02-exercices-regressions-lineraires

Exercice 1 : Notes et présence en cours

Données : Dataframe `attend` *du package* `wooldridge`*.*

Objectif : Estimer l'effet de la présence en cours sur la réussite à un examen.

Question 0

Importer le setup de l'analyse

Question 1

Importer les données et découvrir les variables. Quelles sont les potentielles variables exogènes pour répondre à notre objectif ? Quelle serait la principale variable endogène ?

Question 2

Quel est l'effet attendu de la présence scolaire sur le résultat à l'examen ?

Question 3

Représenter graphiquement la relation entre la réussite et la présence en cours. Quelle est la corrélation entre les deux variables ?

Question 4

Estimer la régression linéaire simple entre la réussite et à la présence en classe. Quel est l'effet estimé de la présence sur la réussite ? Cet effet est-il significatif ? Qu'en est-il de la constante ?

Question 5

Les hypothèses de l'estimation par MCO sont-elles respectées ? Que faut-il en conclure pour notre estimation ?

Question 6

Pourquoi pourrait-il être intéressant d'ajouter l'assiduité dans la remise des devoirs à la régression ? Et pour les anciennes notes à l'université ? Et pour les résultats du test d'entrée à l'université (ACT) ? Quelles sont les relations à attendre avec la réussite à l'examen ?

Question 7

Estimer cette régression linéaire, interpréter les coefficients et tester les hypothèses des estimateurs MCO. Que peut-on en conclure ?

Question 8

L'impact de la présence est-il le même pour un freshman que pour un autre élève ? Le modèle au global est-il le même ?

Exercice2 : Estimation de la valeur d'une maison

Données : `hprice2` du package `wooldridge`.

Vous travailler pour une agence immobilière qui vous demande d'analyser ses données afin d'expliquer rationnellement les prix de vente des maisons actuellement à la vente. Votre agence souhaite savoir si d'une part la criminalité d'un quartier joue sur le prix du bien immobilier. elle souhaite d'autre part avoir le modèle qui explique le plus possible le prix des maisons (prévoyez des preuves que ce modèle est meilleur que les autres) afin que vous puissiez trouver les maisons qui sont les plus sous-évaluées et qui pourraient ainsi être achetées puis revendues plus cher par l'agence.

Exercices 3 : Comprendre le code donné

Vous travaillez dans un laboratoire de recherche. Un de vos collègue vous demande de l'aider à comprendre et interpréter le travail d'un stagiaire qui n'a pas laissé d'indications sur ce qu'il a fait. Lisez le code suivant, commenter ce qui y est fait et interprétez les résultats. Quelle politique publique pourriez-vous recommander sur la base de ces résultats ?

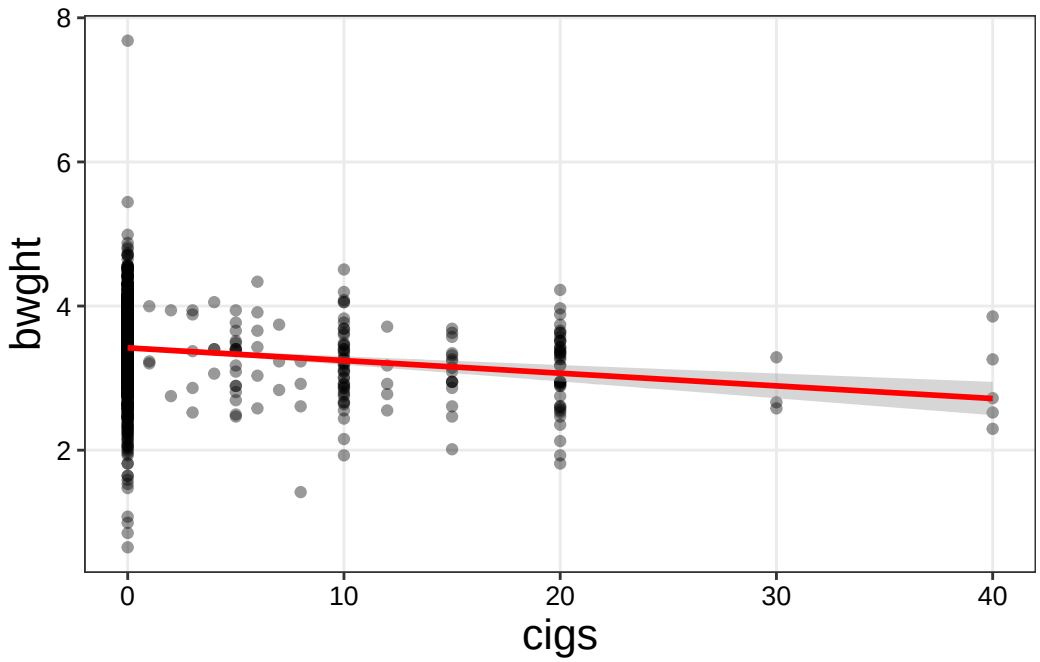
```
source(here::here("02-codes", "utils", "setup.R"))

data("bwght")

df <-
  bwght |>
  tibble() |>
  select(bwght, faminc, fatheduc, motheduc, cigs, male, white) |>
  mutate(bwght = bwght * 0.0283495) |>
  drop_na()
```

```
ggplot(df, aes(x = cigs, y = bwght)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", color = "red")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```r
model_1 <- lm(bwght ~ cigs + faminc + fatheduc + motheduc + male + white,
data = df)
summary(model_1)
```

```
Call:
lm(formula = bwght ~ cigs + faminc + fatheduc + motheduc + male +
    white, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6953 -0.3241  0.0122  0.3698  4.2869

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  3.203637   0.105072  30.490 < 0.0000000000000002 ***
cigs        -0.016751   0.003120  -5.369          0.0000000953 ***
faminc       0.001192   0.001051   1.134               0.25710
fatheduc     0.012363   0.007988   1.548               0.12198
motheduc    -0.011403   0.009011  -1.265               0.20597
male         0.105360   0.032488   3.243               0.00122 **
white        0.128117   0.045684   2.804               0.00512 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5587 on 1184 degrees of freedom
Multiple R-squared:  0.0473,    Adjusted R-squared:  0.04247
F-statistic: 9.798 on 6 and 1184 DF,  p-value: 0.0000000001468
```

```r
shapiro.test(residuals(model_1))
```

```
    Shapiro-Wilk normality test

data:  residuals(model_1)
W = 0.96589, p-value = 0.000000000000000414
```

```r
bptest(model_1)
```

```
    studentized Breusch-Pagan test
```

```
data:  model_1
BP = 1.8036, df = 6, p-value = 0.9368
```

```
check_collinearity(model_1)
```

```
# Check for Multicollinearity

Low Correlation

     Term  VIF     VIF 95% CI adj. VIF Tolerance Tolerance 95% CI
     cigs 1.06 [1.02,   1.17]     1.03      0.94     [0.85, 0.98]
   faminc 1.36 [1.27,   1.47]     1.17      0.74     [0.68, 0.78]
 fatheduc 1.83 [1.69,   1.99]     1.35      0.55     [0.50, 0.59]
 motheduc 1.81 [1.67,   1.97]     1.34      0.55     [0.51, 0.60]
     male 1.01 [1.00, 273.80]     1.00      0.99     [0.00, 1.00]
    white 1.05 [1.01,   1.17]     1.02      0.95     [0.85, 0.99]
```

```
coeftest(model_1, vcov = hccm(model_1, type = "hc0"))
```

```
t test of coefficients:

              Estimate Std. Error t value              Pr(>|t|)
(Intercept)  3.2036371  0.0976107 32.8205 < 0.00000000000000022 ***
cigs        -0.0167509  0.0029566 -5.6656         0.00000001838 ***
faminc       0.0011922  0.0010189  1.1700               0.242232
fatheduc     0.0123627  0.0074523  1.6589               0.097397 .
motheduc    -0.0114030  0.0083059 -1.3729               0.170053
male         0.1053600  0.0324210  3.2497               0.001188 **
white        0.1281166  0.0475235  2.6959               0.007120 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(
  model_1,
  c("fatheduc = 0", "motheduc = 0"),
  vcov = hccm(model_1, type = "hc0")
)
```

```
Linear hypothesis test:
fatheduc = 0
```

```
motheduc = 0

Model 1: restricted model
Model 2: bwght ~ cigs + faminc + fatheduc + motheduc + male + white

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1   1186
2   1184  2 1.5092 0.2215
```

```
linearHypothesis(
  model_1,
  c("fatheduc = 0", "motheduc = 0", "faminc = 0"),
  vcov = hccm(model_1, type = "hc0")
)
```

```
Linear hypothesis test:
fatheduc = 0
motheduc = 0
faminc = 0

Model 1: restricted model
Model 2: bwght ~ cigs + faminc + fatheduc + motheduc + male + white

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1   1187
2   1184  3 1.8516  0.136
```

```
df |>
  select(faminc, fatheduc, motheduc) |>
  cor()
```

```
            faminc  fatheduc  motheduc
faminc   1.0000000 0.4476816 0.4270863
fatheduc 0.4476816 1.0000000 0.6434825
motheduc 0.4270863 0.6434825 1.0000000
```

```r
model_2 <- lm(bwght ~ cigs + log(faminc) + male + white, data = df)
coeftest(model_2, vcov = hccm(model_2, type = "hc0"))
```

```
t test of coefficients:

              Estimate Std. Error t value              Pr(>|t|)
(Intercept)  3.1322935  0.0789844 39.6571 < 0.00000000000000022 ***
cigs        -0.0166752  0.0029214 -5.7080         0.00000001443 ***
log(faminc)  0.0367777  0.0225801  1.6288               0.103628
male         0.1071140  0.0324289  3.3030               0.000985 ***
white        0.1300319  0.0477191  2.7249               0.006526 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model_3 <- lm(bwght ~ cigs + motheduc + male + white, data = df)
coeftest(model_3, vcov = hccm(model_3, type = "hc0"))
```

```
t test of coefficients:

               Estimate  Std. Error t value              Pr(>|t|)
(Intercept)  3.23010849  0.09367819 34.4809 < 0.00000000000000022 ***
cigs        -0.01732508  0.00295393 -5.8651         0.000000005815 ***
motheduc     0.00095794  0.00634372  0.1510               0.879996
male         0.10394789  0.03251299  3.1971               0.001425 **
white        0.14533853  0.04593889  3.1637               0.001597 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model_4 <- lm(bwght ~ cigs  + male + white, data = df)
summary(model_4)
```

```
Call:
lm(formula = bwght ~ cigs + male + white, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6944 -0.3265  0.0137  0.3633  4.2944

Coefficients:
```

```
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  3.242500   0.045087  71.916 < 0.0000000000000002 ***
cigs        -0.017419   0.003035  -5.740          0.000000012 ***
male         0.103921   0.032458   3.202              0.00140 **
white        0.145767   0.044656   3.264              0.00113 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5592 on 1187 degrees of freedom
Multiple R-squared:  0.04345,   Adjusted R-squared:  0.04103
F-statistic: 17.97 on 3 and 1187 DF,  p-value: 0.00000000002072
```

```r
model_5 <- lm(bwght ~ cigs*white  + male + white, data = df)
summary(model_5)
```

```
Call:
lm(formula = bwght ~ cigs * white + male + white, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6869 -0.3282  0.0120  0.3616  4.2927

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)  3.234969   0.046361  69.778 <0.0000000000000002 ***
cigs        -0.012728   0.007353  -1.731              0.0837 .
white        0.154996   0.046568   3.328              0.0009 ***
male         0.103993   0.032465   3.203              0.0014 **
cigs:white  -0.005654   0.008072  -0.701              0.4837
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5593 on 1186 degrees of freedom
Multiple R-squared:  0.04384,   Adjusted R-squared:  0.04062
F-statistic:  13.6 on 4 and 1186 DF,  p-value: 0.00000000007668
```

```r
linearHypothesis(
  model_5,
  c("cigs = 0", "cigs:white = 0"),
  vcov = hccm(model_5, type = "hc0")
)
```

```
Linear hypothesis test:
cigs = 0
cigs:white = 0

Model 1: restricted model
Model 2: bwght ~ cigs * white + male + white

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F        Pr(>F)
1   1188
2   1186  2 18.452 0.00000001284 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(model_5, vcov = hccm(model_5, type = "hc0"))
```

```
t test of coefficients:

              Estimate Std. Error t value              Pr(>|t|)
(Intercept)  3.2349694  0.0481758 67.1493 < 0.00000000000000022 ***
cigs        -0.0127280  0.0063233 -2.0129              0.044353 *
white        0.1549956  0.0481197  3.2210              0.001312 **
male         0.1039935  0.0325190  3.1979              0.001421 **
cigs:white  -0.0056544  0.0070850 -0.7981              0.424990
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
list_exported_models <-
  ls(pattern = "^model_*") |>
  mget() |>
  set_names(ls(pattern = "^model_*"))
```

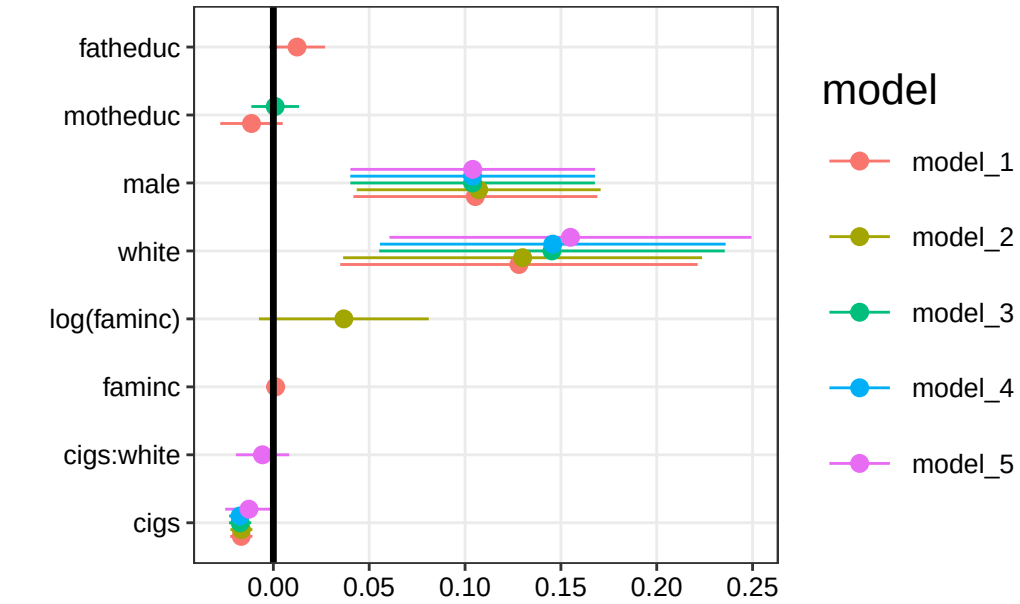
```
modelsummary::modelsummary(
  list_exported_models,
  stars = TRUE,
  gof_map = c("nobs", "adj.r.squared", "aic")
  )
```

	model_1	model_2	model_3	model_4	model_5
(Intercept)	3.204***	3.132***	3.230***	3.243***	3.235***
	(0.105)	(0.084)	(0.100)	(0.045)	(0.046)
cigs	−0.017***	−0.017***	−0.017***	−0.017***	−0.013+
	(0.003)	(0.003)	(0.003)	(0.003)	(0.007)
faminc	0.001				
	(0.001)				
fatheduc	0.012				
	(0.008)				
motheduc	−0.011		0.001		
	(0.009)		(0.007)		
male	0.105**	0.107**	0.104**	0.104**	0.104**
	(0.032)	(0.033)	(0.032)	(0.032)	(0.032)
white	0.128**	0.130**	0.145**	0.146**	0.155***
	(0.046)	(0.046)	(0.045)	(0.045)	(0.047)
log(faminc)		0.037			
		(0.024)			
cigs × white					−0.006
					(0.008)
Num.Obs.	1191	1191	1191	1191	1191
R2 Adj.	0.042	0.042	0.040	0.041	0.041
AIC	2002.4	2000.7	2003.2	2001.2	2002.7

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

```r
cm <- c("cigs", "cigs:white", "faminc", "log(faminc)", "white", "male",
"motheduc",
        "fatheduc")

modelplot(
  list_exported_models,
  coef_omit = 'Interc',
  coef_map = cm,
  vcov = \(x) hccm(x, type = "hc0")
  ) +
geom_vline(xintercept = 0, color = "black", linewidth = 1.2)
```



Coefficient estimates and 95% confidence intervals