

## Resolução 1º Lista de exercícios

**Docente:** Prof. Dr. Daniel Domingues dos Santos

**Monitor:** Matheus Mascioli Berlinger

**Email:** mbmascioli@gmail.com

### 1

Quais as hipóteses para que uma variável seja considerada uma boa variável instrumental? Explique.

Suponha que temos uma variável observável  $z$  que satisfaz as seguintes hipóteses:

- (1)  $z$  é não correlacionado com  $\epsilon$ , isto é,

$$Cov(z, \epsilon) = 0;$$

- (2)  $z$  é correlacionado com  $x$ , isto é,

$$Cov(z, x) \neq 0;$$

- (3)  $z$  afeta  $y$  somente via  $x$

Então, dizemos que  $z$  é uma **variável instrumental** de  $x$ , ou algumas vezes simplesmente um **instrumento** para  $x$ .

A exigência que o instrumento  $z$  satisfaça (1) é resumida dizendo-se “ $z$  é exógena na equação (1)”, e assim nos referimos a (1) como **exogeneidade dos instrumentos**. No contexto de variáveis omitidas, exogeneidade instrumental significa que  $z$  não deve ter efeito parcial  $y$  (após  $x$  e as variáveis omitidas terem sido controladas), e  $z$  deve ser não correlacionada com as variáveis omitidas.

A condição (2) significa que  $z$  deve ser relacionado, positiva ou negativamente, com a variável explicativa endógena  $x$ . Esta condição algumas vezes é referida como uma **relevância dos instrumentos** (como “ $z$  é relevante para explicar a variação em  $x$ ”).

Por fim, a condição (3) significa que  $z$  não pode ter efeito direto em  $y$  (caso contrário, deveria ser ma variável explicativa da regressão). Desta forma,  $z$  só afeta  $y$  através de  $x$ .

## 2

Suponha que queiramos estimar o efeito do uso de maconha na nota média de graduação. Em uma população de alunos veteranos de uma universidade, defina a variável “*diasusados*” como o número de dias do mês anterior que um aluno fumou maconha e considere a equação estrutural

$$supGPA = \beta_0 + \beta_1 diasusados + \beta_2 SAT + \epsilon$$

- a) Defina a variável “*percento*” como a porcentagem de uma sala do ensino médio que informou uso regular de maconha. Se ela for uma candidata a VI de “*diasusados*”, escreva a forma reduzida de “*diasusados*”.

Para ser candidata a VI de “*diasusados*”, a variável “*percento*” deve não ser correlacionada com o termo de erro  $\epsilon$  e ser correlacionada com “*diasusados*”. Para essa última correlação, precisamos de que os efeitos do grupo de colegas do ensino médio se estenda para a faculdade. Quer dizer, para uma determinada nota *SAT*, um estudante que frequentou uma escola de ensino médio onde fumar maconha era mais comum, fumaria mais maconha na faculdade. Mesmo que a condição de identificação se mantenha, a ligação pode ser fraca.

A equação na forma reduzida (onde escrevemos uma variável endógena em termos de variáveis exógenas) fica:

$$diasusados = \pi_0 + \pi_1 SAT + \pi_2 percento + v$$

- b) Você acha que a variável “*percento*” é verdadeiramente exógena na equação estrutural? Que problemas podem surgir nesse caso?

Temos que assumir que a porcentagem de estudantes que fumava maconha em uma escola de ensino médio não está correlacionada com fatores não observados que afetam a nota média no curso superior. Embora estejamos, até certo ponto, controlando a qualidade do ensino médio ao incluirmos *SAT* na equação, isso pode não ser suficiente. Talvez as escolas de ensino médio que tenham feito um melhor trabalho na preparação dos estudantes para a universidade também tenham tido um menor número de alunos que fumavam maconha. Ou, o uso de maconha poderia estar correlacionado com os níveis médios de renda. Naturalmente, essas são questões empíricas que podem ou não ser respondidas.

## 3

Considere um modelo simples para estimar o efeito da propriedade de um computador pessoal (*PC*) na nota média de graduação de formandos de uma grande universidade pública:

$$supGPA = \beta_0 + \beta_1 PC + \epsilon$$

em que *PC* é uma variável binária que indica se o indivíduo tem ou não um computador pessoal.

- a) Por que a propriedade de um  $PC$  pode estar correlacionada com  $\epsilon$ ?

É bem estabelecido que o status socioeconômico afeta o desempenho dos alunos. O termo de erro  $\epsilon$  contém, entre outras coisas, renda familiar, o que tem um efeito positivo no  $GPA$  e também é muito provável que esteja correlacionado com a propriedade do  $PC$ .

- b) Explique por que  $PC$  possivelmente está relacionado à renda anual dos pais. Isso significa que a renda dos pais será uma boa VI de  $PC$ ? Por quê?

As famílias com renda mais alta podem comprar computadores para seus filhos. Portanto, a renda familiar certamente satisfaz o segundo requisito para uma variável instrumental: está correlacionada com a variável explicativa endógena. Mas como sugerimos na parte a), a renda familiar tem um efeito positivo no  $GPA$ , então o primeiro requisito para um bom IV, falha para a renda familiar. Se tivéssemos a renda familiar dos indivíduos, a incluiríamos como uma variável explicativa na equação; Se for a única variável omitida importante correlacionada com  $PC$ , poderíamos então estimar a equação expandida por MQO.

- c) Suponha que, quatro anos atrás, a universidade tenha concedido subsídios para a compra de computadores a aproximadamente metade dos alunos novos, e que os alunos que receberam esses subsídios tenham sido escolhidos aleatoriamente. Explique cuidadosamente como você usaria essa informação para construir uma variável instrumental de  $PC$ .

Este é um experimento natural que afeta a propriedade ou não de computadores. Alguns alunos que compram os computadores quando dada a concessão não o fariam sem a concessão (alunos que não receberam os subsídios ainda podem possuir computadores). Defina uma variável dummy,  $grant$ , igual a 1 se o aluno recebeu uma subvenção e 0 caso contrário. Em seguida, se a concessão foi atribuída aleatoriamente, ela não está correlacionada com  $\epsilon$ . Em particular, não está correlacionada com a renda familiar e outros fatores socioeconômicos em  $\epsilon$ . Além disso, a concessão deve ser correlacionada com  $PC$ : a probabilidade de possuir um  $PC$  deve ser significativamente maior para o estudante que recebe subsídios. Aliás, se a universidade deu prioridade à concessão de estudantes de baixa renda, a concessão seria negativamente correlacionada com  $\epsilon$ , e VI seria inconsistente.

## 4

Explique detalhadamente como você resolveria problema de ter um número de variáveis instrumentais maior do que o número de variáveis endógenas no modelo.

Vamos considerar o caso em que temos um variável explicativa endógena ( $y_2$ ) e duas variáveis exógenas excluídas do modelo ( $z_1$  e  $z_2$ ), candidatas a VI de  $y_2$ . Se  $z_1$  e  $z_2$  forem ambas correlacionadas com  $y_2$  e não correlacionadas com  $\epsilon$ ,

poderemos simplesmente usar cada uma delas como uma VI. Entretanto, nesse caso, teríamos dois estimadores de VI, e nenhum deles seria, de forma geral, eficiente. Como  $z_1$  e  $z_2$  são não correlacionadas com  $\epsilon$ , qualquer combinação linear das variáveis exógenas será uma VI válida. Para encontrar a melhor VI, escolhemos a combinação linear que seja mais altamente correlacionada com  $y_2$ .

## 5

Derive a fórmula do estimador de variáveis instrumentais no caso sobreidentificado explicitando as hipóteses necessárias.

Feito em aula.

## 6

Considere o problema de estimar o efeito causal de faltar às aulas sobre as notas do exame final. Em uma estrutura de regressão simples, temos:

$$nota = \beta_0 + \beta_1 faltas + \epsilon$$

em que *nota* é a nota no exame final e *faltas* é o número total de faltas às aulas durante o semestre. Suponha que você não tenha uma boa candidata a variável instrumental de faltas. Entretanto, você tem duas outras informações sobre os alunos: a nota média ponderada de matemática e habilidade verbal do estudante para ingresso em curso superior (*SAT*) e a nota média acumulada anterior ao semestre (*GPA*). O que você faria em vez da estimação de VI?

Certamente devemos estar preocupados se *faltas* está correlacionado com outros fatores em  $\epsilon$ : alunos mais aptos, altamente motivados, devem ter um menor número de faltas. Assim, uma regressão simples de *nota* sobre *faltas* pode não produzir uma boa estimativa do efeito causal de faltas às aulas. Evidentemente, as informações do *SAT* e do *GPA* são variáveis proxy observáveis da aptidão dos alunos contida em  $\epsilon$ . Desta forma, podemos simplesmente adicioná-las como variáveis explicativas e estimar a equação expandida por MQO.