

ESTIMATION DES LOIS DU REVENU ET DES DÉPENSES DANS UNE BASE DE DONNÉES DE LA PAUVRETÉ.

GANE SAMB LO AND COLLABORATORS

ABSTRACT. Les indicateurs de pauvreté dépendent sûrement de la loi des revenus et des dépenses. Nos travaux actuels ont permis d'estimer indicateurs et leurs variances asymptotiques. Il semble indiqué alors de déterminer la loi des revenus et des dépenses dans la base ESAM I (1996) du Sénégal. Les résultats concernent l'ensemble des dix régions existantes dans la période de l'enquête.

1. INTRODUCTION

L'objectif de cette note est l'estimation de la loi des revenus et des dépenses des ménages sénégalais recensés dans la base de données de l'Enquête Sénégalaises Auprès des Ménages du Sénégal de 1996 (ESAM). Ces revenus jouent un rôle important dans l'appréciation quantitative de la pauvreté. Celle-ci est un phénomène complexe dont l'exploration et l'analyse requièrent l'utilisation simultanée de plusieurs disciplines : l'économie, la sociologie, les statistiques, la médecine, etc. Les aspects quantitatifs, à côté des aspects qualitatifs, jouent un rôle important. L'appréciation quantitative se réalise par plusieurs indicateurs de pauvreté dont naturellement la prévalence de pauvreté, c'est-à-dire la proportion de pauvres dans la population étudiée.

Ces mesures de pauvreté sont basées sur le revenu ou la dépense des ménages. Dans la suite, la variable revenu ou dépense de la population étudiée de N ménages sera notée Y prenant les valeurs ordonnées : Y_1, Y_2, \dots, Y_N . Pour définir un ménage pauvre, on fait appel à un seuil ou ligne de pauvreté Z de sorte qu'un ménage j est déclaré pauvre si son revenu est inférieur à Z , c'est-à-dire

$$(1.1) \quad (j \text{ pauvre}) \iff (Y_j < Z)$$

Le nombre de ménages pauvres est le nombre Q tel que $Y_Q < Z \leq Y_{Q+1}$ les revenus pauvres sont $\{Y_1, Y_2, \dots, Y_Q\}$. La détermination du seuil est aussi une question complexe, ayant fait l'objet d'une multitude de discussion, suivant les approches utilisées : en termes de conditions de vie minimaales, en termes de capacités, en termes de survie, etc. Le lecteur intéressé peut trouver une bonne base de discussion dans Ravallion [3]. Les indicateurs de pauvreté qui nous pré-occupent dans cet article sont des fonctions mathématiques

des revenus des Q pauvres sous la forme globale

$$(1.2) \quad P(Y, N, Q) = \frac{1}{\delta(H)\rho(L)} \sum_{j=1}^Q \beta(M, N, j) \gamma(e_j)$$

où $e_j = (Z - Y_j)/Z$ est le déficit de pauvreté du j -ième pauvre, δ , β , ρ et γ étant des fonctions données; H , L et M étant pris dans $\{N, Q\}$.

La statistique (1.1) est la forme empirique de l'indicateur dans l'échantillon choisi. On pourra alors supposer que les revenus ou les dépenses observées Y_1, Y_2, \dots, Y_n sont des observations indépendantes d'une même variable aléatoire de Loi $G(y) = \mathbb{P}(Y_i \leq y)$. Et (1.2) devient

$$(1.3) \quad p_n = \frac{1}{\delta(H)\rho(L)} \sum_{j=1}^q \beta(M, N, j) \gamma\left(\frac{Z - Y_{j,n}}{Z}\right)$$

où $Y_{1,n} \leq Y_{2,n} \leq \dots \leq Y_{n,n}$ est la statistique d'ordre associée à l'échantillon Y_1, Y_2, \dots, Y_n , $\{h, l, m\} \subset \{q, n\}$, et q est le nombre de pauvres dans l'échantillon, i.e., $q/n = G_n(Z)$ la valeur de Z pour fonction de répartition empirique de l'échantillon, avec bien entendu, par le théorème de Glivenko-Cantelli, $G_n(Z) \rightarrow G(Z)$, quand n tend vers l'infini.

Il devient alors important de faire une théorie asymptotique de la convergence de p_n et de sa loi limite. Lô et al.[1] et G. Dia [6] se sont attelés à cette tâche. Le rôle de la loi de la variable, c'est à dire, de G joue un rôle important, contrairement à ce que pensent Boccanfuso et al.[5].

Dans un avenir à venir, nous nous intéresserons à l'application des résultats de [1] sur la base ESAM I. Mais, en préparation à ce travail et pour son propre intérêt, nous allons dans cette note donner une estimation de G et de ses paramètres éventuels. Les résultats pourront servir à la fois aux chercheurs sur la pauvreté mais aussi toute analyse basée sur ces variables.

Néanmoins, (1.3) montre que les statistiques de la pauvreté concerne la queue inférieure de la distribution et la partie centrale. Or, d'une manière générale, c'est la queue supérieure qui davantage étudiée en statistiques puisqu'on toujours ramener un problème de queue inférieure en un problème de queue supérieure. Nous allons suivre cette tendance en considérant la suite de variables aléatoires

$$(1.4) \quad X_i = 1/(Y_i - y_0), \quad i=1, \dots, n$$

avec $y_0 = \inf\{y \geq 0, G(y) > 1\}$, le revenu ou la dépense théorique minimum. La fonction de répartition à un support infini à droite, i.e., $x_0 = \sup\{x, F(x) < 1\} = +\infty$. Elle est liée à G par

$$(1.5) \quad G(y) = 1 - F(1/(y - y_0)), \quad y \geq y_0.$$

et

$$(1.6) \quad G^{-1}(u) = y_0 + 1/F^{-1}(u), \quad 0 \leq u \leq 1.$$

De manière concrète, nous estimerons y_0 par $Y_{1,n}$ la plus petite observations et la série $X_i = 1/(Y_i - Y_{1,n})$, avec l'exclusion des variables qui réalise ce minimum, sera calibrée.

Dans la section 2 qui suit, nous décrirons les techniques qui serviront à faire cette ajustage que sont : la regression simple, le test de Kolmogorov-Smirnov et les techniques d'ajustage aux lois de type parétien [4]. Dans la section 3, les résultats sur la base ESAM I seront exposés et commentés.

2. OUTILS D'AJUSTAGE

Nous allons décrire deux grandes méthodes : la regression simple et le test de Kolmogorov-Smirnov.

2.1. Méthode des moindres carrés. Soit la fonction de répartition empirique

$$(2.1) \quad F_n(x) = \text{Card}\{j, 1 \leq j \leq n, X_j \leq x\}, \quad x \in R$$

basée sur la série $X_i, i = 1, \dots, n$ décrite ci-haut et soit $X_{1,n} \leq \dots \leq X_n$ la statistique d'ordre qui lui est associée. Sous l'hypothèse

(H) X_1, X_2, \dots, X_n est une suite iid de fonction de répartition commune F

le théorème de Glivenko-Cantelli implique l'approximation uniforme suivante

$$(2.2) \quad F_n(X_{j,n}) = \frac{j}{n} \approx F(X_{j,n})$$

transformée en

$$(2.3) \quad t_{j,n} = F^{-1}\left(\frac{j - 0.5}{n}\right) \approx X_{j,n}$$

Dès lors, les points $(t_{j,n}, X_{j,n})_{1 \leq j \leq n}$ se situent plus ou moins autour de la première bissectrice. Donc une bonne régression de $X_{j,n}$ en $t_{j,n}$, avec une pente proche de l'unité et un coefficient à l'origine proche de zéro, supporte l'hypothèse (H). Cependant, pour chaque fonction de répartition F , la régression sera particularisée selon ses paramètres. Ainsi, ceux-ci seront estimés par la régression et par leur estimation du maximum de vraisemblance ou par la méthode des moments. La règle générale est la suivante.

L'hypothèse (H) est acceptée si le coefficient de régression est proche de l'unité et si les estimations classiques des paramètres sont proches des estimations issues de la régression.

Nous allons illustrer les trois lois utilisées dans cette note.

2.1.1. *Le modèle de Paréto de paramètres $C_2\theta$ et $\alpha > 0$.*

$$(2.4) \quad F(x) = \begin{cases} 1 - (x/c)^{-\alpha}, & \text{si } x \geq c \\ 0 & \text{sinon} \end{cases}$$

Le paramètre α sera estimé selon la méthode de Diop et Lô([2]), qui généralise la méthode de Hill, par

$$(2.5) \quad \hat{\alpha} = 1/(4 \times k^{-4} \sum_{j=l+1}^{j=k} j^4 (\log X_{n-j+1,n} - \log X_{n-j,n}))$$

et C est estimé par la méthode du maximum de vraisemblance qui donne $\hat{C} = X_{1,n}$. De plus (2.3) devient

$$(2.6) \quad y_j = \log X_{j,n} \approx (1/\alpha)x_j + \log C$$

où $x_j = -\log(1 - u_j)$ et $u_j = (j - 0.5)/n$. Nous ferons la régression de y_j en x_j et concluerons que la loi de la série X_i est de type Paréto de paramètres α et C si

le coefficient de régression R^2 est proche de 1, la pente de la régression $\hat{\alpha}$ est proche de $1/\hat{\alpha}$ et si le coefficient à l'origine \hat{b} est proche de $\log X_{1,n}$.

2.1.2. *Le modèle exponentiel de paramètre $\lambda \geq 0$.*

$$(2.7) \quad F(x) = \begin{cases} 1 - \exp(-\lambda x), & \text{si } x \geq 0 \\ 0 & \text{sinon} \end{cases}$$

Le paramètre λ sera estimé selon la méthode du maximum de vraisemblance qui donne $\hat{\lambda} = 1/(\frac{1}{n} \sum_{j=1}^{j=n} X_j)$, l'inverse de la moyenne empirique \bar{X} . De plus (2.3) devient

$$(2.8) \quad y_j = \log X_{j,n} \approx (1/\lambda)x_j$$

où $x_j = -\log(1 - u_j)$ et $u_j = (j - 0.5)/n$. Nous ferons la régression de y_j en x_j et concluerons que la loi de la série X_i est de type Exponentielle λ .

le coefficient de régression R^2 est proche de 1, la pente de la régression $\hat{\lambda}$ est proche de $1/\hat{\lambda} = \bar{X}$ et si le coefficient à l'origine \hat{b} est proche de zéro.

2.1.3. *Le modèle Lognormale de paramètres $\sigma \geq 0$ et m .*

$$(2.9) \quad F(x) = \begin{cases} \Phi((\log x - m)/\sigma), & \text{si } x \geq c \\ 0 & \text{sinon} \end{cases},$$

où

$$(2.10) \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-t^2/2) dt$$

est la fonction de répartition d'une loi normale standard. La variable $X \geq 0$ suit une loi lognormale de paramètres $\sigma \geq 0$ et m si et seulement si $\text{Log}(X)$ suit une loi normale de paramètres $\sigma \geq 0$ et m . Alors m est estimé par la moyenne empirique de la série $\text{Log}(X_i)$

$$\overline{\log X} = \frac{1}{n} \sum_{j=1}^{j=n} \log X_j$$

et σ^2 est estimé par sa variance empirique de la série, i.e.

$$(2.11) \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{j=1}^n (\log X_j - \overline{\log X})^2.$$

Maintenant (2.3) s'écrit

$$(2.12) \quad y_j = \log X_{j,n} \approx \sigma \Phi^{-1}(1 - u_j) + m.$$

Nous ferons la régression de y_j en $x_j = \Phi^{-1}(1 - u_j)$ et concluerons que la loi de la série X_i est de type lognormale de paramètres σ et m si

le coefficient de régression R^2 est proche de 1, la pente de la régression \hat{a} est proche de $\hat{\sigma}$ et si le coefficient à l'origine \hat{b} est proche de $\overline{\log X}$.

Dans cette note, nous nous limiterons à ces trois exemples. Mais la méthode peut être appliquée à toute fonction de répartition. On n'a pas besoin de pouvoir calculer F^{-1} pour l'utiliser. Les moyens de calculs modernes les calculent sans problème.

2.2. Test de Kolmogorov-Smirnov. D'après le Théorème de Kolmogorov-Smirnov, la statistique

$$(2.13) \quad K_n = \sqrt{n} \max_{1 \leq j \leq n} \max\left(\frac{j}{n} - F(X_{j,n}), F(X_{j,n}) - \frac{j-1}{n}\right)$$

converge en loi et la fonction de répartition de la loi limite est

$$(2.14) \quad \mathbb{P}(K_n > x) \rightarrow \xi(x) = \sum_{j=1}^{\infty} (-1)^{j+1} \exp(2j^2 x^2), \quad x > 0.$$

sous l'hypothèse (H). Le test de Kolmogorov-Smirnov se fonde sur la p-value $P = \xi(K_n)$. Une petite valeur ($P < 5\%$) n'est pas favorable à l'hypothèse (H) tandis que qu'une valeur de P de plus de 5% nous fera accepter (H).

Nous allons utiliser ces deux méthodes pour la variable revenu pour les dix régions du Sénégal, et pour l'ensemble du Sénégal.

3. ESTIMATION DE LA LOI DES REVENUS DE LA BASE ESAM

Nous utilisons les données de la base de l'Enquête Sénégalaise Auprès des Ménages (ESAM I) qui a concerné $N=3278$ ménages. La variable REVTOT est constituée du revenu total du ménage. La variable EQADUL représente l'équivalence adulte du ménage. Nous travaillerons $Y = \text{REVTOT}/\text{EQADUL}$, le revenu annuel individualisé. Nous estimerons la fonction de répartition F de $X = 1/(Y - y_0)$ pour chaque région et pour l'ensemble.

Nous constatons que pour l'ensemble du pays et aussi pour l'ensemble des régions, le modèle lognormale est généralement accepté avec tous les constats attendus sur la méthode de la régression. Presque pour toutes les régions, à l'exception de Dakar et Saint-Louis, le modèle lognormal est très bien supporté par les méthodes utilisées. Pour Dakar, Saint-Louis, et le

Sénégal en entier, le modèle est suggéré par la régression et non par le test de Kolmogorov- Smirnov.

Les résultats des calculs sont les suivants.

3.1. Variable revenu.

Région	Sénégal	Kolda	Dakar	Ziguinchor	Diourbel	St-Louis
R ² (%)	99,55	99,47	97,80	99,02	99,126	97,86
σ	1,3	1,2	1,32	1,44	1,39	1,38
m	-11,68	-10,8	-12,22	-11,5	-11,1	-11,52
P(%)	1,41	31,54	4,8710 ⁻³	38,34	38,34	3,73
DN	1,85	0,75	2,28	0,66	0,66	1,396

Région	Tambacounda	Kaolack	Thiès	Louga	Fatick
R ² (%)	99,48	99,64	99,77	99,6	99,75
σ	1,094	1,87	1,4	1,31	1,29
m	-11,03	-11,4	-11,22	-11,38	-11,19
P(%)	46,4	36,5	33,33	37	45,32
DN	0,544	0,69	0,768	0,679	0,56

3.2. Variable dépense.

Région	Sénégal	Kolda	Dakar	Ziguinchor	Diourbel	St-Louis
R ² (%)	99,14	98,87	98,11	98,7	99,36	98,83
σ	0,88	0,81	0,92	1,079	0,86	0,672
m	-11,84	-11,89	-12,23	-11,38	-11,36	-11,61
P(%)	1,4 10 ⁻³	24,72	0,18	8,95	9,4	21,32
DN	2,36	0,831	1,77	1,098	1,08	0,875

Région	Tambacounda	Kaolack	Thiès	Louga	Fatick
R ²	96,84	97,78	98,18	99,27	98,96
σ	0,89	0,95	0,95	0,95	0,939
m	-11,21	-11,21	-11,39	-11,23	-11,35
P(%)	7,35	15,13	1,63	37,33	19,35
DN	1,14	0,971	1,43	0,672	0,905

4. CONCLUSION

Les estimations étant consistantes, il sera question dans un article à venir d'estimer les indicateurs de pauvreté pour ces modèles. Mais il faudra au paravent simuler les lois limites qui seront utilisées pour les lois décrites ci-haut avec des paramètres dans l'ordre de grandeur des estimations trouvées ci-haut. Il sera aussi intéressant d'appliquer les méthodes décrites ici à des bases de même nature

REFERENCES

- [1] **Lo, G.S, Sall S.T. et Seck C.T.**(2004). The asymptotic theory of the poverty measures under the influence of the extremes. Publications de l'Ufr SAT, Université Gaston Berger de Saint-louis, Lerstad n°7.
- [2] **Lo, G. S. et Diop. A** (2004). On a continuous Hill statistic process and its asymptotic normality theory. Rapport technique, Université Gaston Berger de Saint-louis, Lerstad n°4.
- [3] **Ravallion M.(1992)**. Poverty Comparisons. A Guide to Concepts and Methods. Lsms, Working Paper, n°88, *WorldBank*.
- [4] **Barbut M.(1989)**. Distribution de type parétien et représentation des inégalités. In : Mathématiques informatique et sciences humaines. 106, pp.53-69
- [5] **Boccanfuso D., Decaluwé B et Savard L.(2003)**. Poverty, Income Distribution and CGE modeling : Does the Functionnal Form Matter?. Priliminary Draft. Dakar
- [6] **Dia G.(2004)**. Répartition aléatoire des revenus et estimation de l'indice de pauvreté. Publication de l'Ufr SAT, Université Gaston Berger de Saint-louis, Lerstad n°4.
- [7]

LERSTAD, UNIVERSITÉ GASTON BERGER DE SAINT-LOUIS, SÉNÉGAL
E-mail address: gsl0@ugb.sn