O PROBLEMA DAS VARIÁVEIS OMITIDAS E VARIÁVEIS REVERSAS EM PESQUISAS

SILVA, Reginaldo dos Santos[1]

RESUMO: Este artigo procura explorar os problemas advindos da presença de variáveis omitidas e reservas de maneira narrativa com uma abordagem de técnicas para a coleta de dados e tratamento das informações obtidas.

Palavras-Chave: Variáveis omitidas. Variáveis reversas. Correlações espúrias.

SUMMARY: This article looks for to explore the happened problems with the presence of omitted and reverses variables at a narrative way with focusing on tchniques for data collection and information tretment.

Keywords: Omitted variables. Reverses variables. Spurious correlations.

INTRODUÇÃO

O artigo tem como objetivo discutir os problemas de mensuração e credibilidade das medidas em pesquisas nas áreas de ciências sociais aplicadas e biológicas, em função da presença de variáveis omitidas e casualidade reversa, durante o processo de pesquisa. Variável omitida é a variável não avaliada pelo pesquisador, por desconhecimento ou falta de condições de melhorar o modelo. São variáveis estocásticas e seus efeitos estão indicados como o termo de erro. Variável reversa é aquele que, ao ser alterada por influência de uma variável causal, vai atuar como variável causal sobre a mesma, provocando-lhe, também, alterações.

Será uma discussão mais narrativa do que matemática e focada em apenas dois dos muitos possíveis problemas associados com inferência causal. Serão analisados os problemas associados ao viés das variáveis omitidas e da causalidade reversa, ignorando-se as discussões sobre as relações de causa e efeito.

1 CARACTERIZAÇÃO DO PROBLEMA

1.1 PRELIMINARES

As ciências sociais que não trabalham com leis físicas diretas de causa e efeito tentam, através do uso de modelos, relacionar as variações que ocorrem entre fenômenos e suas possíveis causas.

[1] Eng. Agrônomo, mestre em Adm. De Empresas, Prof. Da Faculdade de Agronomia de Ituverava.

As pesquisas tentam avaliar a relação entre uma variável explicativa *X* e uma variável identificada como dependente *Y*, como pode ser mostrado na figura 1. Nessa, o efeito causal[2] estimado de *X* em *Y* é o coeficiente β. Num modelo de regressão linear, β é a mudança em *Y* produzida pela mudança de uma unidade em *X*. Exemplificando, se *X* for medido em quilos/ha de fertilizantes, β é o efeito em *Y* da quantidade de um quilo/ha de fertilizantes. A figura 1 mostra a equação de um modelo genérico e um possível exemplo. A equação de regressão linear apresenta ainda um outro coeficiente β, que representa o intercepto que, por não interessar nesta análise, será omitido no artigo.

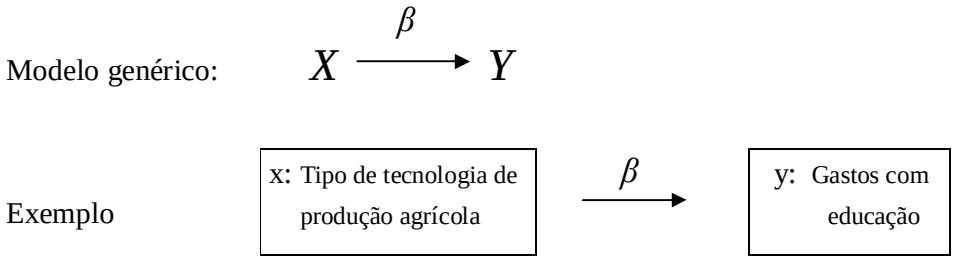

Figura 1. Efeito de *X* em *Y*

No modelo de análise de regressão simples: $Y = \beta X + u$ apresenta um termo de erro *u*, que representa os efeitos de variáveis não observáveis, em *Y*.

A equação 1, a seguir, identificaria a relação entre Gastos com educação (variável dependente) e o Tipo de tecnologia (variável explanatória), utilizado por uma região agrícola.

Equação (1): Gastos com educação = β_{TT} Tipo de tecnologia + *u*.

β_{TT} representa a intensidade de mudança em *Y* (Gastos em educação), em função da variação de uma unidade em *X* (Tipo de tecnologia), e *u* representa o efeito de outras variáveis, omitidas ou desconhecidas.

O problema básico trabalhado no artigo é a estimação dos efeitos de alteração de uma variável explicativa *X* , no valor de uma variável dependente *Y*, causadas por variáveis que influenciam uma quanto a outra.

Serão examinadas as dificuldades associadas em testar uma hipótese em particular: de não haver efeito causa de *X* sobre *Y*, em dados não temporais. A fixação do tema em séries não temporais deve-se ao fato de que a colocação da variável tempo, na análise de uma informação, é uma das técnicas colocadas nos próximos itens deste estudo para que se evite o

[2] Este efeito causal, não necessariamente significa que o fator X cause Y, mas apenas que há uma correlação.

efeito da variável espúria. Gujarati (2000, pág. 231-232) identifica a variável "tempo" como uma variável de tendência, que é introduzida para evitar o problema de correlação espúria e, ao final, tentará ilustrar o uso de teoria e modelos estatísticos para que se possa tratar o problema de inferência em situações onde possam aparecer variáveis omitidas e causalidade reversa.

A figura 2 mostra uma relação que deve estar um pouco mais próxima da realidade do que a imagem apresentada na figura 1. Nela, são encontrada as variáveis *X, Y, W* e *Z* e os vetores *u* das variáveis não observadas, também identificadas com os termos dos erros estocásticos.

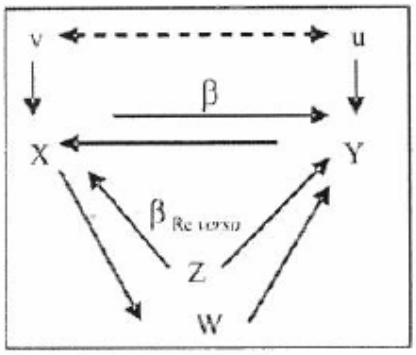

Figura 2 – Outras relações possíveis

Além das variáveis de causa e efeito *X* e *Y*, existem outras que, podem estar relacionadas a *X*, ou a *Y*, ou a ambos (como *W* e *Z*). Se elas não podem ser verificadas, os efeitos da sua omissão dependem da forma com que eles atuam.

Se a variável omitida atua de maneira similar a *W*, ou seja, sendo influenciada por *X* vai influenciar *Y*, seu efeito estará incorporado no efeito estimado causal β. Essa influência representa um problema menor, pois seu efeito estará embutido no erro estocástico.

Se a variável omitida representa uma causa para *X* e também para *Y*, então, sua omissão deverá resultar numa correlação espúria, uma estimação inviesada do efeito causal β.

O segundo problema considerado neste artigo é o da causalidade reversa, ou seja, o efeito que *Y* pode exercer em *X* com a variação do efeito de *X* sobre *Y*. Esta causalidade é identificada pelo coeficiente $\beta_{reverso}$ na figura 2.

A segunda parte do artigo é destinada à análise do que pode ser feito em função desses problemas.

2 DIFICULDADES PARA O PESQUISADOR

Entre a análise de dados e o estabelecimento de causalidade, existe uma série de barreiras para dificultar a análise correta de correlação. Entre elas é possível citar:

a) A dificuldade de se estabelecer a condição ceteris paribus, ou seta: todas as outras condições, ou variáveis devem permanecer constantes (em algumas situações é impossível) e

b) O fato de que uma relação empírica pode ser consistente com várias teorias.

O primeiro problema é identificar se há, e medir o possível viés de estimação de uma variável *Y*, em função da presença de variável omitidas, ou não identificadas que podem influenciar tanto a variável dependente *Y quanto a explicativa X*.

2.2 RELAÇÕES EM QUE VARIÁVEIS OMITIDAS CAUSAM VIÉS

Suponham-se que estejam sendo estimados efeitos causais de associabilidade de gastos com educação de uma região a uma determinada tecnologia de produção agrícola.

A figura 3 dá uma visualização do problema.

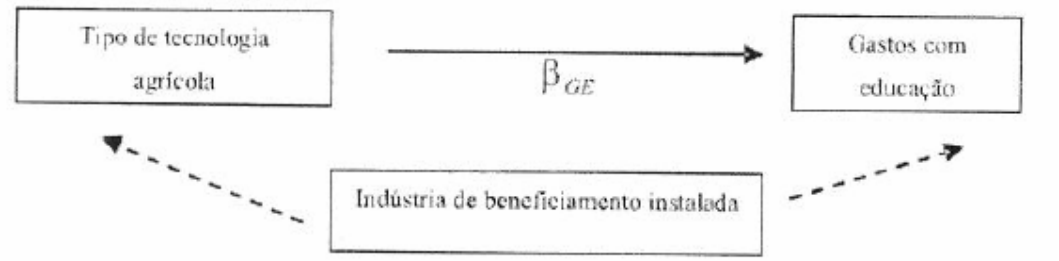

Figura 3 – Correlações espúrias

A variável Indústria de beneficiamento que vai consumir a matéria prima agrícola é aquela que não está sendo considerada pelo analista e, assim, está sendo omitida na análise. É um fator que influencia a escolha de tipo de tecnologia agrícola utilizada e também os gastos com educação, mas, como não está observada, não pode ser incluída nas equações de regressão; então, fica contida no termo de erro, denominado *u*. Supondo-se que "indústria de beneficiamento", *u* seria a sua representação. Entretanto, fosse a única variável omitida, a relação representada por β_{TT} não vai representar a realidade entre Gastos com educação e

Tipo de tecnologia, porque essa variável também influencia "tecnologia utilizada". Este tipo de viés é identificado como uma relação espúria, não por estar no termo do erro *u*, mas por estar influenciando nos dois fatores.

A figura 4 mostra um outro tipo de correlação espúria. Neste caso, o problema não é uma variável comum omitida, mas uma correlação entre uma variável omitida que, junto com outra, interfere na escolha do tipo de tecnologia, sendo que uma delas interfere também diretamente em "Gastos com educação", entretanto, o resultado é o mesmo: uma estimação enviesada do efeito de *X* em *Y*, representado por β_{TT}.

O termo do erro (ν) não observado na equação descrito como "Tipo de tecnologia" ($X = \beta_r R + \nu$), carrega em seu valor esta correlação com o termo de erro da equação descrito como "Gastos com educação" (*u)*.

A variável não observada determinante do valor de *X* (Tipo de tecnologia), identificada por ν, causa mudanças em *X*. Entretanto, *X* também aparece na equação de *Y* (Gastos com educação). Isso por si só não seria problemático, mas a correlação de *u e* ν identifica que *X* e *u* são correlacionados com *Y* na equação que leva a uma problemática, mas a correlação de *u* e *v* identifica que *X* e *u* são correlacionados com *Y* na equação que leva a uma estimação enviesada de β_{GE} como na equação (1)

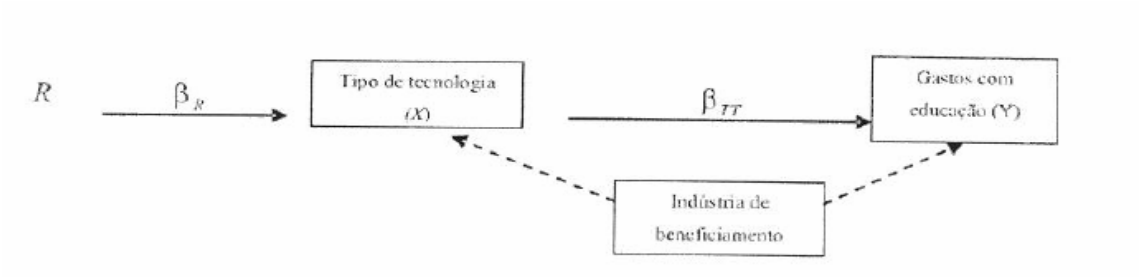

Figura 4 – Correlações espúrias: 2º caso

A figura 5 mostra como acontece o viés de β_{TT}

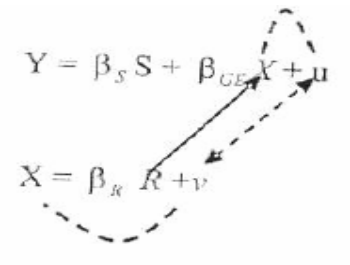

Figura 5 - Forma de relação entre *X* e *u*.

Toda variável não observada resulta na redução da capacidade explicativa da variável dependente que, por sua vez, reduz o poder estatístico da análise, mas não necessariamente em efeito de viés; por isso é importante entender os casos em que a variável omitida resulta em estimação enviesada. Apenas variáveis que afetam tanto a variável dependente quanto a explicativa resultam em estimativa enviesada de β_{GE} ou seja, é uma variável espúria.

A variável *W* da figura 2, por exemplo, não causa viés.

Os riscos de interpretações destas variáveis podem representar enormes custos ao pesquisador e a quem utilizar suas informações; por isso a preocupação em traze-los. Uma análise estatística mais acurada Poe levar à identificação desse problema.

2.3 A HETEROSEDASTICIDADE COMO INDICATIVO DE CORRELAÇÃO ESPÚRIA

A heterosedasticidade pode estar correlacionada a vários fatores, como nos casos em que: datilógrafos que, com o tempo, tendem a ter um menor volume de erros, e também uma menor variância nos erros, como também, a violação da hipótese de que o modelo de regressão esteja corretamente especificado (GUJARATI; 2000, p.56 e 358).

Todas as colocações feitas para a explicação da presença de heterosedasticidade podem ser analisadas do ponto de vista da variável omitida.

Sugere-se que o problema da heterosedasticidade deve ser maior em dados de corte que em séries temporais (GUJARATI; 2000, p. 358). Uma das razões da introdução da variável de tendência (tempo) seria evitar o problema das relações espúrias (GUJARATI; 2000, p.231).

Ao usar artifícios para tratamento da heterosedasticidade, tais como o uso do método dos mínimos quadrados generalizados para determinação dos βs e as propostas de medidas corretivas, pode-se estar incorrendo em erro de não tratamento das variáveis omitidas.

3 CAUSALIDADE REVERSA

Este item será analisado sob o seguinte exemplo:

Suponha que o pesquisador esteja interessado em estimar o efeito da área com fruticultura e nível de renda da população.

A figura 6 mostra que a área plantada com fruticultura pode realmente afetar o nível de renda da população, por promover uma maior rentabilidade por área e uma maior distribuição de renda. Como os investimentos são altos, quanto menor a renda da população, menor a capacidade de investir em fruticultura.

Um aumento na área plantada, leva à melhoria de renda da população que, por sua vez, pode levar a um aumento na área plantada. O inverso também deve acontecer.

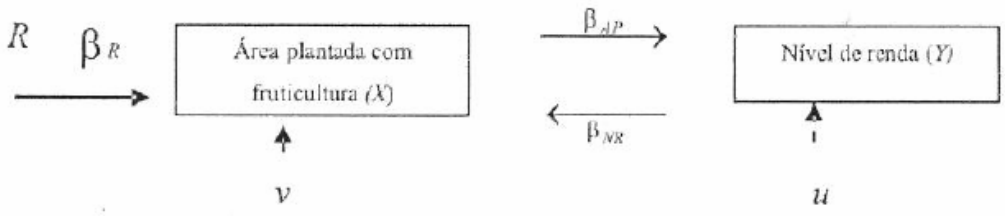

Figura 6 – Causalidade reversa

Num sentido, o problema é similar ao viés da variável omitida, em que a variável estocástica explanatória tem sua própria equação. Por outro lado, há uma diferença na medida em que não há necessariamente uma relação entre u e v (figura 7)

O efeito causa R determinante de X é distinto do efeito causal de X em Y, mas há um efeito causal de X, β_{AP} na determinação de Y. Os termos de erro u e v podem ou não estar correlacionados, mas há uma variável influenciando tanto X quanto Y.

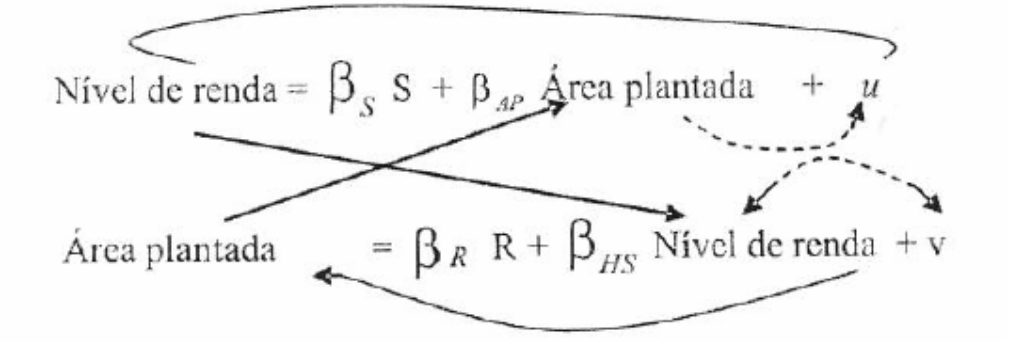

Figura 7 - Desenvolvimento da correlação de X e u no caso de causalidade reversa

4 ABORDAGEM DO PROBLEMA DA VARIÁVEL OMITIDA

As situações, em que o efeito de viés das variáveis omitidas ou das causalidades reversas aparecem, são muitas e de difícil, ou mesmo, impossível solução, dentro das condições do pesquisador. Existem, entretanto, técnicas para a sua abordagem. Algumas são aplicáveis tanto para casos de variáveis omitidas, quanto para causalidade reversa; enquanto outras apenas se aplicam a um problema ou outro.

Existem três abordagens possíveis que permitem ao pesquisador minimizar o efeito da variável omitida.

a) Colher dados adicionais sobre as variáveis não observadas e adiciona-los aos já colhidos para análise,

b) Tentar manipular *X* de maneira que não haja efeito da variável não observada sobre *Y*, sem que isso aconteça de maneira indireta, através de *X*.

c) Modelar a correlação do termo de erro na equação de *X* e *Y* como parte do processo de estimação.

4.1 COLETA DE DADOS ADICIONAIS

O valor da disponibilidade de um número maior de dados não pode ser subestimado, mas também não se pode esquecer que toda operação de coleta implica em custos maiores que podem inviabilizar a pesquisa.

Sendo assim, o foco de discussão será maior nas alternativas b e c de abordagem.

4.2 MANIPULAR *X*

Manipular *X* significa: fazer com que não aconteça nenhum efeito em *Y* que não seja através da variável *X*. Se isso for possível, então, as variações de *Y* associadas com mudanças em *X* serão interpretadas como um efeito causal de *X* sobre *Y*.

4.2.1 O MECANISMO DE RANDOMIZAÇÃO DA AMOSTRA

O mecanismo mais desejável de manipular *X* independente de *Y* é a escolha randômica, ou, ao acaso, dos elementos da amostra. Um exemplo familiar é a forma de escolha de elementos para fazerem parte do grupo de tratamento ou do grupo testemunha.

Na ausência da possibilidade de fazer a amostragem de maneira casuística, o pesquisador pode lanças mão de outro "manipulador".

4.2.2 INSERIR OU UTILIZAR UMA VARIÁVEL QUE ELIMINE O EFEITO DO TERMO DE ERRO *U* SOBRE A VARIÁVEL *X*

O manipulador deve ser algo sobre o qual o pesquisador tenha controle, ou que ocorra naturalmente no dado. A figura 8 é uma representação genérica do enfoque de manipulação de *X*. O objetivo é encontrar uma variável ***I*** ***que seja correlacionada a X,*** mas não atue em *Y* de outra maneira que não seja através de *X*.

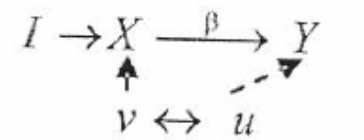

Figura 8 – Manipulador *X*

Exemplos:

a) Com a inserção de uma variável controlável: Suponha que um pesquisador deseje testar uma série de misturas de ração que forneçam a mesma quantidade e tipo de nutrientes, vitaminas e proteínas. Seleciona vários grupos de animais randomicamente.

Existem porém, variáveis omitidas que podem influenciar os animais e a qualidade da ração, como agentes da palatibilidade das mesmas.

O pesquisador pode inserir um complemento que melhore a palatibilidade em doses controladas. Desta maneira, a variável omitida vai ficar sob controle, e o termo de erro *u* da variável *Y* , não mais vai influenciar a variável *X*

b) Com a utilização de uma variável já existente no ambiente: No exemplo de McClellan, Mc Neil e Newhouse, num estudo que avaliava a efetividade de diferentes tratamentos para infarto agudo do miocárdio, os autores perceberam que os pacientes poderiam ter um viés de variáveis não observadas – como severidade da doença – já que não foram escolhidos randomicamente para os diferentes tratamentos. A severidade da doença poderia estar influenciando de maneira enviesada mais um grupo que um outro.

Verificaram que havia uma variável – distância da residência do paciente ao local de tratamento – que não influenciava os resultados dos tratamentos, mas que introduzida

permitiria que o efeito da severidade da doença fosse incorporado aos tipos de tratamento. Os autores demonstraram que a distância das residências aos locais de tratamento não tinha correlação com a severidade da doença, e então concluíram que a distância tinha correlação com o tipo de tratamento, mas não com o resultado.

O que se fez foi uma distribuição dos diversos níveis de severidade da doença entre as diversas distâncias.

4.2.3 AVALIAÇÃO DO EFEITO DO TRATAMENTO

Existem dois enfoques gerais:

4.2.3.1 A diferença nas diferenças

Avaliar o efeito da mudança da variável *Y* em função da variação na variável *X* , por diferentes valores da variável ***I***.

Quando o resultado de diferentes valores de *X* não são função do acaso, mas do efeito da variável ***I*** , é possível separar o efeito da variável não observada, que está em *u* de *Y*. No caso das rações, o efeito palatabilidade foi randomizado nas parcelas em que se colocou um elemento que dava diferentes graus de palatabilidade.

4.2.3.2 Dois estágios dos mínimos quadrados

Uma segunda forma de estimação é um modelo multivariado que usa variáveis ***I*** e qualquer variável pré-determinada no modelo para previsão de *X*.

Suponhamos a equação como: $X = yR + 1 + v$

R é o vetor de todas as outras variáveis predeterminadas no modelo, *y e* α são coeficientes e *v* é o termo do erro.

Esta equação para *X* é conhecida como forma reduzida da equação, e os valores preditos de *X* são:

$$\hat{E}[X|I;R] = R\hat{\gamma} + I\hat{\alpha}$$

O valor predito de *X* não contém o termo do erro, porque o seu valor esperado é zero. Como o valor de *X* foi purgado da correlação de *v* e *u* , pode ser incorporado num instrumento estimador da variável que vai gerar uma estimativa do parâmetro β. Este é o enfoque conhecido como dois estágios dos mínimos quadrados.

4.3 MODELAR CORRELAÇÃO DE u e v

O instrumental de estimar a variável, baseado na manipulação de *X*, elimina a correlação entre *u* e *v*.

Um outro enfoque está ligado ao problema da correlação entre *u* e *v* incorporaando a correlação na estimação do parâmetro causal β. Quando *X* é uma variável discreta, estes modelos são conhecidos na literatura econométrica como "modelo de seleção de amostras"[3].

Suponha-se o seguinte:

- O analista tenta determinar o efeito de associados em planos de prestação de serviços. A (pagamento periódico de um valor) e B (pagamento periódico de um valor menor, mais pagamento de um adicional por serviços) na utilização dos serviços que decidem a que plano se associar por escolha própria.

- Indivíduos com problemas crônicos têm maiores possibilidades de se associarem aos planos A, mas "problemas crônicos" é uma variável não observada pelo analista.

- O valor esperado de utilização por associados aos planos A é:

$E\left(Uso^{A}\right) = \beta_A X + E\left(u \mid \text{escolha}^{A}\right)$**:**

$E\left(Uso^{A}\right)$**:** valor esperado de utilização para pessoas associadas aos planos A

$E\left(u \mid \text{escolha}^{A}\right)$**:** valor esperado do termo de erro condicionado ao fato de a pessoa ter escolhido A, ou neste exemplo: *E(u\maior possibilidade de problemas crônicos)*

São dois os possíveis enfoques de tratamento.

4.3.1 Enfoque dos dois passos

Primeiro: estimar a equação de *X*, ou seja, a equação que explica como as pessoas escolhem os planos.

Da primeira equação, é calculado um termo que representa o valor do termo de erro *u* dada a regra de seleção da amostra. Esse termo é adicionado ao resultado da equação de interesse para correção do fato de que: o termo de erro, condicionado ao fato de que a amostra se auto-selecionou, não tem média zero. Esse procedimento é conhecido como informação limitada, máxima probabilidade.

[3] Neste enfoque é feito um tratamento específico para avaliar o efeito da variável *I* correlacionada à variável *R* na determinação de *X*.

4.3.2 Estimação simultânea

Nesse segundo caso, estima-se, simultaneamente, a seleção da amostra e da equação de resultado de interesse, utilizando-se o estimador da máxima probabilidade "informação total e máxima probabilidade" e pode se aplicado tanto quanto *X* é discreta quanto contínua[4].

Ambos os enfoques são possíveis de críticas. Para Manning et al (1987), a performance do modelo de seleção da amostra depende crucialmente da identificação de pelo menos uma variável que afete a seleção da amostra, mas que seja não associada com a variável de resultado, o que é o mesmo dado requerido pelo enfoque da diferença de diferenças.

5 ABORDAGEM DA CAUSALIDADE REVERSA

Neste caso, embora o resultado seja idêntico ao das variáveis omitidas (ver fig.8), alguns dos enfoques utilizados para a análise das variáveis omitidas não podem ser utilizadas para a causalidade reversa.

Um deles seria o aumento no volume da amostra. Nesse caso, o efeito causal não seria alterado. Mesmo que se fizesse uma coleta de amostra em unidades diferentes de tempo, pode-se observar, pela figura 9, que os dados em investimento em fruticultura (*X*) e nível de renda (*Y*) a questão de causalidade para cada variável, em cada período, é determinada pela influência acontecida no período anterior.

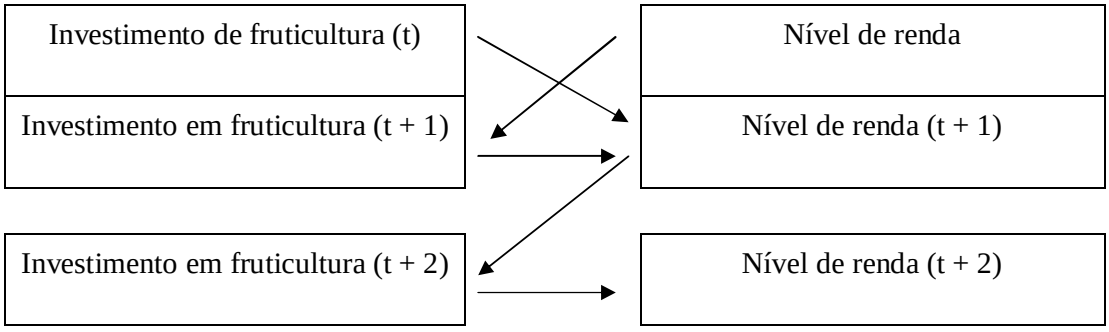

Figura 9 - Efeito de um período adicional no problema de causalidade reversa.

[4] Quando *X* é uma variável contínua, e o termo do erro em cada equação é normalmente distribuído, o estimador de três estágios dos mínimos quadrados é equivalente ao estimador informação total e máxima probabilidade (GRENE, 2000, p.695)

Pela figura 9, o nível de renda não pode influir na variável investimento em fruticultura, no tempo t, mas apenas no período t + 1.

Apenas o enfoque da segunda variável omitida, manipulação de *X*, pode ser utilizado para os casos de causalidade reversa. Na seguinte seqüência:

a – Manipular o valor de *X* de maneira que não haja efeito em *Y*, exceto através de *X*, idealmente através do uso da randomização.

b – Identificação da variável *I* que é correlacionada com *X*, mas não com *u*.

Quando os termos dos erros, nas duas estimações, são correlacionados, é necessário que a estimação seja feita pelo método do terceiro estágio dos mínimos quadrados. Os dois primeiros estágios corrigem o viés em β advindo da causalidade reversa. O terceiro, desenvolve a estimação padrão dos erros dos coeficientes por levar em consideração a correlação dos erros através das equações.

CONCLUSÃO

Como ressaltado no início, o artigo não tinha objetivo de ser um tratado, mas apenas o de ser um alerta aos pesquisadores para os cuidados com as informações coletadas e as conclusões de causalidade obtidas. Foi mostrado que, mesmo em situações de grandes dificuldades, existem técnicas disponíveis que podem melhorar a coleta e interpretação dos dados.

REFERÊNCIAS

DOWD, B.; TOWN, R. **Does X really cause Y?** Academy Health; Advancing research, policy and practice; sep. 2002

GREENE, W. **Econometrics analysis.** New Jersey: Prentice-Hall: Upper Saddle River, 2000

GUJARATI. D. N. **Econometria básica.** São Paulo: Pearson do Brasil; 2000

HECKMAN, J. **Causal parameters and policy anlysis in economics:** a twentieth century perspective; National Bureau of Economic Research Working paper 7333, Cambridge, MA. Disponível em www.nber.org/paperspw7333. Acesso em> sep.1999