L'ajustement déplace l'estimation ; les grappes déplacent l'erreur-type

Ajouter toutes les covariables individuelles du registre fait passer le coefficient de cantine de 4,47 à 4,37 et laisse l'erreur-type à 0,9. Ajouter un terme pour l'école laisse le coefficient tranquille et porte l'erreur-type à 1,5. Ce sont deux problèmes distincts et on les confond couramment.

Pierre Robentz Cassion

Leçon 2 sur 8

Régression appliquée aux données de programme — Un modèle est une comparaison

Ce que couvre cette leçon

- Le tableau autour duquel ce cours est bâti
- Ce que « ajuster sur » fait réellement
- Pourquoi les covariables n'ont presque rien déplacé
- Lire un tableau de coefficients sans se faire tromper
- Quand l'ajustement ne peut pas aider
- Rapportez-le en entier
- La suite

Le tableau autour duquel ce cours est bâti — En Python (suite)

```
import pandas as pd
import statsmodels.formula.api as smf

attendance = pd.read_csv("school-attendance-2024.v1.csv")
roster = pd.read_csv("school-roster-2024.v1.csv").drop_duplicates(
    "student_id", keep="last")
MARKS = {"true": True, "Y": True, "false": False, "N": False}
marked = attendance[attendance["present"].isin(MARKS)].copy()
marked["attended"] = marked["present"].map(MARKS)
per_student = (marked.groupby("student_id")["attended"].mean()
               .rename("rate").reset_index().merge(roster, on="student_id"))

enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
current = enrolment[enrolment["school_year"] == 2024]

d = per_student.merge(
```

Le tableau autour duquel ce cours est bâti — En Python (suite)

```
    current[["student_id", "sex", "age_years", "disability_reported",
             "displacement_status", "admin2"]], on="student_id").dropna()
print(f"complete cases: {len(d)}")

m1 = smf.ols("rate ~ feeding_programme", data=d).fit()
m2 = smf.ols("rate ~ feeding_programme + sex + age_years"
             " + disability_reported + displacement_status", data=d).fit()
m3 = smf.ols("rate ~ feeding_programme + sex + age_years"
             " + disability_reported + displacement_status + admin2", data=d).fit()
m4 = m1.get_robustcov_results(cov_type="cluster", groups=d["school_id"])

for name, m in [("child covariates", m2), ("+ district", m3)]:
    print(f"{name:18} {m.params['feeding_programme[T.True]']:+.4f}")
```

Le tableau autour duquel ce cours est bâti — En R

```r
library(dplyr)

m1 <- lm(rate ~ feeding_programme, data = d)
m2 <- lm(rate ~ feeding_programme + sex + age_years +
           disability_reported + displacement_status, data = d)
m3 <- update(m2, . ~ . + admin2)
```

Le tableau autour duquel ce cours est bâti

Modèle	Coefficient cantine	ET	t
M1 un prédicteur	**+0,0447**	0,0089	5,02
M2 + sexe, âge, handicap, déplacement	**+0,0437**	0,0090	4,85
M3 + district	**+0,0360**	0,0100	3,60
M4 = M1 avec ET groupées par école	**+0,0447**	**0,0154**	**2,90**

Le tableau autour duquel ce cours est bâti

- Ajouter des covariables a changé le coefficient et laissé l'erreur-type tranquille — De M1 à M3, l'estimation passe de…
- Le regroupement a changé l'erreur-type et laissé le coefficient intact — M4 est la *même droite ajustée* que M1 —…
- Ce sont deux problèmes distincts — L'un porte sur le caractère comparable des groupes ; l'autre sur la quantité…

Ce que « ajuster sur » fait réellement

- Un coefficient avec des covariables dans le modèle est la comparaison à l'intérieur des niveaux de ces covariables —. . .

Ce que « ajuster sur » fait réellement — En Python

```
within = (d.groupby("admin2")
          .apply(lambda g: g[g["feeding_programme"]]["rate"].mean()
                 - g[~g["feeding_programme"]]["rate"].mean()))
print(within.round(4))
print(f"unadjusted: {m1.params['feeding_programme[T.True]']:.4f}")
print(f"adjusted:   {m3.params['feeding_programme[T.True]']:.4f}")
```

Ce que « ajuster sur » fait réellement — En R

```
d |> summarise(gap = mean(rate[feeding_programme]) -
                     mean(rate[!feeding_programme]), .by = admin2)
```

Ce que « ajuster sur » fait réellement

District	Écart	Avec cantine	Sans
Artibonite	+3,8 pts	146	141
Centre	**−4,1 pts**	244	**42**
Nord-Ouest	+4,9 pts	251	45
Sud	+7,0 pts	104	178

Ce que « ajuster sur » fait réellement

- Le coefficient ajusté est une moyenne pondérée de ces quatre écarts — et les voir séparément vaut les deux lignes que…
- Un coefficient unique moyenné sur un désaccord reste un nombre, et ce n'est plus un résumé — Avant de rapporter le 3,6…

Pourquoi les covariables n'ont presque rien déplacé

- La cantine a été attribuée par école — Rien concernant un *enfant* n'a décidé s'il recevait des repas scolaires, si...
- Le district, lui, l'a déplacée — de 4,5 à 3,6, et pour la raison inverse : les écoles avec cantine ne sont pas...
- La règle qu'implique le protocole : les facteurs de confusion vivent au niveau de l'attribution — Pour un programme de...

Lire un tableau de coefficients sans se faire tromper

- Rapportez la taille d'échantillon du modèle, non celle du fichier — L'ajustement sur cas complets a écarté 49 élèves…

Lire un tableau de coefficients sans se faire tromper — En Python

```
print(f"file {len(per_student)}, model {int(m3.nobs)}")
```

Lire un tableau de coefficients sans se faire tromper — En R

```
c(file = nrow(per_student), model = nobs(m3))
```

Lire un tableau de coefficients sans se faire tromper

- Ne lisez pas les coefficients des covariables comme des constats — Ils sont ajustés sur un ensemble de variables choisi...
- Vérifiez si une covariable ajoutée a changé l'échantillon — Une covariable avec des valeurs manquantes écarte...
- Rapportez aussi l'estimation non ajustée — La paire est ce qui montre au lecteur le travail qu'a fait l'ajustement, et...

Quand l'ajustement ne peut pas aider

- Le facteur de confusion n'est pas dans le fichier — Ajuster sur ce que vous avez mesuré ne dit rien de ce que vous...
- La covariable est sur le chemin causal — La leçon 5 montre une covariable qui retire 42 % de l'effet en se plaçant...
- Le problème est l'erreur-type, non l'estimation — Aucune covariable ne corrige le regroupement en grappes

Rapportez-le en entier — Exemple (suite)

```
Attendance and school feeding, adjusted analysis

  Linear model, 1,151 students with complete covariates (49 dropped).

  Unadjusted                     +4.47 points   95% CI +2.73 to +6.22
  Adjusted for district          +3.60 points   95% CI +1.64 to +5.56
  Child-level covariates (sex, age, disability, displacement) change the
  estimate by less than 0.1 points and are retained only to show that.

  The district gaps are +3.8, -4.1, +4.9 and +7.0 points. Centre runs the
  other way on 42 unfed students and is not averaged away silently.

  Standard errors above assume independent students and are too small; the
  school-clustered version of the unadjusted model gives +4.47 (95% CI
  +1.45 to +7.49).
```

Rapportez-le en entier — Exemple (suite)

```
Observational. Schools were not randomised into the programme, and
district adjustment does not make the remaining comparison causal.
```

Rapportez-le en entier

- Les deux estimations, et l'intervalle groupé, en six lignes — Un lecteur qui ne voit que le nombre ajusté ne peut pas...

La suite

- Tout jusqu'ici avait un résultat continu.

Lire la leçon complète, avec le code exécutable

https://data-analysis.cassion.dev/fr/cours/regression-for-programmes/reg-02-what-adjustment-moves