Monitoria de Econometria II

Jean Haendell

Variáveis Instrumentais e Eq. Simultâneas

O que são Variáveis Instrumentais e como utilizá-las?

Definição : Variáveis Instrumentais são variáveis que não estão correlacionadas com o erro em uma regressão, mas estão correlacionadas com uma ou mais das variáveis explicativas independentes. Elas são usadas para resolver problemas de endogeneidade, onde as variáveis independentes estão correlacionadas com o erro.

Por que são utilizadas: Existem situações em econometria em que a relação entre duas variáveis é difícil de estabelecer devido à endogeneidade. A endogeneidade pode surgir devido à omissão de variáveis, erro de medição ou simultaneidade. As VIs são uma ferramenta poderosa para corrigir esses problemas e fornecer estimativas consistentes dos parâmetros de interesse.

E como se dão as estimativas com VIs?

Primeiramente, instale e carregue a biblioteca AER:

```
install.packages("AER")
library(AER)
```

Obs.: Durante a prova, apenas a segunda linha de código é necessária, pois a biblioteca já estará instalada no computador.

Suponha que você queira estimar a relação entre Y e X, mas X é endógeno e você tem uma variável instrumental Z. O modelo pode ser estimado da seguinte forma:

```
resultado <- ivreg(Y ~ X | Z, data=seus_dados)
summary(resultado)
```

Variáveis Instrumentais (VI) em Regressão Múltipla

Em uma regressão múltipla, é comum termos várias variáveis independentes. Algumas dessas variáveis podem ser exógenas (não correlacionadas com o erro) e outras podem ser endógenas (correlacionadas com o erro). As VI podem ser usadas para corrigir a endogeneidade das variáveis endógenas.

Modelo Teórico

Suponha que temos a seguinte equação de regressão:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Onde: Y é a variável dependente. X_1 é uma variável independente endógena. X_2 é uma variável independente exógena. ϵ é o termo de erro.

Para corrigir a endogeneidade de X_1, vamos usar uma variável instrumental Z que está correlacionada com X_1, mas não com ϵ.

```
resultado <- ivreg(Y ~ X1 + X2 | Z + X2, data=seus_dados)
summary(resultado)
```

Perceba que repetimos X_2 dos dois lados, pois não mexemos com essa variável. O resultado do `summary` irá retornar os valores da regressão para X_1 e X_2. Interpreta-se da mesma forma que se fazia com o bom e velho `lm()`, basta lembrar que esses valores foram obtidos utilizando-se VIs.

Teste de Hausman Simplificado para Endogeneidade

O teste de Hausman pode ser usado para verificar se uma variável é endógena em um modelo de regressão. O processo é:

1- Estimar um modelo reduzido:

Coloque a variável potencialmente endógena como dependente e estime o modelo. Salve os resíduos deste modelo.

2- Inclua os resíduos no modelo original:

Estime o modelo original novamente, mas agora inclua os resíduos salvos como uma variável adicional.

3- Verifique a significância dos resíduos:

Se os resíduos forem estatisticamente significativos, isso indica que a variável original é endógena.

Como fazer no R:

Estime o modelo reduzido e salve os resíduos:

```
equacao_reduzida <- lm(variavel_suspeita ~ outras_variaveis, data = seus_dados)
res <- equacao_reduzida$residuals
```

Estime o modelo original com os resíduos:

```
modelo_original <- lm(variavel_dependente ~ variavel_suspeita + res + outras_variavei
s, data = seus_dados)
summary(modelo_original)
```

Agora, é só verificar se o resíduo é estatisticamente significante realizando o bom e velho teste de hipóteses. Se for, então a variável suspeita é realmente endógena.

Resolvendo a questão do vinho

Temos então a base de dados `chard`.

```
load("/Users/emilia.franca/Documents/chard.Rdata")
head(chard)
```

```
##          y xper     cap     lab age
## 1  8.4639   10  0.1875  3.9055  42
## 2 10.3116   20  6.3419  7.1255  44
## 3 12.1644   15 11.8028 18.2744  43
## 4  0.0788   11  0.3660  0.9242  58
## 5 11.3100   20  7.3265 10.1505  28
## 6  8.2517   21  8.3297  6.0561  25
```

a)

Equacão revisada significa ter o `xper` no lugar de `mgt` , como a questão sugere.

Aqui, a estimação é como antes.

```
chard.model1<-lm(y ~ xper+cap+lab, data=chard)
summary(chard.model1)
```

```
##
## Call:
## lm(formula = y ~ xper + cap + lab, data = chard)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3447 -1.6842 -0.1289  1.3112  9.4533
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.76226    1.05535   1.670 0.099354 .
## xper         0.14684    0.06343   2.315 0.023517 *
## cap          0.43796    0.11756   3.725 0.000388 ***
## lab          0.23916    0.09980   2.396 0.019195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.756 on 71 degrees of freedom
## Multiple R-squared:  0.5616, Adjusted R-squared:  0.543
## F-statistic: 30.31 on 3 and 71 DF,  p-value: 9.986e-13
```

As interpretações são da mesma maneira de sempre, e portanto ficam por conta do estudante.

b)

O que faremos nessa questão é basicamente criar “novos dados”, que depois passaremos ao modelo estimado no item a).

Vamos por partes.

```
new.chard <- data.frame(xper=c(10,20,30), cap=rep(7.83,3),lab=rep(10.05,3))
```

Esse código cria um novo dataframe chamado new.chard. Esse dataframe tem três colunas:

- *xper*: Contém os valores 10, 20 e 30.
- *cap*: Contém o valor 7.83 repetido três vezes.

- *lab*: Contém o valor 10.05 repetido três vezes. Portanto, o dataframe terá três linhas e três colunas.

Por que repetir 7.83 e 10.05 três vezes? Porque esses valores são a média de `cap` e de `lab`:

```
mean(chard$cap)
```

```
## [1] 7.834691
```

```
mean(chard$lab)
```

```
## [1] 10.04674
```

Agora, vamos usar a função predict para passar esses dados para o modelo do item a:

```
predict(chard.model1, newdata = new.chard, interval = "prediction")
```

```
##         fit      lwr      upr
## 1  9.063415 3.509536 14.61729
## 2 10.531797 4.945711 16.11788
## 3 12.000178 6.104191 17.89617
```

Traduzindo esse código para o português: preveja qual será o valor da variável dependente quando tivermos os valores médios 7.83 para cap, 10.04 para lab, com a experiência variando entre 10, 20 e 30.

- A linha 1 mostra o caso de 10 anos de experiência: o valor médio previsto para a variável dependente é de 9.063415.
- A linha 2 mostra o caso de 20 anos de experiência: o valor médio previsto para a variável dependente é de 10.531797.
- A linha 2 mostra o caso de 20 anos de experiência: o valor médio previsto para a variável dependente é de 12.000178

Em suma, se os valores de cap e lab permanecerem constantes nessa média, a variação em 10 anos da experiência traria, em média, essa variação acima na variável dependente.

c)

Vamos fazer como o Sr. Chardonnay e realizar o teste de Hausmann?

Primeiro, trazemos a variável explicativa que suspeitamos ser endógena para o posto de variável explicada. Depois, trazemos a possível variável instrumental (age) para o posto de variável explicativa e temos esse modelo:

```
chard.model2<-lm(xper ~ age+cap+lab, data=chard)
```

Agora, salvamos os resíduos dessa estimação.

```
res<-chard.model2$residuals
```

Voltamos ao modelo original, mas agora incluímos os resíduos nele. Incluímos um summary e…

```
chard.model3<-lm(y ~ xper+cap+lab+res, data=chard)
summary(chard.model3)
```

```
##
## Call:
## lm(formula = y ~ xper + cap + lab + res, data = chard)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4582 -1.6581 -0.1773  1.4120  8.7115
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.48669    2.18960  -1.136  0.25996
## xper         0.51210    0.17731   2.888  0.00515 **
## cap          0.33213    0.12422   2.674  0.00933 **
## lab          0.23998    0.09721   2.469  0.01601 *
## res         -0.41575    0.18917  -2.198  0.03127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.685 on 70 degrees of freedom
## Multiple R-squared:  0.5899, Adjusted R-squared:  0.5664
## F-statistic: 25.17 on 4 and 70 DF,  p-value: 6.14e-13
```

Temos o mesmo resultado de sempre. O objetivo aqui é verificar se o resíduo (`res`) é estatisticamente significante, realizando o teste de hipóteses já conhecido. Se for, então Sr. Chardonnay está certo, e `xper` está correlacionada com o termo de erro. Melhor partir para uma variável instrumental…

d)

Vimos até aqui que o ideal para essa estimação seria utilizar a variável instrumental `age` , no lugar de `xper` . Como fazer isso? utilizando a funcão `ivreg` , do pacote `AER` .

```
library(AER)
```

```
chard.model4<-ivreg(y ~ xper+cap+lab|cap+lab+age, data=chard)
summary(chard.model4)
```

```
##
## Call:
## ivreg(formula = y ~ xper + cap + lab | cap + lab + age, data = chard)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.40413 -2.17750 -0.09044  2.28339 10.62769
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.4867     2.7226  -0.913   0.3642
## xper          0.5121     0.2205   2.323   0.0231 *
## cap           0.3321     0.1545   2.150   0.0349 *
## lab           0.2400     0.1209   1.985   0.0510 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.338 on 71 degrees of freedom
## Multiple R-Squared: 0.3568,  Adjusted R-squared: 0.3296
## Wald test: 21.24 on 3 and 71 DF,  p-value: 6.307e-10
```

A estrutura do `ivreg` foi explicada acima. Mas, relembrando, escreve-se o modelo dentro da função da mesma forma que no `lm()` no lado esquerdo do "|". Do lado direito, reescreve-se o modelo, substituindo a variável que se suspeita ser endógena (`xper`) por sua variável instrumental (`age`). Depois do `summary` , é só interpretar os resultados como usualmente, inclusive para `xper` .

e)

O mesmo que no item b, mas com o modelo que criamos utilizando o `ivreg` . Simples, né?

```
new.chard2 <- data.frame(xper=c(10,20,30), cap=rep(7.83,3),lab=rep(10.05,3),age=rep(4
2.95,3))

predict(chard.model4, newdata = new.chard2, interval = "prediction")
```

```
##        1        2        3
##  7.64669 12.76771 17.88873
```

Basta comparar esses valores com os do item b, seguindo a mesma ordem.

Equações Simultâneas com R

Introdução

Vamos aprender sobre equações simultâneas e como resolvê-las usando o R. Vamos também discutir a Condição Necessária para Identificação, que é crucial para a análise de equações simultâneas.

O que são Equações Simultâneas?

Equações simultâneas são um conjunto de equações que são determinadas conjuntamente. Em econometria, isso é comum quando estamos lidando com variáveis que são tanto dependentes quanto independentes em diferentes equações. Por exemplo, no sistema de equações do Keynes, temos o investimento como variável explicativa da renda. Entretanto, ao mesmo tempo, o investimento é uma variável explicada pela taxa de juros. Ou seja, o investimento é endógeno (a taxa de juros seria, então, exógena).

Resolução do Exemplo 5 - Questão Keynes

Considere um modelo econômico simples com três equações:

$$\begin{gathered} c_t = \alpha_1 + \alpha_2 y_t + e_{t1} \\ i_t = \beta_1 + \beta_2 r_t + e_{12} \\ y_t = c_t + i_t + g_t \end{gathered}$$

Aqui, c é o consumo, y é o produto, r é a taxa de juros, i o investimento e g os gastos do governo, todos no tempo t.

a)

Na primeira equação: se você estimasse o consumo como função da renda, qual sinal você esperaria que acompanhasse y_t? Certamente seria positivo, pois uma maior renda leva a um maior consumo.

Basta fazer a mesma análise para as outras duas equações.

b)

Colocar as equações na forma reduzida significa simplesmente colocar as variáveis endógenas como função apenas das exógenas. Nesse caso, temos apenas r e g como variáveis endógenas.

Para encontrar a forma reduzida de um sistema de equações simultâneas, você deve isolar cada variável endógena em termos de apenas variáveis exógenas e termos de erro. Vamos isolar y_t, c_t e i_t em relação a r_t e g_t

Substituindo c_t e i_t na equação de y_t,

$$y_t = (\alpha_1 + \alpha_2 y_t + e_{t1}) + (\beta_1 + \beta_2 r_t + e_{t2}) + g_t$$

$$y_t = \alpha_1 + \beta_1 + g_t + \alpha_2 y_t + \beta_2 r_t + e_{t1} + e_{t2}$$

Isolando

$$y_t$$

,

$$y_t(1 - \alpha_2) = \alpha_1 + \beta_1 + g_t + \beta_2 r_t + e_{t1} + e_{t2}$$

$$y_t = \frac{\alpha_1 + \beta_1}{1 - \alpha_2} + \frac{g_t}{1 - \alpha_2} + \frac{\beta_2 r_t}{1 - \alpha_2} + \frac{e_{t1} + e_{t2}}{1 - \alpha_2}$$

Substituindo a nova expressão de y_t em c_t e i_t

$$c_t = \alpha_1 + \alpha_2 \left(\frac{\alpha_1 + \beta_1}{1 - \alpha_2} + \frac{g_t}{1 - \alpha_2} + \frac{\beta_2 r_t}{1 - \alpha_2} + \frac{e_{t1} + e_{t2}}{1 - \alpha_2} \right) + e_{t1}$$

$$i_t = \beta_1 + \beta_2 r_t + e_{t2}$$

c)

Vamos visualizar a base de dados "keynes":

```
load("/Users/emilia.franca/Documents/keynes.Rdata")
head(keynes)
```

```
##        y      c      i      r     g
## 1 43.517 30.217  9.727 19.042 3.573
## 2 50.931 45.836  1.571 18.292 3.524
## 3 56.334 37.275 13.798 16.054 5.261
## 4 67.660 42.143 20.915  9.384 4.602
## 5 63.977 47.028 12.943 10.470 4.006
## 6 77.357 45.185 27.917 18.189 4.255
```

Estimando as formas reduzidas no R (colocando as endógenas como variáveis explicadas e as exógenas como explicativas).

```
fr.y <- lm(y ~ r + g, data=keynes)
fr.c <- lm(c ~ r + g, data=keynes)
fr.i <- lm(i ~ r + g, data=keynes)
```

O stargazer serve para visualizar os resultados das regressões. Embora útil, evitarei aqui.

Seguindo, podemos fazer como na questão do vinho e estimar o mesmo modelo com novos dados.

Criando esses novos dados e definindo $r = 15$ e $g = 20$, temos

```
new.keynes <- data.frame(r=15, g=20)
```

Fazendo as previsões:

```
predict(fr.y, newdata = new.keynes, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 178.9069 131.4448 226.3689
```

```
predict(fr.c, newdata = new.keynes, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 96.83863 62.50169 131.1756
```

```
predict(fr.i, newdata = new.keynes, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 62.06826 39.83607 84.30045
```

Lembrando, o código acima diz: “estime o modelo que foi criado anteriormente para estes novos valores”. Isso é feito para os três modelos. Prevemos, por exemplo, que pra esse valor de juros e gasto do governo, o produto será de 178,90, o consumo de 98,83 e o investimento de 62,07.

d)

A identificação é a capacidade de estimar os parâmetros do modelo de forma única e precisa. A Condição Necessária para Identificação em um sistema de equações simultâneas é que o número de variáveis exógenas excluídas deve ser pelo menos igual ao número de variáveis endógenas incluídas, mais um. (Mais informações na aula “Modelos de Equações Simultâneas”, disponibilizada no Sigaa).

Matematicamente, para uma equação i,

$$K_i \geq G + 1$$

Em que K_i é o número de variáveis exógenas excluídas da equação i e G é o número de variáveis endógenas incluídas na equação i, excluindo a variável dependente.

Equação 1: $c_t = \alpha_1 + \alpha_2 y_t + e_{t1}$

Verificação de identificação:

$$\begin{gathered} K_1 = 2, G = 1 \\ K_1 \geq G + 1 \\ 2 \geq 1 + 1 \\ 2 \geq 2 \end{gathered}$$

Quem são as duas variáveis exógenas que estão excluídas da primeira equação? r e g. E qual é a endógena inclusa? Y_t. Tente fazer a mesma análise para as outras duas equações.

Dizemos que a primeira equação é exatamente identificada.

Equação 2: $i_t = \beta_1 + \beta_2 r_t + e_{12}$

Verificação de identificação:

$$\begin{gathered} K_2 = 2, G = 0 \\ K_2 \geq G + 1 \\ 2 \geq 0 + 1 \\ 2 \geq 1 \end{gathered}$$

A segunda equação é sobreidentificada

Equação 3: $y_t = c_t + i_t + g_t$

Verificação de identificação:

$$
\begin{gathered}
K_3 = 1, G = 2 \\
K_3 \geq G + 1 \\
1 \geq 2 + 1 \\
1 \geq 3
\end{gathered}
$$

Logo, a terceira equação é não identificada.

e) a ser comentada na monitoria.