# 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 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 ?*

```
# Importer les packages
source(here::here("02-codes", "utils", "setup.R"))

# Importer les données attend et les stocker dans la avriable df
df <-
  wooldridge::attend |>
  # Transformer en tibble pour plus de confort
  tibble() |>
  print()
```

```
# A tibble: 680 x 11
   attend termGPA priGPA   ACT final atndrte hwrte frosh  soph missed stndfnl
    <int>   <dbl>  <dbl> <int> <int>   <dbl> <dbl> <int> <int>  <int>   <dbl>
 1     27    3.19   2.64    23    28    84.4 100       0     1      5  0.473
 2     22    2.73   3.52    25    26    68.8  87.5     0     0     10  0.0525
 3     30    3      2.46    24    30    93.8  87.5     0     0      2  0.893
 4     31    2.04   2.61    20    27    96.9 100       0     1      1  0.263
 5     32    3.68   3.32    23    34   100   100       0     1      0  1.73
 6     29    3.23   2.93    26    25    90.6 100       0     1      3 -0.158
 7     30    1.54   1.94    21    10    93.8  75       1     0      2 -3.31
```

```
 8     26    2      2.12    22    34    81.2 100       0     1      6  1.73
 9     24    2.25   2.06    24    26    75   100       1     0      8  0.0525
10     29    3      2.73    21    26    90.6 100       0     1      3  0.0525
# i 670 more rows
```

```
# help(attend) OU ?attend pour avoir une description des variables
```

Il y a plusieurs variables qui mesurent la réussite à l'examen :

- `final`
- `stndfnl`

`final` donne simplement le score à l'examen final, tandis que `stndfnl` correspond à la variable final centrée réduite. Elle permet de comparer plus facilement les élèves entre eux et d'avoir une comparaison qui indique à quel point un élève à bien performé comparé aux autres élèves. On pourrait également utiliser la variable `termGPA` qui est une moyenne de différentes notes si l'on souhaite mesurer la performance continue. Ici, on va utiliser la variable `stndfnl` pour pouvoir estimer si être assidu rend plus compétitif ou non.

Les potentielles variables exogènes pour mesurer l'assiduité seraient :

- `attend` : le nombre de classes sur 32 auxquelles l'élève à participé
- `atndrte` : le pourcentage de classes auxquelles l'élève à participé
- `missed` : le nombre de classes que l'élève à manqué

Ces 3 variables donnent des informations très similaire. Seule l'interprétation va changer. Ici on va choisir la variable `attend` car il semble plus intéressant de voir si les notes augmentent lorsque l'on assiste à un cours de plus plutôt que si l'on assiste à 1% de cours en plus (ça veut dire quoi assister à 1% de cours en plus ?).

## Question 2

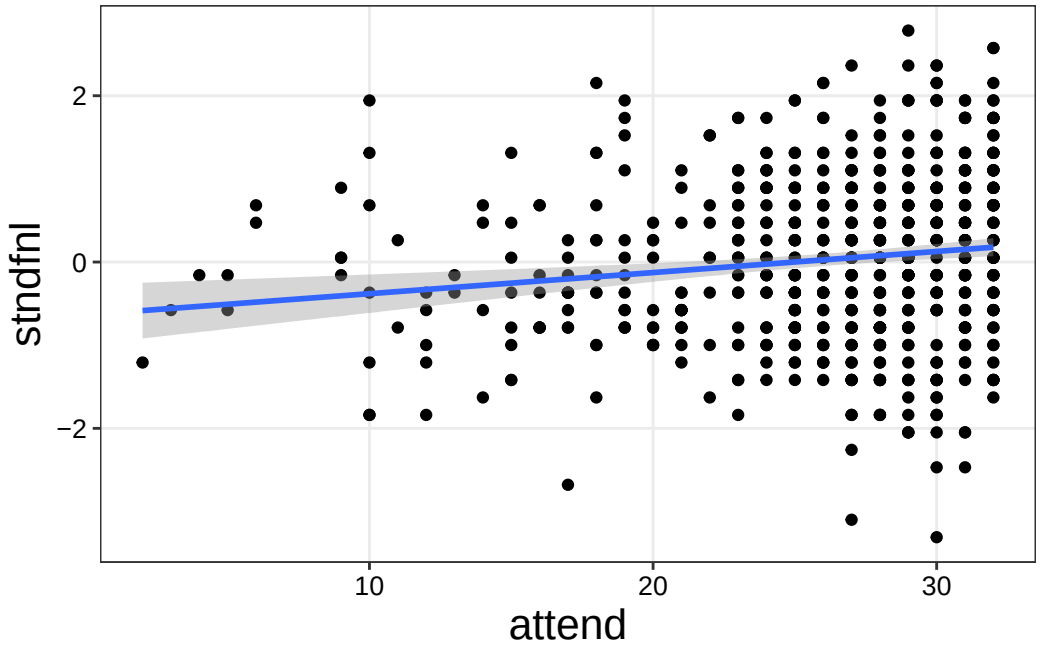
*Quel est l'effet attendu de la présence scolaire sur le résultat à l'examen ?*

On s'attend qu'un élève qui soit davantage présent en cours ait de meilleures notes et soit plus compétitif comparé à un élève qui viendrait moins souvent. En effet, la présence enc ours est un indicateur de la motivation et de sa volonté de travailler. Egalement, il y a souvent des informations importantes données pendant le cours que l'élève n'aura pas ailleurs.

### Question 3

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

```
# utiliser notre jeu de données pour créer un graphique
df |>
  # Créer un graphique avec attend en x et stndfnl en y
  ggplot(aes(x = attend, y = stndfnl)) +
  # Ajouter des points
  geom_point() +
  # Ajouter une droite de régression linéaire
  geom_smooth(
    method = "lm", # On utilise une régression linéaire
    formula = y ~ x, # On modélise y par rapport à x
    se = TRUE # On indique que l'on veut visualiser l'écart-type de beta
  )
```




```
# Calculer la corrélation entre les deux variables
cor.test(df$attend, df$stndfnl)
```

```
    Pearson's product-moment correlation

data:  df$attend and df$stndfnl
t = 3.6825, df = 678, p-value = 0.0002493
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.0655373 0.2129758
sample estimates:
      cor
0.1400327
```

On peut voir qu'on a une corrélation positive bien quelle semble assez faible. Mais il semblerait bien qu'il y ait une corrélation positive entre la présence en cours en la compétitivité à l'examen d'un étudiant. Attention, il ne s'agit que d'une corrélation.

## 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 ?*

```r
# Estimer la régression linéaire simple
reg_simple <- lm(stndfnl ~ attend, data = df)

summary(reg_simple)
```

```
Call:
lm(formula = stndfnl ~ attend, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4363 -0.6799 -0.0242  0.6632  2.6815

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.634471   0.184224  -3.444 0.000608 ***
attend       0.025400   0.006897   3.683 0.000249 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9804 on 678 degrees of freedom
```

```
Multiple R-squared:  0.01961,   Adjusted R-squared:  0.01816
F-statistic: 13.56 on 1 and 678 DF,  p-value: 0.0002493
```

Assister à une classe de plus fait augmenter la note standardisée de 0.02. L'effet est significatif au seuil de 0.1%. On rejette donc l'hypothèse nulle d'absence d'effet de la présence en cours sur la compétitivité à l'examen.

L'effet semble faible, néanmoins il faut bien prendre en compte qu'il s'agit d'une note standardisée. Si l'on regarde la distribution de cette variable :

```r
# Obtenir un résummé statistique de la variable stndfnl
quantile(df$stndfnl, seq(0,1,0.1)) |>
  as_tibble(rownames = "quantile") |>
  print(n = 20)
```

```
# A tibble: 11 x 2
   quantile   value
   <chr>      <dbl>
 1 0%       -3.31
 2 10%      -1.21
 3 20%      -0.788
 4 30%      -0.578
 5 40%      -0.158
 6 50%       0.0525
 7 60%       0.263
 8 70%       0.473
 9 80%       0.893
10 90%       1.31
11 100%      2.78
```

On peut voir que sa distribution est assez resserrée et que 0.02 peut constituer une augmentation intéressante de la compétitivité. Si jamais l'élève assiste assiste à tous les cours, il aura une note $0.02 \times 32 = 0.64$ plus élevée qu'un élève qui n'aura assisté à aucun cours. Une augmentation standardisée de la note de 0.64 représente une augmentation de plusieurs décile dans la performance à l'examen qui ne semble pas négligeable.

On remarque que la constante est significative et négative. Compte tenu du fait que notre variable est standardisée cela est plausible. Un élève qui n'aura assisté à aucun cours aura en moyenne une note standardisée de -0.64 ce qui le place dans les 30% d'élèves les moins performants à l'examen selon la distribution des valeurs réelles. Mais de manière étonnante, on trouve des valeurs bien plus faibles. (Un indice qu'il manque des variables explicatives ?)

## Question 5

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

On va tester la sphéricité des erreurs (homoscédasticité) avec le test de white et de breush-Pagan

```
# Breusch-Pagan test
lmtest::bptest(reg_simple)
```

```
    studentized Breusch-Pagan test

data:  reg_simple
BP = 5.5482, df = 1, p-value = 0.0185
```

```
# White test
whitestrap::white_test(reg_simple)
```

```
White's test results

Null hypothesis: Homoskedasticity of the residuals
Alternative hypothesis: Heteroskedasticity of the residuals
Test Statistic: 7.88
P-value: 0.01949
```

On teste l'hypothèse d'homoscédasticité des résidus -> leur variance doit être constante. On remarque que la p.value pour les deux tests est inférieure à 5%, on rejette donc l'hypothèse nulle d'homoscédasticité au seuil de 5%. Nos résidus semble avoir une variance qui n'est pas constante. Dans ce cas, les estimateurs ne sont plus de variance minimale. On peut corriger ce problème en intégrant une correction de White aux écarts-types.

On teste maintenant la normalité des résidus

```
# test de Jarque-Berra
moments::jarque.test(reg_simple$residuals)
```

```
    Jarque-Bera Normality Test

data:  reg_simple$residuals
JB = 0.081637, p-value = 0.96
alternative hypothesis: greater
```

```r
# test de Shapiro-Wilk
stats::shapiro.test(reg_simple$residuals)
```

```
    Shapiro-Wilk normality test

data:  reg_simple$residuals
W = 0.99764, p-value = 0.4556
```

Le test de Jarque-Berra teste l'hypothèse nulle que la skewness de la série est égale à 0 et que la kurtosis de la série est égale à 3 de manière conjointe. La p.value est supérieure à 0.05. On ne peut pas rejeter l'hypothèse nulle. Il semble que nos résidus aient des moments similaires à ceux d'une loi normale.

Le test de Shapiro-Wilk teste l'hypothèse que la série provient d'un processus générateur d'une loi normale. On ne peut pas rejeter l'hypothèse nulle, il semble que nos résidus proviennent d'une loi normale.

Nos résidus semblant normalement distribués, les tests statistiques restent valident.

Pour l'hypothèse d'exogénéité, il est très probable qu'il existe des variables que l'on n'a pas prise en compte et qui affectent à la fois la présence en cours et la compétitivité à l'examen. On peut penser à :

- Le milieu social : un élève d'un milieu aisé aura en général de meilleures notes et sera plus incité à venir en cours de part la pression sociale
- Le "talent" un élève plus talentueux aura tendance à avoir de meilleur notes. L'effet sur la présence est plus incertain.
- Si un élève à du obtenir un crédit pour aller à la fac, il est probable qu'il va plus venir en cours et que cela peut avoir un effet sur les notes

On peut penser à un grand nombre de variables. Si on ne les prend pas en compte, le coefficient associé à la présence devient biaisé et on ne peut pas parler d'effet causal.

## 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 ?*

L'assiduité dans la remise des devoirs `hwrte` est également une mesure d'assiduité et permet d'introduire un contrôle sur la volonté de travailler de l'élève en dehors des cours ce qui peut influencer sa note.

Les anciennes notes à l'université `priGPA` permettent de contrôler pour la qualité étudiante de l'élève et sa capacité à réussir à la fac. Si on ne contrôle pas pour cette variable, on va comparer des élèves très fort dans leur domaine et des élèves moins fort ce qui risque de biaiser nos résultats.

Il en va de même pour les tests d'entrées à l'université `ACT`

On s'attend pour toutes ces variables à ce que la relation avec la compétitivité à l'examen soit positive.

## Question 7

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

```r
# Estimer la régression multiple
reg_mul <- lm(stndfnl ~ attend + hwrte + priGPA + ACT, data = df)

summary(reg_mul)
```

```
Call:
lm(formula = stndfnl ~ attend + hwrte + priGPA + ACT, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.98362 -0.56827 -0.03468  0.60324  2.32557

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -3.471448   0.308536 -11.251 < 0.0000000000000002 ***
attend       0.009107   0.009046   1.007               0.3144
hwrte        0.003855   0.002282   1.689               0.0917 .
priGPA       0.400493   0.078786   5.083    0.000000482072468 ***
ACT          0.083834   0.011246   7.454    0.000000000000281 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8872 on 669 degrees of freedom
  (6 observations effacées parce que manquantes)
Multiple R-squared:  0.206, Adjusted R-squared:  0.2012
F-statistic: 43.39 on 4 and 669 DF,  p-value: < 0.00000000000000022
```

Le coefficient de la présence en cours en toujours positif mais n'est pas statistiquement significatif. En ajoutant les variables de contrôle, il semblerait que l'effet de la présence en cours sur la compétitivité soit nul. L'assiduité dans la remise des devoirs et positive bien que faible et seulement significative au seuil de 10%. Il semblerait qu'un élève assidu en cours et dans son travail à la maison n'augmente pas sa compétitivité comparé à des élèves moins assidus ayant les mêmes capacités scolaires qu'eux.

En revanche les anciennes notes à l'université ainsi que les notes à l'examen d'entrée sont positives et fortement significatives. Notre modèle explique 20% de la variation des notes et il semblerait que cette variation soit majoritairement expliquée par la capacité scolaire de l'élève plutôt que son assiduité.

```
# Breusch-Pagan test
lmtest::bptest(reg_simple)
```

```
    studentized Breusch-Pagan test

data:  reg_simple
BP = 5.5482, df = 1, p-value = 0.0185
```

```
# White test
whitestrap::white_test(reg_simple)
```

```
White's test results

Null hypothesis: Homoskedasticity of the residuals
Alternative hypothesis: Heteroskedasticity of the residuals
Test Statistic: 7.88
P-value: 0.01949
```

Nos résidus sont toujours hétéroscédastiques. O va donc procéder à une correction de White des écarts-types :

```
reg_mul |>
  lmtest::coeftest(vcov. = car::hccm)
```

```
t test of coefficients:

              Estimate Std. Error  t value              Pr(>|t|)
(Intercept) -3.4714478  0.3085529 -11.2507 < 0.00000000000000022 ***
attend       0.0091075  0.0093380   0.9753                0.3298
```

```
hwrte        0.0038545  0.0026037   1.4804                0.1392
priGPA       0.4004931  0.0851509   4.7033    0.0000031103052657 ***
ACT          0.0838344  0.0112883   7.4267    0.0000000000003404 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lorsque l'on corrige nos écarts-types, l'assiduité dans la remise des devoirs perd sa significativité. Il semblerait bien que l'assiduité ne rende pas compétitif. Mais attention, il y a encore probablement des valeurs manquantes (milieu social par exemple, filière, etc…) qui peuvent biaiser nos coefficients et nous empêcher de parler d'effet causal.

```
# test de Jarque-Berra
moments::jarque.test(reg_mul$residuals)
```

```
    Jarque-Bera Normality Test

data:  reg_mul$residuals
JB = 0.44614, p-value = 0.8001
alternative hypothesis: greater
```

```
# test de Shapiro-Wilk
stats::shapiro.test(reg_mul$residuals)
```

```
    Shapiro-Wilk normality test

data:  reg_mul$residuals
W = 0.99724, p-value = 0.3174
```

Nos résidus semblent toujours normaux, nos tests statistiques restent valides.

```
# Regarder la multicolinéarité
performance::check_collinearity(reg_mul)
```

```
# Check for Multicollinearity

Low Correlation

   Term  VIF   VIF 95% CI adj. VIF Tolerance Tolerance 95% CI
 attend 1.95 [1.76, 2.19]     1.40      0.51     [0.46, 0.57]
  hwrte 1.65 [1.50, 1.85]     1.29      0.60     [0.54, 0.67]
 priGPA 1.57 [1.43, 1.75]     1.25      0.64     [0.57, 0.70]
    ACT 1.31 [1.21, 1.46]     1.15      0.76     [0.69, 0.83]
```

Il ne semble pas que les écarts-types de nos estimateurs soient fortement inflatés par la présence d'autres variables. Mais on va quand même tester si conjointement nos variables d'assiduité ont un effet.

```
# Test de Fisher sur les variables d'assiduité
car::linearHypothesis(
  reg_mul,
  c("attend = 0", "hwrte = 0")
)
```

```
Linear hypothesis test:
attend = 0
hwrte = 0

Model 1: restricted model
Model 2: stndfnl ~ attend + hwrte + priGPA + ACT

  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    671 532.92
2    669 526.53  2    6.3864 4.0572 0.01772 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il semble que prises conjointement nos variables d'assiduité aient un effet significatif sur la compétitivité à l'examen. En revanche le gain estimé d'aller à un cours en plus ou de rendre 1% de devoirs en plus est assez faible : 0.009 et 0.003. Ainsi un élève qui va à tous les cours et rend tous les devoirs aura une note standardisée supérieure de $0.009 \times 32 + 0.003 \times 100 = 0.588$ par rapport à un élève qui ne va jamais en cours et ne rend aucun devoir. Cela n'est pas négligeable mais loin de l'effet attendu.

## 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 ?

Pour tester si l'impact est le même pour un fresman, on va inclure la variable binaire `frosh` qui prend la valeur 1 si l'élève est un freshman et 0 sinon. Puis on va inclure un terme d'interaction avec l'ensemble des variables et faire un test de Fisher pour voir si la relation est différente ou pas.

```r
# Regresion avec simplement la binaire : changement d'ordonnée à l'origine
lm(stndfnl ~ attend + hwrte + priGPA + ACT + frosh, data = df) |>
  lmtest::coeftest(vcov. = car::hccm)
```

```
t test of coefficients:

              Estimate Std. Error  t value              Pr(>|t|)
(Intercept) -3.5234019  0.3144277 -11.2058 < 0.00000000000000022 ***
attend       0.0079210  0.0095046   0.8334                0.4049
hwrte        0.0040260  0.0026251   1.5336                0.1256
priGPA       0.4197687  0.0876529   4.7890    0.0000020660348226 ***
ACT          0.0839211  0.0113030   7.4247    0.0000000000003458 ***
frosh        0.0702423  0.0859891   0.8169                0.4143
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Il ne semble pas qu'être un freshman ait un impact sur la compétitivité à l'examen.

```r
# Regresion avec termes d'interaction
reg_inter <- lm(stndfnl ~ attend * frosh + hwrte* frosh + priGPA* frosh +
ACT* frosh, data = df)

reg_inter |>
  lmtest::coeftest(vcov. = car::hccm)
```

```
t test of coefficients:

               Estimate Std. Error  t value              Pr(>|t|)
(Intercept)  -3.8144649  0.3632413 -10.5012 < 0.00000000000000022 ***
attend        0.0121526  0.0109508   1.1097                0.26751
frosh         1.7321487  0.7352809   2.3558                0.01877 *
hwrte         0.0031417  0.0034408   0.9131                0.36153
priGPA        0.4595858  0.1034610   4.4421     0.00001043399219 ***
ACT           0.0905736  0.0133724   6.7732     0.00000000002778 ***
attend:frosh -0.0198698  0.0241714  -0.8220                0.41135
frosh:hwrte   0.0033445  0.0055420   0.6035                0.54640
frosh:priGPA -0.2609737  0.2176804  -1.1989                0.23100
frosh:ACT    -0.0375410  0.0255554  -1.4690                0.14230
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# test de Fisher
car::linearHypothesis(
  reg_inter,
  c("attend:frosh = 0", "frosh:hwrte = 0", "frosh:priGPA = 0", "frosh:ACT =
  0", "frosh = 0"),
  vcov. = car::hccm
)
```

```
Linear hypothesis test:
attend:frosh = 0
frosh:hwrte = 0
frosh:priGPA = 0
frosh:ACT = 0
frosh = 0

Model 1: restricted model
Model 2: stndfnl ~ attend * frosh + hwrte * frosh + priGPA * frosh + ACT *
    frosh

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1    669
2    664  5 1.3664  0.235
```

La p.value du test contraint sur l'ensemble des termes d'interaction est supérieure à 0.05. On ne peut donc pas rejeter l'hypothèse de non significativité conjointe. Il ne semble pas que la relation soit entièrement différente entre un freshman ou un autre élève. En revanche dans ce modèle un freshman va être plus compétitif qu'un autre élève, ce qui peut s'expliquer par une motivation accrue par exemple.

```r
# Regresion avec termes d'interaction que pour la présence en classe
reg_inter2 <- lm(stndfnl ~ attend * frosh + hwrte + priGPA + ACT, data = df)

reg_inter2 |>
  lmtest::coeftest(vcov. = car::hccm)
```

```
t test of coefficients:

               Estimate Std. Error  t value             Pr(>|t|)
```

```
(Intercept)  -3.6089434  0.3289520 -10.9710 < 0.00000000000000022 ***
attend        0.0110994  0.0098586   1.1259                0.2606
frosh         0.5347897  0.4251179   1.2580                0.2088
hwrte         0.0041471  0.0026584   1.5600                0.1192
priGPA        0.4202478  0.0876262   4.7959    0.0000019987867251 ***
ACT           0.0834788  0.0113128   7.3792    0.0000000000004753 ***
attend:frosh -0.0177016  0.0158197  -1.1190                0.2636
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# test de Fisher
car::linearHypothesis(
  reg_inter,
  c("attend:frosh = 0", "attend = 0"),
  vcov. = car::hccm
)
```

```
Linear hypothesis test:
attend:frosh = 0
attend = 0

Model 1: restricted model
Model 2: stndfnl ~ attend * frosh + hwrte * frosh + priGPA * frosh + ACT *
    frosh

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F Pr(>F)
1    666
2    664  2 0.6799  0.507
```

Il ne semble pas que la présence en cours ait un effet significatif sur la compétitivité à l'examen, freshman ou non.

## 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 ?

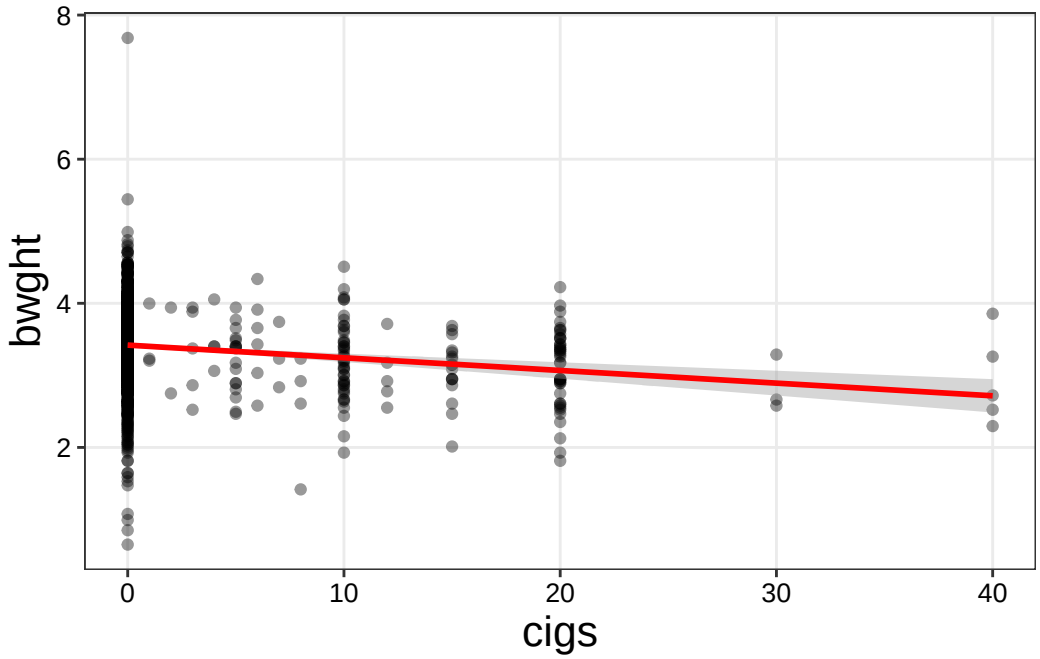
```
source(here::here("02-codes", "utils", "setup.R"))

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'
```




```
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
```

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

```
    Shapiro-Wilk normality test

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

```
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
```

```r
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
```

```r
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
```

```r
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
```

```r
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 < 0.0000000000000022 ***
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_*"))
```

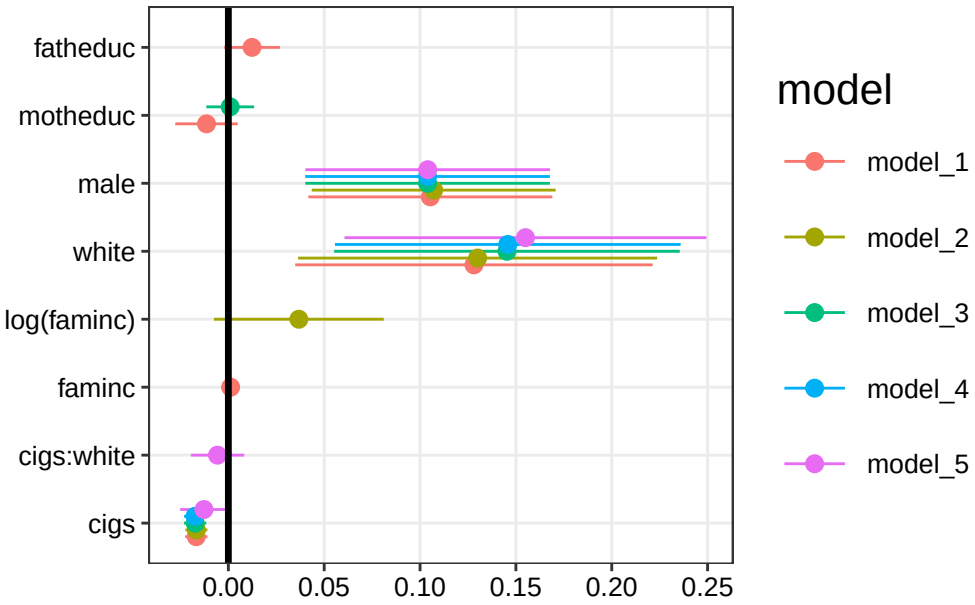
```
modelsummary::modelsummary(
  list_exported_models,
  stars = TRUE,
  gof_map = c("nobs", "adj.r.squared", "aic")
  )
```

```
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)
```

| | 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




Coefficient estimates and 95% confidence intervals