lista 1 Econometria

Miguel Sallum

24/05/2021

Questão 1

Com os dados da tabela abaixo, estime a regressão de Y em função de X2 e X3 e faça os testes da regressão e de cada um dos parâmetros.

```
Y <- c(800, 1160, 1580, 2010, 1890, 2600, 2070, 1890, 1830, 1740, 1380, 1060)

X <- tibble(
  X1 = 1,
  X2 = c(2, 4, 6, 8, 7, 12, 11, 10, 9, 8, 6, 4),
  X3 = c(.8, .7, .5, .4, .2, .2, .8, .7, .6, .1, .5, .4)
)
```

O modelo a ser estimado é:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + v_t$$

A Calcule os parâmetros $\beta_1,\ \beta_2,\ \beta_3$ desse modelo.

```
matX<-as.matrix(X)
Xt<-t(matX)
XtX<-Xt%*%matX
XtX_inv<-solve(XtX)

XtY<-Xt%*%Y
B_hat<-XtX_inv%*%XtY
B_hat
```

```
##         [,1]
## X1  789.3296
## X2  149.5593
## X3 -419.2566
```

B Monte a matriz de resíduos deste modelo. Calcule a soma dos quadrados dos resíduos utilizando método matricial.

```
k<-nrow(B_hat)
n<-nrow(Y)
u_hat<-Y-matX%*%B_hat
u_hat_t<-t(u_hat)
SSR<-u_hat_t%*%u_hat
SSR
```

```
##        [,1]
## [1,] 173444
```

C Calcule o R^2 deste modelo.

```
Yt<-t(Y)
SST<-Yt%*%Y
R2<-1-SSR/SST
R2
```

```
##           [,1]
## [1,] 0.9951975
```

D Monte a matriz de variância e covariância deste modelo.

```
k<-nrow(B_hat)
n<-length(Y)

sigma_hat<-as.numeric(SSR/(n-k))

var_cov<-sigma_hat*XtX_inv
var_cov
```

```
##            X1         X2          X3
## X1  24104.991 -1747.6474 -19990.3416
## X2  -1747.647   202.3422    570.8463
## X3 -19990.342   570.8463  32240.7574
```

E Verifique se os $\beta_1,\ \beta_2,\ \beta_3$ sao significantes ao nível de 5% de significância.

```
##    [,1]
## X1 TRUE
## X2 TRUE
## X3 TRUE
```

Questão 2

A questão anterior adicionamos uma variável *dummy*, que representa a existência ou não de determinado atributo.

```
Xd <- X %>%
  mutate(
    D= c(rep(1, 6), rep(0, 6))
  )
```

O modelo a ser estimado é:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 D + v_t$$

A Calcule os parâmetros $\beta_1,\ \beta_2,\ \beta_3$ desse modelo.

```
Xd<-as.matrix(Xd)
Xtd<-t(Xd)
XtXd<-Xtd%*%Xd
XtXd_inv<-solve(XtXd)

XtdY<-Xtd%*%Y
B_hatd<-XtXd_inv%*%XtdY
B_hatd
```

```
##         [,1]
```

```
## X1  536.0928
## X2  161.8657
## X3 -327.7779
## D   238.0763
```

B Monte a matriz de resíduos deste modelo. Calcule a soma dos quadrados dos resíduos utilizando método matricial.

```
kd<-nrow(B_hatd)
u_hatd<-Y-Xd%*%B_hatd
u_hatd_t<-t(u_hatd)
SSRd<-u_hatd_t%*%u_hatd
SSRd
```

```
##          [,1]
## [1,] 19854.22
```

C Calcule o R^2 deste modelo.

```
R2<-1-SSR/SST
R2
```

```
##           [,1]
## [1,] 0.9951975
```

D Monte a matriz de variância e covariância robusta deste modelo.

```
S<-diag(c(u_hatd))
S2<-S%*%S
var_cov_rob<-(n/n-kd)*XtXd_inv%*%(Xtd%*%S2%*%Xd)%*%XtXd_inv
```

E Verifique se os $\beta_1,\ \beta_2,\ \beta_3$ sao significantes ao nível de 5% de significância.

```
## Warning in sqrt(diag(var_cov_rob)): NaNs produzidos
```

```
##    [,1]
## X1   NA
## X2   NA
## X3   NA
## D    NA
```

Questão 3

Use os valores descritos na tabela abaixo para ilustrar que $E[Y_i(0)] - E[Y_i(1)] = E[Y_i(0) - Y_i(1)]$

```
vilas <- tibble(
  vila = 1:7,
  Y0 = c(10, 15, 20, 20, 10, 15, 15),
  Y1 = c(15, 15, 30, 15, 20, 15, 30),
  tau = Y1- Y0
)

EY0<-mean(vilas$Y0)
EY1<-mean(vilas$Y1)
Edif<- -mean(vilas$tau)
EY0 - EY1 == Edif
```

```
## [1] TRUE
```

Questão 4

Demonstre como chegar nessa igualdade:

$$\frac{1}{N_t}\sum_{i=1}^{n}(y_i|d_i=1) \; + \; \frac{1}{N_C}\sum_{i=1}^{n}(y_i|d_i=0) \; = \; E[Y^1] \; - \; E[Y^0]$$

$$+ \; E[Y^0|D=1] \; - \; E[Y^0|D=0] \; + \; (1 \; - \; \pi)(ATT \; - \; ATU)$$

Questão 5

Em que condições teremos a seguinte igualdade? Justifique

$$\frac{1}{N_t}\sum_{i=1}^{n}(y_i|d_i=1) \; + \; \frac{1}{N_C}\sum_{i=1}^{n}(y_i|d_i=0) \; = \; E[Y^1] \; - \; E[Y^0]$$

Questão 6

Suponha que um laboratório esteja testando um novo medicamento que tem como objetivo prolongar a vida de pacientes com câncer. Para realizar o estudo, os cientistas irão dividir sua amostra em dois grupos. Os indíviduos pares são os indivíduos do grupo tratamento e que tomam o medicamento, enquanto que os indivíduos ímpares pertencem ao grupo controle e tomam um placebo. Os efeitos dos medicamento sao diversos. Suponha que se o paciente tomar o medicamento, então ele terá a expectativa de vida $\mathbf{Y^1_i}$ adicional. De maneira análoga, se o paciente tomar o placebo, então ele terá a expectativa de vida $\mathbf{Y^2_i}$ adicional.

```
pacientes <- tibble(
  paciente = 1:15,
  Y1 = c(8, 9, 8, 4, 7, 1, 5, 7, 5, 4, 5, 10, 5, 10, 2),
  Y2 = c(6, 5, 4, 3, 2, 1, 4, 6, 4, 5, 2, 3, 4, 5, 1)
)
```

A Calcule os Efeitos médios de tratamento.

```
ATE<-pacientes%>%
  mutate(TE=
           Y1-Y2)%>%
  summarise(mean(TE))

ATE
```

```
## # A tibble: 1 x 1
##   `mean(TE)`
##        <dbl>
## 1       2.33
```

B Calcule os efeitos médios do grupo de tratamento. C Calcule os efeitos médios do grupo de controle.

```
AT_<-pacientes%>%
  mutate(
    TE = Y1 - Y2,
    D = (paciente + 1)%%2)%>%
  group_by(D)%>%
  summarise(efeito=mean(TE))#o AT_ entrega o ATU como 0 e o ATT como 1
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
AT_
```

```
## # A tibble: 2 x 2
##       D efeito
##   <dbl>  <dbl>
## 1     0   2.25
## 2     1   2.43
```

D O que voce conclui sobre a eficácia do novo medicamento?

Não teríamos nenhum desses resultados caso fosse feito realmente o experimento. No entanto, podemos ver que o remédio é eficaz (e veríamos isso também pela SDO)

Questão 7

Um determinado grupo de pesquisadores quer analisar a taxa de mortalidade média entre fumantes de cigarro e fumantes de cachimbo/charuto. Os pesquisadores possuem os dados da "tabela 2", em que há informações sobre a classificação etária dos indíviduos e a taxa de mortalidade para cada subgrupo.

```
fumantes <- tibble(
  faixa_etaria = c("20-40", "41-70", "71+"),
  taxa_mortalidade = c(.2, .4, .6),
  composição_cigarro = c(65, 25, 10),
  composição_charuto = c(10, 25, 65)
)
```

Após analisar os dados, responda as seguintes perguntas: A Qual é a taxa média de mortalidade para fumantes de cigarro sem subclassificação?

```
tx_real <- sum( fumantes$taxa_mortalidade * fumantes$composição_cigarro) / sum(fumantes$composição_ciga
tx_real
```

```
## [1] 0.29
```

B Observe que a distribuição etária dos fumantes de cigarros é exatamente o oposto (em termos de construção) dos fumantes de cachimbo e charuto. Portanto, a distribuição de idades é desequilibrada. Ajuste a taxa de mortalidade para fumantes de cigarro para que tenha a mesma distribuição de idade do grupo de comparação, no caso fumantes de cachimbo e charuto. Qual é a nova taxa média de mortalidade? Aumentou ou diminuiu?

```
tx_estimada <- sum( fumantes$taxa_mortalidade * fumantes$composição_charuto) / sum(fumantes$composição_
tx_estimada
```

```
## [1] 0.51
```

Questão 8

A tabela 3 fornece informações sobre idade e rendimento salarial de dois grupos, trainees e non-trainees. Sabendo que o método de *Matched Sample* é o mais adequado para comparação entre esses dois grupos, analise a diferença salarial entre trainees e non-trainees. Há diferença salarial? Monte a tabela de *Matched Sample*.

```
trainees <- tibble(
  unidade = 1:10,
  idade = c(18, 29, 24, 27, 33, 22, 19, 20, 21, 30),
  ganhos = c(9500, 12250, 11000, 11750, 13250, 10500, 9750, 10000, 10250, 12500)
```

```
)
non_trainees <- tibble(
  unidade = 1:20,
  idade = c(20, 27, 21, 39, 38, 29, 39, 33, 24, 30, 33, 36, 22, 18, 43, 39, 19, 30, 51, 48),
  ganhos = c(8500, 10075, 8725, 12775, 12550, 10525, 12775, 11425, 9400, 10750, 11425, 12100, 8950, 805
)

merge(trainees, non_trainees, by = "idade", all.x = T, sort = F)%>%
  distinct(unidade.x, .keep_all = T)%>%
  mutate(
    Dif = ganhos.x - ganhos.y)%>%
  summarise( ganho_matched = mean(Dif))
```

```
##   ganho_matched
## 1        1607.5
```

Há diferença salarial média, de $ 1607.50

Questão 9

Em qual situação o uso de regressão em discontinuidade é recomendado? Dê um exemplo prático e disserte sobre as vantagens desse método.

Questão 10

Imagine dois alunos - o primeiro aluno obteve 1240 e o segundo 1250. Esses dois alunos são realmente tão diferentes um do outro? Bem, claro: esses dois alunos individuais são provavelmente muito diferentes. Mas e se tivéssemos centenas de alunos que tiraram 1240 e centenas mais que fizeram 1250. Voce não acha que esses dois grupos são provavelmente muito semelhantes um ao outro em características observáveis e inobserváveis? Afinal, por que haveria de repente em 1250 uma grande diferença nas características dos alunos em uma grande amostra? Essa é a questão sobre a qual você deve refletir. Se a universidade está escolhendo arbitrariamente um ponto de corte razoável, há motivos para acreditar que ela também esta escolhendo um ponto de corte em que a habilidade natural dos alunos salta exatamente naquele ponto? Para analisar isso , Hoekstra (2009) realizou um estudo, utilizando dados disponibilizados por uma universidade estadual americana, em que realizou a seguinte estimação.

$$ln(earnings) = \psi year + \omega Experience + \theta cohort + \epsilon$$

Em que: • year é um vetor de dummies de anos. • Experience é um vetor de dummies para anos de rendimentos após o colegial, isto é, anos de experiência. • cohort é um vetor de dummies que controlam para o cohort(grupo) em que o aluno se inscreveu na universidade (por exemplo, 1988).

A Interprete os dados. Qual a relação entre a nota do SAT e os ganhos estimados? o quê explica essa descontinuidade no grafico? B "Estimated Discontinuity = 0.095 (z = 3.01)." o que isso significa?

Questão 11

Ainda pensando em RDDs, explique o conceito de "bandwidth". Como o uso de uma largura de banda maior afeta sua estimativa dos efeitos do tratamento?