Modelos de regressão truncada, truncamento ocasional e Poisson

Wooldridge, capítulo 17

Regressão truncada

- Truncamento é um problema mais sério do que censura uma vez que não observamos qualquer informação sobre certo segmento da população
- Modelo de regressão truncada normal

$$y = \beta_0 + \boldsymbol{x\beta} + u, u|\boldsymbol{x} \sim N(0, \sigma^2) \quad (1)$$

- Sob (1) sabemos que, dada uma amostra aleatória da população, o MQO é o procedimento mais eficiente de estimação

Problema

- Não observamos uma amostra aleatória da população; a hipótese RLM.2 é violada

- Uma extração aleatória $(\boldsymbol{x}_i, y_i)$ é observada somente se $y_i \leq c_i$, onde c_i é o valor do truncamento

- Se $\{(\boldsymbol{x}_i . y_i): i = 1, \dots, n\}$ for nossa amostra observada, y_i será necessariamente menor ou igual a c_i. Isso difere do modelo de regressão censurada em que y_i pode ser maior ou igual a c_i: simplesmente não observamos y_i se $y_i > c_i$
- Em um modelo de regressão censurada observamos $\boldsymbol{x}_i$ para qualquer observação extraída aleatoriamente da população; no modelo truncado somente observamos $\boldsymbol{x}_i$ se $y_i \leq c_i$

- Estimar por máxima verossimilhança
- Isso conduz a estimadores consistentes e aproximadamente normais
- Inferência é padrão

Exemplo

- Dados sobre o tempo em meses até que um ex-presidiário seja preso depois de ter sido solto
- 1.445 ex-presos, 893 não foram presos durante o tempo em que foram vigiados
- 893 observações censuradas

- Os dados desse exemplo poderiam ser usados como uma amostra truncada se eliminássemos todos os dados de uma observação sempre que ela tivesse sido censurada. Isso nos daria 552 observações
- O fato de conhecermos um limite inferior de 893 durações é, no entanto, informação útil; a regressão censurada usa essas informações, o que a regressão truncada não faz

- Regressão truncada: caso especial do problema conhecido como seleção de amostra não aleatória
- Outro exemplo: truncamento ocasional

Truncamento ocasional

- Forma de seleção amostral
- Neste caso, não observamos y em razão do resultado de outra variável
- Isto faz com que somente observemos y para um subconjunto da população

Principal exemplo: função oferta de salário

- Interesse: como vários fatores, por exemplo educação, afetam o salário de um indivíduo
- Para as pessoas que estão na força de trabalho, observamos a oferta de salário como o salário corrente. Contudo, para aqueles que estão desempregados, não observamos a oferta de salário
- Como trabalhar pode estar sistematicamente correlacionado a fatores não observáveis que afetam a oferta de salário, usar somente pessoas que estejam trabalhando – como temos feito em todos os exemplos de salários até agora – pode produzir estimadores viesados dos parâmetros na equação de oferta de salário
- O truncamento do salário é ocasional, pois ele depende de outra variável, ou seja, da participação na força de trabalho

- Abordagem habitual para o truncamento ocasional é adicionar uma equação de seleção explicitamente ao modelo populacional de interesse:
- $y = \boldsymbol{x\beta} + u, E(u|x) = 0$ (1)
- $s = 1[\boldsymbol{z\gamma} + v \geq 0]$ (2)
- Equação (1) é de maior interesse
- Assumir que z é exógeno: E(u|$\boldsymbol{x}$,z)=0
- Assumir que v é independente de **z** e, portanto, de $\boldsymbol{x}$
- Assumir que v tem uma distribuição normal padrão

Para demonstração ver Wooldridge pág. 552

- Isto leva à equação

$E(y|z, s = 1) = \boldsymbol{x\beta} + \rho\lambda(\boldsymbol{z\gamma})$ (3)

onde $\lambda(c) = \phi(c)/\Phi(c)$ é a razão de Mills inversa

- Equação (3) mostra que o valor esperado de y é igual a $\boldsymbol{x\beta}$ mais um termo adicional que depende do inverso da razão de Mills avaliado em $\boldsymbol{z\gamma}$
- Essa equação mostra que é possível estimar $\boldsymbol{\beta}$ usando somente a amostra selecionada desde que incluamos $\lambda(\boldsymbol{z\gamma})$ como um regressor adicional

- Se $\rho = 0$, $\lambda(\boldsymbol{z\gamma})$ não aparecerá e o MQO de y sobre $\boldsymbol{x}$ usando a amostra selecionada estima consistentemente $\boldsymbol{\beta}$. Fora isso, omitimos efetivamente uma variável $\lambda(\boldsymbol{z\gamma})$, que geralmente é correlacionada com $\boldsymbol{x}$
- $\rho = 0$ quando u e v são não correlacionados

- Como γ é desconhecido, não podemos avaliar $\lambda(\boldsymbol{z\gamma})$ para cada i
- Porém, com base nas hipóteses que fizemos, s dado $\mathbf{z}$ segue um modelo probit:
- $P(s = 1|\boldsymbol{z}) = \Phi(\boldsymbol{zy})$

- Podemos estimar γ pelo probit de s_i sobre $\boldsymbol{z}_i$ usando a amostra inteira. Em uma segunda etapa, podemos estimar $\boldsymbol{\beta}$
- Procedimento chamado Heckit (Heckman (1976))

- Um teste simples do viés de seleção pode ser feita a partir da regressão de y_i sobre x_i e $\hat{\lambda}_i$
- Fazer um teste t usual de $\hat{\lambda}_i$. Sob H_0:ρ =0, não há problema de seleção de amostra

Importante

- Devemos ter pelo menos um elemento de **z** que não está também em **x,** ou seja, uma variável que afeta a seleção mas que não tenha um efeito parcial sobre y
- Razão para isso: embora a razão de Mills inversa seja uma função não linear de z, ela frequentemente é bem aproximada por uma função linear. Se **z=x**, $\hat{\lambda}_i$ pode ser altamente correlacionado com os elementos de $\boldsymbol{x}_i$. Como sabemos, tal multicolinearidade pode conduzir a erros-padrão muito altos dos $\hat{\beta}_j$

- Alternativa ao método de estimação em duas etapas é a estimação de máxima verossimilhança completa

Variável de contagem

- Outro tipo de variável dependente não negativa {0,1,2,...}
- Exemplos: número de filhos de uma mulher, número de vezes que alguém foi preso durante o ano, o número de patentes solicitadas por uma firma durante um ano

- Normalidade era a hipótese de distribuição padrão da regressão
- Variável de contagem não pode ter uma distribuição normal
- Distribuição de Poisson

- Modelo de regressão de Poisson: modelo mais básico para lidar com dados de contagem
- A probabilidade de que y assuma qualquer valor $h \in \{0,1,2,\dots\}$ para esse modelo pode ser escrita como

$$P(y = h|x) = \exp[-\exp(\boldsymbol{x\beta})][\exp(\boldsymbol{x\beta})]^h/\mathrm{h!} \quad (1)$$

Lembrando

- Seja $y \sim Poisson\ (\lambda)$ então

$$P(y = h) = \frac{e^{-\lambda}\lambda^{h}}{h!}$$

$$E(y) = var(y) = \lambda$$

- Pertence à família exponencial
- λ (o número de ocorrência) é constante ao longo do tempo
- Incrementos independentes: a informação sobre o número de ocorrências em um período nada revela sobre o número em outro período distinto

Os parâmetros do modelo de Poisson são muito mais fáceis de interpretar do que os dos modelos logit e probit

◆média condicional de y é $E(y|\boldsymbol{x}) = e^{\boldsymbol{x\beta}}$, de tal forma que cada parâmetro de inclinação β_j tem a interpretação de semi-elasticidade:

$$\frac{\partial E(y|\boldsymbol{x})}{\partial x_j} = \beta_j . e^{\boldsymbol{x\beta}} = \beta_j . E(y|\boldsymbol{x}) \leftrightarrow$$

$$\beta_j = \frac{1}{E(y|\boldsymbol{x})} . \frac{\partial E(y|\boldsymbol{x})}{\partial x_j}$$

Se x_j aumenta em uma unidade (e os outros regressores permanecem constantes), $E(y|\boldsymbol{x})$ irá aumentar aproximadamente em $100.\beta_j$ por cento

- Estimar por máxima verossimilhança
- A função log-verossimilhança é simples de maximizar , embora as EMVs não sejam obtidas em forma fechada
- Ver livro

- Embora a distribuição de Poisson seja o primeiro passo para dados de contagem, ela é restritiva demais
- Todas as probabilidades e os momentos de ordem mais alta são inteiramente determinados pela média. Em particular, $E(y) = var(y)$

- Se a variância de *y* não é igual à sua média, mas a média condicional ainda é corretamente especificada, as estimativas de Poisson são consistentes, mas os erros-padrão e todas as inferências baseadas neles não são válidas
- Solução simples: Quando usamos a EMV de Poisson, mas não assumimos que a distribuição seja inteiramente correta, chamamos a análise de estimação de quase-máxima verossimilhança (EQMV)
- De forma similar à inferência robusta à heterocedasticidade para MQO, vista em EAE 324, os erros padrão podem ser ajustados

- Um ajuste simples dos erros-padrão está disponível quando assumimos que a variância é proporcional à média:

$Var(y|x) = \sigma^2 E(y|x) \quad (2)$

Onde σ^2>0 é um parâmetro desconhecido

- Quando σ^2=1, obtemos a hipótese de variância de Poisson
- Quando σ^2 >1, a variância será maior que a média para todos os $\boldsymbol{x}$ (superdispersão)
- Quando σ^2<1 (subdispersão)

- Sob a hipótese de distribuição de Poisson, podemos usar a estatística razão de verossimilhança para testar restrições de exclusão
- Sob a hipótese menos restritiva (2) há um ajuste simples e então chamamos a estatística de estatística quase-razão de verossimilhança