Lista Prática 4

Exercício 1. *Neste exercício, usaremos a base de dados* `meap00_01`*, do pacote* `wooldridge`*, cujas observações são escolas de Michigan. Queremos estimar o seguinte modelo:*

$$math4 = \beta_0 + \beta_1 lunch + \beta_2 lenroll + \beta_3 lexppp + \varepsilon \tag{1.1}$$

em que

- `math4`*: percental de alunos de 4º ano com nível de matemática satisfatório*
- `lunch`*: percentual de estudantes elegíveis ao programa de auxílio para almoço*
- `lenroll`*: log de matrículas escolares*
- `lexppp`*: log de gastos por estudante*

a) *Estime por MQO com erros padrão usuais e robustos à heterocedasticidade por White (1980). Obtenha a matriz de variâncias-covariâncias do estimador robusta "na mão", ou seja, não use funções que façam os cálculos para você (como* `vcovHC(...,"HC0")` *ou outras)*[1]*. As estimativas e os erros padrão são distintos? Justifique.*

b) *Faça o teste de White para heterocedasticidade. Quais são os modelos restrito e irrestrito utilizados no teste F? Analise o resultado do teste a partir estatística obtida.*

c) *Faça a regressão de $g_i \equiv \log(\hat{\varepsilon}_i^2)$ em função dos valores preditos $\widehat{math4}$ e $\widehat{math4}^2$, obtidos pelo modelo do item (a). Calcule os pesos $w_i = 1/\exp(\hat{g}_i)$ para estimar o modelo (1.1) por MQGF. Há diferenças em relação aos resultados obtidos por MQO?*

d) *Davidson e MacKinnon (1999, pág. 264) afirmam que é possível iterar o procedimento de MQGF. Para isto usam-se as estimativas obtidas no item (c) para calcular novos pesos $w_i = 1/\exp(\hat{g}_i)$ e novas estimativas, e isso pode ser repetido iteradamente. Faça iterações do procedimento de MQGF enquanto a seguinte condição for verdadeira:*

$$\max\{\hat{\boldsymbol{\beta}}^{it+1} - \hat{\boldsymbol{\beta}}^{it}\} > 1 \times 10^{-15},$$

em que `it` *é o número de uma iteração.*

Informe quantas iterações foram necessárias para a convergência e compare os quatro modelos estimados (MQO, MQO com erros padrão robustos, MQGF, e MQGF com iteração) usando a função `stargazer()` *do pacote de mesmo nome. Há algum modelo melhor para analisar o efeito de gastos (*`lexppp`*) em* `math4`*? Explique.*

[1] Além de obter os resíduos por meio da função `residuals()` sobre o objeto de regressão gerado por `lm()`, é possível obter a matriz $\boldsymbol{X}$ usando `model.matrix()` no mesmo objeto.

Resposta:

a) A estimação robusta à heterocedasticidades resulta em erros padrão maiores do que estimado por MQO.

```r
## MQO
reg.ols = lm(math4 ~ lunch + lenroll + lexppp, meap00_01)

## MQO com erros padrão robustos
# calculando matriz de vcov robusta
ehat = resid(reg.ols)
X = model.matrix(reg.ols)
Sigma = diag(ehat^2)
bread = solve(t(X) %*% X)
meat = t(X) %*% Sigma %*%  X
vcov_robust = bread %*% meat %*% bread
vcov_robust

reg.ols.rob = coeftest(reg.ols, vcov=vcov_robust)
# reg.ols.rob = coeftest(reg.ols, vcov=sandwich::vcovHC(reg.ols,"HC0"))

# Comparando resultados
stargazer::stargazer(reg.ols, reg.ols.rob, type="text")
```

```
=====================================================================
                                 Dependent variable:
                      -----------------------------------------------
                               math4
                                OLS                  coefficient
                                                        test
                                (1)                      (2)
---------------------------------------------------------------------
lunch                        -0.449***                -0.449***
                             (0.015)                  (0.017)

lenroll                      -5.399***                -5.399***
                             (0.940)                  (1.130)

lexppp                        3.525*                   3.525
                             (2.098)                  (2.351)

Constant                     91.932***                91.932***
                             (19.962)                 (23.060)

---------------------------------------------------------------------
Observations                  1,692
R2                            0.373
Adjusted R2                   0.372
Residual Std. Error     15.302 (df = 1688)
F Statistic          334.567*** (df = 3; 1688)
=====================================================================
Note:                                     *p<0.1; **p<0.05; ***p<0.01
```

b) Precisamos estimar o seguinte modelo:

$$\hat{\varepsilon}_i^2 = \alpha + \gamma \widehat{\text{math4}}_i + \delta \widehat{\text{math4}}_i^2 + e$$

O teste de White compara este modelo acima (irrestrito - ur) com o modelo sem nenhuma variável explicativa (restrito - r) que, portanto, possui $R_r^2 = 0$. A hipótese de que ambos modelos são estatisticamente distintos pode ser testado pela seguinte estatística F:

$$F = \frac{R^2_{ur} - R^2_r}{(1 - R^2_{ur})} \frac{N - K - 1}{2}$$

No R, basta usar a função `summary()` no objeto da regressão do modelo estimado:

```
ehat = resid(reg.ols)
yhat = fitted(reg.ols)
reg.resid = lm(ehat^2 ~ yhat + I(yhat^2))
summary(reg.resid)
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept) 1664.12793  315.21438   5.279 1.46e-07 ***
yhat         -28.29903    9.21685  -3.070  0.00217 **
I(yhat^2)      0.11553    0.06591   1.753  0.07982 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

F-statistic: 132.7 on 2 and 1689 DF,  p-value: < 2.2e-16
```

A estatística $F = 132.7$ é uma forte evidência contra a hipótese nula (homocedasticidade).

c) Encontraremos os valores preditos, $\hat{g}_i$, do seguinte modelo:

$$g_i \equiv \log(\hat{\varepsilon}_i^2) = \gamma_0 + \gamma_1 \widehat{\text{math4}}_i + \gamma_2 \widehat{\text{math4}}_i^2 + u$$

```
reg.aux = lm(log(ehat^2) ~ yhat + I(yhat^2))
head(fitted(reg.aux))
```

```
       1        2        3        4        5        6
4.138865 3.194563 3.476371 3.698343 3.739361 3.746300
```

Agora, calcularemos os pesos $w_i = 1/\exp(\hat{g}_i)$ para utilizá-los na estimação por MQGF:

```
w = 1 / exp(fitted(reg.aux))
reg.fgls = lm(math4 ~ lunch + lenroll + lexppp, meap00_01, weights=w)

# Comparativo
stargazer::stargazer(reg.ols, reg.ols.rob, reg.fgls, type="text")
```

```
==========================================================
                          Dependent variable:
                    --------------------------------------
                       math4                    math4
                        OLS      coefficient     OLS
                                    test
                        (1)          (2)          (3)
----------------------------------------------------------
lunch                -0.449***    -0.449***    -0.450***
                      (0.015)      (0.017)      (0.014)

lenroll              -5.399***    -5.399***    -2.819***
                      (0.940)      (1.130)      (0.875)

lexppp                3.525*       3.525        6.438***
                      (2.098)      (2.351)      (1.706)

Constant             91.932***    91.932***    51.896***
                      (19.962)     (23.060)     (16.859)
```

```
----------------------------------------------------------------------------
Observations                          1,692                   1,692
R2                                    0.373                   0.371
Adjusted R2                           0.372                   0.370
Residual Std. Error (df = 1688)       15.302                  1.876
F Statistic (df = 3; 1688)          334.567***              331.527***
============================================================================
Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

d) Foram necessárias 122 iterações para convergência. O resultado do MQGF com iterações foi parecido com o MQGF padrão, sendo que os erros padrão robustos foram um pouco menores.

```r
bhat.ini = coef(reg.fgls)
diff_bhat = max(bhat.ini)
reg.fgls.it = reg.fgls
it=0

while (diff_bhat > 1e-15) {
  ehat.it = resid(reg.fgls.it) # novos resíduos
  yhat.it = fitted(reg.fgls.it) # novos valores preditos
  reg.aux.it = lm(log(ehat.it^2) ~ yhat.it + I(yhat.it^2)) # nova reg auxiliar
  w.it = 1 / exp(fitted(reg.aux.it)) # novos pesos
  reg.fgls.it = lm(math4 ~ lunch + lenroll + lexppp, meap00_01, weights=w.it)
  bhat.it = coef(reg.fgls.it) # novas estimativas
  diff_bhat = max(bhat.it - bhat.ini) # máximo das diferenças de estimativas
  bhat.ini = bhat.it # atualização de valor inicial das estimativas
  it=it+1 # atualização da iteração
}

print(paste0("número de iterações: ", it)) # número de iterações
stargazer::stargazer(reg.ols, reg.ols.rob, reg.fgls, reg.fgls.it, type="text")
```

```
[1] número de iterações: 122
===========================================================================================
                                              Dependent variable:
                                 ----------------------------------------------------------
                                   math4                             math4
                                    OLS       coefficient             OLS
                                                 test
                                    (1)          (2)          (3)          (4)
-------------------------------------------------------------------------------------------
lunch                            -0.449***    -0.449***    -0.450***    -0.453***
                                  (0.015)      (0.017)      (0.014)      (0.015)

lenroll                          -5.399***    -5.399***    -2.819***    -2.758***
                                  (0.940)      (1.130)      (0.875)      (0.845)

lexppp                             3.525*       3.525       6.438***     6.213***
                                  (2.098)      (2.351)      (1.706)      (1.652)

Constant                         91.932***    91.932***    51.896***    53.487***
                                 (19.962)     (23.060)     (16.859)     (16.326)

-------------------------------------------------------------------------------------------
Observations                       1,692                     1,692        1,692
R2                                 0.373                     0.371        0.367
Adjusted R2                        0.372                     0.370        0.366
Residual Std. Error (df = 1688)    15.302                    1.876        1.894
F Statistic (df = 3; 1688)       334.567***                331.527***   326.558***
===========================================================================================
Note:                                                   *p<0.1; **p<0.05; ***p<0.01
```

De modo geral, os resultados de MQGF geraram estimativas maiores para `lenroll` e `lexppp`, inclusive sendo mais eficientes (menores erros padrão). No entanto, deve-se ter cuidado, pois o MQGF pode gerar estimativas viesadas quando a matriz de pesos ($\boldsymbol{W}$) e, consequentemente, a matriz de variâncias-covariâncias dos erros ($\boldsymbol{\Sigma}$) estiverem mal especificadas (por outro lado, o MQO gera estimativas não-viesadas/consistentes, embora ineficientes). No entanto, na subseção "Why Feasible GLS Works", Davidson e MacKinnon (1999, pág. 262) dão uma intuição da prova de que, sob condições de regularidade, o estimador de MQGF é consistente e assintoticamente equivalente ao estimador de MQG, o qual é não-viesado/consistente e eficiente na presença de heterocedasticidade.

□

Exercício 2. *Neste exercício, usaremos a base de dados de Angrist e Evans (1998) que contém quase 32 mil observações de mulheres negras ou hispânicas, todas casadas e com pelo menos dois filhos. Para carregá-la no R, use:*

```
data(labsup, package="wooldridge")
```

Queremos saber como o número de crianças (`kids`) afeta as horas trabalhadas por semana (`hours`). Há suspeita de que `kids` seja um regressor endógeno e, portanto, utilizaremos variáveis instrumentais para gerar variação exógena que permita identificar o efeito do nº de crianças sobre `hours`.

a) *Estime a equação por MQO com erros padrão robustos para heterocedasticidade:*

$$hours = \beta_0 + \beta_1 kids + \beta_2 nonmomi + \beta_3 educ + \beta_4 age + \beta_5 agesq + \beta_6 black + \beta_7 hispan + \varepsilon$$

- `kids`: *nº de crianças*
- `nonmomi`: *renda familiar que não é da mãe*
- `educ`: *anos de educação*
- `age`: *idade*
- `agesq`: *idade ao quadrado*
- `black`: *dummy negra*
- `hispan`: *dummy hispânica*

b) *Um dos instrumentos propostos por Angrist e Evans é a variável dummy `samesex`, que possui valor 1 quando os dois primeiros filhos são do mesmo sexo, e 0 caso contrário. Exponha argumentos (favoráveis ou contrários) para que `samesex` esteja relacionado a `kids`, e que não afete diretamente as horas trabalhadas (`hours`).*

c) *Faça o teste de instrumento fraco no primeiro estágio da regressão. O quão estatisticamente significante é a variável `samesex` em relação a `kids`?*

d) *Use `samesex` como instrumento de `kids`. Estime o modelo estrutural do item (a) pelo estimador VI por meio da* ***função `ivreg()`*** *e* ***analiticamente ("na mão")****. Compare os resultados MQO e VI usando a função `stargazer()` do pacote de mesmo nome.*

e) *Avalie a potencial endogeneidade de* **`kids`** *por meio do Teste de (Durbin-Wu-)Hausman. Analise o seu resultado, explicitando a hipótese nula testada.*

c') *Deste item em diante, considere como um segundo instrumento para* **`kids`** *a variável* **`multi2nd`***, que é uma dummy indicando que os dois primeiros filhos são gêmeos. Faça o teste de instrumentos fracos no primeiro estágio da regressão. O quão conjuntamente significantes são as variáveis* **`samesex`** *e* **`multi2nd`** *em relação a* **`kids`***?*

d') *Use* **`samesex`** *e* **`multi2nd`** *como instrumentos de* **`kids`** *e estime o modelo estrutural do item (a) pelo estimador MQ2E por meio da* ***função `ivreg()`****, da* ***função `lm()`*** *e* ***analiticamente ("na mão")****. Compare os resultados MQO, VI e MQ2E usando a função* **`stargazer()`***.*

e') *Avalie a potencial endogeneidade de* **`kids`** *por meio do Teste de (Durbin-Wu-)Hausman, porém agora comparando* $\hat{\boldsymbol{\beta}}^{MQO}$ *e* $\hat{\boldsymbol{\beta}}^{MQ2E}$*. Analise o seu resultado, explicitando a hipótese nula testada.*

f) *Faça o teste de sobreidentificação de Sargan e analise o resultado, explicitando sua hipótese nula.*

Resposta:

a)
```r
# Pacotes para inferência robusta com heterocedasticidade
library(lmtest)
library(sandwich)

# MQO
reg.ols = lm(hours ~ kids + nonmomi + educ + age + agesq + black + hispan,
         data=labsup)

# MQO com erros padrão robustos
reg.ols.rob = coeftest(reg.ols, vcov=vcovHC(reg.ols, "HC0"))
round(reg.ols.rob, 3)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -10.447      6.588  -1.586    0.113
kids          -2.326      0.116 -20.137   <2e-16 ***
nonmomi       -0.058      0.005 -10.808   <2e-16 ***
educ           0.586      0.037  15.634   <2e-16 ***
age            2.049      0.448   4.570   <2e-16 ***
agesq         -0.028      0.008  -3.602   <2e-16 ***
black          1.058      1.351   0.784    0.433
hispan        -5.114      1.351  -3.784   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b) A ideia é que, caso os pais tenham dois primeiros filhos(as) do mesmo sexo, então é mais provável que o casal tenha mais algum filho. Logo, acredita-se que `samesex` seja positivamente correlacionado com `kids`. Por outro lado, é possível argumentar que `samesex` possa afetar diretamente as horas trabalhadas, pois crianças/adolescentes de diferentes sexos podem não compartilhar roupas e brinquedos, demandando mais recursos financeiros e, consequentemente, mais horas trabalhadas pela mãe.

c) Vamos fazer o teste a partir do modelo de primeiro estágio:

$$kids = \gamma_0 + \gamma_1 samesex + \gamma_2 nonmomi + \gamma_3 educ + \gamma_4 age + \gamma_5 agesq + \gamma_6 black + \gamma_7 hispan + \varepsilon$$

```r
reg.1st = lm(kids ~ samesex + nonmomi + educ + age + agesq + black + hispan,
             data=labsup)
round(coeftest(reg.1st), 3)
```

```
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)     2.010      0.321    6.263   <2e-16 ***
samesex         0.070      0.010    6.847   <2e-16 ***
nonmomi        -0.003      0.000  -10.591   <2e-16 ***
educ           -0.085      0.002  -49.506   <2e-16 ***
age             0.059      0.022    2.694    0.007 **
agesq           0.000      0.000    0.005    0.996
black           0.013      0.066    0.195    0.845
hispan         -0.042      0.066   -0.643    0.520
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para verificar sua significância, só precisamos ver a estatística t ou p-valor da estimativa de `samesex`, $\hat{\gamma}_1$, que é significante a 0,1%, ou seja, rejeitamos a hipótese nula de que é um instrumento fraco ($\gamma_1 = 0$). Alternativamente, seria possível realizar esse teste por meio de estatística F (comparando um modelo irrestrito e outro modelo restrito).

d) $\hat{\boldsymbol{\beta}}^{VI} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{y}$, $V(\hat{\boldsymbol{\beta}}^{VI}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X})^{-1}$ e $\hat{\sigma}^2 = (\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})/(N-K-1)$

```r
## Estimação via ivreg()
library(ivreg)
reg.iv = ivreg(hours ~ kids + nonmomi + educ + age + agesq + black + hispan |
                 samesex + nonmomi + educ + age + agesq + black + hispan,
               data=labsup)

stargazer::stargazer(reg.ols.rob, reg.iv, type="text", digits=3)
```

```
===========================================================
                           Dependent variable:
                   ----------------------------------------
                                           hours
                     coefficient        instrumental
                        test              variable
                         (1)                 (2)
-----------------------------------------------------------
kids                 -2.326***           -4.879
                      (0.116)            (3.013)

nonmomi              -0.058***           -0.065***
                      (0.005)            (0.010)

educ                  0.586***            0.368
                      (0.037)            (0.260)

age                   2.049***            2.201***
                      (0.448)            (0.486)

agesq                -0.028***           -0.028***
                      (0.008)            (0.008)

black                  1.058              1.095
                      (1.351)            (1.361)

hispan               -5.114***           -5.218***
```

```
                         (1.351)          (1.368)

Constant                 −10.447          −5.254
                         (6.588)          (9.024)

─────────────────────────────────────────────────
Observations                              31,857
R2                                         0.058
Adjusted R2                                0.058
Residual Std. Error                 18.924 (df = 31849)
═════════════════════════════════════════════════
Note:                      *p<0.1; **p<0.05; ***p<0.01
```

A estimação por VI fez o erro padrão de `kids` (e de `educ`) "explodir", tornando sua estimativa estatisticamente não-significante. Agora, estimando analiticamente:

```r
## Estimação analítica
y = as.matrix(labsup[,"hours"])
X = as.matrix(cbind(1, labsup[,c("kids", "nonmomi", "educ", "age",
                                 "agesq", "black", "hispan")]))
Z = as.matrix(cbind(1, labsup[,c("samesex", "nonmomi", "educ", "age",
                                 "agesq", "black", "hispan")]))
Pz = Z %*% solve(t(Z) %*% Z) %*% t(Z)

N = nrow(X)
K = ncol(X) - 1

bhat.iv = solve(t(Z) %*% X) %*% t(Z) %*% y
yhat = X %*% bhat.iv
ehat = y - yhat

sig2hat = as.numeric( t(ehat) %*% ehat / (N-K-1) )
Vbhat = sig2hat * solve(t(X) %*% Pz %*% X)

se = sqrt(diag(Vbhat))
t = bhat.iv / se
p = 2 * pt(-abs(t), N-K-1)

round(data.frame(bhat.iv, se, t, p), 3)
```

```
          bhat.iv    se      t      p
1          −5.254 9.024 −0.582 0.560
kids       −4.879 3.013 −1.619 0.105
nonmomi    −0.065 0.010 −6.511 0.000
educ        0.368 0.260  1.417 0.156
age         2.201 0.486  4.531 0.000
agesq      −0.028 0.008 −3.577 0.000
black       1.095 1.361  0.805 0.421
hispan     −5.218 1.368 −3.815 0.000
```

e) O teste de Hausman pode ser feito via estatística de Wald, que avalia a hipótese nula de que as estimativas MQO e VI são estatisticamente iguais:

$$w = (\hat{\boldsymbol{\beta}}^{VI} - \hat{\boldsymbol{\beta}}^{MQO})' \left[V(\hat{\boldsymbol{\beta}}^{VI}) - V(\hat{\boldsymbol{\beta}}^{MQO}) \right]^{-1} (\hat{\boldsymbol{\beta}}^{VI} - \hat{\boldsymbol{\beta}}^{MQO}) \sim \chi^2_J$$

em que $J = 1$ regressor endógeno neste caso.

```r
bhat.ols = coef(reg.ols)
bhat.iv = coef(reg.iv)
```

```
Vbhat.ols = vcov(reg.ols)
Vbhat.iv = vcov(reg.iv)

contrast = bhat.iv - bhat.ols
w = t(contrast) %*% solve(Vbhat.iv - Vbhat.ols) %*% contrast
w
1 - pchisq(abs(w), 1)
```

```
[1,] 0.7188145
[1,] 0.3965331
```

Como o p-valor é igual a 0,39, não rejeitamos a hipótese nula de que o regressor `kids` seja exógeno.

c') Vamos fazer o teste a partir do modelo de primeiro estágio:

$$kids = \gamma_0 + \gamma_1 samesex + \gamma_2 multi2nd + \gamma_3 nonmomi + \gamma_4 educ + \gamma_5 age + \gamma_6 agesq + \gamma_7 black + \gamma_8 hispan + \varepsilon$$

e testar a seguinte hipótese nula conjunta:

$$H_0: \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Uma das formas de fazer esse teste é por meio da estatística F:

$$F = \frac{R^2_{ur} - R^2_r}{(1 - R^2_{ur})} \frac{N - K - 1}{2},$$

que compara o modelo acima com o restrito:

$$kids = \gamma_0 + \tilde{\gamma}_3 nonmomi + \tilde{\gamma}_4 educ + \tilde{\gamma}_5 age + \tilde{\gamma}_6 agesq + \tilde{\gamma}_7 black + \tilde{\gamma}_8 hispan + \varepsilon$$

```
reg.1st = lm(kids ~ samesex + multi2nd +
                nonmomi + educ + age + agesq + black + hispan,
             data=labsup)
reg.1st.r = lm(kids ~ nonmomi + educ + age + agesq + black + hispan,
               data=labsup)
r2ur = summary(reg.1st)$r.squared
r2r = summary(reg.1st.r)$r.squared

F = (r2ur - r2r) / (1 - r2ur) * (N-K-1) / 2
1 - pf(F, 2, N-K-1)
```

```
[1] 0
```

Rejeitamos a hipótese nula de que $\gamma_1 = \gamma_2 = 0$, ou seja, não já indícios de que os instrumentos, de forma conjunta, sejam fracos.

d') $\hat{\boldsymbol{\beta}}^{MQ2E} = (\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{y}$, $V(\hat{\boldsymbol{\beta}}^{MQ2E}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{P_Z}\boldsymbol{X})^{-1}$ e $\hat{\sigma}^2 = (\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}})/(N-K-1)$

```
reg.2sls = ivreg(hours ~
            kids + nonmomi + educ + age + agesq + black + hispan |
            samesex + multi2nd + nonmomi + educ + age + agesq + black + hispan,
                 data=labsup)

stargazer::stargazer(reg.ols.rob, reg.iv, reg.2sls, type="text", digits=3)
```

```
=====================================================================
                                           Dependent variable:
                                    ---------------------------------
                                                  hours
                                    coefficient   instrumental
                                       test         variable
                                       (1)          (2)        (3)
---------------------------------------------------------------------
kids                                -2.326***     -4.879    -2.986**
                                     (0.116)     (3.013)    (1.333)

nonmomi                             -0.058***   -0.065***  -0.060***
                                     (0.005)     (0.010)    (0.007)

educ                                0.586***      0.368     0.530***
                                     (0.037)     (0.260)    (0.119)

age                                 2.049***    2.201***   2.088***
                                     (0.448)     (0.486)    (0.455)

agesq                               -0.028***   -0.028***  -0.028***
                                     (0.008)     (0.008)    (0.008)

black                                 1.058       1.095      1.068
                                     (1.351)     (1.361)    (1.351)

hispan                              -5.114***   -5.218***  -5.141***
                                     (1.351)     (1.368)    (1.354)

Constant                             -10.447     -5.254     -9.104
                                     (6.588)     (9.024)    (7.112)

---------------------------------------------------------------------
Observations                                     31,857     31,857
R2                                                0.058      0.072
Adjusted R2                                       0.058      0.072
Residual Std. Error (df = 31849)                 18.924     18.789
=====================================================================
Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

A estimativa de **kids** diminuiu, mas o erro padrão diminuiu consideravelmente. Agora, analiticamente:

```r
## Estimação analítica
Z = as.matrix(cbind(1, labsup[,c("samesex", "multi2nd", "nonmomi", "educ",
                                  "age", "agesq", "black", "hispan")]))
Pz = Z %*% solve(t(Z) %*% Z) %*% t(Z)

bhat.2sls = solve(t(X) %*% Pz %*% X) %*% t(X) %*% Pz %*% y
yhat = X %*% bhat.2sls
ehat = y - yhat

sig2hat = as.numeric( t(ehat) %*% ehat / (N-K-1) )
Vbhat = sig2hat * solve(t(X) %*% Pz %*% X)

se = sqrt(diag(Vbhat))
t = bhat.2sls / se
p = 2 * pt(-abs(t), N-K-1)

round(data.frame(bhat.2sls, se, t, p), 3)
```

```
         bhat.2sls    se      t     p
1           -9.104 7.112 -1.280 0.201
kids        -2.986 1.333 -2.241 0.025
```

```
nonmomi       −0.060  0.007  −9.127  0.000
educ           0.530  0.119   4.446  0.000
age            2.088  0.455   4.588  0.000
agesq         −0.028  0.008  −3.600  0.000
black          1.068  1.351   0.791  0.429
hispan        −5.141  1.354  −3.798  0.000
```

e') O teste de Hausman pode ser feito via estatística de Wald, que avalia a hipótese nula de que as estimativas MQO e MQ2E são estatisticamente iguais:

$$w = (\hat{\boldsymbol{\beta}}^{MQ2E} - \hat{\boldsymbol{\beta}}^{MQO})' \left[V(\hat{\boldsymbol{\beta}}^{MQ2E}) - V(\hat{\boldsymbol{\beta}}^{MQO}) \right]^{-1} (\hat{\boldsymbol{\beta}}^{MQ2E} - \hat{\boldsymbol{\beta}}^{MQO}) \sim \chi^2_J$$

em que $J = 1$ regressor endógeno neste caso.

```r
bhat.2sls = coef(reg.2sls)
Vbhat.2sls = vcov(reg.2sls)

contrast = bhat.2sls - bhat.ols
w = t(contrast) %*% solve(Vbhat.2sls - Vbhat.ols) %*% contrast
1 - pchisq(abs(w), 1)
```

```
[1,] 0.6189672
```

Com p-valor é igual a 0,62, tornou-se ainda mais fraca a evidência de que o regressor seja endógeno. Note que, mesmo mais eficientes em relação ao modelo exatamente identificado (IV), as estimativas MQ2E são bastante parecidas com MQO, e o teste de Hausman testa exatamente a significância estatística da diferença entre MQ2E e MQO (vetor de constrastes).

f) O teste de sobreidentificação pode ser feito a partir da estatística de Sargan:

$$SARG = NR^2_{resid} \sim \chi^2_{(L-J)}$$

```r
L = 2 # núm. instrumentos
J = 1 # núm. regressores endógenos

# Estimação nos resíduos
reg.resid = lm(resid(reg.2sls) ~ samesex + multi2nd + nonmomi + educ +
                  age + agesq + black + hispan, data=labsup)
round(coeftest(reg.r2resid), 3)
```

```
              Estimate Std. Error  t value  Pr(>|t|)
(Intercept)     0.061       6.575    0.009     0.993
samesex        −0.133       0.211   −0.632     0.527
multi2nd        0.358       1.136    0.315     0.753
nonmomi         0.000       0.005    0.004     0.997
educ            0.000       0.035    0.000     1.000
age             0.000       0.448    0.000     1.000
agesq           0.000       0.008    0.000     1.000
black           0.002       1.351    0.001     0.999
hispan          0.004       1.353    0.003     0.998
```

Note que, pelas estatísticas t da regressão nos resíduos, já podemos ver que, individualmente, ambos instrumentos não possuem correlação estatisticamente significante com o termo de erro. Basta verificar isto conjuntamente:

```r
# Estatística SARG
r2resid = summary(reg.resid)$r.squared
sarg = N * r2resid # sempre positivo
1 - pchisq(sarg, df=L-J) # p-valor
```

```
[1] 0.4797525
```

Não rejeitamos os hipótese nula de que os instrumentos são exógenos em relação ao termo de erro.

□