Censura e Seletividade

Alexandre Gori Maia

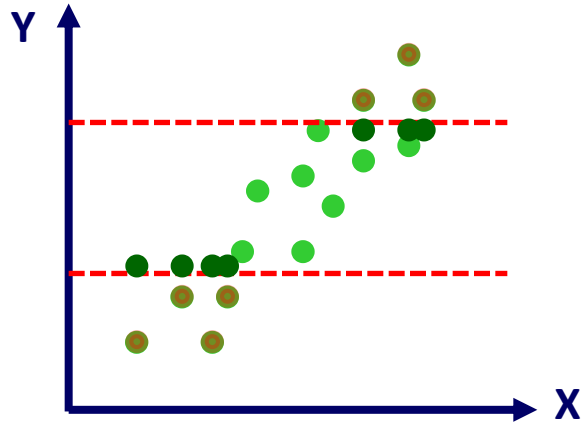
Instituto de Economia - UNICAMP

Ementa

- Censura e truncamento nos dados
- Modelo para variável dependente censurada
- Modelo com correção de seletividade

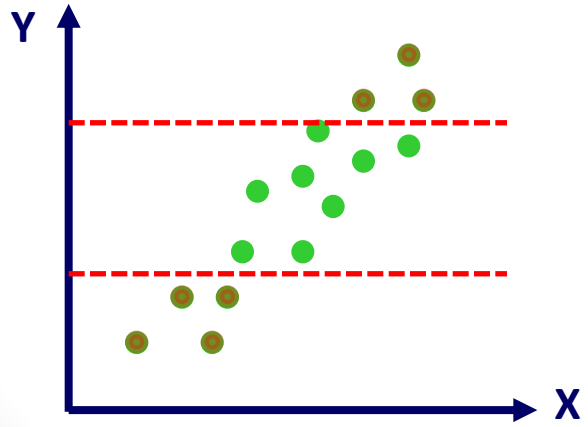
Censura e Truncamento

Censura



No caso da censura, há alguma limitação imposta à mensuração da variável dependente, impedindo que observemos valores inferiores (censura à esquerda) ou superiores (censura à direita) da variável dependente. **As observações permanecem na amostra**.

Truncamento



No caso do truncamento, há alguma limitação imposta à mensuração da variável dependente, impedindo que observações com valores inferiores (truncamento à esquerda) ou superiores (truncamento à direita) da variável dependente sejam observadas. Assim, **essas observações não pertencem à amostra**.

Esperança Condicional Zero

Seja o modelo de regressão:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{e}$$

Um pressuposto importante dos estimadores de MQO é que os erros do modelo (**e**) não se relacionem com os regressores (vetor **x**).

$$E(\mathbf{e}|\mathbf{x}) = 0 \quad \Rightarrow \quad E(\mathbf{y}|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$$

Caso esse pressuposto seja quebrado, os estimadores de MQO serão tendenciosos e inconsistentes, ou seja:

$$E(\mathbf{e}|\mathbf{x}) \neq 0 \quad \Rightarrow \quad E(\mathbf{y}|\mathbf{x}) \neq \mathbf{x}\boldsymbol{\beta}$$

O pressuposto de esperança condicional zero dos erros pode ser quebrado em diversas situações, entre esses, a **censura** na variável dependente ou **seletividade** na amostra;

Censura em Y - Definição

Suponha que a relação linear seja dada pelas variáveis:

$$Y^* = \mathbf{x}\boldsymbol{\beta} + e^* \Rightarrow E(Y^*|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$$

Onde Y^* representa, por exemplo, a produção agrícola. O problema é que Y^* nem sempre é observado. Por exemplo, Y^* não será observado para produtores que não cultivaram na safra agrícola.

A variável observada Y, que é a produção declarada pelo produtor, será neste caso dada por:

$$Y_i = Y_i^* \ para \ Y_i^* > 0$$

$$Y_i = 0 \ para \ Y_i^* \leq 0$$

Se assumirmos simplificadamente que a relação entre Y e $\mathbf{x}$ será linear, poderemos ter estimativas tendenciosas e inconsistentes para $\boldsymbol{\beta}$, ou seja:

$$E(Y|\mathbf{x}) \neq \mathbf{x}\boldsymbol{\beta}$$

Modelo Tobit – Definição

O modelo Tobit assume que a relação linear para a variável latente Y^* (não observada) seja dada por:

$$Y^* = \mathbf{x}\boldsymbol{\beta} + e^* \quad \Rightarrow \quad E(Y^*|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$$

Como Y^* é não observado, os estimadores de $\boldsymbol{\beta}$ derivam-se das relações observadas para (1) valores censurados de Y e (2) valores não censurados de Y. Caso, por exemplo, a censura ocorra à esquerda de 0, a equação para os valores censurados de Y será dada por:

$$P(Y = 0|\mathbf{x}) = 1 - \Phi(\mathbf{x}\boldsymbol{\beta})$$

A função não linear Φ define a relação entre os valores não censurados de Y e $\mathbf{x}$. A partir desta relação, obtemos um termo denominado Razão Inversa de Mills (RIM).

Por sua vez, a função para os valores não censurados de Y será definida por:

$$E(Y|Y > 0, \mathbf{x}) = \mathbf{x}\boldsymbol{\beta}\text{+RIM}$$

Assumindo essas duas equações, os estimadores são então obtidos por Máxima Verossimilhança.

Modelo Tobit - Exemplo

Suponha que o valor da produção agrícola (*Y*, em R$ no ano) seja uma função linear da área do estabelecimento (A, em hectares) e número membros familiares ocupados na produção (*F*):

$$Y = \beta_0 + \beta_1 A + \beta_2 F + \varepsilon$$

As estimativas de MQO para uma amostra de 333 produtores seriam:

$$Y = -3527^{***} + 576^{***} A + 3{,}976 F + \hat{\varepsilon}$$

Entretanto, há censura à esquerda no valor da produção, que assume zero para 24 dos 333 produtores. As estimativas de MQO podem ser viesadas e inconsistentes.

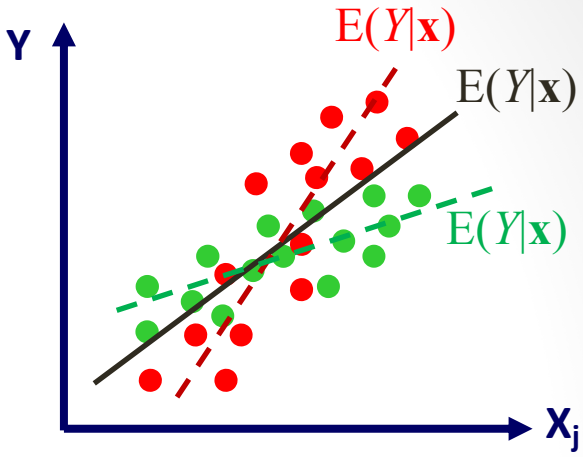
As estimativas do modelo Tobit seriam:

$$Y = -50233^{***} + 593^{***} A + 14918 F + \hat{\varepsilon}$$

As estimativas do modelo Tobit não são diretamente comparáveis às de MQO, pois limitam-se ao subconjunto de observações não censuradas (*Y*>0).

Viés de Seleção - Definição

Seja a relação entre Y e $\mathbf{x}$: $E(Y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$



Suponha, por exemplo, que haja dois grupos na população, vermelho e verde, cada um com relações distintas entre Y e **x**.

Caso a seleção não seja aleatória e a amostra seja majoritariamente representada por um dos grupos, as estimativas de MQO serão viesadas e inconsistentes (**viés de seleção**).

Por exemplo, se a amostra conter exclusivamente observações vermelhas, os estimadores de MQO superestimarão a real relação entre Y e ***x***.

O viés de seleção pode ser causado por diversos fatores, por exemplo:

- Truncamento dos dados;
- Falhas no planejamento da pesquisa;
- Erros na declaração dos valores das variáveis, resultando em valores nulos;

Modelo Heckit– Definição

Seja a relação entre Y e $\mathbf{x}$: $E(Y|\mathbf{x}) = \mathbf{x\beta}$

Em que há seleção na amostra, ou seja, os valores de Y não são observados para uma parcela da população.

O modelo de Heckit pode ser desenvolvido em duas etapas. Na primeira etapa, a equação de seleção define a probabilidade de o valor de Y ser observado (D=1):

$$P(D = 1|\mathbf{z}) = \Phi(\mathbf{z})$$

O vetor $\mathbf{z}$ pode ser o mesmo que $\mathbf{x}$, mas o ideal é que contenha mais regressores (variáveis instrumentais).

A partir da equação de seleção, define-se o termo denominado Razão Inversa de Mills (RIM). Então, na segunda etapa, define-se a equação para Y com correção da seletividade:

$$E(Y|D = 1, \mathbf{x}) = \mathbf{x\beta}+\text{RIM}$$

O modelo Heckit pode ser estimado pelo método de máxima verossimilhança.

Modelo Heckit - Exemplo

Seja agora um modelo logarítmica para o valor da produção agrícola como uma função da área:

$$\ln Y = \beta_0 + \beta_1 \ln A + \beta_2 F + \varepsilon$$

Ao aplicar a transformação logarítmica, eu elimino 24 observações da amostra com valores zero para *Y*. As estimativas de MQO para esse subconjunto da amostra seriam:

$$\ln Y = 7.381^{***} + 0.433^{***} \ln A + 0.127 F + \hat{\varepsilon}$$

O truncamento da variável dependente (*Y*>0) pode introduzir viés de seletividade na amostra, ou seja, as estimativas de MQO podem ser tendenciosas.

As estimativas do modelo Heckit seriam:

$$\ln Y = 7.619^{***} + 0.434^{***} \ln A + 0.070 F + \hat{\varepsilon}$$

As estimativas do modelo Heckit são comparáveis às de MQO, pois ambas referem-se ao subconjunto de produtores com valores positivos para a produção agrícola (*Y*>0).

Exercício

1) O arquivo Data_AgriculturalCensus06.csv contém uma amostra de 333 produtores agrícolas (Referência: Maia et al. Climate change and farm-level adaptation: the Braziian Sertão. International Journal of Climate Change Strategies and Management, v.10, n. 5, pp. 729-751, 2018). Pede-se:
   a) Compare as estimativas de MQO e de um modelo Tobit para o valor da produção (*VT_PRODUCAO*) como uma função linear da áera (*AREA_TOTAL*) e do número de empregados familiares (TOTAL_FAMILIARES);
   b) Compare as estimativas de MQO e de um modelo Heckit para o log do valor da produção como uma função do log da áera e do número de empregados familiares (sem log). Há evidências de viés de seleção na amostra?