Une interrogation sur la validité de l'utilisation de « dummy variables » pour prédire la croissance dans l'étude *A Reassessment of the Relationship between Inequality and Growth* (2000)

29 mars 2016

Le consensus sur les liens entre la croissance et les inégalités de revenus semble difficile à obtenir parmi les spécialiste du domaine. Un nombre important de méthodes et de résultats s'opposent (voir par exemple [2] pour une revue de l'état de l'art en 2003). Soucieux de reproduire certaines de ces études pour mieux en saisir les détails, nous nous sommes attardés sur une étude écrite par Kristin Forbes (MIT) en 2000 [1]. L'auteure obtient des résultats en apparente contradiction avec le courant dominant au moment de la publication en se basant pourtant sur les mêmes données. Elle trouve un lien positif entre la croissance et les inégalités alors que les autres études semblaient s'accorder sur une corrélation nulle ou négative. L'originalité de sa méthode réside dans l'utilisation de variables « dummy » *i.e.* une variable qui a une valeur constante pour chaque pays et une pour chaque période afin de se focaliser sur les variations au sein de chaque pays. Si l'hypothèse semble théoriquement correcte, la confrontation aux données nous a amené à douter des résultats que peut produire une telle méthode. L'interrogation pourrait être résumée en les termes suivant :

Chaque pays possédant en moyenne 4 observations, est-ce que l'ajout d'une variable constante pour chaque pays (plus une autre pour chaque période) ne réduit pas très significativement la variance de l'ensemble des observations ? Si tel est le cas, quel crédit peut-on accorder au signe d'un coefficient qui n'expliquerait qu'une part très réduite de la variance ?

1 Résumé des points clés de Forbes 2000 [1]

Cette étude se définit comme en opposition avec les modèles actuels. Alors qu'un consensus semble s'accorder pour un coefficient négatif (« These studies generally find a negative and just-significant coefficient on inequality », p.869), l'auteure met en avant les différents problèmes que rencontrent ces études pour estimer ces coefficients : manque de robustesse des études, données imprécises, problème de variables omises et le fait que les études précédentes se sont focalisées sur les différences entre pays au lieu de s'intéresser à l'impact qu'aurait un changement des inégalités au sein de chaque pays.

Kristin Forbes propose l'utilisation des données Deininger & Squire pour s'assurer d'avoir des données de bonne qualité et d'utiliser des variables « dummy » pour prendre en compte les variables manquantes et s'assurer que l'étude réponde à la question de l'influence des inégalités au sein de chaque pays. Le modèle de régression est le suivant :

$$\begin{aligned} Growth_{i,t} = \beta_1 Inequality_{i,t-1} + \beta_2 Income_{i,t-1} + \beta_3 MaleEducation_{i,t-1} \\ + \beta_4 MaleEducation_{i,t-1} + \beta_5 PPPI_{i,t-1} + \alpha_i + \eta_t + u_{i,t} \end{aligned} \tag{1}$$

Où i fait référence à un pays, t à une période de 5 années et avec : $Growth_{i,t}$ la croissance de i pendant t, $Inequality$ le coefficient de Gini, $Income$ le PIB par habitant, $MaleEducation/FemaleEducation$

le nombre moyen d'année d'étude secondaire chez les hommes/femmes de plus de 25 ans, $PPPI$ le « price level of investment » , α_i la variable constante par pays, η_t la variable constante par période de temps et $u_{i,t}$ le terme d'erreur. La variable $PPPI$ est ajoutée dans cette régression pour mesurer les distorsions du marché : elle mesure comment varient les coûts des investissements (taxes, régulation, corruption, coût de devises étrangères) entre un pays quelconque et les États Unis.

Différentes méthodes d'estimation des coefficients sont ensuite testées et comparées. La difficulté de cette régression réside dans la présence d'un terme retardé : $Growth_{i,t} = Income_{i,t} - Income_{i,t-1}$ que l'on tente d'expliquer avec $Income_{i,t-1}$. L'auteure compare donc plusieurs méthodes et adopte finalement les résultats obtenus par la méthode d'Arellano et Bond basée sur la méthode des moments généralisée. Les résultats obtenus grâce aux différentes méthodes sont résumés en figure 1. Chaque simulation utilise 180 points sur 45 pays soit 4 observations par pays, chiffre qui descend à 3 avec l'utilisation des premières différences dans Arellano et Bond. On voit que le coefficient devant le terme d'inégalités est positif quelque soit la méthode d'estimation utilisée. Ces résultats sont en désaccord avec ceux des études précédentes, l'auteure décide de faire une comparaison méthodique et minutieuse des différences avec les autres études. Elle résume ses résultats dans le tableau présenté en figure 2.

Table 3—Regression Results: Alternate Estimation Techniques

	Five-year periods				Ten-year periods: fixed effects (5)
Estimation method	Fixed effects (1)	Random effects (2)	Chamberlain's π-matrix (3)	Arellano and Bond (4)	
Inequality	0.0036	0.0013	0.0016	0.0013	0.0013
	(0.0015)	(0.0006)	(0.0002)	(0.0006)	(0.0011)
Income	−0.076	0.017	−0.027	−0.047	−0.071
	(0.020)	(0.006)	(0.004)	(0.008)	(0.016)
Male Education	−0.014	0.047	0.018	−0.008	−0.002
	(0.031)	(0.015)	(0.010)	(0.022)	(0.028)
Female Education	0.070	−0.038	0.054	0.074	0.031
	(0.032)	(0.016)	(0.006)	(0.018)	(0.030)
PPP	−0.0008	−0.0009	−0.0013	−0.0013	−0.0003
	(0.0003)	(0.0002)	(0.0000)	(0.0001)	(0.0003)
R^2	0.67	0.49			0.71
Countries	45	45	45	45	45
Observations	180	180	135	135	112
Period	1965–1995[a]	1965–1995[a]	1970–1995	1970–1995	1965–1995

Notes: Dependent variable is average annual per capita growth. Standard errors are in parentheses. R^2 is the within-R^2 for fixed effects and the overall-R^2 for random effects.

[a] Estimates are virtually identical for the period 1970–1995 (with 135 observations).

Figure 1 – Tableau 3 de [1]

Sur le tableau Figure 2, il est clair que l'un des facteurs déterminants du changement de signe du coefficient des inégalités est l'utilisation des variables « dummy » . L'auteure met également en avant sur le tableau 4 de son étude (non présenté ici) que le calcul de la moyenne sur une période de 5 années et non sur un période plus longue (comme dans l'étude de Perotti (1996) [3] par exemple) peut en faire changer le signe. Nous attacherons moins d'importance à ce point car il est en apparence sans lien avec les remarques à venir.

2 Reproduction des calculs

Intéressés par ces résultats, nous avons tenté de reproduire le plus fidèlement possible les traitements décrits. Nous avons donc téléchargé et arrangé les bases de données citées en référence en restant aussi proche que possible des indications de l'étude. L'ensemble des données, traitements et calculs (faits en

Table 6—Sensitivity Analysis: Alternate Specifications

Specification source	Independent variables other than *Inequality* and *Income*	Coefficient on inequality[a] X-country OLS[b]	Pooled OLS[c]	Panel FE[d]	Countries	Observations	R^2
(1) This paper & Perotti (1996)	*FemaleEducation*, *Male Education*, *PPPI*	−0.0004 (0.0003)	0.0004 (0.0005)	0.0048 (0.0017)	45	144	0.73
(2) Alesina & Perotti (1994)	*Prim*, *Pstab*	−0.0005 (0.0004)	−0.0000 (0.0006)	0.0034 (0.0016)	40	104	0.82
(3) Birdsall et al. (1995)	*Assa*, *Gcons*, *PPPI*, *Prim*, *Revo*, *Sec*	−0.0021 (0.0005)	−0.0001 (0.0008)	0.0041 (0.0017)	38	102	0.83
(4) Deininger & Squire (1998)	*Bmp*, *FemaleEducation*, *Inv*, *MaleEducation*, *PPPI*	−0.0007 (0.0003)	0.0002 (0.0005)	0.0038 (0.0017)	43	141	0.75
(5) Perotti (1996)	*FemaleEducation*, *MaleEducation*, *Pop* > 65, *PPPI*	−0.0005 (0.0005)	0.0006 (0.0007)	0.0044 (0.0016)	42	140	0.74
(6) Levine & Renelt (1992)	*Gcons*, *Inv*, *Popgr*, *Prim*, *Revcp*, *Sec*	−0.0015 (0.0005)	0.0001 (0.0008)	0.0035 (0.0018)	38	102	0.83
(7) Levine & Renelt (1992)	*Bmp*, *Exp*, *Gcons*, *Inv*, *Popgr*, *Prim*, *Revcp*, *Sec*	−0.0013 (0.0004)	0.0006 (0.0008)	0.0026 (0.0017)	37	100	0.87
(8) Barro & Sala-i-Martin (1995)	*Bmp*, *FemaleEducation*, *Fhigh*, *Gcons*, *GDP*HM*, *Goved*, *Inv*, *Lifex*, *MaleEducation*, *Mhigh*, *Pstab*, *Tot*	−0.0007 (0.0004)	0.0016 (0.0006)	0.0037 (0.0018)	38	102	0.86
(9) Caselli et al. (1996)	*Assa*, *Bmp*, *FemaleEducation*, *Gcons*, *Inv*, *Lifex*, *MaleEducation*	−0.0008 (0.0003)	0.0010 (0.0007)	0.0026 (0.0017)	38	102	0.84
(10) Caselli et al. (1996)	*Assa*, *Bmp*, *FemaleEducation*, *Gcons*, *Inv*, *MaleEducation*, *Tot*	−0.0008 (0.0003)	0.0008 (0.0006)	0.0028 (0.0017)	38	102	0.84

[a] Dependent variable is average annual growth from 1965–1990. Standard errors are in parentheses.
[b] Cross-country estimation. Independent variables are from 1965 or the earliest available year thereafter.
[c] Data divided into five-year panels. Estimation obtained using OLS on this pooled data. Country and period dummies are not included.
[d] Data divided into five-year panels. Estimation obtained using fixed effects (including both country and period dummies).

Figure 2 – Tableau 6 de [1]

python via jupyter notebook) sont disponibles en ligne [1].

Malgré nos efforts, il a été impossible de reproduire exactement la base de données utilisée dans l'étude. Certaines données, notamment celle de la worldbank ne sont plus disponibles dans l'état exact. Cependant nous somme parvenus à obtenir une base raisonnablement proche de la base de l'étude. La figure 3 compare les valeurs moyennes, les écarts quadratiques, les valeurs minimales et maximales des deux bases de données.

Ne souhaitant pas s'attaquer au problème des variables retardées, nous nous sommes contentés de réaliser une régression linéaire ordinaire. C'est un algorithme que nous connaissons et que nous maîtrisons bien, à l'inverse de la méthode d'Arellano et Bond. Nous n'avons pas pris le temps d'investiguer en grand détail les différences fines entre les deux méthodes mais il nous semble raisonnable de dire que la critique formulée sur la régression linéaire s'applique aussi à la méthode plus complexe d'Arellano et Bond. Cette dernière méthode est introduite pour résoudre le problème de la variable retardée (la variable croissance contient la variable PIB) alors que notre critique se situe en amont dans la part de la variance absorbée par l'utilisation de « dummies » . La régression linéaire a par ailleurs l'avantage d'expliciter l'utilisation des ces « dummies » alors qu'elle est cachée par le calcul des premières différences dans la méthode d'Arellano et Bond.

En premier lieu, commençons par tracer sur la figure 4 la croissance en fonction de chacune des variables explicatives. Au vu de ces graphiques, force est de constater que les corrélations sont loin d'être évidentes.

Nous ajoutons ensuite une « dummy » pour chaque pays et une autre pour chaque période de temps. Nous ajoutons, par exemple, une variable "France" qui vaudra 1 pour toutes les observations concernant la France et 0 pour toutes les autres observations. On normalise ensuite les données en

1. https ://github.com/klemnnain/stage_M2/tree/master/forbes

Table 1—Summary Statistics: High-Quality Data

Variable	Definition	Source	Year	Mean	Standard deviation	Minimum	Maximum
Female Education	Average years of secondary schooling in the female population aged over 25	Barro & Lee	1965	0.90	0.95	0.04	3.10
			1970	0.95	0.94	0.04	3.36
			1975	1.11	0.94	0.05	3.62
			1980	1.40	1.10	0.14	5.11
			1985	1.54	0.99	0.20	4.84
			1990	1.76	1.02	0.21	4.69
Income	Ln of Real GNP per capita, in 1987 $US, calculated using the Atlas method	World Bank	1965	7.62	1.46	5.49	9.45
			1970	7.68	1.31	5.63	9.54
			1975	8.19	1.23	5.63	9.81
			1980	8.38	1.34	5.33	9.96
			1985	8.00	1.27	5.07	9.75
			1990	8.28	1.51	5.23	10.04
			1995	8.30	1.55	5.17	10.22
Inequality	Inequality, measured by the gini coefficient. As in Deininger and Squire, I have added 6.6 to gini coefficients based on expenditure (instead of income)	Deininger & Squire	1965	37.8	8.37	24.3	55.5
			1970	40.3	9.45	25.1	57.7
			1975	39.9	9.03	23.3	61.9
			1980	38.1	8.36	21.5	57.8
			1985	37.4	8.59	21.0	61.8
			1990	38.0	9.03	23.3	59.6
Male Education	Average years of secondary schooling in the male population aged over 25	Barro & Lee	1965	1.13	0.85	0.18	2.94
			1970	1.27	0.86	0.35	3.27
			1975	1.47	0.92	0.37	3.55
			1980	1.79	1.06	0.57	5.07
			1985	1.90	0.99	0.65	4.81
			1990	2.16	1.02	0.73	4.85
PPPI	Price level of investment, measured as the PPP of investment/exchange rate relative to the United States	Heston & Summers	1965	76.7	22.7	40.8	119.2
			1970	68.1	18.9	41.2	107.1
			1975	86.4	24.6	36.5	130.7
			1980	93.5	28.5	44.4	140.7
			1985	61.2	16.3	31.9	94.3
			1990	75.7	31.4	27.9	129.3

Note: If the gini coefficient is not available for a given year, the observation is taken from the closest year in the five-year period ending in the stated year.
Sources: Barro & Lee, the data set compiled in Barro and Lee (1996). Deininger & Squire, the data set compiled in Deininger and Squire (1996). Heston & Summers, the "Penn World Tables" version 5.6 described in Alan Heston and Robert Summers (1991). World Bank, "World*Data 1995" published by the World Bank and available on CD-ROM.

(a) Tableau 1 de [1]

categorize	year	mean	std	min	max
PPPI	1960	77.385	21.762	40.750	119.160
	1965	67.948	18.928	41.150	107.060
	1970	85.604	24.082	36.450	139.580
	1975	95.598	30.612	35.330	187.260
	1980	66.842	23.813	31.860	162.850
	1985	79.609	32.627	27.910	136.140
gini	1960	38.587	8.197	24.300	55.500
	1965	40.824	9.558	25.100	57.700
	1970	40.665	9.464	23.300	61.940
	1975	38.794	9.021	24.900	63.180
	1980	38.545	8.070	23.420	61.760
	1985	39.211	8.374	24.530	59.600
growth	1960	0.058	0.038	-0.012	0.126
	1965	0.140	0.088	-0.061	0.320
	1970	0.120	0.063	-0.075	0.211
	1975	0.002	0.054	-0.084	0.135
	1980	0.096	0.071	-0.084	0.224
	1985	0.031	0.062	-0.146	0.131
log(GNI_PC)	1960	6.312	1.395	4.661	8.250
	1965	6.578	1.326	4.742	8.565
	1970	7.465	1.151	5.082	9.012
	1975	8.070	1.355	5.207	9.666
	1980	7.969	1.367	5.276	9.813
	1985	8.361	1.600	5.698	10.314
sch_female	1960	1.024	1.155	0.030	3.790
	1965	1.024	1.056	0.040	4.300
	1970	1.256	1.145	0.060	4.760
	1975	1.643	1.220	0.110	5.110
	1980	1.860	1.224	0.220	5.260
	1985	2.162	1.303	0.290	5.370
sch_male	1960	1.283	1.023	0.270	3.710
	1965	1.344	0.963	0.320	4.260
	1970	1.623	1.128	0.100	4.750
	1975	2.060	1.197	0.370	5.140
	1980	2.311	1.160	0.670	4.880
	1985	2.584	1.214	0.780	4.800

(b) Recapitulatif du jeu de données reproduit

Figure 3 – Comparaison de la base de Forbes avec notre base

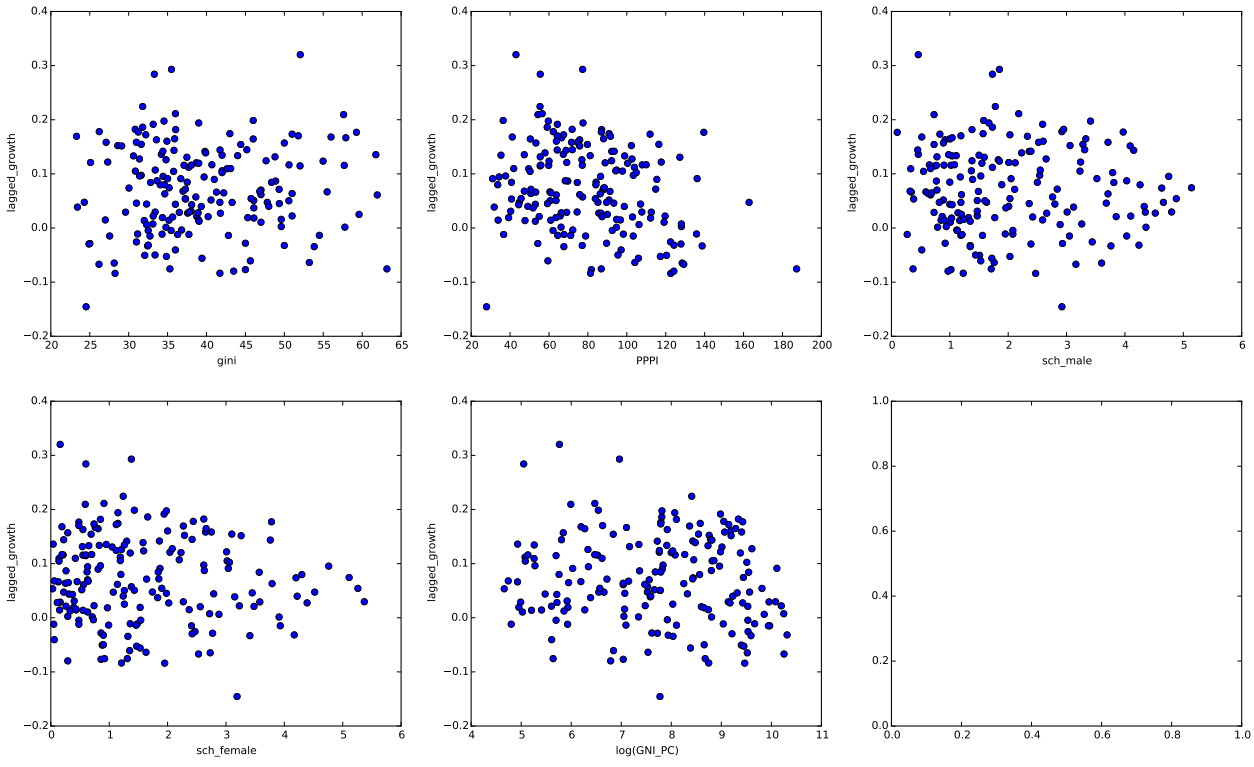

Figure 4 – Scatter plot de chaque variable explicative

retranchant à chaque variable sa moyenne et en la divisant par sa variance. Grâce à cela on peut comparer directement l'influence des variables en regardant la valeur du coefficient de la régression linéaire.

Le tableau 3 donne les coefficients obtenus grâce à la régression linéaire ordinaire. Nous affichons

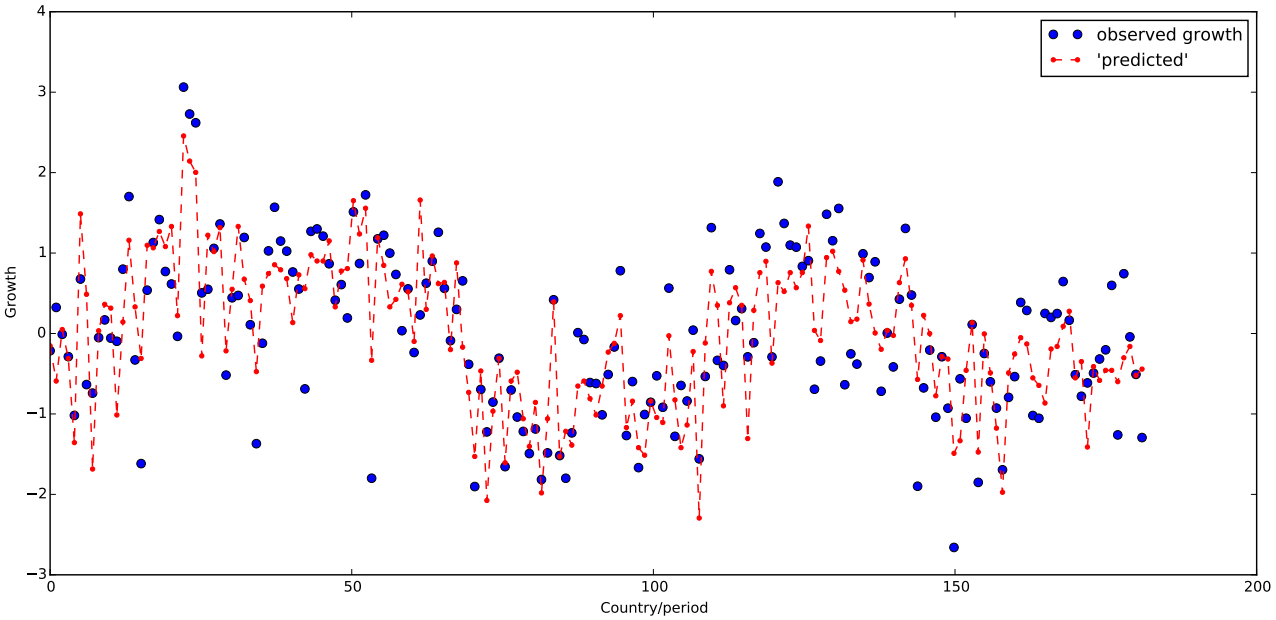

Figure 5 – Résultat de la régression linéaire ordinaire

deux résultats, ceux avec les données normalisées afin de pouvoir comparer l'importance relative des différents coefficients dans la régression linéaire et des coefficients non-normalisés afin de comparer avec les résultats de Forbes et d'avoir une idée de l'impact réel des différentes variables. On voit que les résultats pour les données non-normalisées sont assez proches des résultats de Forbes. La seule différence est un facteur 2 dans le coefficient devant les inégalités. Cette différence provient sans doute de petites différences dans les données, elle n'est pas problématique car elle va à l'encontre des résultats que nous cherchons à montrer : elle rend notre démonstration plus difficile.

	Normalized coef	Normalized std	Coefficients	std
gini	0,6495	0,188	0,0060	0,0018
PPPI	-0,4525	0,126	-0,0014	0,0004
sch_male	0,9552	0,568	0,0637	0,0381
sch_female	0,1607	0,588	0,0197	0,0383
log(GNI_PC)	-1,4715	0,352	-0,0820	0,0191
R-squared :	0.725			

Table 1 – Coefficient de la régression linéaire simple

Les points bleus de la figure 5 représentent les données de croissance à prédire et la courbe rouge en pointillée est la prédiction que nous sommes parvenus à faire avec les variables explicatives. La prédiction semble performante et la courbe rouge est proche des données, signe que le nombre de variables explicatives est suffisant. Environ 3/4 de la variance est expliquée.

3 Une répartition inégale de la variance ?

La reproduction des résultats précédents nous a amené à nous interroger sur la part explicative de chaque variable dans la régression précédente. Une fois que l'on s'est focalisé sur les variations autour de la valeur moyenne de chaque pays, que reste-t-il à prédire et quelle part est réellement prédite par les variables ? Une première approche très simple pour aborder ce problème est de réaliser la régression linéaire précédente en deux temps. Tout d'abord nous allons effectuer la régression en

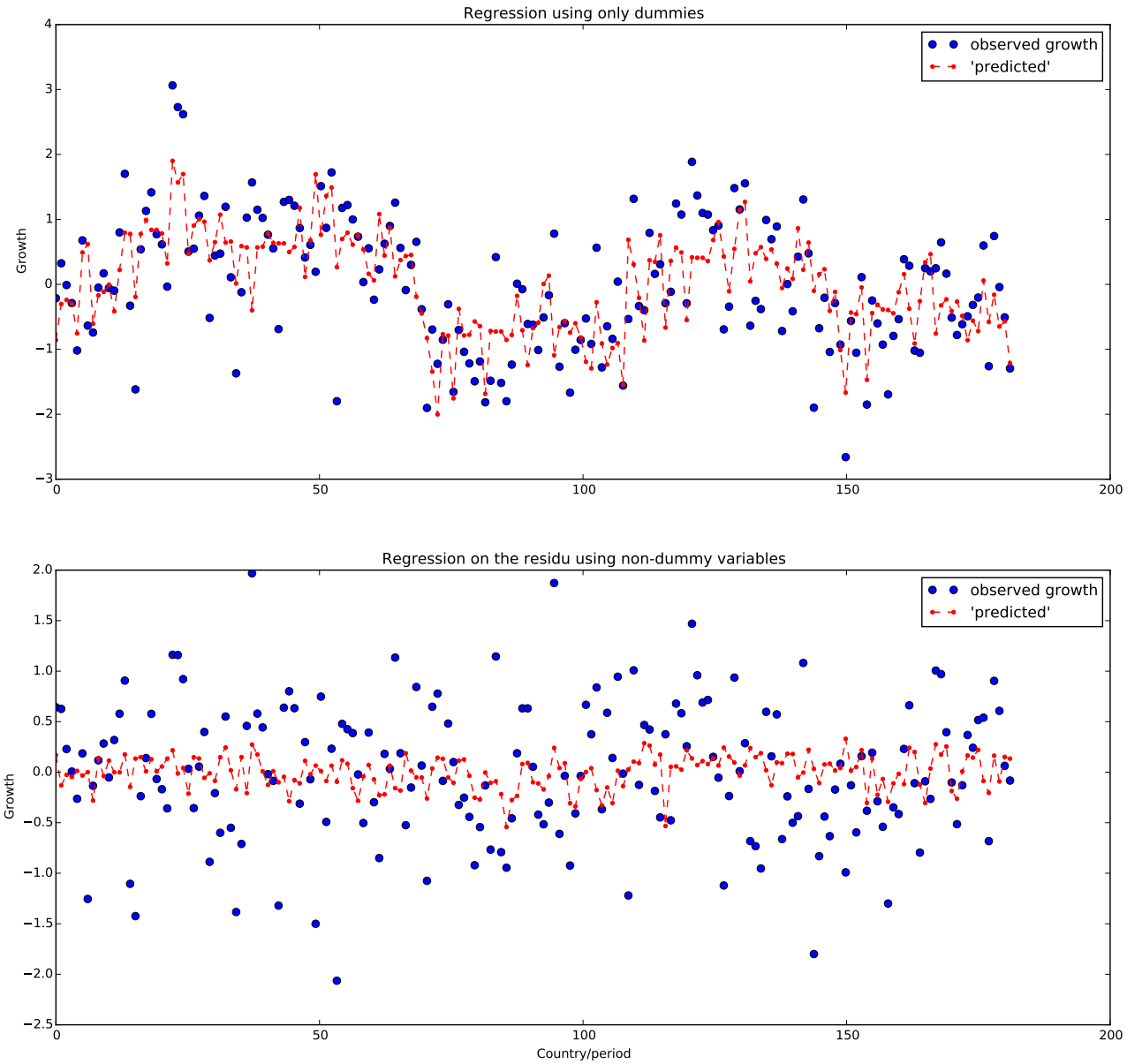

Figure 6 – Résultat de la régression linéaire ordinaire en deux étapes

utilisant uniquement les variables « dummy » ensuite nous effectuerons une régression avec les variables restantes en prenant le résidu de la régression précédente comme variable à prédire.

Explained variance share : first step	54.7 %
Explained variance share : second step	2.8 %
gini	0.03
PPPI	-0.18
sch_male	-0.03
sch_female	0.05
log(GNI_PC)	0.03

Table 2 – Coefficient de la régression linéaire ordinaire

On peut voir sur la figure 6 que lorsque la régression est effectuée avec les variables « dummy » en premières, une part très importante de la variance est absorbée par cette étape. Qui plus est, le

résidu de cette première régression semble très difficile à prédire puisqu'on voit que la part de la variance expliquée lors de la deuxième étape de la régression est quasiment nulle. Il est alors légitime de questionner l'utilisation de ces « dummies » avec si peu de données.

Il semble clair qu'avec nettement plus de données le problème ne se poserait pas. S'il on avait 300 points par pays, on pourrait *a priori* prédire beaucoup moins de variance avec la simple valeur moyenne. Il pourrait donc s'agir d'un problème de sur-optimisation des coefficients. Pour vérifier cette hypothèse, nous utiliserons une procédure de validation croisée. La validation croisée donne une image réaliste du pouvoir prédictif d'un modèle.

Les modèles économiques empiriques sont basés sur des données historiques et ont pour but de guider les politiques économiques futures. En d'autres termes, les modèles doivent s'entraîner (*i.e.* ajuster leurs coefficients) sur certaines données et tester leurs précisions sur d'autres données que le modèle n'a jamais vu et qui représenterait le futur, l'inconnu. La validation croisée permet de reproduite artificiellement ce schéma d'apprentissage à partir d'un seul jeu de données (voir [6] pour plus d'informations).

Pour ce faire, nous avons divisé le jeu de données en 5 parties de tailles égales composées d'observations tirées aléatoirement dans le jeu de données. Nous avons ensuite entraîné nos modèles sur 4/5 des données (*i.e.* calculé les coefficients) et testé leurs validités sur le 1/5 restant en calculant l'écart des prédictions aux données. Cela nous permet de connaître la capacité d'un modèle à produire une prédiction sur des données qu'il n'a jamais vu. Cette méthode nous permet de comparer deux modèles, le premier est la régression linéaire ordinaire le deuxième est un modèle de Lasso. Ce dernier est un modèle de régression linéaire ordinaire avec un terme de pénalisation des coefficients qui autorise à écarter certaines variables du modèle prédictif lorsqu'elles participent seulement à la sur-optimisation des des coefficients aux données et qu'elles n'apportent rien à la prédiction sur les données nouvelles (voir [7] pour plus de détails). 1/5 des données représentant peu de points, il se peut que le cinquième choisi ne soit pas distribué de manière représentative du jeu complet. Pour palier à ce problème nous avons effectué la procédure un grand nombre de fois (2000 * 5) et avons moyenné les résultats obtenus.

Les résultats de ce test sont très instructifs. En premier lieu, on peut constater sur la figure 7 que le modèle Lasso a diminué la valeur des coefficients de la régression, on peut voir cela à l'apparent lissage de la courbe. En regardant les résultats sur la table 3, on peut voir que la part de la variance expliquée est nettement plus faible pour la validation croisée ce qui est normal puisque le modèle doit maintenant prédire une valeur sur des données qu'il n'a jamais vu ce qui est plus difficile mais aussi plus réaliste. Deuxièmement, ce test valide bien l'hypothèse de la sur-optimisation des coefficients. En effet, le modèle de Lasso, dont les coefficients ont été artificiellement réduits par pénalisation, obtient en moyenne, de meilleurs résultats (cf. tableau 3).

	mean	std
Explained variance share : OLS	0.312676319641	0.120855814149
Explained variance share : Lasso	0.264664473665	0.237392947754

Table 3 – Résultats du test de sur-optimisation

Le résultat le plus instructif de cette expérience vient de l'étude des coefficients de la régression de Lasso. Ces coefficients peuvent être interprétés exactement de la même manière que les coefficients d'une régression linéaire. On peut voir sur le tableau 4. que seule la variable $PPPI$ a un coefficient non-négligeable et que le reste de la prédiction est en fait assurée par les « dummies ». On voit en particulier que le coefficient pour la variable *gini* est de l'ordre de 10^{-6} en moyenne sur toutes les simulations. **Grâce à ces résultats on vient de prouver que la prédiction la plus précise que nous puissions faire de la croissance à partir de ces données ne prend pas en compte le coefficient de Gin !**

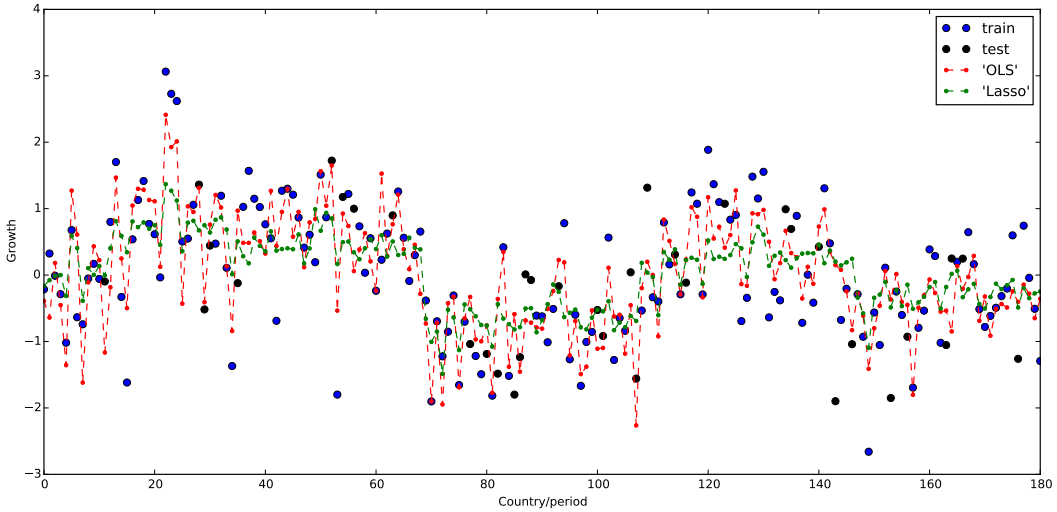

Figure 7 – Comparaison de la prédiction par le Lasso et par le modèle linéaire

1975	-3.16e-01	CIV	-1.54e-02	MYS	3.69e-05
1985	-1.99e-01	TTO	-1.10e-02	GBR	2.84e-05
PPPI	-1.64e-01	ITA	8.25e-03	BRA	2.41e-05
1965	1.44e-01	log(GNI_PC)	6.79e-03	USA	-1.80e-05
1970	1.33e-01	NOR	6.21e-03	TUR	-6.02e-06
BGR	-1.16e-01	FIN	5.03e-03	1980	3.88e-06
JPN	9.97e-02	DEU	4.45e-03	gini	-3.88e-06
KOR	7.26e-02	PRT	-4.14e-03	DNK	1.18e-06
IND	-5.62e-02	sch_male	4.02e-03	CAN	-8.98e-08
IRN	5.61e-02	GAB	3.54e-03	AUS	0.00e+00
CHL	-5.47e-02	FRA	2.44e-03	sch_female	0.00e+00
BGD	-4.86e-02	NLD	9.42e-04	COL	0.00e+00
VEN	-4.58e-02	ESP	7.38e-04	TUN	0.00e+00
HKG	4.17e-02	BEL	-2.40e-04	CHN	0.00e+00
SGP	3.99e-02	PHL	-2.30e-04	GRC	0.00e+00
DOM	-3.65e-02	IDN	-9.85e-05	PER	0.00e+00
1960	-3.10e-02	SWE	8.51e-05	CRI	0.00e+00
PAK	-2.83e-02	THA	7.56e-05	NZL	0.00e+00
LKA	-2.40e-02	MEX	-3.70e-05	IRL	0.00e+00

Table 4 – Coefficients de la régression de Lasso. Ces coefficients sont la moyenne des coefficients sur 2000 * 5 simulations.

Cette nouvelle approche met en doute la méthode de Forbes et la possibilité d'ajouter des « dummy » variables sans précautions particulières. La même méthode a été reprise dans d'autres études [4], [5]. Nous ne pouvons aucunement remettre en doute la validité de ces études à partir des simples remarques faites pour Forbes car les données utilisées sont très différente (le nombre de points par pays est notamment beaucoup plus élevé). Cependant, si les remarques faites dans cette note sont pertinentes, alors il serait légitime d'appliquer la même méthode d'analyse à ces études pour tester la validité des mêmes hypothèses dans les autres études.

Conclusion :

Cette note n'a pas vocation à remettre en question les résultats produits par l'étude de Forbes mais plutôt de formaliser nos interrogations sur la méthode utilisée. Même si nos intuitions se révèlent pertinentes, cette note ne constitue pas une preuve de la fausseté des résultats de Forbes. De nombreux points pouvant avoir une influence cruciale sur les résultats restent non-traités dans cette note. Premièrement, une preuve sérieuse doit s'assurer que les données utilisées sont rigoureusement les mêmes que celles de Forbes. Même s'il nous semble que le jeu que nous avons réussi à créer est de bonne qualité qu'une étude se doit d'être robuste à quelques différences minimes dans les données, il serait plus rigoureux d'obtenir (par exemple en demandant à l'auteure) le jeu de données utilisé pour l'étude. Le deuxième point limite de l'étude est l'utilisation de la régression linéaire au lieu de la méthode d'Arellano et Bond. Si nos intuitions se révèlent exactes, l'étape suivante serait d'appliquer la même procédure de validation croisée pour tester la méthode Arellano et Bond et particulièrement la robustesse du signe du coefficient de gini. Enfin cette étude, réalisée en marge d'un stage de M2, a été faite en peu de temps (d'où notre préférence pour des algorithme connus) et sans concertation avec l'auteure de l'étude, la validation des nos hypothèses pourrait commencer par un contact avec cette dernière pour avoir son avis sur nos remarques.

Références

[1] Forbes, Kristin J. 2000. *A Reassessment of the Relationship between Inequality and Growth.* American Economic Review, 90(4) : 869-887.

[2] Banerjee, Abhijit V. and Duflo, Esther, *Inequality And Growth : What Can The Data Say ?* (June 2000). MIT Dept. of Economics Working Paper No. 00-09. Available at SSRN : http ://ssrn.com/abstract=232731 or http ://dx.doi.org/10.2139/ssrn.232731

[3] Perotti, Roberto. *Growth, Income Distribution, and Democracy : What the Data Say.* Journal of Economic Growth 1.2 (1996) : 149–187. Web...

[4] Jonathan D. Ostry, Andrew Berg, and Charalambos G. Tsangarides, *Redistribution, Inequality, and Growth.* IMF Staff Discussion Note. February 2014.

[5] OECD. (2015), In It Together : Why Less Inequality Benefits All, OECD Publishing, Paris. DOI : http ://dx.doi.org/10.1787/9789264235120-en

[6] Trevor Hastie, Robert Tibshirani, Jerome Friedman *The Elements of Statistical Learning : Data Mining, Inference, and Prediction.* Second Edition February 2009, p. 241 - 245, disponible en ligne : http ://statweb.stanford.edu/ tibs/ElemStatLearn/

[7] Trevor Hastie, Robert Tibshirani, Jerome Friedman *The Elements of Statistical Learning : Data Mining, Inference, and Prediction.* Second Edition February 2009, p. 68 - 69, disponible en ligne : http ://statweb.stanford.edu/ tibs/ElemStatLearn/