Econometria 2

Miguel Sallum

23/06/2021

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Regressão Linear e Estimação

A regressão linear é um modelo de relação entre variáveis, e pode ser chamado também de função de esperança condicional. Ela é tradicionalmente estimada com método dos mínimos quadrados ordinários, mas é equivalente por método dos momentos e (se não me engano) por máxima verossimilhança.

Caso tenhamos somente um regressor, estamos estimando a esperança condicional da forma

$$E[Y|X] \ = \ \beta_0 \ + \ \beta_1 X$$

Tradicionalmente, no entanto, representamos o modelo como

$$Y \ = \ \beta_0 \ + \ \beta_1 X + \mu$$

Onde μ são os fatores não observados. Para estimarmos as regressões, são necessárias algumas hipóteses: 1. É importante lembrar que a regressão *não é* um modelo causal. Para ser causal, são necessárias algumas hipóteses a mais. Sendo flexível com notação, em geral nosso interesse é estimar o modelo causal

$$E[Y|do(X)] \ = \ \beta_0 \ + \ \beta_1 X$$

Queremos encontrar então o valor adequado de β_1, que seria o efeito médio de X sobre Y. para isso, precisamos da hipótese **(não-observável)** 4. $Corr(X, \mu) = 0$

Estimação matricial

```r
ols <- function(Y, X){
  X <- as.matrix(X)
  Y <- as.matrix(Y)
  Xt <- t(X)
  XtX <- Xt %*% X
  XtX_inv <- solve(XtX)
```

```
  XtX_inv %*% Xt %*% Y
}

SST <- function(Y) {Y - mean(Y) %>% t(.) %*% .}

res <- function(Y, X) {Y - (X %*% ols(Y, X))}

SSR <- function(Y, X) {t(res(Y, X)) %*% res(Y, X)}

R2 <- function(Y, X){
  X <- as.matrix(X)
  Y <- as.matrix(Y)

  1 - SSR(Y,X) /SST(Y)
}

vcov<- function(Y, X, robust = FALSE){

  k <- nrow(ols(Y, X))
  n <- length(Y)
  if (robust) {
    sigma2

  } else {
    sigma2 <- as.numeric(SSR(Y, X)/(n-k))*diag(nrow = k)
  }
  var_cov <- sigma2 %*% XtX_inv

}
```

Métodos do R e Bibliotecas

Para rodar as regressões,podemos usar a função nativa *lm()*, em geral já em conjunção com *summary()* para termos os algumas caracteristicas importantes, como R^2, erros-padrão e significância assumindo homoscedasticidade. Caso estejamos usando tidyverse, podemos usar as funções assim:

```
#lm(dependente ~ regressores, data = dados)%>%
 # summary()
```

No entanto, na maior parte dos casos, a hipótese da homoscedasticidade é muito fort

Potencial Outcomes

Potencial Outcomes é uma forma de pensar sobre causalidade usando de contrafactuais. A ideia é que teríamos a informação do resultado de cada indivíduo para cada nível de intervenção de X. Os exemplos do tema em geral são binários, **mas o método não se restringe a isso**. Em casos binários, podemos representar o resultado do individuo i caso ele receba o tratameto ($X = 1$) como Y_i^1, e como Y_i^0 caso ele não seja tratado .

Trabalhando com esses dados hipotéticos surgem alguns conceitos interessantes:

$$\begin{aligned}
Treatment\ Effect_i\ (TE_i) &= Y_i^1\ -\ Y_i^0 \\
Average\ Treatment\ Effect\ (ATE) &= E[Y^1 - Y^0]\ =\ E[Y^1] - E[Y^0] \\
Treatment\ Effect\ on\ the\ Treated\ (ATT) &= E[Y^1 - Y^0|\ Treatment = 1] \\
Treatment\ Effect\ on\ the\ Unreated\ (ATU) &= E[Y^1 - Y^0|\ Treatment = 0]
\end{aligned}$$

Nenhum desses efeitos pode der extraido diretamente dos dados, mas eles ajudam a entender quais são as hipóteses necessárias para nossas estimativas valerem em cada caso. Caso quisermos estimar o efeito a partir da diferença simples de médias, por exemplo, chegamos que:

$$\begin{aligned}
SDO &= E[Y^1|D=1] - E[Y^0|D=0] \\
&= (E[Y^1] - E[Y^0]) + (E[Y^0|D=1] - E[Y^0|D=0]) + (1-\pi)(ATT - ATU) \\
&= ATE + viés\ de\ seleção + (1-\pi)(viés\ de\ efeito\ heterogêneo)
\end{aligned}$$

Considerando que o que nós buscamos conhecer são os efeitos de fato, ATE, então devemos nos preocupar com métodos para que os vieses sejam nulos. A forma mais simples é a randomização, fazendo com que as esperanças de ambos os grupos sejam as mesmas.

Tendo uma tabela com os resultados potenciais e os tratamentos binários, podemos estimar os efeitos com

```r
tibble(
  Y1 = c(8, 5, 9, 8, 4, 6, 10, 11, 8, 6),
  Y0 = c(4, 5, 6, 7, 5, 6, 4, 3, 2, 8),
  D = rep(c(1, 0), 5)) %>%
  mutate(
    epsilon = Y1-Y0) %>%
  summarise(
    ATE = mean(epsilon),
    ATT = sum(epsilon*D)/sum(D),
    ATU = sum(epsilon*(1-D))/sum((1-D)),
    SDO = sum(Y1*D)/sum(D) -sum(Y0*(1-D))/sum((1-D)))
```

```
## # A tibble: 1 x 4
##     ATE   ATT   ATU   SDO
##   <dbl> <dbl> <dbl> <dbl>
## 1   2.5   3.6   1.4     2
```

Regression Discontinuity Design

Variável Instrumental

Em muitos dos casos em que queremos estimar os efeitos causais de uma variavel sobre a outra, uma regressão simples não conseguiria estimar, pois a hipótese 4 não valeria. Isto é, existe alguma correlação entre X e μ. A ideia por trás de uma variável instrumental é, então, uma forma de encontrar variações exógenas, "descorrelacionadas" com os *fatores não-observados*. Para fazer isso, você encontra uma variável Z que afeta o tratamento, e só afeta o resultado através do tratamento. Isto é

1. $corr(X, Z) \neq 0$
2. $corr(\mu, Z) = 0$

Homogenous Treatment Effects (Cunningham Cap. 7)

Em *Homogenous Treatment Effects*, você supõe que todas as pessoas do grupo que receberam o tratamento terão uma mudança na variável de interesse com a mesma intensidade. Ou seja, se fazer universidade aumento minha renda em 10%, então aumentou em 10% para todos que fizeram universidade.

Portanto, suponha um modelo onde você deseja estimar o quanto um aumento de educação causa aumento na renda.

$$Y_i = \alpha + \delta S_i + \gamma A_i + \varepsilon_i$$

Onde, Y_i é a renda de cada individuo, S_i os anos de educação e A_i uma variável não observada que representa abilidade. Desse modo, o modelo que conseguiremos estimar é o seguinte:

$$Y_i = \alpha + \delta S_i + \eta_i$$

onde η_i é o erro composto equivalente a $\gamma A_i + \varepsilon_i$. Como assumimos que "abilidade" está correlacionada com a variável de "educação", então apenas ε_i está descorrelacionado com os regressores.

utilizando o valor estimado de $\hat{\delta}$ da OLS tradicional temos que

$$\hat{\delta} = \frac{Cov(Y, S)}{Var(S)} = \frac{E[YS] - E[Y]E[S]}{Var(S)}$$

Se utilizarmos o valor de Y da regressão onde A é observável teremos que

$$\hat{\delta} = \frac{E[S(\alpha + \delta S + \gamma A + \epsilon)] - E[\alpha + \delta S + \gamma A + \epsilon]E[S]}{Var(S)}$$

$$\hat{\delta} = \frac{\delta E(S^2) - \delta E(S)^2 + \gamma E(AS) - \gamma E(S)E(A) + E(\varepsilon S) + E(S)E(\varepsilon)}{Var(S)}$$

$$\hat{\delta} = \delta + \gamma \frac{Cov(A, S)}{Var(S)}$$

Logo, se $\gamma > 0$ e $Cov(A, S) > 0$, então $\hat{\delta}$ será viesado para cima.E como deve ser positivamente correlacionada com eduação, então isso é o que deve acontecer.

Mas se encontrarmos uma nova variável Z_i que causa as pessoas a ter mais anos de estudos e que é descorrelacionada com abilidade (as variáveis não observáveis), podemos utilizar ela como uma variável instrumental para estimar δ.

Para isso precisamos primeiro calcular a covariancia de Y e Z

$$Cov(Y, Z) = Cov(\alpha + \delta S + \gamma A + \varepsilon, Z)$$

$$= E[Z(\alpha + \delta S + \gamma A + \varepsilon)] - E[\alpha + \delta S + \gamma A + \varepsilon] \quad E[Z]$$
$$= E[\alpha Z + \delta SZ + \gamma AZ + \varepsilon Z] - \ \{\alpha + \delta E(S) + \gamma E(A) + E(\varepsilon) \ \} \ E[Z]$$

$$= \{\alpha E(Z) + \delta E(SZ) + \gamma E(AZ) + E(\varepsilon Z)\} - \{\alpha E(Z) + \delta E(S)E(Z) + \gamma E(A)E(Z) + E(\varepsilon)E(Z) \ \}$$

$$= \{ \ \alpha E(Z) - \alpha E(Z) \ \} + \delta\{ \ E(SZ) - E(S) \ E(Z) \ \} + \gamma \ \{E(AZ) - E(A) \ E(Z) \ \} + \{E(\varepsilon Z) - E(\varepsilon) \ E(Z)\}$$

$$= \delta \ Cov(S, Z) + \gamma \ Cov(A, Z) + Cov \ (\varepsilon, Z)$$

Como sabemos que $Cov(A, Z) = 0$ e $Cov(\varepsilon, Z) = 0$, uma vez que não existe essa relação entre os instrumentos, podemos estimar $\hat{\delta}$.

$$\hat{\delta} = \frac{Cov(Y, Z)}{Cov(S, Z)}$$

Dessa forma, podemos usar a variável instrumental Z para estimar $\hat{\delta}$ caso Z seja independente da variável oculta e do erro estrutural da regressão. Ou seja, o instrumento deve ser independente das duas partes do erro composto η_i citado no início.

Two-stage least squares

Uma forma mais intuitiva de trabalhar com Variáveis Instrumentais é através das *Two-stage least squares* (ou $2SLS$). Seguindo o raciocínio de antes, suponha que temos dados de Y, S e Z para cada observação i. Nesse caso o processo de geração de dados é dado por:

$$Y_i = \alpha + \delta S_i + \varepsilon_i$$

$$S_i = \gamma + \beta Z_i + \epsilon_i$$

onde $Cov(Z, \varepsilon) = 0$, $\beta \neq 0$. \ Sabendo que $\sum_{i=1}^{n}(X - i\bar{x}) = 0$ podemos reescrever o estimador de variável instrumental como

$$\hat{\delta} = \frac{Cov(Y, Z)}{Cov(S, Z)}$$
$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(S_i - \bar{S})}$$
$$= \frac{\frac{1}{n}\sum}{\frac{1}{n}}$$

Panel Data

Diferenças-em-Diferenças

Event Study, Two-Way Fixed Effects e Generalização de DiD