\beta}$.

Conclusion (**Takeaway 1**). **with heteroskedasticity in η, log-linear estimates are inconsistent.**

Problem 3. Show that $\hat{\beta}$ is consistent when $E(\ln\eta|x) = 0$.
solution:
obvious due to OLS formula: $\hat{\beta} = (\sum x_i^2)^{-1}(\sum x_i \log(y_i)) = \beta + (\sum x_i^2)^{-1}(\sum x_i \log(\eta_i))$ and the law of iterated expectations.
implication:

- $E(\ln\eta|x) = 0$ is a sufficient condition for log-linear regression to be consistent.

Problem 4. Show that $E(\eta|x) = 1$ does not necessarily imply $E(\ln\eta|x) = 0$.
solution:
Just take a counter example when distribution of η does not depend on x.
When η is discrete with two values $\frac{1}{2}$,2 with probability $\frac{2}{3}$ and $\frac{1}{3}$ separately then

$$E(\ln\eta|x) = \frac{2}{3}\ln(\frac{1}{2}) + \frac{1}{3}\ln(2) = -\frac{1}{3}\ln(2) \neq 0.$$

implication:

- $E(\eta|x) = 1$ is not enough to ensure $E(\ln \eta|x) = 0$, the condition for consistency of log-linear regression.

Problem 5. Show that $E(\ln \eta)$ depends on higher order of moments of η, i.e., $E(\eta^2), E(\eta^3), ...$

solution:

To see this, consider the expansion of $\ln x = \sum_{n=1}^{\infty}(-1)^{n-1}\frac{(x-1)^n}{n}$ therefore

$$\begin{aligned} E(\ln \eta) &= \int \ln(\eta) f(\eta) d\eta \\ &= \int \sum_{n=1}^{\infty}(-1)^{n-1}\frac{(\eta-1)^n}{n} f(\eta) d\eta \end{aligned}$$

where $f(\eta)$ is the density of η.

implication:

- expectation of logarithm of a random variable X will depend on its higher order moments.

Problem 6. Show that when $E(\eta|x) = 1$ and $E(\eta^2|x) = g(x) > 0$ where $g(x)$ is a non-constant function depend on x, then $E(\ln \eta|x)$ is a function in x.

solution:

To see this, use the conclusion of the previous problem, $E(\ln \eta|x)$ will depend on higher order of moments of $\eta|x$ thus depend on $E(\eta^2|x)$ therefore, $E(\ln \eta|x)$ will be a function in x.

implication:

- when there is heteroskedasticity in η, $E(\ln \eta|x) \neq 0$ thus log-linear estimate is not consistent.

Conclusion (**Takeaway 2**). **Bias due to heteroskedasticity in η can cause $\hat{\beta}$ to have the wrong sign.**

Under the conditions given in Exercise 3, we further assume $x \sim N(0, \sigma_x^2)$ and η is log-normally distributed with mean 1 and standard deviation $\sigma_\eta(x) = e^{\delta x}$.

Problem 7. Derive the expectation and variance of $\ln(\eta)|x$.

solution:

$\ln(\eta)|x$ is a normally distributed random variable and we use the formulas in Problem 1.

$$\begin{aligned} E[\ln(\eta)|x] &= -\frac{1}{2}\ln(e^{2\delta x} + 1) \\ Var[\ln(\eta)|x] &= \ln(e^{2\delta x} + 1). \end{aligned}$$

Problem 8. Show that $\hat{\beta}$ is not consistent.

Solution:

There is heteroskedasticity in η thus we can use the conclusion of Problem (6).

Or we can try to show this directly by using

$$\hat{\beta} = (\sum x_i^2)^{-1}(\sum x_i \log(y_i)) = \beta + (\sum x_i^2)^{-1}(\sum x_i \log(\eta_i)).$$

We only need to show $E(x \ln(\eta)) \neq 0$ and this is what we will do in the next problem.

Problem 9. Show that $E(x\ln(\eta)) = -\frac{1}{2}\delta\sigma_x^2$.

Solution:

Since $\ln(\eta)|x$ is normally distributed with mean $E[\ln(\eta)|x] = -\frac{1}{2}\ln(e^{2\delta x}+1)$, $x\ln(\eta)|x$ is also normally distributed with mean

$$E[x\ln(\eta)|x] = -\frac{1}{2}x\ln(e^{2\delta x}+1).$$

Therefore by law of iterated expectation

$$\begin{aligned} E(x\ln(\eta)) &= E\left[E[x\ln(\eta)|x]\right] \\ &= \int\left\{-\frac{1}{2}x\ln(e^{2\delta x}+1)\right\}\phi(x)dx \end{aligned}$$

where $\phi(x)$ is the PDF of x.

We construct a new function

$$g(x) = -\frac{1}{2}x\ln(e^{2\delta x}+1) + \frac{1}{2}\delta x^2$$

and since

$$\begin{aligned} g(-x) &= \frac{1}{2}x\ln(\frac{e^{2\delta x}+1}{e^{2\delta x}}) + \frac{1}{2}\delta x^2 \\ &= \frac{1}{2}x\ln(e^{2\delta x}+1) - \frac{1}{2}\delta x^2 \\ &= -g(x), \end{aligned}$$

$g(x)$ is odd.

Therefore

$$\begin{aligned} E(x\ln(\eta)) &= \int\left\{-\frac{1}{2}x\ln(e^{2\delta x}+1)\right\}\phi(x)dx \\ &= \int\left\{-\frac{1}{2}x\ln(e^{2\delta x}+1) + \frac{1}{2}\delta x^2 - \frac{1}{2}\delta x^2\right\}\phi(x)dx \\ &= \int g(x)\phi(x)dx - \int\frac{1}{2}\delta x^2\phi(x)dx \\ &= -\int\frac{1}{2}\delta x^2\phi(x)dx \\ &= -\frac{1}{2}\delta E(x^2) \\ &= -\frac{1}{2}\delta\sigma_x^2. \end{aligned}$$

Remark 1. **There are some errors in the proof provided in Appendix A.**

Problem 10. Show that $\text{plim}_{n\to\infty}\hat{\beta} = \beta - \frac{\delta}{2}$.

Solution:

The bias is the limit of $(\sum x_i^2/N)^{-1}(\sum x_i\log(\eta_i)/N)$, with $E(x_i^2) = \sigma_x^2$, the conclusion follows $E(x\ln(\eta)) = -\frac{1}{2}\delta\sigma_x^2$ derived in Problem 9.

Conclusion (**Takeaway 3**). **All else equal when $\delta > 0$ ($\delta < 0$),then the bias is downward (upward).**

This is a straightforward application of the result in Problem 10.

- when $\delta > 0$, $\text{plim}_{n\to\infty}\hat{\beta} - \beta = \frac{\delta}{2} > 0$.
- when $\delta < 0$, $\text{plim}_{n\to\infty}\hat{\beta} - \beta = \frac{\delta}{2} < 0$.

Remark 2. In the paper, they derived this conclusion in a more general context using concept of second order stochastic dominance. If you are interested, please refer to their equation (6).

3.3 log1plus regression

Exercise 4 (log1plus regression). Suppose the true model is

$$y = e^{x\beta}\eta$$

with homoskedastic η and with observable data $(x_i, y_i)_{i=1}^N$, we can estimate a log1plus regression

$$\ln(y+1) = x\lambda + \phi$$

and obtain estimate $\hat{\lambda} = (\sum x_i^2)^{-1}(\sum x_i \log(y_i+1))$.

Problem 11. Show that $\lambda \neq \beta$.

Solution:

λ is

$$\begin{aligned}\lambda &= \frac{\partial E[\ln(y+1)|x]}{\partial x} \\ &= \frac{1}{E[(y+1)|x]}\frac{\partial E[(y+1)|x]}{\partial x} \\ &= \frac{1}{1+E(y|x)}\frac{\partial E(y|x)}{\partial x}\end{aligned}$$

while

$$\beta = \frac{\partial E(\ln(y)|x)}{\partial x} = \frac{1}{E(y|x)}\frac{\partial E(y|x)}{\partial x}.$$

Conclusion (**Takeaway 4**). **log1plus regression coefficient are not interpretable as semi-elasticities nor can any economically meaningful relationship.**

Problem 12. Derive the relationship between λ and β.

Solution:

From previous problem, we have

$$\lambda\,[1+E(y|x)] = \beta\,[E(y|x)]$$

therefore

$$\lambda = \frac{E(y|x)}{1+E(y|x)}\beta.$$

Implications:

1. β cannot be recovered from λ as $E(y|x)$ is not observable.
2. When $E(y|x)$ is large, $\lambda \approx \beta$. But when $E(y|x)$ is large, y has few observations with zero values and no need to use log1plus.
3. When $E(y|x)$ is small, λ and β has a big difference.
4. λ and β has the same sign as $E(y|x) > 0$.

Conclusion (**Takeaway 5**). **log1plus regression coefficient is almost certain to suffer from two forms of bias that make even the sign of β difficult to infer from λ estimates.**

I will just provide some intuition.

1. nonlinear relationship between $\ln(1+y)$ and x under any reasonable economic model between y and x.
2. nothing special to add 1 instead of another constant c.

Problem 13 (**bias in log1plus regression**)**.** Suppose the true model is

$$y = e^{x\beta}\eta$$

with homoskedastic η and with observable data $(x_i, y_i)_{i=1}^N$, we can estimate a log1plus regression

$$\ln(y+1) = x\lambda + \phi.$$

Show that $\ln(y+1)$ is nonlinear in x.

solution:$\ln(y+1) = \ln(e^{x\beta}\eta + 1)$ is a nonlinear function in x.

3.4 IHS regression

Conclusion (**Takeaway 6**). **Takeaway 4 and 5 holds for linear regression of an IHS-transformed outcome variable.**

3.5 Poisson regression

Definition 3 (**Poisson regression**)**.** The model assumes the dependent variable has a Poisson distribution conditional on x with density

$$f(y|x) = \frac{e^{-\mu(x)}[\mu(x)]^y}{y!}$$

with conditional expectation

$$\mu(x) = E(y|x) = e^{x\beta}.$$

Note 1 (**Poisson Pseudo Maximum likelihood estimator**). The likelihood of an observed data $\{x_i, y_i\}_{i=1}^N$ is

$$L(y_1, \cdots, y_N|x_1, \cdots, x_N; \beta) = \prod_{i=1}^{N} \frac{e^{-e^{x_i\beta}}[e^{x_i\beta}]^{y_i}}{y_i!},$$

and then log-likelihood is

$$\ln L = \sum_{i=1}^{N} \left[y_i(x_i\beta) - e^{x_i\beta} - (y_i!) \right].$$

Take first order condition with respect to β

$$\sum_{i=1}^{N} \left[y_i - e^{x_i\beta} \right] x_i = 0.$$

Example 2 (**Fixed effects Poisson regression**)**.** Let α_i be the fixed effects for group i, then the fixed effects Poisson model condition expectation is

$$E(y_{it}|x_{it}) = e^{\alpha_i + x_{it}\beta} = e^{\alpha_i} e^{x_{it}\beta}.$$

The likelihood of an observed data $\{x_{it}, y_{it}\}_{i=1,t=1}^{N,T}$ is

$$L(\beta,\alpha) = \prod_{i=1}^{N}\prod_{t=1}^{T} \frac{e^{-e^{\alpha_i + x_{it}\beta}} [e^{\alpha_i + x_{it}\beta}]^{y_{it}}}{y_{it}!},$$

The log-likelihood is

$$\ln L(\beta,\alpha) = \sum_i \left[-\sum_t e^{\alpha_i + x_{it}\beta} + \sum_t y_{it}(x_{it}\beta + \alpha_i) - \sum_t \ln y_{it}! \right]$$

differentiating with respect to α_i yields

$$\hat{\alpha}_i = \frac{\sum_t y_{it}}{\sum_t e^{x_{it}\beta}}$$

and substitute into the log-likelihood we obtain the concentrated likelihood function

$$\ln L_{conc}(\beta) \propto \sum_i \sum_t \left[y_{it} x_{it}\beta - y_{it} \ln \left(\sum_s e^{x_{it}\beta} \right) \right].$$

Conclusion (**Takeaway 7,9,10,11**)**.** Poisson regression has the following advantages:

1. clear interpretation of semi-elasticity.(Takeaway 7)
2. do not require the relationship between higher-order moments of error and covariate for consistent estimation thus allow heteroskedasticity in the error η (compared to log-linear regression).(Takeaway 7)
3. admit separable group fixed effects.(Takeaway 9)
4. fixed effects Poisson requires excluding any group which the outcomes are all zero. These groups contain no information about regression coefficient.(Takeaway 10)
5. produce valid estimates when y is continuous.(Takeaway 11)
6. admits an exposure variable that acts as a scaling variable and can be used in IV regression.(Takeaway 11)

Example 3 (**conditional mean-variance equality and overdispersion**). We are interested in modelling the following relationship: y is the number of accidents at an intersection and $x$is the average daily traffic volume of this intersection.Suppose we use Poisson model and $E(y|x) = e^{x\beta}$ with $\beta > 0$.

conditional mean-variance equality holds when

- small x, small $E(y|x) = e^{x\beta}$ and small variance of $y|x$ and $E(y|x) = Var(y|x)$.
- larger x, larger $E(y|x) = e^{x\beta}$ and larger variance of $y|x$ but *still* $E(y|x) = Var(y|x)$.

Overdispersion happens when

- larger x, larger $E(y|x) = e^{x\beta}$ and larger variance of $y|x$ but $E(y|x) < Var(y|x)$.

When the daily traffic volume increases, the variance of number of accidents increases more, thus called "overdispersion".

Conclusion (**Takeaway 8**). **Poisson regression imposes conditional mean-variance equality restriction and violations of this resctriction reduces efficiency but does not cause any bias.**

3.6 Other count-based regression models

Definition 4 (**Negative binomial model**). Binomial model: $B(n,p)$, the probability distribution of the number of success in n experiments with success probability p. The support is $1, \cdots, n$.

Negative binomial model: $NB(r,p)$, the probability of the number of failures before the the number of success reaches r. The support is $1, \cdots, n, \cdots$.

The PDF of $NB(r,p)$ is

$$P(Y = k) = p \times C_{k+r-1}^{k}(1-p)^k p^{r-1} = C_{k+r-1}^{k}(1-p)^k p^r.$$

Since $C_{k+r-1}^{k} = \frac{(k+r-1)!}{k!(r-1)!} = \frac{\Gamma(k+r)}{k!\Gamma(r)}$, an alternative of PDF is

$$P(Y = k) = \frac{\Gamma(k+r)}{k!\Gamma(r)}(1-p)^k p^r$$

where $\Gamma(\cdot)$ is the gamma function.

The mean of $NB(r,p)$ is $\frac{r(1-p)}{p}$ and variance is $\frac{r(1-p)}{p^2}$.

Example 4 (**Fixed effects negative binomial model**). Hausman, Hall and Griliches (1984) has proposed a fixed effects negative binomial model for panel data but this is criticized by Allison and Waterman (2002). They argued that HHG model is not a true fixed effects model.

The PDF of HHG model is

$$f\left(y_{it} \mid \lambda_{it}, \theta_i\right) = \frac{\Gamma\left(\lambda_{it} + y_{it}\right)}{\Gamma\left(\lambda_{it}\right)\Gamma\left(y_{it}+1\right)}\left(\frac{\theta_i}{1+\theta_i}\right)^{y_{it}}\left(\frac{1}{1+\theta_i}\right)^{\lambda_{it}}$$

where $\lambda_{it} = e^{x_{it}\beta}$ and θ_i is the fixed effects.

The mean and variance of y_{it} are given by

$$E\left(y_{it}\right) = \theta_i \lambda_{it}$$
$$Var\left(y_{it}\right) = \left(1+\theta_i\right)\theta_i \lambda_{it}.$$

If we write $\theta_i = e^{\delta_i}$ then

$$E\left(y_{it}\right) = \exp\left(\delta_i + \beta x_{it}\right)$$
$$Var\left(y_{it}\right) = \left(1 + e^{\delta_i}\right) E\left(y_{it}\right).$$

The problem is that δ_i plays a different role than x_{it}.

- x_{it} only affects variance through $E(y_{it})$
- δ_i affects the variance indirectly through $E(y_{it})$ and directly through e^{δ_i}.

Definition 5 (**Zero inflated Poisson (ZIP)**)**.** Zero-inflated Poisson distribution has the following PDF

$$P(Y = 0) = \pi + (1 - \pi)e^{-\lambda}$$

$$P\left(Y = y_i\right) = (1 - \pi)\frac{\lambda^{y_i} e^{-\lambda}}{y_i!}, \quad y_i = 1, 2, 3, \ldots$$

The mean is $(1 - \pi)\lambda$ and the variance is $\lambda(1 - \pi)(1 + \pi\lambda)$.

Conclusion (**Takeaway 12**)*.* **Negative binomial or zero-inflated Poisson regression may be more efficient than Poisson regression but do not admit separable group fixed effects.**