HEC LIÈGE
Management School - Liège Université

Chapitre 4 : Regression linéaire multivariée (RLM)

Économétrie (ECON0212)

Malka Guillot
HEC Liège

Recap - 3-causalité

04:00

Wooclap

lien de participation (code : $RJPLYF$)

Aujourd'hui - Régression linéaire multivariée

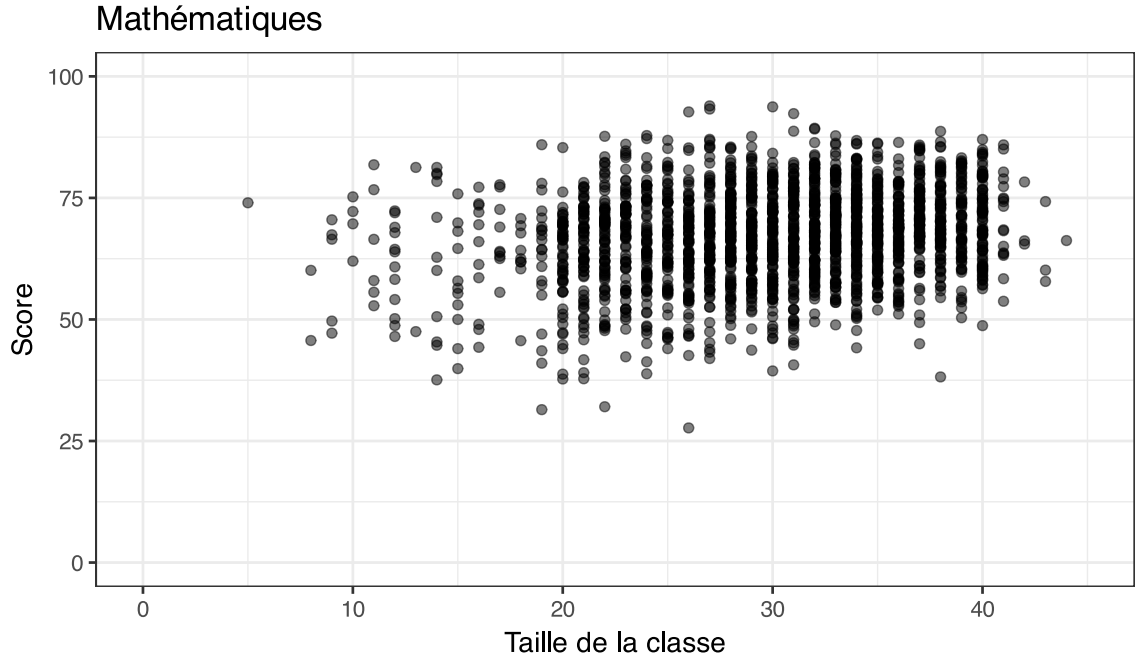
- Plusieurs variables indépendantes dans le modèle
- Interprétation pour les régresseurs continus et indicatrice
  - Le piège de la variable indicatrice
- Biais des variables omises
- R^2 ajusté
- Applications empiriques :
  - *Taille de la classe* et *performance des élèves*

Taille de la classe et performance des étudiants

- Revenons à l'analyse d'Angrist et Lavy (1999) sur l'effet de la taille des classes sur les performances des élèves en Israël.
- Avec une **régression linéaire univarié**, nous avons trouvé que la taille de la classe était positivement ***associée*** aux scores des élèves en mathématiques et en lecture.

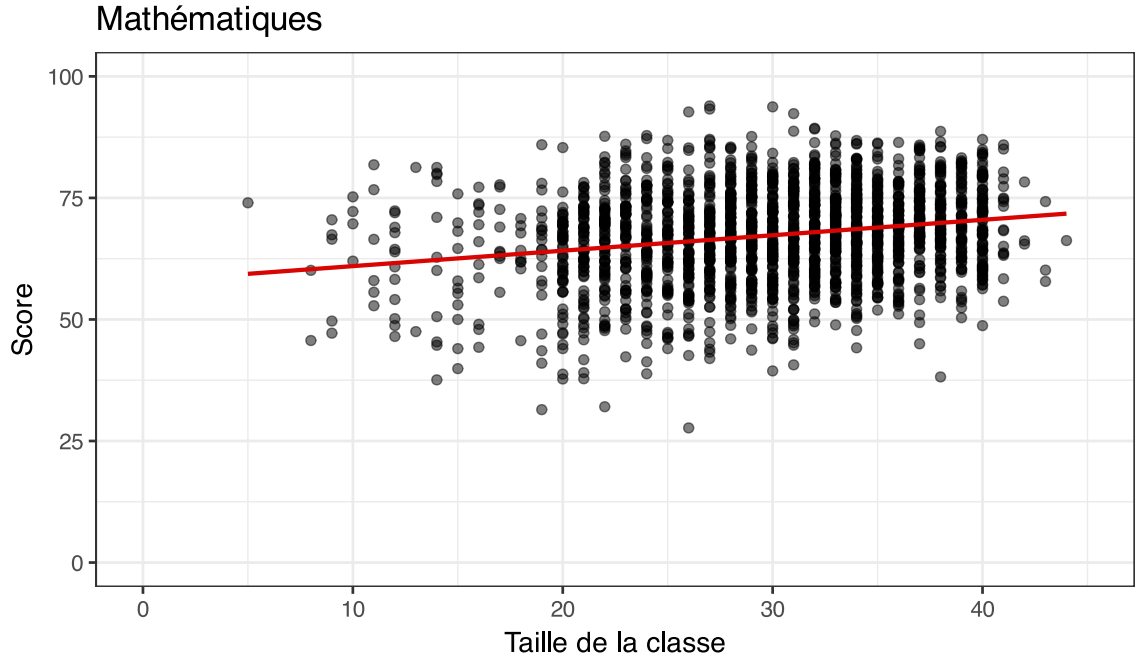
Taille de la classe et performance des étudiants

Régression univariée



Taille de la classe et performance des étudiants

Régression univariée



Taille de la classe et performance des étudiants

Régression univariée

```
reg avgmath classize
```

Source	SS	df	MS		
Model	8764.35423	1	8764.35423	Number of obs	= 2,019
Residual	177092.425	2,017	87.7999133	F(1, 2017)	= 99.82
				Prob > F	= 0.0000
				R-squared	= 0.0472
				Adj R-squared	= 0.0467
Total	185856.779	2,018	92.0994943	Root MSE	= 9.3702

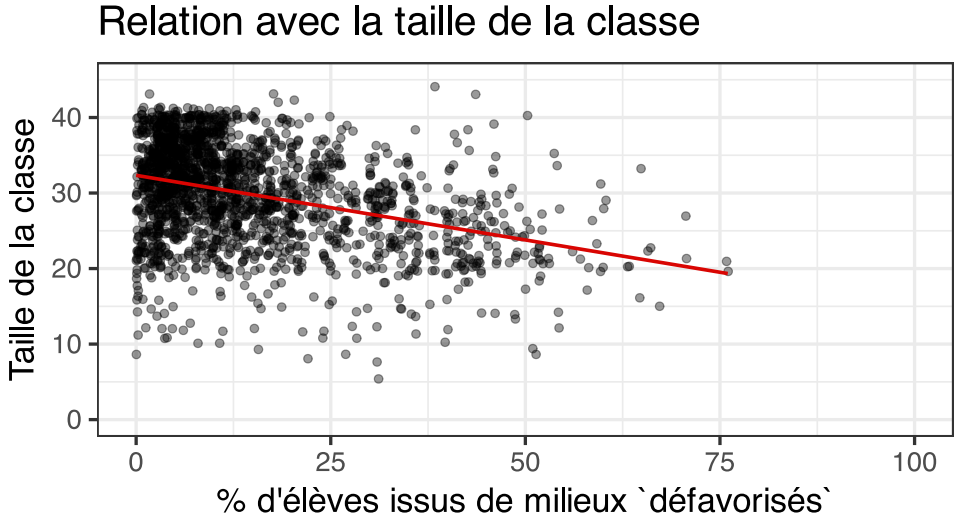
avgmath	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
classize	.3174906	.0317774	9.99	0.000	.2551707	.3798105
_cons	57.79392	.9736486	59.36	0.000	55.88445	59.70338

Taille de la classe et performance des étudiants

- Angrist et Lavy (1999) étudie l'effet de la taille des classes sur les performances des élèves en Israël.
- **Régression linéaire univarié** : la taille de la classe est ***positivement associée*** aux scores des élèves en mathématiques et en lecture.
- Cela est intuitivement inattendu et contraste avec les résultats simples de l'expérience randomizée *STAR [cf. 3-causalité]*.
- Est-il possible qu'une autre variable puisse être liée à la taille de la classe ***ET*** aux performances des élèves ?
- En particulier, nous avons mentionné l'**effet de localisation** :
  - les grandes classes peuvent être plus courantes dans les villes plus riches et plus grandes,
  - tandis que les petites classes peuvent être plus fréquentes dans les régions rurales plus pauvres.
- Examinons cette hypothèse.

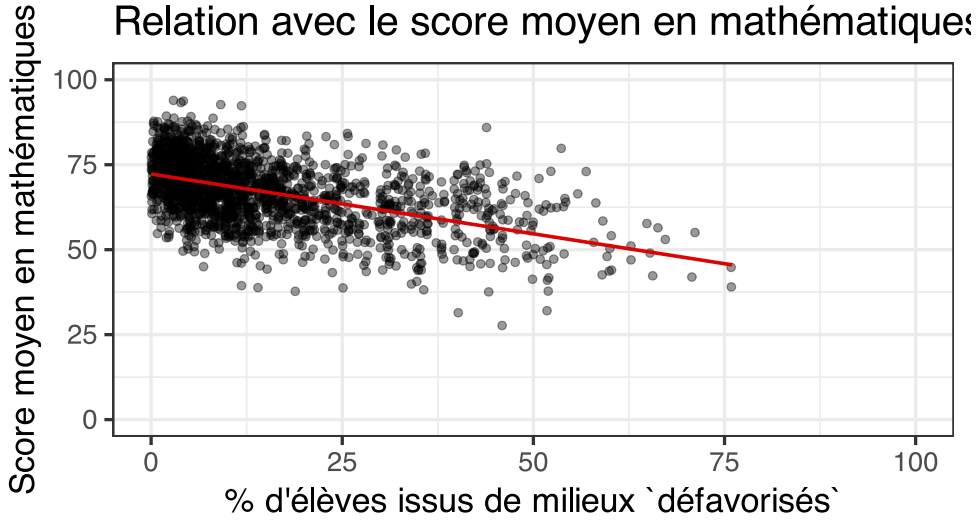
Facteurs confondants (*confounders*)

Lien entre la **taille de la classe** et **la part d'élèves issus de milieux `défavorisés`** dans la classe.



👉 En moyenne, il y a un plus grand % d'élèves issus de milieux `défavorisés` dans les classes plus petites.

Lien entre **le score moyen en mathématiques** et **la part d'élèves issus de milieux `défavorisés`** dans la classe.



👉 En moyenne, plus le % d'élèves issus de milieux `défavorisés` est élevé, plus le score moyen en maths est bas.

Taille de la classe et performance des étudiants : Régression multivariée

Supposons que nous voulions connaître l'effet de la taille de la classe sur les scores moyens en maths, **en contrôlant pour** le fait qu'il existe une *relation négative entre* :

- le % d'élèves issus de milieux défavorisés et la taille de la classe
- **ET** le score moyen en mathématiques.

$\Rightarrow$ Il faut inclure à la fois les variables `classize` et `disadvantaged` en tant que *régresseurs* dans la régression.

On obtient une estimation de l'effet de la taille de la classe sur le score moyen en mathématiques, ***purgée* de l'effet de la variable `disadvantaged`**.

Le modèle à estimer devient :

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \backslash\% \text{ disadvantaged}_i + e_i$$

- C'est une ***régression multivariée*** !

Plan du cours

1. Modèle de Régression Linéaire Multiple (RLM)
2. Variables catégorielles
3. Biais de variables omises

Plan du cours

[Objectif] Régression Linéaire Multiple

- Rappel d'il y a 2 cours, le modèle de **Régression Linéaire Simple** (ou univarié) peut être écrit comme

$$y_i = \beta_0 + \beta_1 x_{1,i} + \epsilon_i,$$

où y_i est la ***variable dépendante*** et x_i est la ***variable indépendante***.

- Rappelez-vous : Nous disons que `X` *cause* `Y` lorsque si nous devions intervenir et changer la valeur de `X` ***sans changer rien d'autre***, alors `Y` changerait également en conséquence.

⚠ À moins que tous les autres facteurs affectant y_i ne soient pas corrélés avec x_i, β_1 **ne peut pas être interprété comme un effet causal**.

Nous devons **enrichir le modèle** et prendre en compte les facteurs qui sont simultanément liés à y_i **et** x_i.

[Définition] Régression Linéaire Multivariée (RLM)

On cherche à étudier la relation entre :

- Une **variable dépendante** y_i
  - (par exemple, le score moyen en mathématiques)
- **Plusieurs variables indépendantes** $x_{1,i}$, $x_{2,i}$, ..., $x_{k,i}$
  - (par exemple, la taille de la classe et le % d'élèves défavorisés)

Le modèle de régression multivariée suppose que la relation linéaire suivante est valide dans la population :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \epsilon_i,$$

où x_1, x_2, ..., x_k sont les k régresseurs, et β_1, β_2, ..., β_k sont les k coefficients associés.

[Définition] Régression Linéaire Multivariée (RLM)

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \epsilon_i,$$

- β_0 est la constante et β_1, β_2, ..., β_k sont les pentes
- ϵ_i est le terme d'erreur qui capture l'effet des variables qui ne sont pas incluses dans le modèle.
- Le modèle est linéaire en ses paramètres (β_0, β_1, ..., β_k).
- Le modèle reflète une relation dans la **population**

[Estimation] Modèle de Régression Linéaire Multivariée

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \epsilon_i,$$

où x_1, x_2, ..., x_k sont les k régresseurs, et β_1, β_2, ..., β_k sont les k coefficients associés.

Nous obtenons les valeurs de $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ de la même manière qu'auparavant, en utilisant la méthode des **MCO**.

- $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ sont les valeurs qui minimisent la **Somme du Carré des Résidus** (SCR).
- Autrement dit, elles minimisent

$$\begin{aligned}\sum_i \epsilon_i^2 &= \sum_i (y_i - \hat{y}_i)^2 \\ &= \sum_i \left[y_i - (\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i})\right]^2\end{aligned}$$

Cas $k = 2$: interprétation en terme d'effet net

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$$

On se concentre sur $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{1,i} y_i}{\sum_{i=1}^{n} \hat{r}_{1,i}^2}$$

- $\hat{r}_{1,i}$ est le résidu de la régression de x_1 sur les x_2 et une constante : $x_{1,i} = \gamma_0 + \gamma_1 x_{2,i} + r_{1,i}$
- $\hat{r}_{1,i}$ est la part de $x_{1,i}$ qui n'est pas expliquée par $x_{2,i}$ (ie. le résidu).
- $\hat{\beta}_1$ est l'estimateur de la pente dans la régression de y sur une constante et $\hat{r}_{1,i}$:

$$y_i = \beta_0 + \hat{\beta}_1 \hat{r}_{1,i} + \epsilon_i$$

- $\hat{\beta}_1$ peut s'interpréter comme l'effet de x_1 sur y en maintenant x_2 constant.

Moindres Carrés Ordinaires (MCO): formule de coefficients

Des conditions du premier ordre du problème de minimisation, on déduit les estimateurs MCO :

> **Constante :** $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} - \cdots - \hat{\beta}_k \bar{x}_k$
>
> **Pentes :** $\hat{\beta}_k = \dfrac{\sum_{i=1}^{n} \hat{r}_{k,i} y_i}{\sum_{i=1}^{n} \hat{r}_{k,i}^2}, \quad j = 1, 2, \cdots, k$

Où $\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{k,i}$

Et $\hat{r}_{k,i}$ est le résidu de la régression de x_k sur les autres régresseurs (et une constante) ($x_k = \gamma_0 + \gamma_1 x_{1,i} + \cdots + \gamma_{k-1} x_{k-1,i} + r_{k,i}$) :

$$\hat{r}_{k,i} = x_k - \hat{\beta}_0 - \hat{\beta}_1 x_{1,i} - \cdots - \hat{\beta}_{k-1} x_{k-1,i}$$

Estimateur des Moindres Carrés Ordinaires (MCO)

On peut réécrire les $\hat{\beta}_k$ ainsi :

$$\hat{\beta}_k = \frac{\sum_{i=1}^{n} \hat{r}_{k,i} y_i}{\sum_{i=1}^{n} \hat{r}_{k,i}^2} = \frac{\sum_{i=1}^{n} (\hat{r}_{k,i} - \bar{\hat{r}_k})(y_i - \bar{y})}{\sum_{i=1}^{n} (\hat{r}_{k,i} - \bar{\hat{r}_k})^2}, \quad j = 1, 2, \cdots, k$$

car les résidus sont nuls en moyenne ($\bar{\hat{r}_k} = \frac{1}{n} \sum_{i=1}^{n} \hat{r}_{k,i} = 0$)

Cf. définition de l'estimateur par MCO de la pente dans le RLS :

- $\hat{\beta}_k$ est l'estimateur de la pente dans la régression de y sur une constante et $\hat{r}_k$.
- $\hat{\beta}_k$ peut s'interprétêr comme l'effet de x_k sur y en maintenant tous les autres régresseurs constants
  - $\hat{r}_k$ est le résidu, ie. ce qui reste de x_k après avoir pris en compte la variation des $x_1, x_2, \cdots, x_{k-1}$.

Moindres Carrés Ordinaires (MCO): 2 variables explicatives

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$$

On considère le RLS de x_1 sur x_2 :

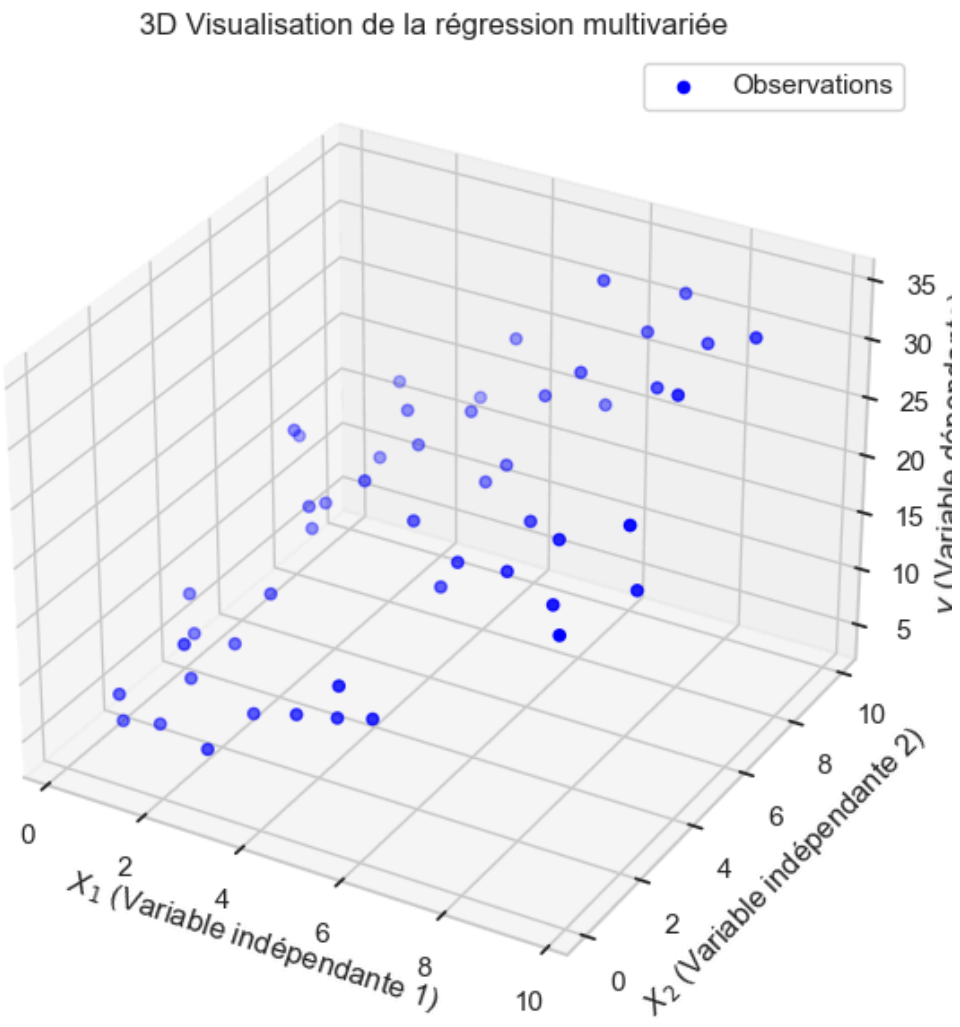
$$x_{i,1} = \gamma_0 + \gamma_1 x_{2,i} + r_{i,1}$$

dont l'estimation par MCO donne :

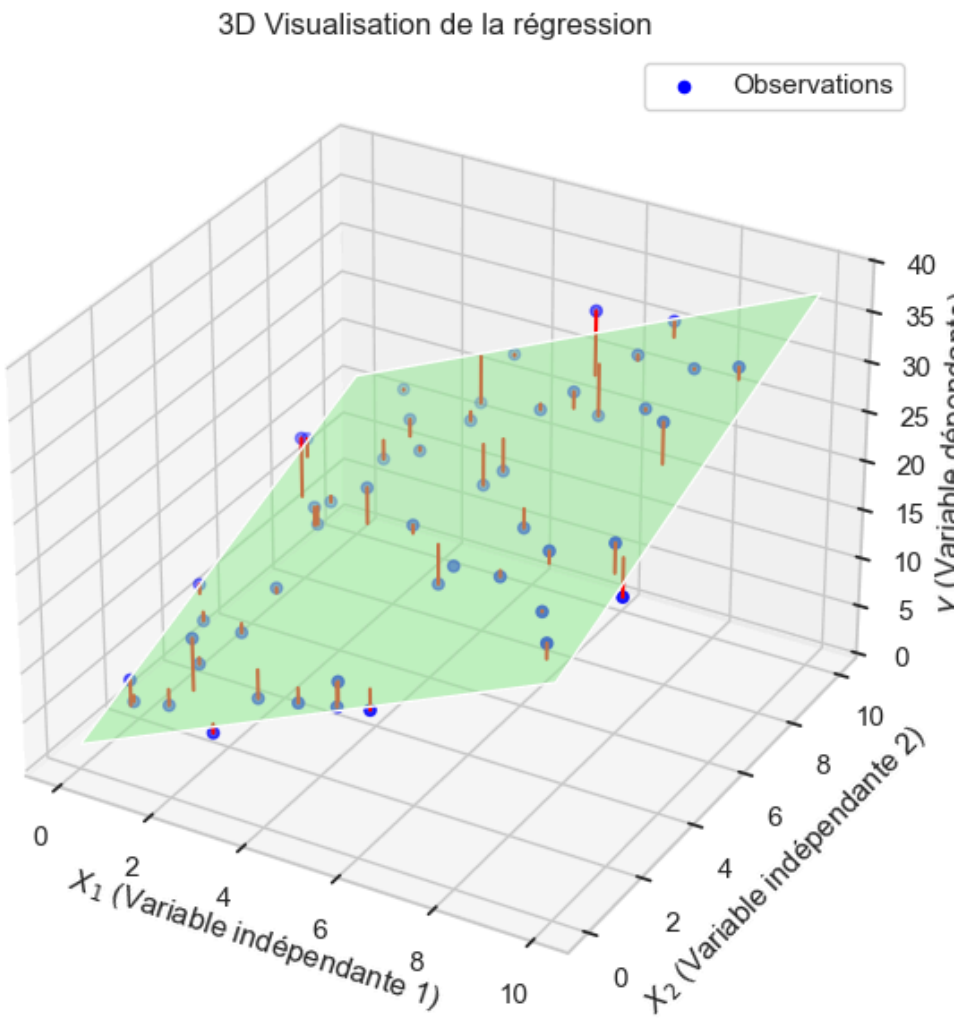
$$\hat{r}_{i,1} = x_{i,1} - \hat{\gamma}_0 - \hat{\gamma}_1 x_{2,i}$$

C'est la part de x_1 qui n'est pas expliquée par x_2.

Modèle de Régression Multiple : Géométriquement



Modèle de Régression Multiple : Géométriquement



Valeur prédites, Résidus, et qualité de l'ajustement

Comme dans le cas du RLM, on peut calculer

- les valeurs prédites $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i}$
- et les résidus $\hat{\epsilon}_i = y_i - \hat{y}_i$.
- La qualité de l'ajustement est mesurée par le coefficient de détermination R^2

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \in [0, 1]$$

Avec $SCE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, $SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$ et $SCR = \sum_{i=1}^{n} \hat{\epsilon}_i^2$

[Remarques] Qualité de l'ajustement

- Le R^2 augmente mécaniquement quand on ajoute des variables dans la régression.
  - Le R^2 n'est pas en tant que tel une bonne manière de décider si on ajoute des variables dans le modèle.
- Un R^2 faible :
  - n'indique pas que le modèle est mauvais.
  - ne dit rien de si les coefficients estimés peuvent s'interprétés de manière **causale**
  - indique le modèle n'est pas très utile pour faire une **prédiction**

Modèle de Régression Multiple : Interprétation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i}$$

Pour des variables y et x_k numériques :

> Constante $(\hat{\beta}_0)$: **La valeur estimée (ou prédite) de y $(\hat{y})$ quand tous les régresseurs $(x_1, x_2, x_3, \dots)$ sont égaux à 0.**
>
> Pente $(\hat{\beta}_k)$: **Le changement prédit, en moyenne, dans la valeur de y *associé* à une augmentation d'une unité de x_k...**
> **... en maintenant tous les autres régresseurs constants !**

Modèle de Régression Multiple : Interprétation

Les estimations $\hat{\beta}_k$ (ie b_k) s'interprètent comme des **effets marginaux** ou ou **Toutes Choses Égales Par Ailleurs (TCEPA)** :

$$\hat{y}_i = b_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i}$$

devient

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

Lorsque x_2 est constant, $\Delta x_2 = 0$, et donc

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1$$

De même, pour x_1 est constant, $\Delta x_1 = 0$, et donc

$$\Delta\hat{y} = \hat{\beta}_2 \Delta x_2$$

Modèle de Régression Multiple : Interprétation

Remarque :

- le *maintien de tous les autres régresseurs constants* est la seule partie qui change par rapport à la régression univariée
- Autrement dit, on considère l'effet individuel de la variable x_k sur y comme **isolé** de l'effet que les autres régresseurs pourraient avoir sur y.
    - effet **Toutes Choses Égales Par Ailleurs (TCEPA)**

Lien avec l'inférence causale :

- Seuls les régresseurs inclus dans le modèle sont maintenus constants, ceux qui ne sont pas dans le modèle peuvent encore varier et "biaiser" les estimations.

Application : Taille de la classe et performance des étudiants

Regression multivariée avec `Stata`

- Très similaire avec la régression linéaire univariée

```
regress dependent_variable  independent_variable_1 independent_variable_2 ...
```

Taille de la classe et performance des étudiants : Régression multivariée

Estimons le modèle précédent par OLS :
$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \backslash\% \text{ disadvantaged}_i + e_i$

```
regress avgmath  classize  disadvantaged
```

Taille de la classe et performance : Régression multivariée

```
regress avgmath  classize  disadvantaged
```

Source	SS	df	MS	Number of obs	=	2,019
				F(2, 2016)	=	330.87
Model	45929.6883	2	22964.8442	Prob > F	=	0.0000
Residual	139927.091	2,016	69.4082793	R-squared	=	0.2471
				Adj R-squared	=	0.2464
Total	185856.779	2,018	92.0994943	Root MSE	=	8.3312

avgmath	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
classize	.0716782	.0301848	2.37	0.018	.0124816	.1308748
disadvantaged	-.3395788	.014675	-23.14	0.000	-.3683585	-.3107991
_cons	69.94438	1.012486	69.08	0.000	67.95875	71.93001

2 Questions

1. Quelle est l'interprétation de chaque coefficient ?
2. Comment expliquer le changement du coeficient associé à la variable `classize` comparé au cas univarié?

Taille de la classe et performance : Régression multivariée

avgmath	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
classize	.0716782	.0301848	2.37	0.018	.0124816	.1308748
disadvantaged	-.3395788	.014675	-23.14	0.000	-.3683585	-.3107991
_cons	69.94438	1.012486	69.08	0.000	67.95875	71.93001

Réponse 1 : Quelle est l'interprétation de chaque coefficient ?

- $b_0 = 69.94$: Quand `class size` et `disadvantaged` sont égaux à 0 et 0, la valeur *predite* de la note moyenne en maths est égale à 69.94.
- $b_1 = 0.07$: En gardant le pourcentage *d'élèves défavorisés* constant dans la classe, l'augmentation de la taille de la classe d'un élève est ***associée, en moyenne,*** à une augmentation de 0.07 point du score moyen en maths.
- $b_2 = -0.34$: En gardant la *taille de la classe* constante, une augmentation de 1 ***point de pourcentage*** du pourcentage *d'élève défavroisés* est ***associée, en moyenne,*** à une diminution de 0.34 point du score moyen en maths.

Taille de la classe et performance : Régression multivariée

avgmath	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
classize	.0716782	.0301848	2.37	0.018	.0124816	.1308748
disadvantaged	-.3395788	.014675	-23.14	0.000	-.3683585	-.3107991
_cons	69.94438	1.012486	69.08	0.000	67.95875	71.93001

Réponse 2 : Comment expliquer le changement du coefficient associé à la variable `classize` comparé au cas univarié?

- b_1 décroit quand quand on introduit la variable `disadvantaged` dans la régression.
  - $0.31 \rightarrow 0.07$
- Ceci était attendu, puisque une partie de l'effet positif de la taille de la classe était en fait dû au fait qu'il y a une part d'élèves défavorisés plus faible dans les grandes classes.

Pourcentage (%) vs. point de pourcentage (ppt) [apparté]

Exemple : le % des élèves désavantagés dans la classe augmente de 10 à 25 %

Questions:

1. Quel est le changement en *points de pourcentage* ?
2. Quel est le changement en *pourcentage* ?

Pourcentage (%) vs. point de pourcentage (ppt) [apparté]

Exemple: le % des élèves désavantagés dans la classe augmente de 10 à 25 %

Réponses :

1. Il y a une augmentation de $25 - 10 = 15$ ***points de pourcentage*** (ppt).
2. Il y a une augmentation de $\frac{25-10}{10}\% = 150$ ***pour cent*** (%).

Vous ***devez*** faire attention à savoir si vous parlez de changements en termes de ***points de pourcentage*** ou en ***pourcentage !***. Ils impliquent des magnitudes très différentes !

Exercice 1

10:00

Nous allons réaliser des régressions avec **reading** (note en lecture) comme variable dépendante.

1. Ouvrir la base `grade5.dta` dans stata.
2. Regresser `avgverb` sur `classize` et `disadvantaged`. Interpreter les coefficients, et comparez les avec ceux la régression sur la note en maths.
3. Quelles esont les autres variables disponibles qu'on pourrait vouloir ajouter dans la régression ?
    - Estimer cette régression, incluant toutes les variables qui vous intéressent.
    - Discuter de la valeur des coefficients : signe et magnitude.

Plan du cours

1. Modèle de Régression Linéaire Multiple (RLM)
2. Variables catégorielles
3. Biais de variables omises

Variables numérique ou indicatrice : interprétation

> - In English, ***variable indicatrice*** = ***dummy variable***
> - Parfois aussi appelé ***variable binaire*** (ou potentiellement ***variable catégorielle***)

Vous savez comment interpréter les coefficients lorsque la variable est numérique (c'est-à-dire continue).

Et si l'un des régresseurs est une ***variable indicatrice***, c'est-à-dire qu'elle prend la valeur 1 si une condition est `VRAIE` et 0 sinon ?

Exemple : Comment interpréter les coefficients dans le modèle suivant

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

`religious` est une variable indicatrice égale à 1 si l'école est une école religieuse, 0 si ce n'est pas le cas.

Variables numérique ou indicatrice : interprétation

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

```
regress avgmath  classize  religious
```

Source	SS	df	MS		
Model	13469.157	2	6734.57851	Number of obs	= 2,019
Residual	172387.622	2,016	85.5097333	F(2, 2016)	= 78.76
				Prob > F	= 0.0000
Total	185856.779	2,018	92.0994943	R-squared	= 0.0725
				Adj R-squared	= 0.0716
				Root MSE	= 9.2471

avgmath	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
classize	.2311258	.0334519	6.91	0.000	.165522	.2967296
religious	-3.780027	.5096029	-7.42	0.000	-4.77943	-2.780623
_cons	61.3092	1.07138	57.22	0.000	59.20807	63.41033

Variables numérique ou indicatrice : formellement

Notre modèle est le suivant :

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

Nous avons les égalités suivantes :

$$\begin{aligned}\mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} = 0) &= b_0 + b_1 \times 0 + b_2 \times 0 \\ &= b_0\end{aligned}$$

$\rightarrow b_0$ correspond à la valeur de l'espérance moyenne du score en mathématiques quand la taille de la classe est de 0 et que l'école n'est pas religieuse.

Variables numérique ou indicatrice : formellement

Notre modèle est le suivant :

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

Nous avons les égalités suivantes :

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0,1\} \ \& \ \text{class size} = n_1) = b_0 + b_1 \times n_1 + b_2 \times \text{religious}$$

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0,1\} \ \& \ \text{class size} = n_1 + 1) = b_0 + b_1 \times (n_1 + 1) + b_2 \times \text{religious}$$

$$\mathbb{E}(\text{average math score} | \text{religious} \in \{0,1\} \ \& \ \text{class size} = n_1 + 1) - \mathbb{E}(\text{average math score} | \text{religious} \in \{0,1\} \ \& \ \text{class size} = n_1)$$

$$= b_0 + b_1 \times (n_1 + 1) + b_2 \times \text{religious} - (b_0 + b_1 \times n_1 + b_2 \times \text{religious}) = b_1$$

$\rightarrow b_1$ correspond à la variation attendue du score moyen en maths associée, en moyenne, à une classe dont la taille augmente d'un étudiant, en contrôlant le statut religieux de l'école (= en maintenant le statut religieux constant).

Variables numérique ou indicatrice : formellement

Notre modèle est le suivant :

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

Nous avons les égalités suivantes :

$$\begin{aligned}\mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) &= b_0 + b_1 \times \text{class size} + b_2 \times 1 \\ &= b_0 + b_1 \times \text{class size} + b_2\end{aligned}$$

$$\begin{aligned}\mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N}) &= b_0 + b_1 \times \text{class size} + b_2 \times 0 \\ &= b_0 + b_1 \times \text{class size}\end{aligned}$$

$$\begin{aligned}&\mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) - \\ &\quad \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N}) \\ &\quad = b_0 + b_1 \times \text{class size} + b_2 - (b_0 + b_1 \times \text{class size}) = b_2\end{aligned}$$

$\rightarrow b_2$ correspond à la différence attendue du score moyen en mathématiques entre les écoles religieuses et non religieuses, en maintenant la taille de la classe constante.

Variables numérique ou indicatrice : Résumé

Notre modèle est le suivant :

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$

Nous avons les égalités suivantes :

$$b_0 = \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} = 0)$$

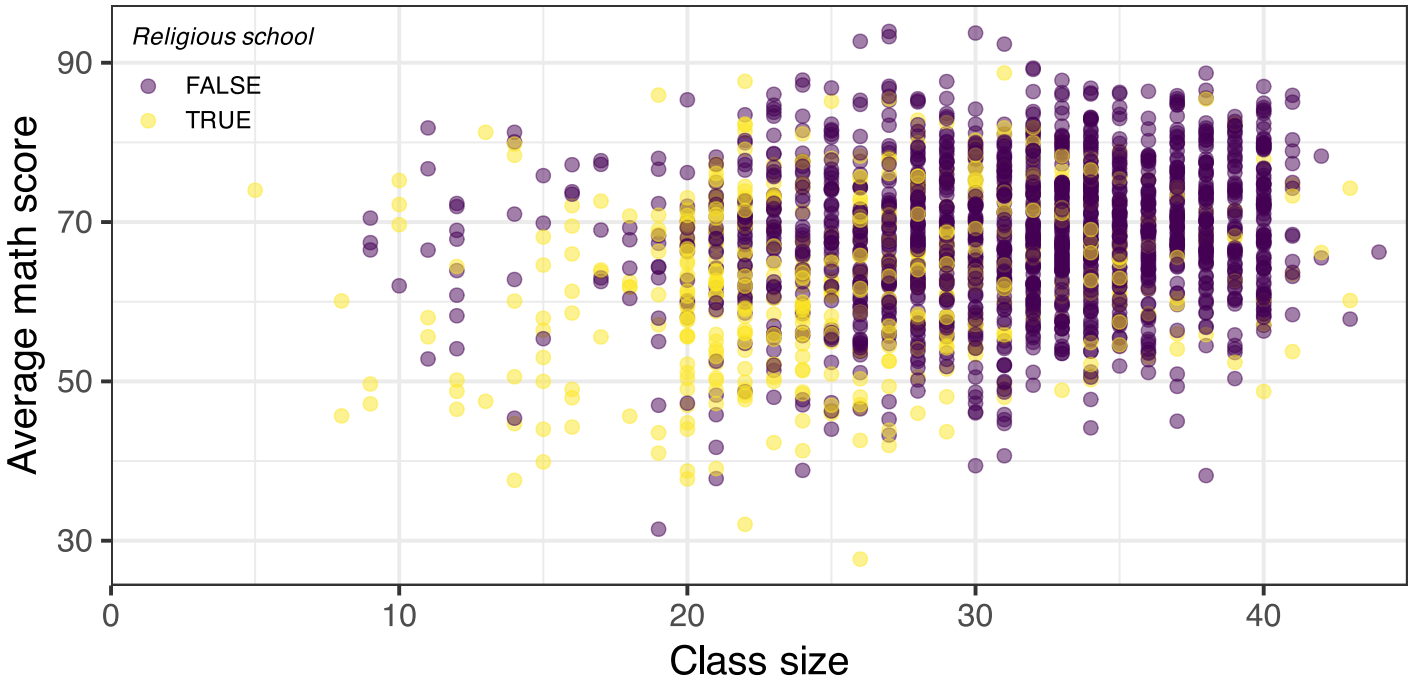
$$b_1 = \mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1 + 1) - \mathbb{E}(\text{average math score} | \text{religious} \in \{0, 1\} \ \& \ \text{class size} = n_1)$$

$$b_2 = \mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) - \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N})$$

$$b_0 + b_2 = \mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} = 0)$$

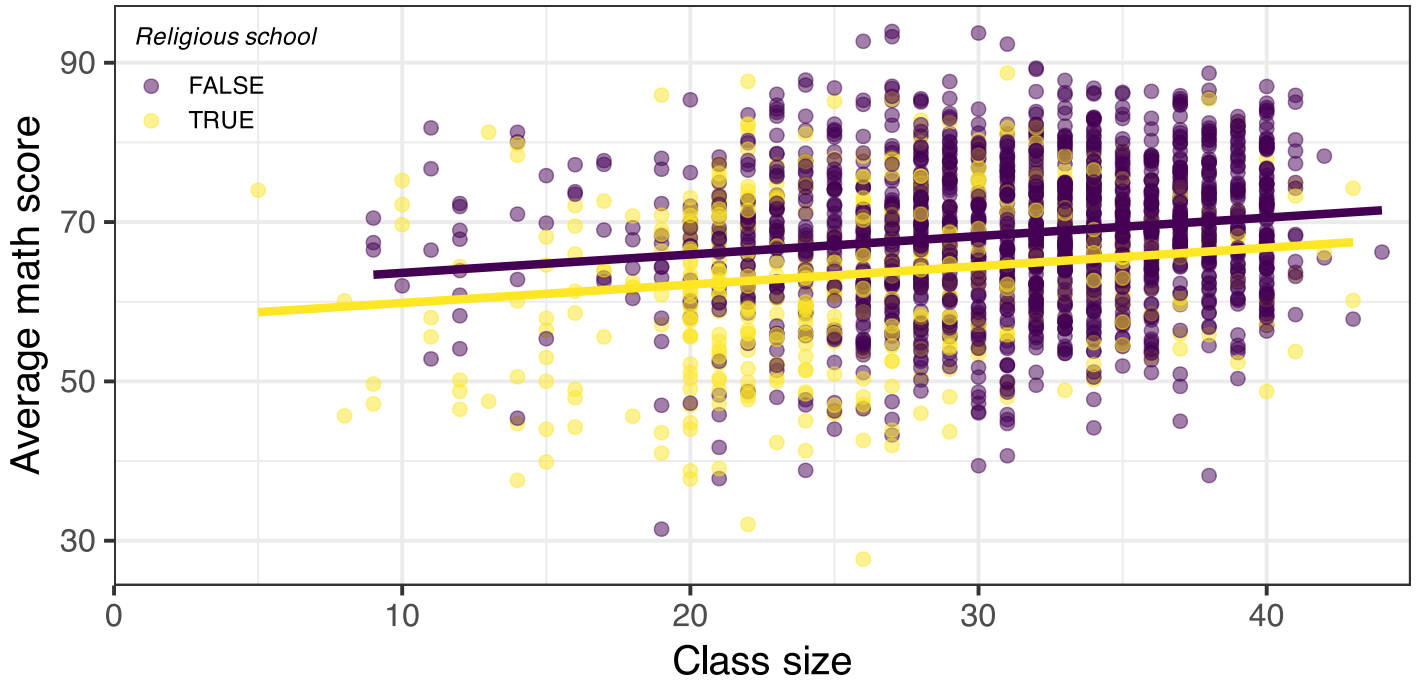
Variables numérique ou indicatrice : Graphiquement

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$



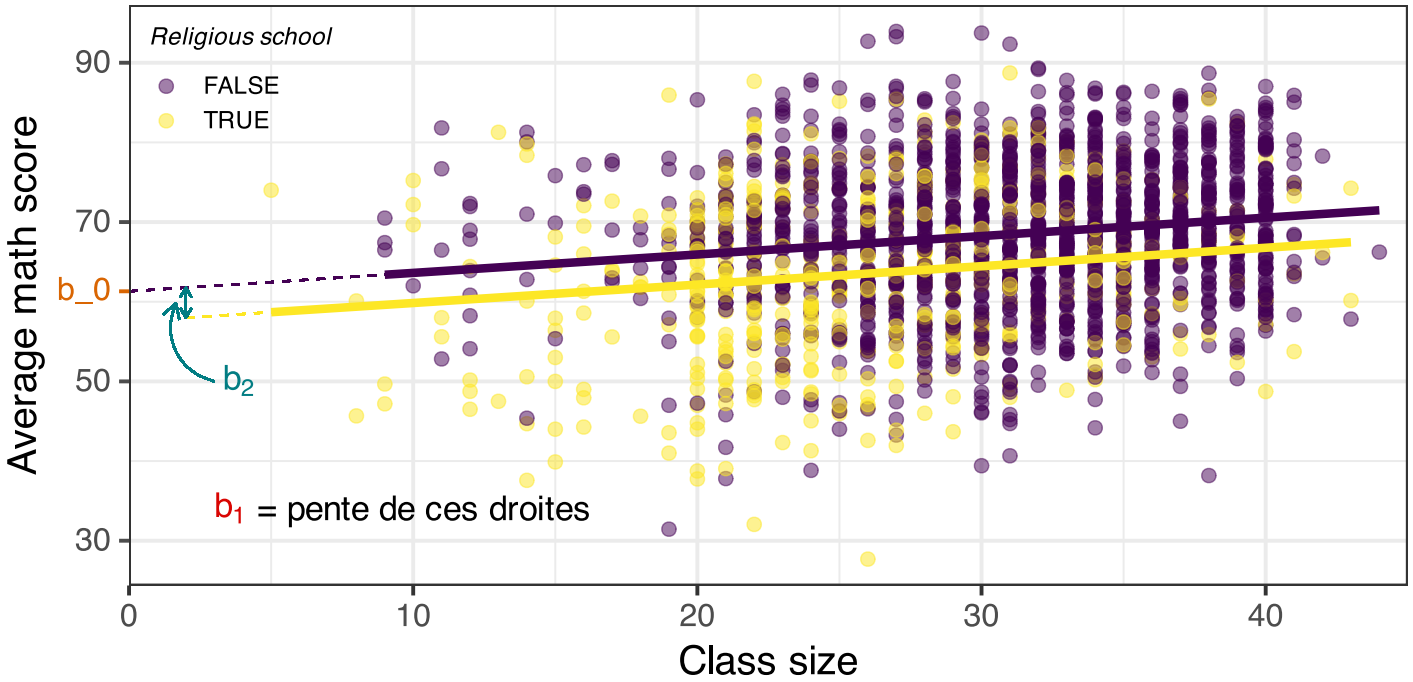
Variables numérique ou indicatrice : Graphiquement

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$



Variables numérique ou indicatrice : Graphiquement

$$\text{average math score}_i = b_0 + b_1 \text{class size}_i + b_2 \text{religious}_i + e_i$$



Pas de collinéarité parfaite

Il existe une condition à satisfaire pour ajouter des régresseurs au modèle :

> Toute variable additionnelle doit apporter **un quantité non-nulle de nouvelle information *a minima*** .

En d'autres termes, les régresseurs **ne peuvent pas être en collinéarité parfaite**, c'est-à-dire qu'ils ne sont pas des combinaisons linéaires les uns des autres :

$$x_2 \neq ax_1 + b$$

Même s'ils ne sont pas parfaitement corrélés, les effets individuels de régresseurs fortement corrélés sont difficiles à démêler.

Notez que cela implique que le nombre d'observations doit être supérieur au nombre de variables indépendantes.

Pas de collinéarité parfaite : Le piège des variables indicatrices

Cette condition est particulièrement pertinente pour les ***variables catégorielles*** :

- c'est-à-dire les variables qui prennent un nombre limité de "niveaux" possibles
- par exemple : le genre, les saisons, la race, les niveaux d'éducation, etc.

Revenons à notre régression sur les écoles `religieuses` :

avgmath	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
classize	.2311258	.0334519	6.91	0.000	.165522	.2967296
religious	-3.780027	.5096029	-7.42	0.000	-4.77943	-2.780623
_cons	61.3092	1.07138	57.22	0.000	59.20807	63.41033

Et si je crée une variable `is_religious` et une variable `is_notreligious` et que je régresse `avgmath` sur les deux (et `classize`) ?

Pas de collinéarité parfaite : variables indicatrices = Piège !

Et si je crée une variable `is_religious` et une variable `is_notreligious` et que je régresse `avgmath` sur les deux (et `classize`) ?

```
gen is_religious    = religious == 1
gen is_notreligious = religious == 0
reg avgmath is_religious is_notreligious classize
```

avgmath	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
is_religious	-3.780027	.5096029	-7.42	0.000	-4.77943	-2.780623
is_notreligious	0	(omitted)				
classize	.2311258	.0334519	6.91	0.000	.165522	.2967296
_cons	61.3092	1.07138	57.22	0.000	59.20807	63.41033

Seul l'un des deux a un coefficient ! Pourquoi ?

```
tabulate is_religious is_notreligious
```

is_religio us	is_notreligious 0	1	Total
0	0	1,522	1,522
1	497	0	497
Total	497	1,522	2,019

Pas de collinéarité parfaite : variables indicatrices = Piège !

$\rightarrow$ `Stata` détecte automatiquement la collinéarité parfaite entre deux variables et supprime automatiquement une des deux de la régression

⚠ vous devez faire attention à la ***catégorie omise/référence*** : la catégorie "de base" à partir de laquelle les coefficients sont interprétés.

- les coefficients des variables indicatrices sont interprétés par rapport à la catégorie omise.

$$b_2 = \mathbb{E}(\text{average math score} | \text{religious} = 1 \ \& \ \text{class size} \in \mathbb{N}) - \mathbb{E}(\text{average math score} | \text{religious} = 0 \ \& \ \text{class size} \in \mathbb{N})$$

- Particulièrement important pour les variables ayant plus de 2 catégories.
- Pas besoin de créer une variable binaire pour chaque possibilité, `Stata` détecte la variable(s) catégorielle(s)
  - (*à condition qu'elles soient stockées en tant que* `entier` *ou* `catégorielle`)
  - Cependant, il faut vérifier quelle catégorie a été omise.

[Stata] Variables catégorielles

- **Pb**: une variable catégorielle peut être représentée par des catégories numériques
  - Ex: `school`={1:Inner city, 2:rural, 3:suburban, 4:urban}
  - Si la variable catégorielle est `string`, la transformer en numérique
- Il faut préciser si la variable est catégorielle ou continue dans la régression :
  - `i.school` pour une variable catégorielle (`i` pour *indicatrice*)
  - `c.school` pour une variable continue (`c` pour *continue*)
- Par exemple, estimer les 2 modèles suivants
  - `reg math c.school` et `reg math i.school`
  - et observer la différence

Exercice 2 : Le piège des variables binaires, illustration

10:00

Estimons une régression où il y a une dépendance linéaire parfaite entre les variables indépendantes.

1. Ouvrer la base de données *star_data.dta*. Supprimer les observations ayant des valeurs manquantes.

2. Créer trois variables binaires : (i) `small` égal à `1` si les élèves sont dans une petite classe et à `0` sinon; (ii) `regular` égal à `1` si les élèves sont dans une classe de taille normale et à `0` sinon; (iii) `regular_plus` égal à `1` si les élèves sont dans une classe de taille normale+aide et à `0` sinon.

3. Créer une dernière variable `sum` égale à la somme de `small`, `regular` et `regular_plus`. A quoi `sum` est elle égale ? Que cela signifie-t-il?

4. Regresser `math` sur `small`, puis sur `regular` et enfin sur `regular_plus`. Quelle est la note moyenne en `math` prédite pour les élèves dans une classe normal+aide class?

5. Regresser `math` sur `small`, `regular` et `regular_plus`. Que remarquez vous ? Quelle est la catégorie de référence? Est ce que c'est cohérent avec la question précédente ?

Plan du cours

1. Modèle de Régression Linéaire Multiple (RLM)
2. Variables catégorielles
3. Biais de variables omises

Biais de variables omises [*Omitted Variable Bias* i.e. OVB]

Biais de variables omises: ne pas inclure des variables de contrôle importantes dans la régression

Cela rend le coefficient de la variable indépendante d'intérêt peu fiable (*biaisé*).

Soit y notre variable dépendante, x notre régresseur et z la variable omise. Soit les 3 modèles suivants :

1. Le **"vrai" modèle** : $y = \gamma_0 + \gamma_1 x + \gamma_2 z + \epsilon$ $\Rightarrow$ ***Régression multivariée***
2. Modèle estimé : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ $\Rightarrow$ ***Régression univariée***
   - z est **omise** de la régression
3. ***Variable omise sur le régresseur***: $z = \delta_0 + \delta_1 x + \eta$

La formule de l'OVB est: $$\text{OVB} = \gamma_2 \times \delta_1$$

[Démonstration] Biais de variable omise [1]

On montre que $\hat{\beta}_1 \rightarrow \gamma_1 + \gamma_2 \times \delta_1$

On a

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

En remplaçant y_i par le "vrai" modèle on obtient :

$$\hat{\beta}_1 = \gamma_1 + \gamma_2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})z_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})e_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

En prenant l'espérance de $\hat{\beta}_1$ on obtient :

$$\mathbb{E}(\hat{\beta}_1 | x_i, z_i, \forall i) = \gamma_1 + \gamma_2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})z_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

[Démonstration] Biais de variable omise [2]

On a donc :

$$\mathbb{E}(\hat{\beta}_1 | x_i, z_i, \forall i) = \gamma_1 + \gamma_2 \times \delta_1$$

Notre estimation β_1 du vrai coefficient γ_1 est donc biaisée par $\gamma_2 \times \delta_1$:

Donc

$$\text{OVB} = \mathbb{E}(\hat{\beta}_1 - \gamma_1 | x_i, z_i, \forall i) = \gamma_2 \times \delta_1$$

Biais de variable omise (OVB)

$$\text{OVB} = \underbrace{\text{coefficient de la régression multivariée sur la variable omise}}_{\gamma_2} \times \underbrace{\frac{Cov(x,z)}{Var(x)}}_{\delta_1}$$

Le biais dépend de γ_2 et de la corrélation entre x et z. Cette formule permet d'obtenir :

- La ***magnitude*** du biais (si z est observé),
- Le ***signe*** du biais (positif/negatif): puisqu'en pratique z n'est pas observé (sinon, on l'inclurait dans la régression), c'est le cas le plus pertinent

	$corr(x,z) > 0$	$corr(x,z) < 0$
$\gamma_2 > 0$	bias positif	bias negatif
$\gamma_2 < 0$	bias negatif	bias positif

Bias de variable omise en pratique

Question:

Imaginez que vous souhaitiez découvrir la relation entre le revenu et le nombre d'années d'études.

- Pourquoi une simple régression du revenu sur le nombre d'années d'études ne produirait-elle pas une estimation fiable ?
- Quelle pourrait être la variable omise ? Quel est le signe attendu de l'OVB ?

Bias de variable omise en pratique

Revenons à notre modèle initial. On avait :

Régression univariée : $\text{avg. math score} = b_0 + b_1 \text{class size} + e$

avgmath	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
classize	.3174906	.0317774	9.99	0.000	.2551707	.3798105
_cons	57.79392	.9736486	59.36	0.000	55.88445	59.70338

Régression multivariée : $\text{avg. math score} = c_0 + c_1 \text{class size} + c_2 \backslash\% \text{ disadvantaged} + e$

avgmath	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
classize	.0716782	.0301848	2.37	0.018	.0124816	.1308748
disadvantaged	-.3395788	.014675	-23.14	0.000	-.3683585	-.3107991
_cons	69.94438	1.012486	69.08	0.000	67.95875	71.93001

Variable omise & classize: $\backslash\% \text{ disadvantaged} = d_0 + d_1 \text{class size} + e$

disadvanta~d	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
classize	-.7238744	.0428693	-16.89	0.000	-.8079472	-.6398017
_cons	35.781	1.313502	27.24	0.000	33.20504	38.35696

Bias de variable omise en pratique

Régression univariée : $\text{average math score} = b_0 + b_1 \text{class size} + e$

Régression multivariée :
$\text{average math score} = c_0 + c_1 \text{class size} + c_2 \backslash\% \text{ disadvantaged} + e$

Variable omise & classize: $\backslash\% \text{ disadvantaged} = d_0 + d_1 \text{class size} + e$

On obtient :

$$b_1 = 0.317 = \underbrace{0.072}_{c_1} + \underbrace{(-0.34)}_{c_2} \times \underbrace{(-0.724)}_{d_1} = c_1 + OVB$$

Le biais est positif : on avait initialement surestimé l'effet de la taille de la classe.

Question de compréhension [groupe de 2]

06:00

Une chercheuse estime l'effet du revenu du ménage sur la consommation de biens culturels Y, mesurée par le montant annuel dépensé en cinéma, musées, concerts.

Elle estime le modèle suivant : $Y_i = \beta_0 + \beta_1 Revenu_i + e_i$ et trouve que $\beta_1 > 0$: les ménages plus riches consomment davantage de biens culturels.

On soupçonne cependant que le modèle souffre d'un biais de variable omise.
Proposez une variable omise, et proposez :

1. Le lien attendu entre la variable omise et le revenu (+/-)
2. Le lien attendu entre la variable omise et la consommation culturelle (+/-)
3. En déduisez le signe du biais sur l'estimateur de β_1 (positif ou négatif).

lien

R^2 Ajusté

Concept pas fondamentalement important en soi, mais tellement utilisé qu'il faut absolument le connaître.

- Par construction, le R^2 augmentera toujours lorsqu'un nouveau régresseur est ajouté à la régression.
- Le R^2 ajusté impose une pénalité pour l'ajout de régresseurs au modèle.

$$R^2_{\text{Ajusté}} = 1 - \frac{n-1}{n-k-1}\frac{SCR}{SCT} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

Où

- k est le nombre de régresseurs
- n est le nombre d'observations

[Propriétés] R^2 Ajusté

$$R^2_{\text{Ajusté}} = 1 - \frac{n-1}{n-k-1}\frac{SCR}{SCT} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

1. $R^2_{\text{Ajusté}} < R^2$: le R^2 ajusté est toujours inférieur ou égal au R^2.
   - Car $\frac{n-1}{n-k-1} > 1$.
2. Ajouter une variable au modèle a 2 effets sur $R^2_{Ajusté}$:
   - SCR doit baisser, ce qui fait augmenter $R^2_{Ajusté}$
   - $\frac{n-1}{n-k-1}$ augmente, ce qui fait baisser $R^2_{Ajusté} \rightarrow$ l'effet net est ambigu
3. Le $R^2_{\text{Ajusté}}$ peut être négatif.

Exercice 3: Recap

10:00

On utilise les données STAR pour illustrer les points précédents.

1. Sélectionner les élèves de 2eme année (`grade`).

2. Regresser `math` sur `school` (use `tab school` avant pour regarder le contenu de la variable). Interpreter les coefficients. Quelle est la catégorie de référence ? Les résultats sont-ils surprenants ? Quelle pourrait être une variable omise ?

3. Calculez la part d'étudiants bénéficiant d'un repas gratuit (i.e `lunch` est égal à "free") par `school`. Qu'observez-vous ? Ajoutez `lunch` à la régression de la question précédente. Comment les coefficients changent-ils ?

4. Régresser `math` sur `star_num`. Interpréter les coefficients, puis regresser `math` sur `star`, `gender`, `ethnicity`, `lunch`, `degree`, `experience` et `school`. Rappelons nous que c'est une expérimentation aléatoire. Peut on dire que la randomisation a été bien faite ?

5. Quel est le R^2 ajusté de la régression multiple précédente ? Comment l'interprétez vous ? Que pouvez-vous en déduire quant à l'importance des caractéristiques observables des individus, des enseignants et des écoles dans l'explication des résultats scolaires ?

Où en sommes nous de notre quête de la causalité

✅ Comment gérer les données? Lisez-les, ordonnez-les, visualisez-les...

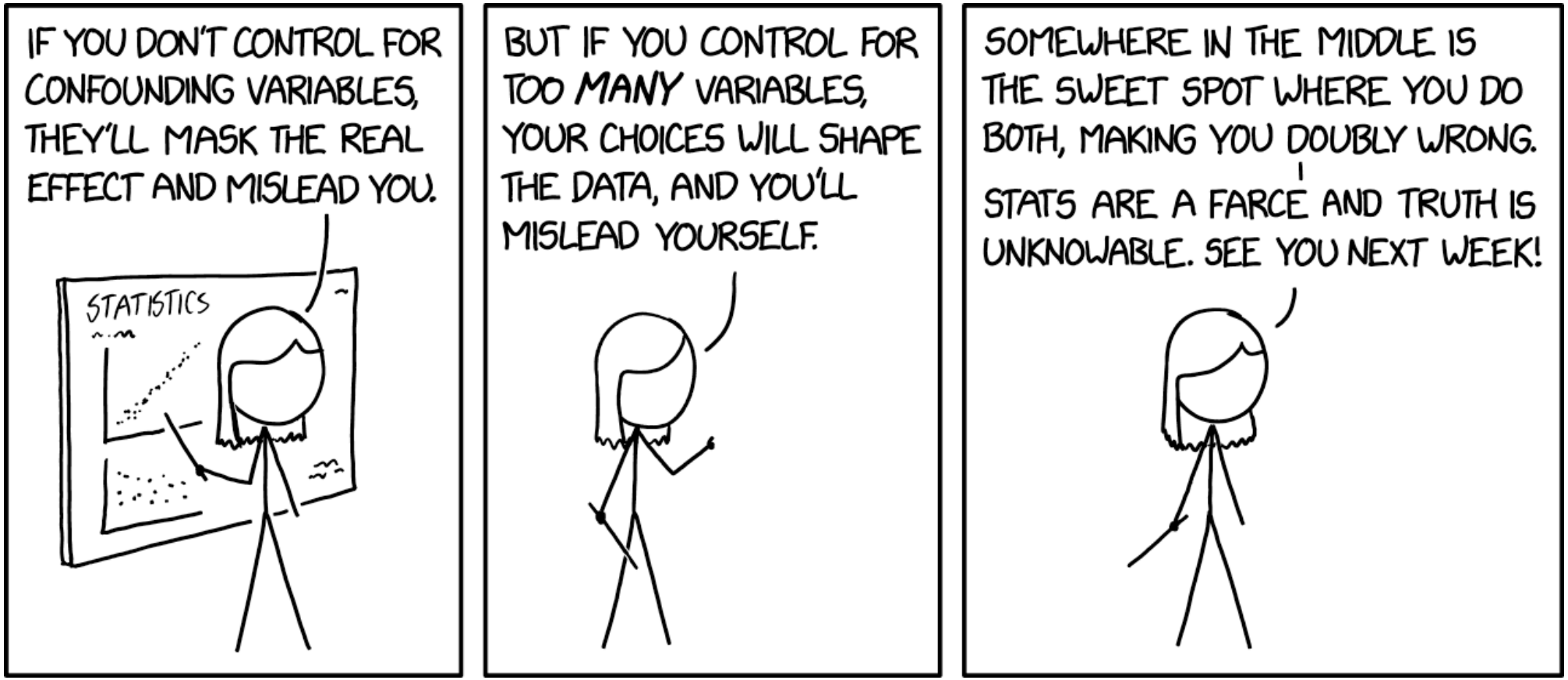
🚧 **Comment résumer une relation entre plusieurs variables?** Régression linéaire univariée et multivariée... *to be continued*

✅ Qu'est ce que la causalité ?

❌ Comment faire si nous n'observons qu'une partie de la population ?

❌ Nos résultats sont ils uniquement dus au hasard?

❌ Comment trouver de l'exogénéité en pratique ?


IF YOU DON'T CONTROL FOR CONFOUNDING VARIABLES, THEY'LL MASK THE REAL EFFECT AND MISLEAD YOU.
STATISTICS
BUT IF YOU CONTROL FOR TOO MANY VARIABLES, YOUR CHOICES WILL SHAPE THE DATA, AND YOU'LL MISLEAD YOURSELF.
SOMEWHERE IN THE MIDDLE IS THE SWEET SPOT WHERE YOU DO BOTH, MAKING YOU DOUBLY WRONG.
STATS ARE A FARCE AND TRUTH IS UNKNOWABLE. SEE YOU NEXT WEEK!

À LA SEMAINE PROCHAINE !

mguillot@uliege.be

MERCI À

Florian Oswald et à toute l'équipe de ScPoEconometrics pour le livre et leurs ressources