

Le coefficient est l'écart

Ajustez la présence sur une seule indicatrice de cantine. La constante vaut 85,21 % — la moyenne des écoles sans programme. Le coefficient vaut 4,94 points — l'écart entre les deux moyennes, à la quatrième décimale. Une régression à une indicatrice est une comparaison de deux groupes, rien de plus.

Pierre Robentz Cassion

Leçon 1 sur 8

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

Ce que couvre cette leçon

- La même comparaison, écrite deux fois
- Ce que signifie chaque terme
- Où les deux chemins diffèrent, et de combien
- Trois prédicteurs, trois lectures
- Ce que le modèle ne dit pas
- Rapportez-le en entier
- La suite

La même comparaison, écrite deux fois — En Python (suite)

```
import pandas as pd
import numpy as np
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")          # two never-de-registered transfers

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"))
```

La même comparaison, écrite deux fois — En Python (suite)

```
model = smf.ols("rate ~ feeding_programme", data=per_student).fit()  
print(model.summary().tables[1])
```

La même comparaison, écrite deux fois — En R

```
library(dplyr)
```

```
per_student |> lm(rate ~ feeding_programme, data = _) |> summary()
```

La même comparaison, écrite deux fois

Terme	Coefficient	ET	t
Constante	0,8521	0,0070	121,30
feeding_programme [T0,049]	0,0494	0,0088	5,63

La même comparaison, écrite deux fois — En Python

```
means = per_student.groupby("feeding_programme")["rate"].mean()
print(means)
print(f"difference: {means[True] - means[False]:.4f}")
```

La même comparaison, écrite deux fois — En R

```
per_student |> summarise(rate = mean(rate), .by = feeding_programme)
```

La même comparaison, écrite deux fois

- Moyenne sans programme : 0,8521. Avec : 0,9015. Écart : 0,0494

Ce que signifie chaque terme

- La constante est la valeur ajustée quand tous les prédicteurs valent zéro — Ici c'est `feeding_programme = False`, donc...
- Un coefficient est la variation du résultat pour une variation d'une unité de ce prédicteur, les autres étant maintenus...
- Une erreur-type est la même erreur-type que celle calculée au cours précédent — La colonne `t` vaut le coefficient sur...

Ce que signifie chaque terme — En Python

```
coef = model.params["feeding_programme[T.True]"]  
se = model.bse["feeding_programme[T.True]"]  
print(f"{coef:+.4f} 95% CI [{coef - 1.96*se:+.4f}, {coef + 1.96*se:+.4f}]")
```

Ce que signifie chaque terme — En R

```
confint(model)
```

Ce que signifie chaque terme

- $+0,0494$, IC à 95 % $+0,0322$ à $+0,0665$ — le même intervalle, atteint depuis les mêmes données par un autre chemin

Où les deux chemins diffèrent, et de combien

- Les moindres carrés ordinaires supposent une seule variance résiduelle pour les deux groupes — C'est le test t à...

Où les deux chemins diffèrent, et de combien — En Python

```
from scipy import stats

fed = per_student[per_student["feeding_programme"]]["rate"]
unfed = per_student[~per_student["feeding_programme"]]["rate"]
print(f"variances: {fed.var():.5f} and {unfed.var():.5f}")
print("pooled t:", stats.ttest_ind(fed, unfed).statistic.round(2))
print("Welch t:", stats.ttest_ind(fed, unfed, equal_var=False).statistic.round(2))
```

Où les deux chemins diffèrent, et de combien — En R

```
t.test(rate ~ feeding_programme, data = per_student, var.equal = TRUE)
t.test(rate ~ feeding_programme, data = per_student)
```

Où les deux chemins diffèrent, et de combien

- La régression reproduit exactement le t à variance commune — Quand les variances de groupe sont proches les deux. . .

Trois prédicteurs, trois lectures

Prédicteur	Le coefficient se lit
Binaire (feeding_programme)	L'écart entre les deux groupes
Continu (age_years)	La variation par année
Catégoriel à k niveaux	$k-1$ coefficients, chacun un écart contre le niveau omis

Trois prédicteurs, trois lectures — En Python

```
enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
current = enrolment[enrolment["school_year"] == 2024]
joined = per_student.merge(
    current[["student_id", "sex", "age_years", "admin2"]], on="student_id")

print(smf.ols("rate ~ admin2", data=joined).fit().params.round(4))
print(joined.groupby("admin2")["rate"].mean().round(4))
```

Trois prédicteurs, trois lectures — En R

```
lm(rate ~ admin2, data = joined) |> coef()  
joined |> summarise(rate = mean(rate), .by = admin2)
```

Trois prédicteurs, trois lectures

- Un prédicteur catégoriel produit un coefficient par niveau sauf un — et chaque coefficient est une comparaison contre. . .
- Énoncez le niveau de référence dans la légende du tableau — « Contre Artibonite » fait deux mots et c'est la différence. . .

Ce que le modèle ne dit pas

- Ce n'est pas un effet — Le coefficient de 4,9 points est l'écart entre deux ensembles d'écoles qui n'ont pas été. . .
- Ce n'est pas une prédiction qui vaille la peine — La régression est enseignée ailleurs comme une machine à prédire, et. . .

Rapportez-le en entier — Exemple

Attendance and school feeding, February to April 2024

Linear model, one predictor, 1,200 students.

Intercept (no programme) 85.21%

Feeding programme +4.94 points 95% CI +3.22 to +6.65

The coefficient is the difference between the two group means and is reported as such. Schools were not randomised into the programme; the comparison is observational.

The standard error assumes independent students. It is not, and lesson 6 gives the corrected version.

- La dernière ligne est une reconnaissance de dette que le cours honore — et l'écrire vaut mieux que d'ignorer qu'elle. . .

- Un prédicteur reproduit une comparaison que vous auriez pu faire sans modèle.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/regression-for-programmes/
reg-01-the-coefficient-is-the-difference](https://data-analysis.cassion.dev/fr/cours/regression-for-programmes/reg-01-the-coefficient-is-the-difference)