

## P2 - Econometria III

Pedro Mendes

### 15, C2)

Os dados em `fertil2` incluem informações sobre o número de filhos, anos de escolaridade, idade, e variáveis de religião e status econômico de mulheres de Botsuana durante 1988.

```
fertil2 <- tibble::as_tibble(wooldridge::fertil2)
```

```
fertil2 |>  
  head()
```

```
## # A tibble: 6 x 27  
##   mnthborn yearborn   age electric radio   tv bicycle  educ   ceb agefbrth  
##   <int>    <int> <int>    <int> <int> <int>  <int> <int> <int>    <int>  
## 1         5      64   24         1     1     1         1    12     0      NA  
## 2         1      56   32         1     1     1         1    13     3     25  
## 3         7      58   30         1     0     0         0     5     1     27  
## 4        11      45   42         1     0     1         0     4     3     17  
## 5         5      45   43         1     1     1         1    11     2     24  
## 6         8      52   36         1     0     0         0     7     1     26  
## # ... with 17 more variables: children <int>, knowmeth <int>, usemeth <int>,  
## #   monthfm <int>, yearfm <int>, agefm <int>, idlnchld <int>, heduc <int>,  
## #   agesq <int>, urban <int>, urb_educ <int>, spirit <int>, protest <int>,  
## #   catholic <int>, frsthalf <int>, educ0 <int>, evermarr <int>
```

(i)

Estime o modelo

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

por OLS e interprete as estimativas. Em particular, mantendo **age** fixo, qual é o efeito estimado de mais um ano de escolaridade em fertilidade? Se 100 mulheres completassem mais um ano de escolaridade, haveria uma diminuição na quantidade de filhos (representados pela variável `children`)?

```
ols_model <- fertil2 |>  
  fixest::feols(children ~ educ + age + I(age^2))
```

```
summary(ols_model)
```

```
## OLS estimation, Dep. Var.: children
## Observations: 4,361
## Standard-errors: IID
##           Estimate Std. Error   t value   Pr(>|t|)
## (Intercept) -4.138307   0.240594 -17.20036 < 2.2e-16 ***
## educ        -0.090575   0.005921 -15.29813 < 2.2e-16 ***
## age         0.332449   0.016549  20.08815 < 2.2e-16 ***
## I(age^2)     -0.002631   0.000273  -9.65113 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.45908   Adj. R2: 0.568427
```

- Caso haja um aumento de um ano na educação de 100 mulheres, em média, ocorreria uma diminuição em 9 na quantidade de filhos.
- O aumento de um ano na idade leva a um aumento de 33 na quantidade de filhos.
- Aparentemente existe um efeito não linear significativo na idade, o que pode expressar a questão da tendência decrescente da fertilidade das mulheres com o passar dos anos, o que mostra que o efeito do aumento da idade na quantidade de filhos não é o mesmo para todas as idades.

## (ii)

A variável **frsthalf** é uma variável *dummy* igual a um, caso a mulher tenha nascido durante os primeiros seis meses do ano. Presumindo que **frsthalf** não seja correlacionada com o termo de erro do item (i), mostre que **frsthalf** é um candidato VI razoável a **educ** (Dica: é preciso fazer uma regressão).

```
forma_reduzida <- fertil2 |>
  fixest::feols(educ ~ age + I(age^2) + frsthalf)
```

```
summary(forma_reduzida)
```

```
## OLS estimation, Dep. Var.: educ
## Observations: 4,361
## Standard-errors: IID
##           Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)   9.692864   0.598069 16.206945 < 2.2e-16 ***
## age          -0.107950   0.042040 -2.567789 1.0268e-02 *
## I(age^2)      -0.000506   0.000693 -0.729597 4.6568e-01
## frsthalf      -0.852285   0.112830 -7.553742 5.1227e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.70926   Adj. R2: 0.107037
```

```
theta <- forma_reduzida$coefficients['frsthalf']
r <- residuals(forma_reduzida)
z <- fertil2$frsthalf
age <- fertil2$age

cov_r_z <- cov(r, z)
```

1.  $cov(u, z) = 0$  por suposição do item (ii)
2.  $\theta \neq 0 \rightarrow -0.8522854$
3.  $cov(r, z) \approx 0 \rightarrow -3.5624827 \times 10^{-15}$
4.  $cov(r, x_j), j = 1, 2 \approx 0 \rightarrow$  rodando a regressão  $r = \delta_0 + \delta_1 age + \delta_2 age^2 + e$ , percebe-se que todos os coeficientes são aproximadamente iguais a 0, além de nenhuma ser significativa.

```
model_r <- fertil2 |>
  dplyr::bind_cols(r = r) |>
  fixest::feols(r ~ age + I(age^2))

summary(model_r)
```

```
## OLS estimation, Dep. Var.: r
## Observations: 4,361
## Standard-errors: IID
##               Estimate Std. Error      t value Pr(>|t|)
## (Intercept)  2.4020e-15   0.595427  4.0300e-15      1
## age         -3.8754e-14   0.042035 -9.2195e-13      1
## I(age^2)     2.1200e-16   0.000693  3.0657e-13      1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.70926   Adj. R2: -4.589e-4
```

Logo, `frsthalf` é um candidato IV razoável para `age`.

### (iii)

**Estime o modelo do item (i) usando `frsthalf` como IV para `educ`. Compare o efeito estimado de educação com a estimativa OLS do item (i).**

```
iv_model <- AER::ivreg(
  children ~ age + I(age^2) + educ | age + I(age^2) + frsthalf,
  data = fertil2
)

summary(iv_model)
```

```
##
## Call:
## AER::ivreg(formula = children ~ age + I(age^2) + educ | age +
##      I(age^2) + frsthalf, data = fertil2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.05272 -0.71481  0.06224  0.76236  7.23693
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.3878054  0.5481502  -6.180 6.98e-10 ***
## age          0.3236052  0.0178596  18.119 < 2e-16 ***
## I(age^2)     -0.0026723  0.0002797  -9.555 < 2e-16 ***
## educ         -0.1714989  0.0531796  -3.225  0.00127 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.491 on 4357 degrees of freedom
## Multiple R-Squared:  0.5502,   Adjusted R-squared:  0.5499
## Wald test:  1765 on 3 and 4357 DF,   p-value: < 2.2e-16
```

O efeito da educação sobre o número de filhos é maior quando estimado pelo modelo de variável instrumental usando a variável `frsthalf`, já que em tal modelo, o efeito de um ano adicional de educação sobre o número de filhos é 1.89 vezes maior.

#### (iv)

**Adicione as variáveis binárias `electric`, `tv` e `bicycle` ao modelo e presuma que elas sejam exógenas. Estime a equação por OLS e 2SLS e compare os coeficientes estimados em `educ`. Interprete o coeficiente em `tv` e explique por que a posse de televisão tem efeito negativo sobre a fertilidade.**

```
ols_model_2 <- fertil2 |>
  fixest::feols(children ~ age + I(age^2) + electric + tv + bicycle + educ)
```

```
## NOTE: 5 observations removed because of NA values (RHS: 5).
```

```
summary(ols_model_2)
```

```
## OLS estimation, Dep. Var.: children
## Observations: 4,356
## Standard-errors: IID
##              Estimate Std. Error   t value   Pr(>|t|)
## (Intercept) -4.389784   0.240317 -18.26662 < 2.2e-16 ***
```

```
## age          0.340204    0.016442   20.69153   < 2.2e-16 ***
## I(age^2)     -0.002708    0.000271  -10.00951   < 2.2e-16 ***
## electric     -0.302729    0.076187   -3.97351   7.1969e-05 ***
## tv           -0.253144    0.091437   -2.76850   5.6554e-03 **
## bicycle      0.317895    0.049366    6.43954   1.3283e-10 ***
## educ        -0.076709    0.006353  -12.07528   < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.44664    Adj. R2: 0.575475
```

```
iv_model_2 <- AER::ivreg(
  children ~ age + I(age^2) + electric + tv + bicycle + educ | age + I(age^2) +
  data = fertil2
)

summary(iv_model_2)
```

```
##
## Call:
## AER::ivreg(formula = children ~ age + I(age^2) + electric + tv +
##           bicycle + educ | age + I(age^2) + electric + tv + bicycle +
##           frsthalf, data = fertil2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.9519 -0.7184  0.0290  0.7384  7.3372
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.5913324   0.6450889  -5.567 2.74e-08 ***
## age          0.3281451   0.0190587  17.218 < 2e-16 ***
## I(age^2)     -0.0027222   0.0002766  -9.843 < 2e-16 ***
## electric     -0.1065314   0.1659650   -0.642  0.5210
## tv           -0.0025550   0.2092301   -0.012  0.9903
## bicycle      0.3320724   0.0515264    6.445 1.28e-10 ***
## educ        -0.1639814   0.0655269   -2.503  0.0124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.479 on 4349 degrees of freedom
## Multiple R-Squared:  0.5577,    Adjusted R-squared:  0.5571
## Wald test: 921.7 on 6 and 4349 DF,  p-value: < 2.2e-16
```

Porque ter tv te faz transar menos.

## Cap. 17, C8)

O arquivo `jtrain2` contém dados sobre um experimento de treinamento profissional para um grupo de homens. O programa começaria em janeiro de 1976 e se estenderia até meados de 1977. O programa acabou em dezembro de 1977. A ideia é testar se a participação no programa de treinamento profissional teve um efeito nas probabilidades de desemprego e rendimentos de 1978.

```
jtrain2 <- tibble::as_tibble(wooldridge::jtrain2)
```

```
jtrain2 |> head()
```

```
## # A tibble: 6 x 19
##   train  age  educ black  hisp married nodegree mosinex  re74  re75  re78
##   <int> <int> <int> <int> <int>   <int>   <int>   <int> <dbl> <dbl> <dbl>
## 1     1    37    11     1     0       1       1     13     0     0  9.93
## 2     1    22     9     0     1       0       1     13     0     0  3.60
## 3     1    30    12     1     0       0       0     13     0     0 24.9
## 4     1    27    11     1     0       0       1     13     0     0  7.51
## 5     1    33     8     1     0       0       1     13     0     0  0.290
## 6     1    22     9     1     0       0       1     13     0     0  4.06
## # ... with 8 more variables: unem74 <int>, unem75 <int>, unem78 <int>,
## #   lre74 <dbl>, lre75 <dbl>, lre78 <dbl>, agesq <int>, mostrn <int>
```

(ii)

Estabeleça uma regressão linear de treino em muitas variáveis demográficas e pré-treino: `unem74`, `unem75`, `age`, `educ`, `black`, `hisp` e `married`. Essas variáveis são significativas conjuntamente ao nível de 5%?

```
ols_model_train <- lm(train ~ unem74 + unem75 + age + educ + black + hisp + married, data = jtrain2)
summary(ols_model_train)
```

```
##
## Call:
## lm(formula = train ~ unem74 + unem75 + age + educ + black + hisp + married, data = jtrain2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.6024 -0.4196 -0.3437  0.5537  0.7669
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.338022   0.189445   1.784   0.0751 .
```

```
## unem74      0.020880    0.077294    0.270    0.7872
## unem75     -0.095571    0.071902   -1.329    0.1845
## age         0.003206    0.003403    0.942    0.3467
## educ        0.012013    0.013342    0.900    0.3684
## black       -0.081666    0.087732   -0.931    0.3524
## hisp        -0.200017    0.116971   -1.710    0.0880 .
## married     0.037289    0.064404    0.579    0.5629
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4917 on 437 degrees of freedom
## Multiple R-squared:  0.02238,    Adjusted R-squared:  0.006722
## F-statistic: 1.429 on 7 and 437 DF,  p-value: 0.1915
```

As variáveis não são conjuntamente significantes ao nível de 5% (p-valor = 0,1915).

### (iii)

**Estime uma versão probit do modelo linear do item (ii). Calcule o teste de razão de verossimilhança para a significância conjunta de todas as variáveis. O que você conclui?**

```
probit_model_train <- glm(
  train ~ unem74 + unem75 + age + educ + black + hisp + married,
  family = binomial(link = "probit"),
  data = jtrain2
)

summary(probit_model_train)

##
## Call:
## glm(formula = train ~ unem74 + unem75 + age + educ + black +
##      hisp + married, family = binomial(link = "probit"), data = jtrain2)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3620  -1.0421  -0.9159   1.2702   1.6962
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.424107   0.489506  -0.866   0.3863
## unem74       0.053026   0.198834   0.267   0.7897
## unem75      -0.247725   0.184806  -1.340   0.1801
## age          0.008344   0.008780   0.950   0.3419
```

```
## educ          0.031443    0.034657    0.907    0.3643
## black        -0.206930    0.224614   -0.921    0.3569
## hisp         -0.539777    0.307947   -1.753    0.0796 .
## married      0.096625    0.165503    0.584    0.5593
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 604.20  on 444  degrees of freedom
## Residual deviance: 594.02  on 437  degrees of freedom
## AIC: 610.02
##
## Number of Fisher Scoring iterations: 4
```

```
lmtest::lrtest(probit_model_train)
```

```
## Likelihood ratio test
##
## Model 1: train ~ unem74 + unem75 + age + educ + black + hisp + married
## Model 2: train ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1    8 -297.01
## 2    1 -302.10 -7 10.182    0.1785
```

As variáveis não são conjuntamente significantes ao nível de 5%.

#### (iv)

**Com base em suas respostas aos itens (ii) e (iii), parece-lhe que a participação em treinamento profissional possa ser tratada como exógena como forma de explicar o status de desemprego de 1978? Explique.**

Sim, pois a variável não apresenta correlação relevante em relação às demais variáveis do modelo, ou seja, a variável não apresenta endogeneidade.

#### (v)

**Estabeleça uma regressão simples de unem78 em train e reporte os resultados em forma de equação. Qual é o efeito estimado de participar do programa de treinamento na probabilidade de estar desempregado em 1978? Isso é estatisticamente significativo?**

```
ols_model_unem78 <- lm(unem78 ~ train, data = jtrain2)
summary(ols_model_unem78)
```



```
##
## Call:
## lm(formula = unem78 ~ train, data = jtrain2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3538 -0.3538 -0.2432  0.6462  0.7568
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.35385     0.02849  12.419  <2e-16 ***
## train       -0.11060     0.04419  -2.503   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4594 on 443 degrees of freedom
## Multiple R-squared:  0.01394,    Adjusted R-squared:  0.01172
## F-statistic: 6.265 on 1 and 443 DF,  p-value: 0.01267
```

$$unem78 = -0.11 \cdot train$$

**(vi)**

**Estabeleça um probit de unem78 em train. Faz sentido comparar o coeficiente probit em train com o coeficiente obtido do modelo linear do item (v)?**

```
probit_model_unem78 <- glm(unem78 ~ train, data = jtrain2, family = binomial(link = "probit"),
summary(probit_model_unem78)
```

```
##
## Call:
## glm(formula = unem78 ~ train, family = binomial(link = "probit"),
##      data = jtrain2)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.9346 -0.9346 -0.7466  1.4414  1.6815
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.37496     0.07975  -4.702 2.58e-06 ***
## train       -0.32095     0.12848  -2.498  0.0125 *
##
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 549.47  on 444  degrees of freedom
## Residual deviance: 543.17  on 443  degrees of freedom
## AIC: 547.17
##
## Number of Fisher Scoring iterations: 4
```

Não, pois a interpretação do efeito marginal dado pelo coeficiente do modelo probit depende dos valores de nível da variável independente em questão, nesse caso, `train`.

**(vii)**

**Encontre as probabilidades apropriadas dos itens (v) e (vi). Explique por que elas são idênticas. Qual abordagem você usaria para medir o efeito e a significância estatística do programa de treinamento profissional?**

```
cat("=== COEFICIENTES PROBIT ===\n")
```

```
## === COEFICIENTES PROBIT ===
```

```
pnorm(probit_model_unem78$coefficients)
```

```
## (Intercept)      train
##  0.3538462    0.3741239
```

```
cat("\n=== VALORES AJUSTADOS PROBIT (train) ===\n")
```

```
##
## === VALORES AJUSTADOS PROBIT (train) ===
```

```
predict.glm(probit_model_unem78, type = "response") |>
  dplyr::as_tibble() |>
  dplyr::distinct(value) |>
  purrr::flatten_dbl()
```

```
## [1] 0.2432432 0.3538462
```

```
cat("\n=== COEFICIENTES OLS ===\n")
```

```
##
## === COEFICIENTES OLS ===
```

```
ols_model_unem78$coefficients
```

```
## (Intercept)      train  
##    0.3538462   -0.1106029
```

```
cat("\n=== VALORES AJUSTADOS OLS (train) ===\n")
```

```
##  
## === VALORES AJUSTADOS OLS (train) ===
```

```
predict(ols_model_unem78, type = "response") |>  
  dplyr::as_tibble() |>  
  dplyr::distinct(value) |>  
  purrr::flatten_dbl()
```

```
## [1] 0.2432432 0.3538462
```