

$$PROGRES = a + b * RURAL + c * T35PL + e$$

où a , b et c sont trois coefficients qu'il faut estimer, e est un terme d'erreur qui montre que la relation n'est pas déterministe, mais qu'on essaie de minimiser.

L'objectif des calculs que la régression multivariée effectue, c'est de déterminer les valeurs de b et de c qui permettent de prédire le *PROGRES* des élèves le mieux possible lorsqu'on sait pour chaque élève s'il étudie en milieu rural et s'il est dans une classe de plus de 35 élèves ou non.

Peu important ici les calculs. Ce qui nous intéresse, c'est de montrer le type d'information utilisée par ces calculs pour identifier ce qui vient du milieu rural et ce qui vient de la taille de la classe. Fondamentalement, la régression multivariée utilise l'existence de quatre catégories d'élèves figurées dans le tableau suivant (en prenant l'exemple de la Côte d'Ivoire en CM1) :

	Milieu rural	Milieu urbain	Tous milieux confondus
Classe de moins de 35 élèves	Groupe A 832 élèves Progression moyenne : -0,13 points	Groupe B 249 élèves Progression moyenne : +0,07 points	Groupe A+B 1081 élèves Progression moyenne : -0,08 points
Classe de plus de 35 élèves	Groupe C 233 élèves Progression moyenne : -0,03 points	Groupe D 705 élèves Progression moyenne : +0,06 points	Groupe C+D 938 élèves Progression moyenne : +0,04 points

La dernière colonne est là seulement pour mémoire. Elle rappelle que, si on néglige la dimension géographique rural/urbain, l'estimation de l'effet des classes est $+0,04 - (-0,08) = +0,12$. C'est l'estimation qui serait donnée par la régression simple du progrès des élèves sur la taille de la classe, car la régression simple équivaut dans ce cas à une simple comparaison de moyennes.

Tout l'intérêt de cet exemple est de montrer pourquoi la régression multiple donne un résultat sensiblement différent, et pourquoi c'est ce résultat qu'il faut préférer.

Comment, en effet, neutraliser l'influence du milieu environnant dans notre comparaison des classes de moins de 35 élèves et de plus de 35 élèves ? Le plus simple est d'effectuer des comparaisons distinctes dans chacun des deux milieux. En milieu rural, apprend-on plus vite dans les classes plus nombreuses ? La réponse est obtenue en comparant les groupes A et C. Les élèves des classes plus nombreuses progressent plus vite de 0,10 points (en effet : $-0,03 - (-0,13) = +0,10$). Et en milieu urbain, apprend-on plus vite dans les classes plus nombreuses ? Cette fois, il s'agit de comparer les groupes B et D. Les élèves des classes plus nombreuses progressent légèrement moins vite, de -0,01 points ($0,06 - 0,07 = -0,01$).

De la même façon, comment neutraliser l'influence de la taille des classes pour mesurer l'effet du milieu environnant ? On compare les groupes A et B, où les classes sont toujours de moins de 35 élèves, et on constate un avantage de 0,20 points pour le milieu urbain ($0,07 - (-0,13) = +0,20$). Dans les classes de plus de 35 élèves (groupes C et D), l'avantage est légèrement moindre : +0,09 (en effet, $0,06 - (-0,03) = +0,09$).

Ce que nous cherchons à estimer, c'est l'effet moyen des classes de plus de 35 élèves. C'est nécessairement une moyenne pondérée de l'effet mesuré en milieu rural et l'effet mesuré en milieu urbain. Il est donc forcément situé entre +0,10 et -0,01. Or que constate-t-on ? L'estimation obtenue sans tenir compte du milieu environnant (+0,12) est en dehors de cette fourchette ! On voit bien ce que cette estimation par simple comparaison de moyennes a d'insuffisant : on risque d'affirmer que les classes de plus de 35 élèves font progresser les élèves de 0,12 points de plus, alors que ni en milieu rural, ni en milieu urbain, on n'a constaté une telle différence ! L'estimation de +0,12 est en fait entraînée par un « effet de contexte » qui vient fausser la comparaison.

La régression multivariée, elle, utilise bien l'information détaillée contenue dans les cases A, B, C et D. C'est le principe de son fonctionnement : elle s'appuie sur la variété des situations existantes pour attribuer simultanément tous leurs effets aux différentes variables. Le résultat est le suivant :

Source	SS	df	MS	Number of obs = 2019		
Model	14.7921955	2	7.39609773	F(2, 2016)	=	9.92
Residual	1503.71035	2016	.745888071	Prob > F	=	0.0001
				R-squared	=	0.0097
Total	1518.50255	2018	.752478963	Adj R-squared	=	0.0088
				Root MSE	=	.86365

PROGRES	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
T35PLUS	.0441819	.0451415	0.979	0.328	-.044347	.13271
RURAL	-.1442731	.0450964	-3.199	0.001	-.2327135	-.05583
_cons	.0289829	.0435281	0.666	0.506	-.0563819	.11434

L'effet estimé d'une classe de plus de 35 élèves vaut +0,04 points. Il s'agit bien d'une valeur moyenne entre les effets mesurés d'une part en milieu rural (+0,10) et l'effet mesuré en milieu urbain (-0,01). Cette estimation est donc nettement préférable à l'estimation de +0,12.

Résumons la différence. La simple comparaison de moyennes mêle en fait deux effets : l'effet du milieu environnant et l'effet de la taille de classe. La régression multiple procède, elle, à des comparaisons toutes choses égales par ailleurs : elle compare les progressions en milieu rural d'un côté, en milieu urbain de l'autre. Elle fournit ensuite une estimation de l'effet moyen de la taille des classes sur l'ensemble des deux milieux possibles. Ce faisant, elle isole bien un effet dû à la taille de la classe, et non au milieu géographique.

En résumé, la régression multivariée est une technique pour tirer parti de la variété qui existe au sein de notre échantillon. Si on croise les multiples dimensions possibles (rural/urbain, classes nombreuses/peu nombreuses, maîtres de diplômes variés,...), on a autant de catégories possibles, dont une bonne partie sont représentées dans notre échantillon. En supposant que les effets de toutes ces dimensions s'ajoutent, la régression multivariée propose une identification simultanée de tous ces effets de façon à maximiser le pouvoir prédictif de notre modèle.

Remarque : identification des effets et problèmes de collinéarité approchée :

On l'a vu, milieu et taille de classe sont liés. Les classes nombreuses sont souvent en milieu urbain, et les classes moins nombreuses en milieu rural. (Dans le tableau, cela apparaît dans le fait que les groupes A et D sont plus nombreux que les groupes B et C).

Mais le lien entre ces deux variables n'est pas parfait : il existe des classes peu nombreuses en milieu urbain, et des classes nombreuses en milieu rural. Heureusement, car c'est là-dessus qu'a reposé notre identification des effets! En particulier, pour mesurer l'effet de la taille des classes, on était bien content de pouvoir utiliser le groupe C (les classes nombreuses en milieu rural) pour le comparer au groupe A et le groupe B (les classes peu nombreuses en milieu urbain) pour le comparer au groupe D. Sans ces classes des groupes B et D, l'identification des effets aurait été impossible. Toutes les classes nombreuses auraient été en milieu urbain, et toutes les classes peu nombreuses en milieu rural, et les deux choses auraient été indiscernables. Ici, les groupes B et C sont relativement importants, et on peut donc faire des comparaisons assez fiables. Il n'y a pas de problème de collinéarité approchée.

On parle de **collinéarité approchée** lorsque deux variables du modèle sont fortement corrélées (ou fortement collinéaires), c'est-à-dire lorsque les catégories qu'elles définissent se recoupent très largement (mais pas complètement). Par exemple, l'âge du maître et le nombre d'années depuis lequel il enseigne risquent d'être fortement collinéaires. Ou encore, le nombre d'années de formation professionnelle reçues par le maître et le type de formation seront très fortement corrélés.

Tant que la corrélation n'est pas totale, cependant, la régression multiple peut fonctionner et donner des résultats numériques. C'est comme si elle s'appuyait sur des groupes de comparaison C et B comprenant chacun quelques élèves : l'identification est possible, mais devient de plus en plus fragile puisqu'elle dépend d'un tout petit nombre d'individus. Lorsque la collinéarité approchée est vraiment forte, cependant, les estimations n'ont plus de sens : elles dépendent trop étroitement d'un petit nombre d'individus spécifiques, d'une erreur de mesure,...

Une conséquence de la collinéarité approchée est que les effets ne peuvent plus être estimés de façon fiable. Il ne faut pas demander l'impossible : si les données ne permettent pas de bien distinguer âge du maître et nombre d'années d'enseignement effectuées, vouloir les introduire tous les deux dans la régression va rendre les résultats hyper-sensibles à un petit nombre de maîtres (ceux qui se sont mis à enseigner à un âge avancé, par exemple), et le risque est que les coefficients sur l'âge du maître comme sur son expérience prennent des valeurs très peu fiables.

Pour éviter ce problème de collinéarité approchée, il faut prendre conscience du lien entre certaines variables explicatives, et n'en garder qu'une lorsque ce lien est trop fort. Mais plus généralement, les limites des possibilités d'identification conduisent à recommander de **limiter le nombre de variables du modèle, en ne gardant que celles dont le rôle est a priori important et qui sont bien distinctes deux à deux.**

Revenons à notre problème d'évaluation des facteurs des apprentissages scolaires. Comment utiliser la régression multivariée ? On peut proposer une démarche en deux étapes :

1. Etape 7 : la sélection des variables du modèle :

Dans cet étape, le fil directeur est donné a priori par le modèle conceptuel présenté au chapitre 2. On a une idée préalable des facteurs qui expliquent les apprentissages scolaires, et on introduira les variables nécessaires pour tester si ces effets existent de façon significative, et quelle est leur valeur. On introduira les variables nécessaires, ni plus, ni moins :

- pas moins, car oublier une variable pertinente c'est se heurter à l'objection de la variable manquante qui vient fausser l'estimation des effets des autres variables. Ce qui précède a pu persuader de l'importance du problème.
- pas plus, car ce serait demander trop des données existantes que de permettre d'identifier simultanément les effets de dizaines de variables à partir de configurations existantes forcément en nombre limité. En particulier, quand deux variables visent le même concept ou semblent étroitement liées, on n'introduira que l'une des deux dans le modèle¹ : *par exemple, on choisira soit NIVEAUVI, soit RICHESSE (toutes les deux sont des indicateurs du niveau de vie, en se référant à la possession ou non de respectivement 3 et 13 objets).*

Concrètement, le modèle se présente comme une grande équation où figurent, à gauche, la progression à expliquer, et à droites les variables explicatives : score initial (pour tenir compte des effets de rappel ou de régression à la moyenne), variables contextuelles et variables de politique éducative.

On part d'un modèle préalable où interviennent, par exemple, huit variables contextuelles et seize variables de politique éducative. On demande au logi-

¹ Voir pour plus de détail l'encadré ci-dessus et la remarque sur la collinéarité approchée.

ciel statistique dont on dispose d'effectuer la régression de la progression de l'élève sur le score initial et ces dix-huit variables, selon la méthode des moindres carrés ordinaires¹. Le résultat obtenu en Côte d'Ivoire au CM1 est le suivant :

Source	SS	df	MS	Number of obs = 2019		
Model	429.62223	26	16.5239319	F(26, 1992) =	30.23	
Residual	1088.88032	1992	.546626665	Prob > F =	0.0000	
				R-squared =	0.2829	
				Adj R-squared =	0.2736	
Total	1518.50255	2018	.752478963	Root MSE =	.73934	

PROGRES	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
STINI5FM	-.4135175	.0180611	-22.895	0.000	-.4489381	-.3780968
FILLE	.0383344	.0340661	1.125	0.261	-.0284745	.1051432
AGEMOINS	-.0204688	.0588729	-0.348	0.728	-.1359278	.0949902
AGEPLUS	-.062001	.0372578	-1.664	0.096	-.1350693	.0110674
NIVEAUVI	.0681624	.0221777	3.073	0.002	.0246685	.1116564
REDOUBLE	-.2032126	.03878	-5.240	0.000	-.2792662	-.127159
DEJEUNER	.1056927	.0456052	2.318	0.021	.0162538	.1951316
LIV_FR_1	.0082163	.0719368	0.114	0.909	-.132863	.1492955
LIV_FRMT	.0167746	.0515525	0.325	0.745	-.0843279	.1178771
AIDE	.0572849	.0355844	1.610	0.108	-.0125016	.1270714
DOMFRANC	.0548826	.0355311	1.545	0.123	-.0147994	.1245646
SERVICE	-.0106939	.0046907	-2.280	0.023	-.0198931	-.0014948
MAITRFEM	-.2419585	.1101303	-2.197	0.028	-.4579411	-.0259758
NIVCYCLB	.0889051	.044332	2.005	0.045	.0019632	.1758469
DIPCYCLB	-.2403068	.0669545	-3.589	0.000	-.3716149	-.1089986
FPI1_3M	-.2691863	.0905656	-2.972	0.003	-.4467996	-.0915729
FPI6M	.1667868	.1265257	1.318	0.188	-.0813498	.4149234
FPI1AN	.0904249	.0807612	1.120	0.263	-.0679604	.2488102
FPI1ANPL	.2904415	.0776921	3.738	0.000	.1380753	.4428077
FORCON1	-.2484907	.050796	-4.892	0.000	-.3481095	-.1488719
FORCONPL	.0798364	.0410005	1.947	0.052	-.0005721	.1602448
DOUBLFLX	-.3818666	.0896833	-4.258	0.000	-.5577496	-.2059837
MULTIGRA	-.0119611	.0945516	-0.127	0.899	-.1973915	.1734693
T35PLUS	-.0600618	.0446154	-1.346	0.178	-.1475596	.0274359
ELECLASS	-.0691888	.0403977	-1.713	0.087	-.1484149	.0100374
RURAL	-.2794755	.0485941	-5.751	0.000	-.374776	-.1841749
_cons	.2331635	.1218173	1.914	0.056	-.005739	.4720661

A partir de ce premier modèle, diverses attitudes sont possibles. Certains sont tentés d'introduire d'autres variables, d'essayer d'autres formulations (par exemple, distinguer trois catégories de taille de classe plutôt que deux,...). Le problème est que, si l'on ne se restreint pas et si on essaie des dizaines de spécification, on finira toujours par arriver au résultat qu'on sou-

¹ Ce n'est pas le meilleur modèle statistique. On devrait tenir compte de l'existence de deux niveaux de données (les niveaux élève et classe) au moyen de modèles des moindres carrés généralisés ou de modèles multiniveaux. Nous ne les présentons pas ici en raison de leur plus grande technicité. Mais il faut les utiliser si on le peut. Ils permettront de constater que les effets de niveau classe sont estimés moins précisément que ne le laissait croire, de façon illusoire, le modèle des moindres carrés ordinaires.

haite... mais on n'aura rien appris. L'essentiel est donc de respecter une certaine déontologie. Le modèle, pour l'essentiel, est acquis dès ce stade là. On essaie d'en tirer les enseignements, avec prudence mais sans rejeter ce qui ne correspond pas aux attentes!

Ce qu'on va faire, c'est vérifier simplement qu'il n'y a pas de problème technique. Si le coefficient d'une variable semble contre-intuitif, on va vérifier la construction de cette variable, voir si son effet peut vraiment être identifié (par exemple, si on a introduit l'âge du maître et son expérience et qu'on a trouvé un effet très négatif à l'expérience et très positif à l'âge, cela peut signifier que les données n'ont pas vraiment permis d'identifier séparément les deux effets, et donnent des résultats sans signification. Dans ce cas, on ôtera la variable âge, pour garder seulement la variable expérience). On peut également vérifier que les différentes catégories distinguées reposent sur un nombre suffisant d'observations : si une catégorie n'est représentée que par moins de cinq classes, l'effet correspond peut-être à des traits particuliers de ces cinq classes, comme par exemple à un charisme spécifique de leurs maîtres,...

Le but en tout cas n'est pas d'introduire n'importe quelle variable pour maximiser le pouvoir prédictif du modèle (le fameux R^2)¹. Le but est de bien tester un modèle raisonnable de prédiction de la progression des élèves.

2. Etape 8 : la lecture du modèle statistique :

On a lu l'information donnée sur l'effet d'une variable par le modèle statistique lorsqu'on parvient à un énoncé de la forme suivante : «Une modification de *tant* de la valeur prise par telle variable a un effet moyen positif (ou négatif) de *tant* sur les résultats des élèves, cet effet étant mesuré avec une précision de *tant*.»

¹ Le R^2 , ou coefficient de détermination, est la part de variance expliquée par les variables du modèle. Il décrit le pouvoir prédictif du modèle. Plus il est élevé, plus le modèle permet de prédire précisément la progression qu'on peut attendre des élèves en fonction des conditions d'enseignement. Le R^2 augmente forcément avec le nombre de variables qu'on introduit dans le modèle. Mais introduire trop de variables n'est pas bon, car on se heurte à des problèmes de collinéarité approchée.

Cela suppose d'abord de choisir l'unité de mesure des résultats des élèves (la variable expliquée). On peut prendre tout simplement la note obtenue au test, par exemple en la ramenant à une note sur 100. Cet échelle étant quelque peu arbitraire, on choisit souvent pour unité l'écart type de la distribution des scores. On a vu que cette statistique mesure la dispersion moyenne des scores autour de la moyenne. C'est une unité commode qui permet des comparaisons entre études, puisque cette unité peut être adoptée quel que soit le test. L'unité de mesure de la variable explicative découle de sa construction. Par exemple, la variable SERVICE, qui mesure l'ancienneté du maître se lit en années. On dira : «l'augmentation de l'expérience du maître d'une année a pour effet...». Cas particulier, quand il s'agit de variable binaire (oui/non), on ne dira plus «l'augmentation de la variable a pour effet» mais «le passage de l'état *non* à l'état *oui*».

Nous disposons du nécessaire pour une première lecture de la colonne des coefficients (notée «coef»). Par exemple :

- *L'augmentation de l'expérience du maître d'une année a pour effet moyen de diminuer les résultats de ses élèves de 1% d'écart type (coefficient de -0,01).*
- *Le passage du milieu urbain au milieu rural a pour effet moyen de diminuer les résultats des élèves de 28% d'écart type (coefficient de -0,28).*

Mais une lecture plus complète doit prendre en compte la précision avec laquelle les effets ont été estimés. C'est poser la question de la significativité statistique des coefficients de la régression.

La significativité en statistiques

En statistiques, les estimations ne conduisent jamais à une valeur sûre du paramètre qu'on veut estimer. La valeur retenue est toujours la valeur la plus probable, mais on sait que la mesure ou l'estimation est imparfaite.

Il convient dès lors de ne pas considérer simplement l'estimation la plus probable de la vraie valeur, mais de donner une fourchette dans laquelle on peut garantir, par exemple à 95%, que la vraie valeur se trouve ; c'est cette fourchette qu'on appelle l'**intervalle de confiance**. Dans l'exemple de résultat de régression présenté ci-dessus, ce sont les deux colonnes de droite. Elles signifient que la vraie valeur du paramètre étudié a 95% de chances de se situer, au vu de l'information disponible, entre la valeur de gauche

(borne inférieure) et la valeur de droite (borne supérieure). Les intervalles de confiance sont de lecture assez simple ; c'est peut-être le premier point de repère à utiliser lorsqu'on veut juger de la précision avec laquelle une variable est estimée.

Quand dira-t-on qu'une variable est «**significative à 5%**» ? En fait, cela signifie simplement qu'on a 95% de chances de ne pas se tromper en affirmant que l'effet de la variable est différent de 0. C'est le cas lorsque la statistique de Student (ici la colonne intitulée «t») vaut plus de 2 en valeur absolue.

La théorie ne dit pas quel seuil de significativité il faut retenir (5% de chances de se tromper, est-ce beaucoup ou peu ?). Le choix du seuil repose plutôt sur une convention (on retient souvent les seuils de 5 ou 10%).

Plus généralement, la **probabilité limite** nous permet de préciser exactement le risque de se tromper en considérant que l'effet est nul. Cette probabilité est donnée ici dans la colonne «P>|t|». Par exemple, la valeur estimée du coefficient «RURAL» d'une part, et la précision de l'estimation d'autre part, rendent très improbable que la vraie valeur puisse être 0 (probabilité limite inférieure à un pour mille : $Pr=0,000$). Par contre, le coefficient estimé pour la variable FPI1AN n'est pas assez précisément estimé pour qu'on puisse rejeter l'hypothèse d'un coefficient nul. Le risque de se tromper serait assez élevé ($Pr=0,26$).

Le plus intéressant, d'ailleurs, n'est pas toujours de dire qu'on est sûr que l'effet est différent de 0. Un effet non nul peut être pourtant négligeable ; simplement, les données auront permis de l'estimer très précisément et on est sûr que l'effet est petit mais différent de 0. Par conséquent, c'est bien l'ampleur de l'effet et la précision de son estimation qui sont, conjointement, nécessaires à l'interprétation. C'est pourquoi il nous paraît recommandé de **prendre pour repère l'intervalle de confiance**.

La lecture complète du résultat est alors la suivante :

· *L'augmentation de l'expérience du maître d'une année a pour effet moyen le plus probable de diminuer les résultats de ses élèves de 1% d'écart type (coefficient de -0,01). On est sûr à plus de 95% que l'effet est différent de 0 ; et on est sûr à 95% que l'ampleur de l'effet se situe entre -1,99% et -0,15%.*

- *Le passage du milieu urbain au milieu rural a pour effet moyen le plus probable de diminuer les résultats des élèves de 28% d'écart type (coefficient de -0,28). On est sûr à plus de 95% que l'effet est différent de 0, et on est sûr à 95% que l'ampleur de l'effet se situe entre -37,5% et -18,4%.*

Qu'est devenu notre problème de l'évaluation de l'effet des classes nombreuses ?

On l'a vu, la comparaison bivariée faisait apparaître de meilleurs résultats dans les classes de plus de 35 élèves. Mais la régression multivariée conduit à dire que les résultats y sont plutôt moins bons (ils ont 95% de chances de se trouver entre 15% d'écart type en dessous et 3% d'écart type au dessus, comme permet de le lire l'intervalle de confiance, et le plus probable est qu'ils se situent 6% d'écart type en dessous). Autrement dit, il n'y a pas d'effet significatif de la taille de classe en Côte d'Ivoire selon les seuils de significativité généralement retenus, mais il est en tout cas presque certain qu'il n'y a pas un effet positif. En particulier, l'hypothèse d'un effet de +0,12 (ou +12% d'écart type), hypothèse qui découle de la simple comparaison bivariée, est ici rejetée au terme de la régression multivariée.

Comment en est-on arrivé là ? Comme nous l'avions senti, il se trouve que les classes plus nombreuses ont d'autres caractéristiques qui induisent des effets favorables sur les résultats des élèves. Une fois décomptés tous ces effets, reste l'effet net des classes plus nombreuses : celui-là est plutôt négatif.

Pour récapituler tout cela, on peut décomposer la différence de progrès entre classes de plus de 35 élèves et classes de moins de 35 élèves en deux types d'effets :

- *Le premier type, ce sont les effets de contexte : les caractéristiques de ces deux catégories de classes diffèrent par d'autres choses que par la taille des classes. Ces différences de contexte se traduisent par des différences de progrès.*
- *Le second type, c'est l'effet propre de la taille de classe.*

Ces deux types d'effets vont s'additionner pour donner la différence de progrès moyen entre classes de moins de 35 élèves et classes de plus de 35 élèves.

Décomposition de la différence de résultats entre classes nombreuses et classes peu nombreuses.

Différence de progrès allant avec une classe plus nombreuse :

	progrès des classes peu nombreuses	progrès des classes nombreuses	Différence de Progrès
PROGRES	-0,08	0,04	0,12

Premier type d'effets : effets de contexte allant avec une classe plus nombreuse :

	contexte des classes peu nombreuses (A)	contexte des classes nombreuses (B)	Différence de contexte (B-A)	effet du contexte (k)	Effet de la différence de contexte (k*(B-A))
STINI5FM	-0,08	0,13	0,21	-0,41	-0,09
FILLE	39%	44%	6%	0,04	0,00
AGEMOINS	10%	11%	1%	-0,02	0,00
AGEPLUS	48%	40%	-8%	-0,06	0,00
NIVEAUVI	39%	83%	43%	0,07	0,03
REDOUBLE	68%	66%	-3%	-0,20	0,01
DEJEUNER	80%	86%	6%	0,11	0,01
LIV_FR_1	11%	8%	-3%	0,01	0,00
LIV_FRMT	75%	82%	7%	0,02	0,00
AIDE	57%	68%	11%	0,06	0,01
DOMFRANC	33%	44%	11%	0,05	0,01
SERVICE	12,74	13,86	1,12	-0,01	-0,01
MAITRFEM	2%	4%	2%	-0,24	-0,01
NIVCYCLB	36%	37%	1%	0,09	0,00
DIPCYCLB	33%	33%	0%	-0,24	0,00
FPI1_3M	15%	3%	-12%	-0,27	0,03
FPI6M	0%	6%	6%	0,17	0,01
FPI1AN	50%	35%	-15%	0,09	-0,01
FPI1ANPL	28%	47%	19%	0,29	0,05
FORCON1	16%	12%	-5%	-0,25	0,01
FORCONPL	22%	35%	13%	0,08	0,01
DOUBLFLX	3%	6%	3%	-0,38	-0,01
MULTIGRA	6%	1%	-5%	-0,01	0,00
ELECLASS	21%	48%	28%	-0,07	-0,02
RURAL	77%	25%	-52%	-0,28	0,15
TOTAL DES EFFETS DE CONTEXTE					0,18

Second type d'effets : effet propre d'une classe plus nombreuse :

T35PL	-0,06	-0,06
-------	-------	-------

Note : les différentes variables utilisées sont présentées en annexe 1.

Le tableau ci-dessus donne la décomposition de cette différence de progrès. Pour comprendre la logique de ce tableau, commençons par la colonne de droite : la première ligne donne la différence de progrès total observée entre classes de plus de 35 élèves et classes de moins de 35 élèves : +12% d'écart type. Les deux dernières lignes donnent la décomposition de cette différence en deux composantes : le contexte dans son ensemble a un effet favorable de +18% pour les classes plus nombreuses ; mais la taille de ces classes en elle-même est un handicap de -6%.

Le tableau détaille également les différents effets de contexte. Pour qu'il y ait un effet de contexte qui aille avec la taille de classe, deux conditions doivent être réunies :

- 1. Les classes de moins de 35 élèves et les classes de plus de 35 élèves diffèrent en moyenne pour la caractéristique considérée (c'est la différence B-A, sur la quatrième colonne);*
- 2. Cette caractéristique a un effet causal sur les progrès des élèves (c'est le coefficient k de la cinquième colonne).*

L'effet de contexte est le produit de ces deux choses.

Pour bien comprendre pourquoi il s'agit du produit, prenons l'exemple de la proportion de filles dans les deux groupes de classes.

Prenons pour référence des classes qui auraient 100% de garçons, et supposons que leur score moyen est de 0. Par rapport à de telles classes, les classes de moins de 35 élèves ont 39% de filles en plus. Or chaque fille progresse de 0,04 écart type de plus qu'un garçon. En moyenne, les classes de moins de 35 élèves gagnent donc $0,04 \times (39/100)$ écart type sur les classes de référence (qui n'ont que des garçons). Les classes de plus de 35 élèves ont, elles, 44% de filles. Leur progression moyenne est donc de $0,04 \times (44/100)$ plus élevée que les classes de référence.

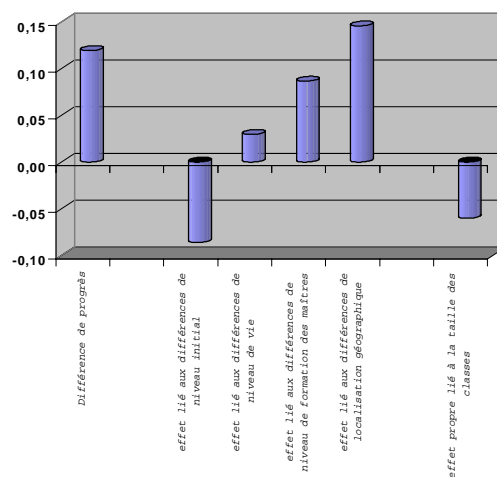
Si on compare les classes de moins de 35 élèves et de plus de 35 élèves maintenant, les premières gagnent $0,04 \times (39/100)$ et les secondes gagnent $0,04 \times (44/100)$. La différence entre les deux qui s'explique par la composition garçons/filles est donc de $0,04 \times (44/100 - 39/100)$.

On retrouve bien la formule du tableau : $k \times (B-A)$ où B-A vaut 44%-39% et k vaut 0,04. La ligne dans son ensemble se résume de la façon suivante : les classes nombreuses ont 6% de filles de plus que les classes peu nombreuses ; or les filles progressent plus que les garçons de 0,04 écart type. De ce fait, les classes plus nombreuses gagnent $0,04 \times 0,06$ c'est-à-dire près de 0% d'écart type en moyenne (0,24%).



- On peut suivant le même principe lire chaque ligne du tableau. Par exemple :*
- La progression des élèves est ralentie lorsqu'ils partent d'un score élevé (en effet, $k=-0,41$). Or les élèves des classes plus nombreuses partent avec un score moyen plus élevé de 0,21 ($B-A=0,21$). Leur progression est donc en moyenne réduite de -0,09 écart type.*
 - La progression des élèves est ralentie en milieu rural ($k=-0,28$). Or les élèves des classes plus nombreuses sont nettement moins souvent en milieu rural (la différence, $B-A$, est de 52%). De ce fait, leur progression est donc accélérée par rapport aux classes moins nombreuses de 0,15 écart type.*

Ont été portés en grisé les effets de contexte qui pèsent le plus. Ils sont conformes à ce qu'on pouvait anticiper intuitivement, même si l'important ici est qu'on les a quantifiés, afin de pouvoir mesurer l'effet propre des classes : les classes plus nombreuses progressent moins dans la mesure où elles partent d'un niveau initial plus élevé ; mais elles progressent davantage dans la mesure où elles ont un public d'élèves davantage urbain, au niveau de vie moyen plus élevé, et bénéficient de maîtres qui ont reçu des formations plus longues. Ces principaux effets sont portés graphiquement ci-dessous : la différence totale apparaît bien comme la somme des effets de contexte (ici, les quatre principaux d'entre eux) et de l'effet propre de la taille de classe. Le graphique rend aussi manifeste la méprise qu'il y aurait à confondre l'effet causal de la taille des classes (-0,06) et la différence totale apparente (+0,12), puisque ces deux grandeurs sont de signes opposés.



Récapitulatif : l'étude de l'effet de la taille des classes en Côte d'Ivoire en 5^{ème} année :

Dans ce chapitre, nous nous étions donné un objectif concret : mesurer l'effet de la taille des classes sur les progrès des élèves, dans un échantillon de classes de CM1 représentatif de la situation en Côte d'Ivoire. Nous avons dû entamer une démarche complexe qu'il est bon de résumer ici :

- Nous avons d'abord décrit la situation existante en termes de taille de classe en Côte d'Ivoire. Nous avons retenu que la taille des classes varie relativement moins que dans d'autres pays de la sous-région, et qu'il pouvait être intéressant pour l'analyse de distinguer deux groupes de classes représentant à peu près deux moitiés égales : les classes de moins de 35 élèves, et les classes de plus de 35 élèves.
- Nous nous sommes demandés si nous pouvions répondre à notre question des effets de la taille des classes en comparant les progrès moyens des élèves dans ces deux groupes. Concrètement, il est apparu que les élèves progressent sensiblement moins vite dans les classes moins nombreuses (une différence de 12% d'écart type, statistiquement significative à 5% et au delà).
- Cependant, il n'était pas possible d'interpréter cette différence comme causale. On pressentait en particulier que les classes moins nombreuses étaient plus souvent en milieu rural, ce qui devait influencer sur leurs résultats et brouillait par conséquent la comparaison entre classes plus ou moins nombreuses. La description bivariée des variables qualitatives «classes nombreuses» (T35PL) et «milieu rural» (RURAL) nous a permis de vérifier l'existence d'une corrélation en constatant que les classes moins nombreuses sont significativement concentrées en milieu rural.
- Le problème qui se posait à nous était donc nécessairement d'évaluer simultanément les effets causaux de la taille de la classe et du milieu rural. Plus généralement, il nous fallait contrôler pour toute variable qui avait un impact sur les scores des élèves et qui pouvait être corrélée avec la taille de la classe. Il nous fallait donc un modèle complet de prédiction des apprentissages des élèves. Nous avons pour cela utilisé la régression multivariée,

qui a consisté à modéliser simultanément en les tenant pour additifs les effets de 25 variables susceptibles d'influer sur les progrès des élèves.

Nous avons obtenu un effet causal négatif des classes plus nombreuses, même si cet effet n'était pas suffisamment précisément estimé pour être jugé statistiquement significatif.

Mieux, la régression multivariée nous a permis de rendre compte de la comparaison des moyennes des classes de moins de 35 élèves et des classes de plus de 35 élèves. L'avantage final des classes moins nombreuses tient à des effets de contexte ainsi qu'à un effet propre de la taille de la classe. Les effets de contexte défavorisent, dans leur ensemble, les classes moins nombreuses : en particulier, les classes moins nombreuses sont plus souvent en milieu rural, bénéficient moins souvent de maîtres qui ont reçu une longue formation initiale, et ont un public d'élèves moins aisé et tout cela a pour effet de retarder les apprentissages de leurs élèves. Cependant, simultanément, la petite taille de la classe est en soi un facteur favorable. Mais ce facteur favorable ne suffit pas à compenser les handicaps liés au contexte. Cela explique la différence de progression au total défavorable aux élèves des classes moins nombreuses.

- La conclusion de cette étude de l'effet de la taille des classes est double :
 1. Les classes plus nombreuses ont, en moyenne en Côte d'Ivoire, un effet négatif sur les apprentissages des élèves, mais non significatif ; toutes choses égales par ailleurs, augmenter la taille des classes réduit donc les progrès des élèves, mais pas de façon prononcée. Cet effet défavorable doit être pris en compte, même si d'autres objectifs (par exemple d'extension de la scolarisation) peuvent légitimer des classes plus nombreuses dans certains cas.
 2. Cependant, ces élèves des classes plus nombreuses bénéficient en moyenne par ailleurs d'autres atouts qui font qu'ils progressent significativement plus vite que les élèves des classes moins nombreuses. Ils ne constituent donc pas en tant que tel un groupe particulièrement désavantagé qu'il serait nécessaire de cibler dans une politique de remédiation spécifique.

D. Résumé de la méthode d'analyse des données proposée :

Ce chapitre a permis de proposer une méthode d'analyse des données, organisée autour de la régression multivariée. Cette méthode est recommandée lorsqu'il s'agit de mettre en évidence les conditions qui rendent certaines écoles plus efficaces dans le système scolaire étudié. On peut en déduire autant de «leviers» pour la politique éducative, c'est-à-dire autant de variables qui ont un impact causal mesurable sur les apprentissages des élèves.

Mais la régression multiple elle-même n'est qu'une phase de l'analyse, autour de laquelle d'autres phases s'organisent. Il est peut-être bon de résumer ici les différentes étapes de ce travail :

1. La construction des variables : A partir des centaines de variables élémentaires dont on dispose, on crée un nombre restreint de variables (quelques dizaines) pour l'analyse. Les choix, inévitables, se font selon deux critères : la **pertinence** de la variable s'établit par rapport à un modèle théorique de prédiction des progrès des élèves qu'on veut tester et quantifier ; le caractère **opératoire** d'une variable est garanti par une mesure simple fiable sur un nombre suffisant d'individus distincts.

2. La fusion des différents fichiers en un unique fichier d'analyse : en général, l'essentiel de l'analyse se fait au niveau des élèves, puisque ce sont leurs apprentissages qui mesurent l'efficacité d'un système scolaire ; cela nécessite de rassembler trois niveaux d'information (élève, classe et directeur).

3. La vérification des données : Quelques statistiques simples permettent de vérifier un peu plus les données, et d'éliminer éventuellement certains problèmes dus à la collecte des données, au codage ou à la création des variables.

4. L'analyse des scores : L'analyse des résultats aux tests doit tenir compte d'une limite évidente : il n'y a **pas de mesure absolue du niveau des élèves**. Mais deux choses restent possibles. L'une consiste à évaluer le degré auquel certains objectifs définis au préalable sont atteints. L'autre consiste à comparer entre elles des populations distinctes, sur des tests standardisés.

¹ Voir le chapitre 4 qui explore en détail ces questions de représentativité.

5. L'analyse des variables explicatives : La description des variables explicatives permet d'explorer les **moyens mis en œuvre** par le système éducatif étudié. Le seul réel problème qui se pose est celui de la représentativité de l'échantillon¹. En effet, certaines formules pédagogiques ont pu être sur-représentées ; il faut en tenir compte pour une description qui veut refléter la situation du pays.

6. Les descriptions bivariées : La description bivariée ne permet pas de mettre en évidence des relations causales. Mais elle permet de mettre en évidence des **publics cibles**.

7. La sélection des variables du modèle explicatif : La sélection des variables du modèle explicatif est essentiellement commandée par les hypothèses que l'on fait sur le processus d'apprentissage (le modèle théorique). Mais il faut aussi s'interroger sur la robustesse du modèle statistique obtenu en réfléchissant à la formulation des variables, au nombre d'observations distinctes pour chaque variable et aux problèmes d'identification des effets lorsque des variables sont presque collinéaires.

8. La lecture du modèle statistique : La lecture du modèle statistique doit permettre non seulement de quantifier les différents effets, mais de aussi de lire leur significativité statistique.

La difficulté des politiques éducatives n'est pas d'avoir des idées sur ce qui peut améliorer le système éducatif : les idées sur l'école sont légion ! La difficulté est de trouver des cadres de réflexion qui permettent de **tester ces hypothèses**, souvent plausibles mais parfois contradictoires², et de **définir des priorités** parmi les mesures souhaitables, en tenant compte des limitations de ressources.

L'analyse multivariée offre un cadre utile pour une telle réflexion. Elle constitue un effort pour mesurer des relations causales exploitables, en reliant les résultats des élèves aux moyens d'enseignement mis en œuvre. Les questions auxquelles elle s'efforce de répondre sont extrêmement concrètes : quels ni-

¹ Cette question est abordée davantage au chapitre 4 sur la constitution de l'échantillon.

² Un exemple simple : «Les femmes enseignantes sont moins performantes parce qu'elles ont de lourdes charges familiales et sont souvent absentes (congé de maternité,...)» ou «Les femmes enseignantes sont plus performantes parce qu'elles sont davantage proches des jeunes enfants et favorisent la bonne adaptation des filles à l'école». Comment choisir, sinon en comparant *toutes choses égales par ailleurs* les apprentissages dans des classes ou enseignent des hommes et des femmes ?

veaux de recrutement de maîtres et quelles formations se sont-ils avérés les plus performants ? Quel est l'effet des différents modes d'organisation des classes ? C'est un point de départ pour définir des leviers d'action de la politique éducative.

En même temps, la régression multivariée ne permet pas tous les raffinements. Comme toute méthode fondée sur la comparaison, elle exige beaucoup de données : 120 écoles, par exemple, c'est peu pour évaluer simultanément les effets des différents types d'organisation des classes, des caractéristiques du maître, du milieu géographique,... Il n'est par conséquent guère possible de mettre en évidence des interactions et des chaînes de causalité complexes. D'autres approches plus qualitatives permettent des éclairages complémentaires. Entretiens et observation sont des outils utiles pour comprendre les processus à l'œuvre. Mais les résultats obtenus par un modèle statistique constituent un point d'ancrage empirique fort : nous croyons qu'en cherchant à les interpréter (les nuancer, les compléter), l'enquête qualitative sera mieux à même de sortir des discours convenus pour proposer des éclairages nouveaux.

Annexe 1 : variables utilisées dans le modèle en Côte d'Ivoire :

La variable expliquée est la progression aux tests de français et de mathématiques (**PROGRES**) ; elle a été centrée réduite. L'unité de lecture des résultats est donc le % d'écart type.

La première variable explicative est le score initial : **STINI5FM**
 Une variable dit si l'élève a déjà redoublé **REDOUBLE**

Huit variables extra-scolaires disent :

- | | |
|--|-----------------|
| 1. Si l'élève est en retard sur l'âge normal ; | AGEPLUS |
| 2. S'il est en avance ; | AGEMOINS |
| 3. Si le niveau de vie de sa famille est relativement élevé ³ ; | NIVEAUVI |
| 4. Si l'élève mange un repas de midi ; | DEJEUNER |
| 5. Si l'élève reçoit une aide pour ses devoirs à la maison ; | AIDE |
| 6. Si on parle français au domicile de l'élève ; | DOMFRANC |
| 7. Si l'élève est une fille ; | FILLE |
| 8. Si l'élève vit en milieu rural. | RURAL |

Seize variables scolaires disent :

- | | |
|---|-----------------|
| 1. Si l'élève possède un livre de français ; | LIV_FR_1 |
| 2. Si l'élève possède livre de mathématiques et livre de français ; | LIV_FRMT |
| 3. Le nombre d'années d'expérience du maître ; | SERVICE |
| 4. Si le maître est une femme ; | MAITRFEM |
| 5. Si le maître a suivi au moins une année de lycée ; | NIVCYCLB |
| 6. Si le maître est titulaire du baccalauréat ; | DIPCYCLB |
| 7. Si le maître a reçu un à trois mois de formation prof. initiale, | FPI1_3M |
| 8. ou six mois, | FPI6M |
| 9. ou un an, | FPI1AN |
| 10. ou plus d'un an ; | FPI1ANPL |
| 11. Si le maître a bénéficié d'une session de formation continue, | FORCON1 |
| 12. ou de plusieurs ; | FORCONPL |
| 13. Si la classe est à double flux ; | DOUBLFLX |
| 14. Si la classe est multigrade ; | MULTIGRA |
| 15. Si la classe a plus de 35 élèves ; | T35PL |
| 16. Si la classe a l'électricité. | ELECLASS |

³ On a choisi comme critère de niveau de vie le nombre d'objets présents à la maison parmi trois: vidéo, réfrigérateur et voiture. NIVEAUVI peut donc prendre les valeurs 0, 1, 2 ou 3.

Annexe 2 : programme statistique utilisé dans l'exemple suivi
(étude de l'effet de la taille des classes en Côte d'Ivoire en CM1)

Pour ceux qui souhaiteraient refaire les analyses à titre d'exercice à partir des données du CD PASEC, voici le programme qui a été élaboré dans Stata et qui peut servir d'exemple... à améliorer!

```
capture log close
log using C:\CI\stata\log\TCLASSE.log, replace
clearset maxvar 300set matsize 130
use C:\CD_PASEC\CI\stata\dta\CI5.dta

/* *****

I.PRELIMINAIRES :

***** */

/* Le programme commence par examiner la variable de taille de
classe et crée quelques variables utiles */

gen PROGRES=STFIN5FM-STINI5FM
tab TCLASSE
sum TCLASSE
sum TCLASSE, detail
gen T35PLUS=0
replace T35PLUS=1 if TCLASSE>35

/* On crée ensuite une variable NOMISS qui identifie les élèves
pour lesquels on a tous les éléments nécessaires en vue d'une
analyse complète des résultats, c'est-à-dire les élèves pour
lesquels aucune variable de notre modèle n'est manquante, que ce
soit une variable de niveau élève, classe ou école */

gen NOMISS=0
replace NOMISS=1 if
/* *****

REGRESSEURS DE NIVEAU ELEVE

***** */
PROGRES~=. &FILLE~=. &AGEMOINS~=. &AGEPLUS~=. &NIVEAUVI~=. & /*
*/ REDOUBLE~=. &DEJEUNER~=. &LIV_FR_1~=. &LIV_FRMT~=. & /*
```

```

*/  AIDE~=.&DOMFRANC~=.&          /*
*****
REGRESSEURS DE NIVEAU MAITRE
*****
*/  SERVICE~=.&MAITRFEM~=.&NIVCYCLB~=.&DIPCYCLB~=.& /*
*/  FPI1_3M~=.&FPI6M~=.&FPI1AN~=.&FPI1ANPL~=.& /*
*/  FORCON1~=.&FORCONPL~=.&DOUBLFLX~=.&MULTIGRA~=.& /*
*/  T35PLUS~=.&ELECLASS~=.& /*
*****
REGRESSEURS DE NIVEAU ECOLE
*****
RURAL~=.
```



```

/* *****

II. COMPARAISONS DE MOYENNES;
EFFET DE CONTEXTE LIE AU MILIEU RURAL

***** */

/* On commence par la comparaison naïve des progrès sans distin-
guer par milieu géographique */
sort T35PLUS
by T35PLUS: sum PROGRES if NOMISS==1
/* On réalise le test de Student pour voir si la différence de
moyenne est significative */
ttest PROGRES if NOMISS==1, by(T35PLUS)
/* On affine la comparaison de progrès en distinguant milieu
rural et milieu urbain */
by T35PLUS: sum PROGRES if NOMISS==1 & RURAL=1

/* *****

III. REGRESSION POUR TENIR COMPTE DE L'EFFET
DE CONTEXTE RURAL

***** */

reg PROGRES T35PL RURAL if NOMISS==1
```

```

/* *****
      IV. EVALUATION DES DIFFERENCES DE CONTEXTE
      EN FONCTION DE LA TAILLE DES CLASSES
***** */

/* On résume les caractéristiques des classes de moins de 35 et
de plus de 35 élèves dans toutes les dimensions qui nous sem-
blent pouvoir influencer sur les progrès des élèves, et risquent
donc de troubler la comparaison */
by T35PLUS: sum /*
*****
                        REGRESSEURS DE NIVEAU ELEVE
*****
*/ STINI5FM FILLE AGEMOINS AGEPLUS NIVEAUVI REDOUBLE /*
*/ DEJEUNER LIV_FR_1 LIV_FRMT /*
*/ AIDE DOMFRANC /*

*****
                        REGRESSEURS DE NIVEAU MAITRE
*****

*/ SERVICE MAITRFEM NIVCYCLB DIPCYCLB FPI1_3M FPI6M /*
*/ FPI1AN FPI1ANPL FORCON1 FORCONPL DOUBLFLX MULTIGRA /*
*/ TCLASSE ELECLASS /*

*****
                        REGRESSEURS DE NIVEAU ECOLE
*****
*/ RURAL /*
*/ if NOMISS==1

/* *****

      V. AMPLEUR DES EFFETS DE CONTEXTE ET EFFET PROPRE
      DES EFFETS DE CONTEXTE ET EFFET PROPRE
      DE LA TAILLE DES CLASSES
***** */
/* La régression permet de mettre en évidence l'impact causal
des différents contextes, ainsi que de la taille de la classe
elle-même, sur les apprentissages des élèves */

```

```

reg PROGRES /*
*****
                REGRESSEURS DE NIVEAU ELEVE
*****
*/  STINI5FM FILLE AGE MOINS AGE PLUS NIVEAU VI REDOUBLE /*
*/  DEJEUNER LIV_FR_1 LIV_FRMT /*
*/  AIDE DOMFRANC /*

*****
                REGRESSEURS DE NIVEAU MAITRE
*****

*/  SERVICE MAITRFEM NIVCYCLB DIPCYCLB FPI1_3M FPI6M /*
*/  FORCON1 FORCONPL DOUBLFLX MULTIGRA T35PLUS /*
*/  FPI1AN FPI1ANPL ELECLASS /*
*****
                REGRESSEURS DE NIVEAU ECOLE
*****
*/  RURAL if NOMISS==1

log close

```