Robust Methods for Personal Income Distribution Models

MARIA-PIA VICTORIA-FESER
UNIVERSITÉ DE GENÈVE

1 Introduction

C'est un très grand honneur pour moi de pouvoir aujourd'hui présenter mon travail de doctorat intitulé Robust Methods for Personal Income Distribution Models qui a reçu cette année à Genève le Prix Latsis Universitaire.

Ma thèse traite un sujet à la fois économique (pause) et statistique, ou pour être plus précis, un sujet économétrique. Il s'agit de l'étude de la mesure de la distribution des revenus vue sous l'angle du statisticien.

La statistique dite *classique* bien qu'essentielle, souffre d'un désavantage considérable. La théorie ne tient pas compte du fait que les données à disposition peuvent contenir des erreurs. Ces erreurs ou données aberrantes (par exemple celles qui sont enregistrées à dix fois leur vraie valeur) sortent du cadre proposé par la théorie et en général biaisent considérablement les résultats. Par contre, la statistique dite *robuste*, par définition tient compte du fait que certaines données peuvent ne pas être représentatives, situation très fréquente avec les données sur les revenus. Une analyse robuste devient donc plus réaliste car elle n'est pas ou peu influencée par une minorité d'observations extrêmes.

2 La mesure de la distribution des revenus sous l'angle classique

La mesure de la distribution des revenus est un sujet ancien: il suffit de mentionner ici les oeuvres de Pareto et de Lorenz développées entre la fin du 19e et le début du 20e siècle (Pareto 1896, Pareto 1897, Lorenz 1905). La mesure de la distribution des revenus reste néanmoins un sujet important puisqu'elle est au centre des débats politiques et économiques de notre temps. Les problèmes d'inégalité et de pauvreté y sont aussi étroitement reliés.

La statistique est un outil indispensable à l'étude de ces sujets car l'analyse des données sur les revenus joue un rôle central. De nos jours, les banques de données sont abondantes et peuvent même être sur-abondantes laissant les chercheurs devant une tempête d'informations difficiles à digérer sans l'aide des techniques statistiques.

Un des premiers pas dans l'analyse est la concentration de l'information: les milliers de données sur les revenus qui caractérisent une distribution dans une ville, un pays ou une région donnés, pour une certaine période, devraient pouvoir être représentées par un modèle mathématique adéquat. Pareto (1896) fut le premier à proposer ce type de modèle. Depuis ses travaux, un nombre considérable de modèles ont été développés (March 1898, Amoroso 1925, Gibrat 1931, Aitchison and Brown 1957, Davis 1941, Weibull 1951, Fisk 1961, Thurow 1970, McDonald and Ransom 1979, MacDonald 1984, Singh and Maddala 1976, Dagum 1977, Dagum 1980, Majumder and Chakravarty 1990).

Les données peuvent être représentées par des histogrammes et les modèles mathématiques formalisent en fait ces histogrammes en une formule contenant des paramètres. Pour chaque jeu de données, il existe un modèle capable de décrire ce jeu de données de telle sorte que si les données ne sont plus à disposition, le modèle estimé est

suffisant pour caractériser les données en question. Par exemple, un modèle souvent utilisé est la distribution de Gamma qui dépend de deux paramètres inconnus. Pour différentes combinaisons de valeurs pour ces paramètres, la discription de la distribution des revenus est différente. Sur la figure 1, on peut voir un histogramme de la distribution des revenus au Royaume-Uni en 1979, pour les ménages recevant une aide de l'état. L'information contenue dans ces 700 observations, peut être résumée par

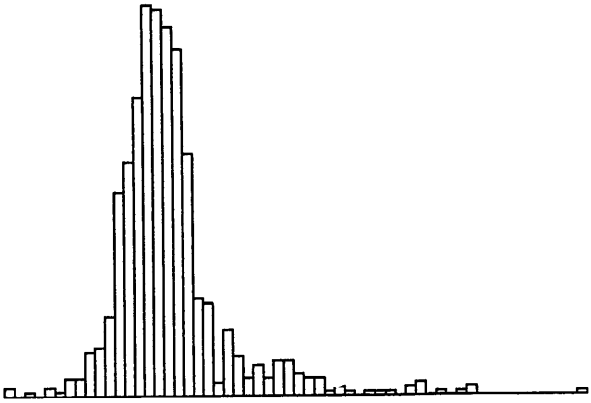


Figure 1: Distribution des revenus au Royaume-Uni en 1979

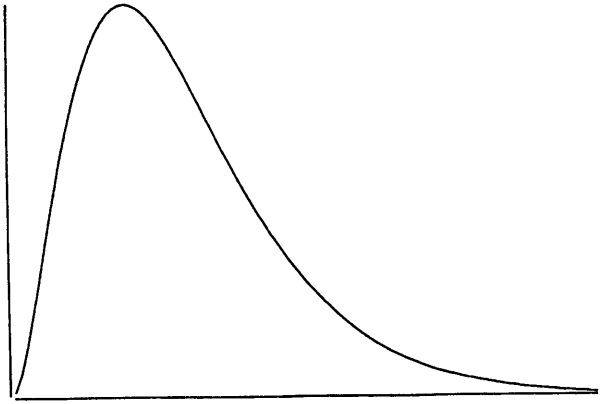


Figure 2: Une distribution Gamma

un modèle Gamma décrit sur la figure 2. Le seul problème est de pouvoir trouver la combinaison adéquate de valeurs pour les paramètres qui décrit au mieux la distribution observée. Ce problème statistique est normalement résolu par une méthode appelée le maximum de vraisemblance. Les valeurs "estimées" des paramètres sont celles qui maximisent la vraisemblance entre les données observées et le modèle mathématique.

Si nous adoptons cette démarche classique, le résultat obtenu est représenté par la courbe sur la figure 2. Comme vous pouvez le constater, ce que nous obtenons est plutôt décevant. L'ajustement entre le modèle mathématique estimé et les données est loin d'être convainquant. Alors que s'est-il passé?

Pour répondre à cette question, il faut replacer la théorie dite classique dans son contexte. Le maximum de vraisemblance, comme d'autres méthodes similaires, est une théorie qui se base sur l'hypothèse que le modèle représente exactement les données ou, en d'autres termes, que les données sont parfaites, sans erreurs. Cela implique que les données reçoivent la même attention, indépendemment de leur valeur. Malheureusement, les valeurs extrêmes pèsent très lourd dans les calculs, et les résultats sont souvent considérablement biaisés . . . Reprenons l'exemple sur les données britanniques . . . Certaines valeurs sont sonsidérées extrêmes car la probabilité d'observer des données ayant ces valeurs est très faible. Le maximum de vraisemblance essaye d'accomoder toutes les observations jusqu'à en oublier les plus importantes, à savoir la majorité d'entre-elles au centre de la distribution.

Que nous reste-t-il à faire? Certains chercheurs ont proposé des modèles compliqués qui tiennent compte de ces éventualités. Cependant je crois que de cette manière l'analyse perd de sa force qui est non seulement d'avoir des paramètres interprétables mais aussi de rester simple en utilisant des modèles simples.

3 Les méthodes robustes

Dans ma thèse de doctorat, je propose d'utiliser la statistique dite robuste pour remédier à ces problèmes. La statistique robuste remonte au moins en tant qu'idée philosophique à l'enfance de la statistique. Bon nombre de statisticiens travaillant avec des données ont su trouver leurs méthodes ad-hoc pour éviter des résultats catastrophiques lorsque les données contiennent des valeurs aberrantes. Cependant, sa formalisation dans un cadre mathématique et statistique est due aux professeurs Huber et Hampel dans les années soixante (Huber 1964, Huber 1965, Huber 1981, Hampel 1968, Hampel 1971, Hampel1974). Ont suivi ensuite un bon nombre de jeunes chercheurs issus entre autres de l'école polytechnique fédérale de Zurich où les deux maîtres ont enseigné (voir par exemple Hampel, Ronchetti, Rousseeuw et Stahel 1986). La statistique robuste est maintenant une science bien répandue, et j'en suis une modeste preuve aujourd'hui.

Contrairement à la théorie classique, la statistique robuste fait l'hypothèse plus souple que les modèles proposés pour décrire les données qui peuvent ne pas être totalement exacts ou, en d'autres termes, qu'il peut y avoir des erreurs dans les données ou d'autres types de déviations du modèle. Une procédure robuste permet non seulement de détecter les valeurs influençant démesurément les résultats, mais aussi de traiter ces données en fonction de la majorité des autres. Par exemple, lorsque l'on calcule une moyenne de salaires, la moyenne peut être à un certain niveau même si disons 75% des salaires sont en dessous de ce niveau. En fait, il suffit d'un salaire disproportionnellement plus élevé que les autres pour augmenter le niveau de la moyenne alors que les autres salaires ne changent pas. Une procédure robuste considère ce dernier salaire comme une valeur extrême et le résultat est moins biaisé.

Si l'on prend l'exemple des revenus britanniques, une estimation robuste des paramètres de la distribution Gamma est représentée par

la courbe sur la figure 4. Le résultat est maintenant plus convainquant puisque cette fois la vraisemblance entre le modèle et les données est satisfaisante. Alors que l'estimateur classique modélise bien les valeurs exceptionnelles, l'estimateur robuste capture principalement la distribution de la majorité des observations.

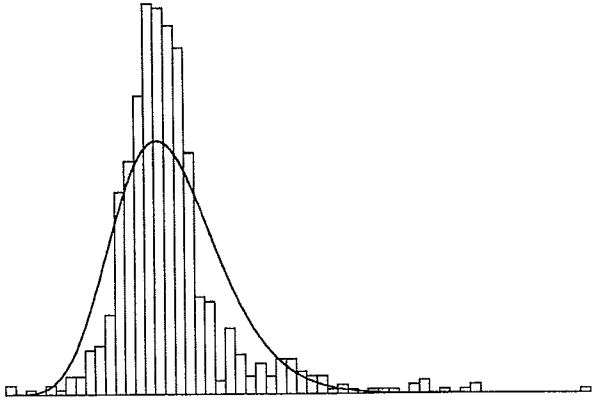
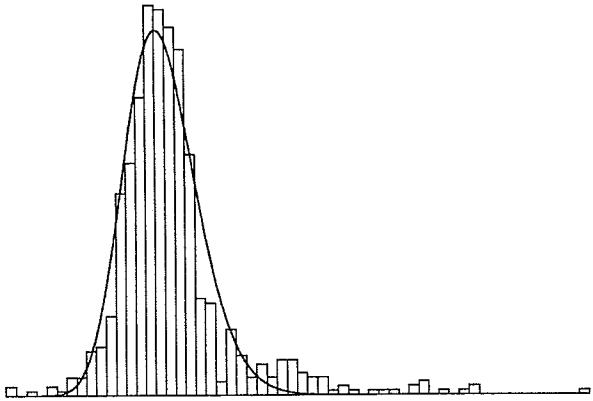


Figure 3: Estimation classique d'une distribution Gamma pour les données du Royaume-Uni



Fibure 4: Estimation robuste d'une distribution Gamma pour les données du Royaume-Uni

Je vous épargnerai ici les détails techniques qui sont nécessaires pour le calcul des méthodes robustes puisqu'ils impliquent des pages, voir des chapitres de formules mathématiques. Cependant l'idée de base est simple: lors de la procédure d'estimation, les données extrêmes reçoivent moins de poids que le reste des données et donc ne peuvent plus influencer démesurément les résultats. Ces poids sont déterminés d'une façon objective par l'ensemble des données.

Certains pourraient avoir un regard encore très sceptique vis-à-vis des méthodes robustes car, après tout, si les observations ayant une valeur dite extdrême ne sont pas des erreurs, mais représentent bel et bien une donnée légitime, alors pourquoi devrions-nous leur donner moins de poids lors de la procédure d'estimation? Le fait est que lorsque l'on veut concentrer l'information, depuis les milliers de données jusqu'au modèle mathématique plus quelques paramètres, nous sommes obligés de perdre l'aspect individuel des observations et de les voir à travers une formule les amalgamant. La question est donc: faut-il choisir une méthode classique qui considère chaque observation comme également importante, ou une méthode robuste qui se concentre sur la majorité des données? Certainement que dépendant du choix, le regard porté aux données à travers le modèle est différent!

Si votre choix est le même que le mien, faut-il alors rejeter complètement les méthodes classiques? La réponse est non! En effet, le maximum de vraisemblance, méthode classique d'estimation, est la méthode la plus efficace lorsque le modèle proposé est exact ou, en d'autres termes, lorsque les données ne contiennent pas de valeurs extrêmes. Plus efficace s'entend ici comme la méthode qui permet d'estimer les paramètres avec le plus de précision. Pour expliquer ce dernier point, supposons que nous tirions des échantillons d'une population dont nous avons que la moyenne est 100 par exemple. Pour chaque échantillon, nous calculons un estimateur classique et un robuste de la moyenne. La distribution des valeurs de ces estimateurs peut nous indiquer lequel des deux est le plus souvent le plus

proche de la vraie valeur de 100. Celui qui est le plus souvent le plus proche est appelé le plus efficace et, lorsque les données sont bonnes, c'est le cas de l'estimateur classique du maximum de vraisemblance. N'oubliez pas pour autant que dans la réalité, les données ne sont pas toujours comme nous aimerions qu'elles soient...

Dans ma thèse, je traite des problèmes un peu plus complexes que ceux qui je vous ai exposés jusqu'ici. Les données à disposition, en plus de contenir des erreurs, peuvent aussi être sous formes moins directement tractables, comme groupées en classes de revenus, ou pire avec une partie de l'information manquante. Je développe des méthodes qui sont aussi applicables en général pour d'autres types de problèmes.

L'étude de la répartition des revenus ne s'arrête pas à une partie descriptive. L'étude de l'inégalité dans la distribution des revenus et l'étude de la pauvreté y sont étroitement reliés. Dans ma thèse, j'étudie ces problèmes et démontre par exemple que les méthodes classiques d'estimation de l'inégalité peuvent considérablement être influencées par des valeurs extrêmes.

4 Conclusion

Pour conclure, laissez-moi répéter l'importance des méthodes robustes non seulement pour le problème de la répartition des revenus, de l'inégalité et de la pauvreté, mais aussi pour tout autre problème impliquant une analyse basée sur des modèles statistiques. Depuis mon travail de doctorat, j'ai eu la chance de pouvoir poursuivre la recherche dans la statistique robuste et publier des nouveaux résultats en travaillant aussi avec des spécialistes dans les domaines de l'économie, la finance, les sciences sociales et médicales (Cowell and Victoria-Feser 1994, Cowell and Victoria-Feser

1995, Heritier and Victoria-Feser 1995, Victoria-Feser 1995a, Victoria-Feser 1995b, Victoria-Feser and Ronchetti 1994, Victoria-Feser and Ronchetti 1995). Je suis fermement convaincue qu'une analyse robuste devrait accompagner toute analyse statistique, pour au moins laisser les données, si nécessaire, exprimer une autre version de la réalité.

Références

Aitchison, J. and J.C. Brown (1957). *The Log-Normal Distribution, Cambridge,* Massachussets: Cambridge University Press.

Amoroso, L. (1925). Ricerche intorno alla curva dei redditi. *Annali di Matematica Pura et Applicata* 4-21, 123-157.

Covelle, F. A. and M.-P. Victoria-Feser (1994). Poverty measurement with contaminated data: A robust approach. *European Economic Revue.* Forthcoming.

Cowell, F. A. and M.-P. Victoria-Feser (1995). Robustness properties of inequality measures. *Econometrica.* Forthcoming.

Dagum, C. (1977). A new model of personal income distribution: Spécification and estimation. *Economie Appliquée* 30, 413-436.

Dagum, C. (1980). Generating systems and properties of income distribution models. *Metron* 38(3-4), 3-26.

Davis, H. T. (1941). *The Analysis of Economic Time Series. Bloomington,* Indiana: the Principia Press.

Fisk, P. R. (1961). The graduation of income distribution. *Econometrica 29,* 171-185.

Gibrat, R. (1931). *Les Inégalités Economiques. Paris: Sirey.*

Hampel, F. R. (1968). *Contribution to the Theory of Robust Estimation.* Ph. D. thesis, University of California, Berkeley.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics* 42, 1887-1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69, 383-393.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). Robust Statistics: *The Approache Based on Influence Functions.* New York: John Wiley.

Heritier, S. and M.-P. Victoria-Feser (1995). Some practical applications of bounded-influence tets. In C. Rao (Ed.), *Handbook of Statistics Vol 15: Robust Inference.* Forthcoming.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35, 73-101.

Huber, P. J. (1965). A robust version of the probability ration test. *Annals of Mathematical Statistics* 36, 1753-1758.

Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley.

Lorenz, M. O. (1905). Methods for measuring concentration of wealth. *Journal of the American Statistical Association 9,* 209-219.

Majumder, A. and S. R. Chakravarty (1990). Distribution of personal income: Development of a new model and its application to US income data. *Journal of Applied Econometrics 5,* 189-196.

March, L. (1898). Quelques exemples de distribution de salaires. *Journal de la Société Statistique de Paris,* 193-206.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica 52,* 647-664.

Mc Donald J. B. and M. R. Ransom (1979). Functional forms, estimation techniques and the distribution of income. *Econometrica 47,* 1513-1525.

Pareto, V. (1896). Ecrits sur la courbe de la répartition de la richesse. In *Oeuvres complètes de Vilfredo Pareto.* Giovanni Busino. Librairie Droz, Genève, 1965.

Pareto, V. (1897). Cours d'économie politique. Lausanne, Switzerland. Vol. 2, part I, chapter I.

Singh, S. K. and G. S. Maddala (1976). A fonction for the size distribution of income. *Econometrica 44,* 963-970.

Thurow, L. C. (1970). Analysing the American income distribution. American *Economic Review 60,* 261-269.

Victoria-Feser, M.-P. (1995a). Robust methods for personal income distribution models with application to Dagum's model. In C. Dagum and A. Lemmi (Eds.), *Income Distribution, Social Welfare, Inequality and Poverty.* CT USA: JAI-Press of Greenwich, To appear.

Victoria-Feser, M.-P. (1995b). Robuste model choice text for non-nested hypothesis. *Journal of the Royal Statistical Society, Serie B.* Under revision.

Victoria-Feser, M.-P. and E. Ronchetti (1994). Robust methods for personal income distribution models. *The Canadian Journal of Statistics 22,* 247-258.

Victoria-Feser, M.-P. and E. Ronchetti (1995). Robust estimation for grouped data. *Journal of the American Statistical Association.* Under revision.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics 18,* 293-297