Journées de méthodologie statistique de l'Insee 2022

TROIS PETITS PAPIERS EN QUÊTE D'AUTEUR

Papier n°3 : Du revenu déclaré en tranches au vrai revenu fiscal,
ou comment estimer un modèle économétrique sans disposer d'observations

Marc CHRISTINE (*)

(Luigi PIRANDELLO[1])

(***) *Insee, Direction de la méthodologie et de la coordination statistique et internationale*

mchristine7577@gmail.com

Mots-clés : Échantillonnage, calage, imputation, modèle linéaire, revenu

Domaine concerné : calage sur marges, imputation, modélisation stochastique

Résumé

Une variable essentielle explicative du comportement des ménages dans de multiples domaines est le revenu du ménage. La difficulté est de disposer d'une observation correcte de cette variable. Lorsque les enquêtes sont échantillonnées dans les bases fiscales, on peut disposer d'une donnée *vraie*, le revenu déclaré dans l'IRPP.

Mais des enquêtes, soit plus anciennes, échantillonnées dans les bases issues du recensement de la population (systèmes Octopusse, puis Nautile) ou sur des sujets spécifiques nécessitant des bases de sondage appropriées à la thématique ne permettent pas de disposer de cette information.

Celle-ci est alors collectée dans l'enquête elle-même ou dans le cadre commun qui l'enserre, tel que le Tronc Commun des Enquêtes Ménages (TCM).

Dans ce cadre, la donnée *observée* est évidemment entachée de plusieurs erreurs ou imperfections.

- le revenu déclaré dans le TCM ou dans l'enquête présente en général des sous-estimations systématiques par rapport au revenu vrai (non observé).
- il pose des questions de concept, de champ, de définition, de périmètre (inclut-il ou non les revenus de redistribution, mobiliers, sans parler de la prise en compte ou non de la retenue à la source..), quel contour du ménage utilise-t-il ?
- il peut y avoir des décalages temporels entre l'information collectée dans l'enquête et celle utilisée comme référence pour le calage.

[1] Ce papier a été présenté aux JMS2022 sous le pseudonyme de Luigi PIRANDELLO.

- Il est sujet à de la non-réponse.
- **le déclaratif est incomplet** ; les ménages ont la possibilité de répondre en clair ou, à défaut, en tranches ou seulement sur le second mode.

Ce papier se propose de fournir une approche pour répondre à deux questions :

- peut-on, au niveau individuel, à partir de données déclaratives observées sur le revenu (soit en valeur, soit en tranches), « reconstituer » un revenu vrai au sens du revenu fiscal (supposé *inobservable*) ? On montrera que, si l'on dispose seulement de la *distribution* des revenus vrais et de l'observation des revenus déclarés en tranches, il est possible, à partir d'un modèle linéaire, d'imputer un revenu vrai estimé à chaque unité statistique.

- si l'on dispose de sources fiscales constituant une donnée exogène exacte sur la distribution vraie des revenus, peut-on et comment utiliser ces informations pour les incorporer dans le processus de calage des enquêtes ? De surcroît, on s'intéresse en général non seulement au revenu total (ou revenu moyen) des ménages, mais aussi à la *distribution* des revenus et le calage doit intégrer ces deux dimensions, ce que montrera le papier.

On n'abordera dans ce papier ni les questions conceptuelles ou de champ (on supposera que données collectées et données vraies correspondent au même concept et au même champ), ni celles relatives au traitement de la non-réponse et encore moins celles de l'analyse et la correction des points aberrants : on se placera ici en aval de ce processus d'imputation.

Short abstracts in English and French

This paper deals with the use of income data in household surveys. Most of the time, income is reported by the households surveyed, either in value or by group, and is often different from the true income as measured in tax sources, and underestimated. This paper shows how survey data can nevertheless be calibrated to true total income. This is done by imputing a theoretical income to each household by means of linear models using only observations on reported income and assuming the probabilistic distribution of true income, but without knowing the true income individually.

Ce papier traite de l'utilisation des données sur le revenu dans les enquêtes auprès des ménages. La plupart du temps, le revenu est déclaré par les ménages interrogés, soit en valeur, soit en tranches, et il est souvent différent du vrai revenu tel que mesuré dans les sources fiscales, et sous-estimé. Ce papier montre comment on peut néanmoins caler les données recueillies dans l'enquête sur le vrai revenu total. Pour cela, on impute un revenu théorique à chaque ménage au moyen de modèles linéaires utilisant les seules observations sur les revenus déclarés et en supposant connue la distribution probabiliste du vrai revenu, mais sans connaître les vrais revenus individuellement.

---

1. Introduction

Dans le cadre des travaux du groupe Marges de l'Insee[2], on s'est posé la question suivante : si l'on dispose, d'une part, d'un revenu déclaré de chaque ménage interrogé dans le Tronc commun des ménages (TCM) d'une enquête, et, d'autre part, de sources fiscales (essentiellement les données relatives au revenu déclaré dans l'IRPP) constituant une donnée exogène exacte sur la distribution vraie des revenus, peut-on, et comment, utiliser ces informations pour les incorporer dans le processus de calage des enquêtes ?

La question est compliquée par les circonstances suivantes :

- on s'intéresse en général non seulement au revenu total (ou revenu moyen) des ménages, mais aussi à la *distribution* des revenus et le calage doit intégrer ces deux dimensions.

- le revenu déclaré dans le TCM ou directement dans l'enquête (cf. [3] et annexe 1), à supposer qu'il corresponde conceptuellement au revenu des sources fiscales, considéré comme la donnée de référence (vraie), présente en général des sous-estimations systématiques par rapport au revenu vrai (non observé) (cf. [4] et annexe 2).

- il peut y avoir des décalages temporels entre l'information collectée dans l'enquête et celle utilisée comme référence pour le calage.

- Il peut y avoir des confusions, même de bonne foi, de la part du répondant, depuis le prélèvement à la source de l'IRPP.

À cela s'ajoute une 5ème difficulté : **le déclaratif est tronqué** ; les ménages ont la possibilité, dans le TCM ou dans les enquêtes, de répondre en clair ou, à défaut, en tranches.

Récemment, par exemple, l'auteur s'est penché sur les enquêtes sur le Suivi de la demande Touristique, actuellement sous-traitées à un prestataire privé, tant au niveau de l'échantillonnage que de celui de la collecte, et tirées dans des bases de sondages tout à fait imparfaites. Dans ces enquêtes, réalisées auprès d'un panel de ménages (constitué à partir d'un "vivier" de répondants potentiels), le prestataire collecte actuellement « la somme des ressources mensuelles nettes du foyer avant prélèvement à la source (incluant les pensions, loyers, allocations, 13è mois, primes etc..) », telle que déclarée par le panéliste. En principe, la donnée, ***déclarée en douze tranches***[3], est mise à jour en début de chaque année, mais il n'est pas sûr que ce soit rigoureusement fait.

Enfin, il peut subsister un certain % de non-réponses.

À l'inverse, il est vrai que les perspectives offertes par les ***appariements sur les sources fiscales*** permettent de résoudre une grande partie des problèmes rencontrés.

Ce papier ne remet évidemment pas en cause les travaux et les méthodes déjà mis en œuvre par les experts métiers pour résoudre la question. Il se propose simplement de donner quelques pistes de réflexion qu'il espère nouvelles, comme modeste contribution au sujet.

[2] Ce groupe de travail avait pour objectif de définir les variables et les sources de calage à utiliser de manière standardisée pour le traitement de la plupart des enquêtes auprès des ménages.

[33] Moins de 300 / 300 à 600 / 601 à 900 / 901 à 1200 / 1201 à 1500 / 1501 à 1900 / 1901 à 2300 / 2301 à 2700 / 2701 à 3000 / 3001 à 3800 / 3801 à 5300 : 5301 à 6900 / 6901 ou plus.

2. Position du problème

Dans la méthode classique de calage sur marges, on calcule de nouveaux poids proches des poids initiaux de sondage et astreints à vérifier différentes équations de calage sur des totaux connus de manière exogène. Or, en matière de revenus, compte tenu des difficultés évoquées ci-dessus, il n'est pas licite ni suffisant d'utiliser les valeurs déclarées pour les caler sur les vrais totaux fiscaux.

En préalable, on suppose effectuées différentes étapes de traitement sur l'information collectée dans l'enquête :

- Correction de la non-réponse
- Analyse et une correction ou élimination des points aberrants.
- Éventuellement, transformation des revenus déclarés en tranches, en valeurs[4].

On se place ici en aval de ce processus de traitement.

Le présent papier propose des solutions pour résoudre cette question du calage en passant par la reconstitution d'approximations du revenu *vrai* à partir du revenu *déclaré*.

2.1. Approche non paramétrique

Soit Y_i le revenu *déclaré* du ménage i en valeur (éventuellement *imputé* à partir d'une déclaration en tranches, ou *imputé* pour les non-répondants à partir de modèles utilisant les déclarations des répondants, cf. *supra*).

2.1.1. Formalisation : approche théorique de la transformation des revenus.

Supposons que les revenus observés Y_i soient des variables aléatoires de même loi, définie par sa fonction de répartition F.

On cherche à définir les revenus *vrais* X_i^* comme une *transformation des revenus observés*, c'est-à-dire à construire une transformation T, de telle sorte que la distribution des $X_i^* = T(Y_i)$ ait une fonction de répartition donnée F_0. On doit donc avoir :

$$P\{T(Y_i) < x\} = F_0(x),$$

soit, T étant supposée *continue strictement croissante* (donc bijective) :

$$P\{Y_i < T^{-1}(x)\} = F_0(x).$$

D'où : $F \circ T^{-1} = F_0$, ce qui définit la transformation adéquate : $\boxed{T = F_0^{-1} \circ F}$.

[4] Un processus spécifique est en général opéré sur les données d'enquête au moyen d'équations économétriques (à l'aide de la méthode des *résidus simulés*) mettant en jeu diverses caractéristiques socio-démographiques des ménages, estimées sur les répondants en clair et appliquées aux autres ménages : le programme réalise une **imputation**, à partir des données déclarées par les répondants en clair, à la fois pour les répondants en tranches et pour les non-répondants. Dans le cas d'une réponse en tranches, on s'assure que la valeur simulée se situe bien dans la tranche déclarée, si bien qu'il faut parfois faire plusieurs simulations avant que cela convienne.

2.1.2. **Mise en œuvre sur un échantillon**

La fonction de répartition théorique est supposée définie sur une population finie de taille N, sous forme de moyenne empirique :

$$F_0(x) = \frac{1}{N}\sum_{i=1}^{N} 1_{X_i<x}.$$

Telle que définie, c'est une fonction en escalier. On peut cependant l'ajuster sur une distribution paramétrique continue de forme donnée.

On suppose que l'on dispose d'un échantillon *s* de logements-ménages *i*, le plan de sondage étant caractérisé seulement par les probabilités d'inclusion Π_i (sans précision sur la manière dont il a été tiré).

→ Si F est inconnue, on peut la remplacer par la *fonction de répartition empirique* estimée à partir de l'échantillon :

$\hat{F}_Y(y)$ = estimation de la proportion de ménages dont le revenu déclaré est strictement inférieur à *y* :

$$\hat{F}_y(y) = \frac{\sum_{i\in s} \frac{1_{Y_i<y}}{\Pi_i}}{\sum_{i\in s} \frac{1}{\Pi_i}}.$$

On utilisera alors une transformation approchée : $\boxed{\hat{T} = F_0^{-1} \circ \hat{F}_Y}$.

On définira alors, pour tout ménage *j*, un *revenu vrai estimé* :

$$\boxed{\hat{X}_j = \hat{T}(Y_j) = F_0^{-1}[\hat{F}_Y(Y_j)] = F_0^{-1}\left[\frac{\sum_{i\in s} \frac{1_{Y_i<Y_j}}{\Pi_i}}{\sum_{i\in s}\frac{1}{\Pi_i}}\right]}.$$

► Ordonnons les revenus observés dans l'enquête pour les individus de l'échantillon en les renumérotant de telle sorte que :

$$Y_{(1)} < Y_{(2)} < \ldots < Y_{(n)},$$

où *n* est la taille de l'échantillon (on suppose pour simplifier *qu'il n'y a pas d'ex-æquo dans les valeurs des* Y_i).

Alors : $\hat{F}_Y(Y_{(j)}) = \frac{\sum_{i\in s}\frac{1_{Y_i<Y_{(j)}}}{\Pi_i}}{\sum_{i\in s}\frac{1}{\Pi_i}} = \frac{\sum_{i=1}^{j-1}\frac{1}{\Pi_{(i)}}}{\sum_{i\in s}\frac{1}{\Pi_i}}$ (pour $j \geq 2$) et : $\hat{F}_Y(Y_{(1)}) = 0$, en notant $\Pi_{(i)}$ la probabilité d'inclusion du ménage classé *i*-ième lorsqu'on ordonne les revenus observés.

Alors, la solution précédente s'écrit : $\hat{X}_{(j)} = \hat{T}(Y_{(j)}) = F_0^{-1}[\hat{F}_Y(Y_{(j)})]$, soit :

$$\boxed{\begin{cases} X_{(1)} = F_0^{-1}(0) \\ \hat{X}_{(j)} = F_0^{-1}\left(\frac{\sum_{i=1}^{j-1}\frac{1}{\Pi_{(i)}}}{\sum_{i\in s}\frac{1}{\Pi_i}}\right) \; pour \; j \geq 2 \end{cases}}.$$

2.1.3. Discussion

L'inconvénient de cette solution est qu'elle ne prend pas en compte explicitement les valeurs des revenus observés pour construire des revenus « vrais », mais seulement ***leurs rangs*** dans la séquence des revenus observés et les probabilités d'inclusion des logements (ménages) correspondants.

2.2. Approche paramétrique

2.2.1. Cadre théorique

Tout en se plaçant dans le même cadre que précédemment, on peut spécifier une forme particulière de relation entre les fonctions de répartition F_0 et F. La forme la plus simple consiste à supposer que **les deux distributions ne diffèrent que par des paramètres de position et d'échelle, soit :**

$$F(y) = F_0(\alpha + \beta y).$$

Les revenus vrais seront alors définis par :

$$\boxed{X_i = F_0^{-1} \circ F(Y_i) = \alpha + \beta Y_i}.$$

Ces paramètres (α et β) sont évidemment inconnus. On suppose a priori $\beta > 0$ (T ***croissante***).

On va les estimer à partir de l'échantillon de l'enquête ($\hat{\alpha}$ et $\hat{\beta}$ étant les valeurs estimées) en imposant les conditions suivantes :

- l'estimation du revenu total des ménages fondée sur l'échantillon de l'enquête est conforme à la vraie valeur R_T connue par les sources externes (fiscales). **(C1)**

- la répartition des revenus dans l'enquête est conforme à la vraie répartition, celle-ci étant approchée par les *quantiles* de la distribution exacte (exogène)[5]. **(C2)**

Ces conditions vont conduire aux équations suivantes :

(C1) : $\sum_{i \in s} \frac{X_i}{\Pi_i} = R_T$, soit : $\hat{\alpha} \sum_{i \in s} \frac{1}{\Pi_i} + \hat{\beta} \sum_{i \in s} \frac{Y_i}{\Pi_i} = R_T$.

(C2) : l'ajustement sur le quantile d'ordre τ, soit q_τ, de la vraie distribution va s'écrire :

$$\sum_{i \in s} \frac{1_{X_i < q_\tau}}{\Pi_i} = \tau \sum_{i \in s} \frac{1}{\Pi_i},$$

Dans cette dernière égalité :

- la somme de gauche représente l'estimateur de HORVITZ-THOMSON, fondé sur l'échantillon de l'enquête, du nombre de ménages dont le revenu réel est strictement inférieur à q_τ

- celle de droite, l'estimateur du nombre total de ménages (estimateur de HÁJEK) .

Le quotient de ces deux sommes représente donc l'estimation de la proportion de ménages dont le revenu réel est strictement inférieur à q_τ. Cette estimation doit être ajustée sur la probabilité τ.

[5] Pour simplifier, on propose ici un calage sur des quantiles. Des travaux théoriques sur le calage sur une fonction de répartition pourraient être appliqués dans ce cadre.

La condition obtenue s'écrit aussi :

$$\sum_{i \in s} \frac{1_{Y_i < \frac{q_\tau - \alpha}{\beta}}}{\Pi_i} = \tau \sum_{i \in s} \frac{1}{\Pi_i}.$$

Notons $\hat{q}_\tau(Y)$ le **quantile empirique d'ordre τ de la distribution des Y_i dans l'échantillon**, définie par la relation[6] : $\sum_{i \in s} \frac{1_{Y_i < \hat{q}_\tau(Y)}}{\Pi_i} = \tau \sum_{i \in s} \frac{1}{\Pi_i}$. La condition **(C2)** devient alors tout simplement :

$$\hat{q}_\tau(Y) = \frac{q_\tau - \alpha}{\beta}, \text{ ou : } \boxed{\alpha + \beta \hat{q}_\tau(Y) = q_\tau}.$$

La condition **(C2)** va en fait être déclinée pour différentes valeurs de τ, par exemple : $\tau = \frac{1}{p}, \frac{2}{p}, \ldots, \frac{p-1}{p}$, où p est un entier fixé (= 2, 4, 10, 100…).

On aura donc $p - 1$ équations :

$$\alpha + \beta \hat{q}_{\frac{k}{p}}(Y) = q_{\frac{k}{p}}, \text{ pour } k = 1, 2, \ldots, p-1,$$

où $q_{\frac{k}{p}}$ est le quantile exact d'ordre $\frac{k}{p}$ de la vraie distribution des revenus et $\hat{q}_{\frac{k}{p}}(Y)$ le quantile empirique correspondant, pour la distribution, dans l'échantillon, des revenus déclarés Y_i.

Finalement, les paramètres estimés $\hat{\alpha}$ et $\hat{\beta}$ doivent vérifier le système :

$$\begin{cases} \hat{\alpha} \sum_{i \in s} \frac{1}{\Pi_i} + \hat{\beta} \sum_{i \in s} \frac{Y_i}{\Pi_i} = R_T \\ \alpha + \beta \hat{q}_{\frac{k}{p}}(Y) = q_{\frac{k}{p}}, \text{pour } k = 1, 2, \ldots, p-1. \end{cases}$$

2.2.2. Résolution approchée

Mis à part le cas $p = 2$ (ajustement sur la seule médiane de la vraie distribution des revenus, où l'on obtient un système de deux équations à deux inconnues), ce système (à *p* équations) ne possède en général pas de solution.

On peut convenir, par exemple, de chercher des solutions approchées :

- satisfaisant la contrainte $\hat{\alpha} \sum_{i \in s} \frac{1}{\Pi_i} + \hat{\beta} \sum_{i \in s} \frac{Y_i}{\Pi_i} = R_T$
- et minimisant :

$$\sum_{k=1}^{p-1} [\alpha + \beta \hat{q}_{\frac{k}{p}}(Y) - q_{\frac{k}{p}}]^2$$

On obtient alors aisément : $\hat{\alpha} = \frac{R_T - \hat{\beta} \sum_{i \in s} \frac{Y_i}{\Pi_i}}{\sum_{i \in s} \frac{1}{\Pi_i}}$ et on doit ensuite minimiser :

$$\sum_{k=1}^{p-1} [\hat{\alpha} + \hat{\beta} \hat{q}_{\frac{k}{p}}(Y) - q_{\frac{k}{p}}]^2$$

[6] On suppose provisoirement que cette égalité est vérifiée pour un $\hat{q}_\tau(Y)$ unique. Voir discussion technique plus loin.

$$= \sum_{k=1}^{p-1} \left[\frac{R_T - \hat{\beta} \sum_{i \in s} \frac{Y_i}{\Pi_i}}{\sum_{i \in s} \frac{1}{\Pi_i}} + \hat{\beta}\, \hat{q}_{\frac{k}{p}}(Y) - q_{\frac{k}{p}} \right]^2 = \sum_{k=1}^{p-1} \left[\hat{\beta}\, [\hat{q}_{\frac{k}{p}}(Y) - \frac{\sum_{i \in s} \frac{Y_i}{\Pi_i}}{\sum_{i \in s} \frac{1}{\Pi_i}}] + \frac{R_T}{\sum_{i \in s} \frac{1}{\Pi_i}} - q_{\frac{k}{p}} \right]^2 .$$

La solution est alors, par un calcul simple :

$$\hat{\beta} = \frac{\sum_{k=1}^{p-1} \left[\hat{q}_{\frac{k}{p}}(Y) - \frac{\sum_{i \in s} \frac{Y_i}{\Pi_i}}{\sum_{i \in s} \frac{1}{\Pi_i}} \right] \left[\frac{R_T}{\sum_{i \in s} \frac{1}{\Pi_i}} - q_{\frac{k}{p}} \right]}{\sum_{k=1}^{p-1} \left[\hat{q}_{\frac{k}{p}}(Y) - \frac{\sum_{i \in s} \frac{Y_i}{\Pi_i}}{\sum_{i \in s} \frac{1}{\Pi_i}} \right]^2}.$$

On peut simplifier cette expression en remarquant que $\frac{\sum_{i \in s} \frac{Y_i}{\Pi_i}}{\sum_{i \in s} \frac{1}{\Pi_i}}$ est l'estimateur, issu de l'échantillon, du revenu moyen *déclaré*, soit : $\hat{R}_m(Y)$, et que $\frac{R_T}{\sum_{i \in s} \frac{1}{\Pi_i}}$ est un estimateur du revenu moyen *réel*, soit : $\hat{R}_m$. On obtient alors :

$$\hat{\beta} = \frac{\sum_{k=1}^{p-1} \left[\hat{q}_{\frac{k}{p}}(Y) - \hat{R}_m(Y) \right] \left[q_{\frac{k}{p}} - \hat{R}_m \right]}{\sum_{k=1}^{p-1} \left[\hat{q}_{\frac{k}{p}}(Y) - \hat{R}_m(Y) \right]^2}.$$

2.2.3. Discussion technique

La fonction de répartition empirique $\hat{F}_Y(y)$ est une fonction en escalier, discontinue en chaque valeur Y_i. Si l'on ordonne les revenus observés dans l'enquête en les renumérotant $Y_{(i)}$, il est facile de voir que cette fonction est **constante sur les intervalles]$Y_{(i)}$, $Y_{(i+1)}$[**.

La définition du quantile empirique d'ordre τ de la distribution des Y_i, soit $\hat{q}_\tau(Y)$, donnée précédemment par l'équation $\hat{F}_Y[\hat{q}_\tau(Y)] = \tau$, n'est donc pas tout à fait correcte puisqu'elle n'assure ni l'existence ni l'unicité d'un tel quantile. Il faut en fait introduire une définition adaptée à la non-bijectivité de la fonction $\hat{F}_Y$, soit :

$$\hat{F}_Y[\hat{q}_\tau(Y)] \leq \tau \leq \hat{F}_Y^+[\hat{q}_\tau(Y)],$$
où $\hat{F}_Y^+$ est la limite à droite de $\hat{F}_Y$.

On observe alors que, si $\hat{F}_Y(Y_{(j)}) \leq \tau < \hat{F}_Y(Y_{(j+1)})$, **tout réel q_τ tel que : $Y_{(j)} < q_\tau \leq Y_{(j+1)}$ répond à cette définition et peut être considéré comme quantile empirique d'ordre τ**.

Compte tenu des valeurs écrites précédemment des quantités $\hat{F}_Y(Y_{(j)})$, cette condition peut s'écrire :

Pour τ donné et $j \geq 2$ tel que : $\frac{\sum_{i=1}^{j-1} \frac{1}{\Pi_{(i)}}}{\sum_{i \in s} \frac{1}{\Pi_i}} \leq \tau < \frac{\sum_{i=1}^{j} \frac{1}{\Pi_{(i)}}}{\sum_{i \in s} \frac{1}{\Pi_i}}$, tous les points de]$Y_{(j)}$, $Y_{(j+1)}$] peuvent convenir comme quantile empirique $\hat{q}_\tau(Y)$.

- On peut lever cette difficulté simplement en convenant qu'alors : $\hat{q}_\tau(Y) = \frac{Y_{(j)} + Y_{(j+1)}}{2}$.

- On peut au contraire s'en servir explicitement pour relâcher certaines contraintes. En effet, l'équation **(C2)**, dont la formulation précédente équivalait à : $\hat{F}_Y(\frac{q_\tau - \alpha}{\beta}) = \tau$, en assimilant la fonction $\hat{F}_Y$ a une fonction bijective, doit être remplacée par : $\hat{F}_Y(\frac{q_\tau - \alpha}{\beta}) \leq \tau \leq \hat{F}_Y^+(\frac{q_\tau - \alpha}{\beta})$, ce qui équivaut à :

$$Y_{(j_\tau)} < \frac{q_\tau - \alpha}{\beta} \leq Y_{(j_\tau+1)}, \text{ où } j_\tau \geq 2 \text{ est défini par : } \frac{\sum_{i=1}^{j_\tau - 1} \frac{1}{\Pi_{(i)}}}{\sum_{i \in s} \frac{1}{\Pi_i}} < \tau < \frac{\sum_{i=1}^{j_\tau} \frac{1}{\Pi_{(i)}}}{\sum_{i \in s} \frac{1}{\Pi_i}}.$$

L'équation **(C2)** est alors transformée en un système de doubles contraintes à l'inégalité :

$$\{\alpha + \beta Y_{(j_\tau)} < q_\tau \leq \alpha + \beta Y_{(j_\tau+1)}.$$

Ces différentes contraintes (correspondant à différentes valeurs de τ) ne sont pas nécessairement compatibles. On peut chercher à en satisfaire le plus grand nombre possible, ce qui va conduire au critère d'optimisation suivant : si une contrainte de type $\delta_k \leq 0$ est satisfaite, on lui attribue un score nul ; sinon, on lui attribue un score d'autant plus élevé que δ_k est plus éloigné (positivement) de 0. On peut ainsi prendre comme expression du score, par exemple : $\delta_k 1_{\delta_k > 0}$ et on va chercher à minimiser le score (fonction positive ou nulle).

En prenant $\tau = \frac{k}{p}$ pour $k = 1, 2, \ldots, p-1$ et en notant $j_{\frac{k}{p}}$ les indices correspondant aux quantiles $q_{\frac{k}{p}}$ d'ordre $\frac{k}{p}$ pour la vraie distribution des revenus (mêmes notations que dans le § précédent), le score à minimiser s'écrira :

$$\sum_{k=1}^{p-1} [[\alpha + \beta Y_{(j_{\frac{k}{p}})} - q_{\frac{k}{p}}] 1_{\alpha + \beta Y_{(j_{\frac{k}{p}})} - q_k \geq 0} + [q_{\frac{k}{p}} - \alpha - \beta\, Y_{(j_{\frac{k}{p}}+1)}] 1_{q_{\frac{k}{p}} - \alpha - \beta\, Y_{(j_{\frac{k}{p}}+1)}}],$$

sous la contrainte **(C1)** :

$$\alpha \sum_{i \in s} \frac{1}{\Pi_i} + \beta \sum_{i \in s} \frac{Y_i}{\Pi_i} = R_T.$$

2.2.4. Amélioration possible de la solution

L'inconvénient de la solution précédente est qu'elle suppose qu'il existe un modèle ***unique*** de relation entre le revenu déclaré et le revenu vrai.

On peut améliorer la procédure en postulant différents modèles selon différentes catégories de logements, pourvu qu'on puisse définir et identifier celles-ci de manière cohérente dans le TCM ou dans l'enquête et dans la source fiscale de référence : on peut ainsi déterminer des coefficients spécifiques pour chaque catégorie de logements (par exemple selon la taille du ménage, son type, l'âge de la personne de référence …).

Pour que la distribution d'ensemble soit respectée, il convient toutefois que les poids des diverses catégories dans l'enquête soient conformes à leur poids réel, ce que l'on pourra obtenir au travers de la phase de calage sur ces structures.

On peut aussi ajuster différents modèles par **classe de revenus**. Pour simplifier et par souci de cohérence, on prend des classes définies à partir des quantiles.

Par exemple, supposons qu'on distingue deux catégories de revenus : au-dessus et en dessous de la médiane vraie μ.

Pour les revenus X_i inférieurs à la médiane, on postule un modèle : $X_i = \alpha_1 + \beta_1 Y_i$ et pour les revenus supérieurs, un modèle : $X_i = \alpha_2 + \beta_2 Y_i$.

On va écrire des équations de calage sur chaque catégorie de revenus séparément. Ainsi, pour les revenus de la classe « inférieure » :

- l'équation de conformité au revenu total vrai s'écrira : $\sum_{i \in s} \frac{X_i \, 1_{X_i < \mu}}{\Pi_i} = R_1$, où R_1 est la vraie valeur (connue par les sources externes) du total des revenus des ménages de cette catégorie. Cette équation devient :

$$\alpha_1 \sum_{i \in s} \frac{1_{Y_i < \frac{\mu - \alpha_1}{\beta_1}}}{\Pi_i} + \beta_1 \sum_{i \in s} \frac{Y_i \, 1_{Y_i < \frac{\mu - \alpha_1}{\beta_1}}}{\Pi_i} = R_1. \textbf{ (C1)}$$

- l'équation d'ajustement sur le quantile d'ordre τ, soit q_τ, de la vraie distribution s'écrira comme précédemment : $\sum_{i \in s} \frac{1_{X_i < q_\tau}}{\Pi_i} = \sum_{i \in s} \frac{1_{Y_i < \frac{q_\tau - \alpha_1}{\beta_1}}}{\Pi_i} = \tau \sum_{i \in s} \frac{1}{\Pi_i}$ pour des valeurs $\tau \leq \frac{1}{2}$, soit encore (sous réserve de l'hypothèse simplificatrice d'unicité des quantiles) :

$$\alpha_1 + \beta_1 \hat{q}_\tau(Y) = q_\tau.$$

En particulier, pour $\tau = \frac{1}{2}$, on obtient : $\sum_{i \in s} \frac{1_{Y_i < \frac{\mu - \alpha_1}{\beta_1}}}{\Pi_i} = \frac{1}{2} \sum_{i \in s} \frac{1}{\Pi_i}$, ce qui permet de récrire l'équation **(C1)** sous la forme :

$$\frac{\alpha_1}{2} \sum_{i \in s} \frac{1}{\Pi_i} + \beta_1 \sum_{i \in s} \frac{Y_i \, 1_{Y_i < \frac{\mu - \alpha_1}{\beta_1}}}{\Pi_i} = R_1.$$

En considérant différentes valeurs de τ, par exemple : $\tau = \frac{1}{p}, \frac{2}{p}, \ldots, \frac{1}{2}$, où p est un entier fixé *qu'on peut supposer **pair** pour simplifier* (= 2, 4, 10, 100…) et en notant comme précédemment $q_{\frac{k}{p}}$ le quantile exact d'ordre $\frac{k}{p}$ de la vraie distribution des revenus et $\hat{q}_{\frac{k}{p}}(Y)$ le quantile empirique correspondant pour la distribution des revenus déclarés Y_i, on obtiendra les équations :

$$\begin{cases} \alpha_1 + \beta_1 \hat{q}_{\frac{k}{p}}(Y) = q_{\frac{k}{p}}, \text{ pour } k = 1, 2, \ldots, \dfrac{p-1}{2} \\ \alpha_1 + \beta_1 \, \hat{\mu}(Y) = \mu. \end{cases}$$

où $\hat{\mu}(Y)$ est la médiane empirique de la distribution des Y_i.

On peut alors chercher, comme précédemment, à déterminer les paramètres estimés $\hat{\alpha}_1$ et $\hat{b}_1$:

satisfaisant la contrainte $\frac{\hat{\alpha}_1}{2} \sum_{i \in s} \frac{1}{\Pi_i} + \hat{\beta}_1 \sum_{i \in s} \frac{Y_i \, 1_{Y_i < \frac{\mu - \hat{\alpha}_1}{\hat{\beta}_1}}}{\Pi_i} = R_1$

et minimisant

$$\sum_{k=1}^{p/2} [\hat{\alpha}_1 + \hat{\beta}_1 \hat{q}_{\frac{k}{p}}(Y) - q_{\frac{k}{p}}]^2 + [\hat{a}_1 \hat{\mu}(Y) + \hat{b}_1 - \mu]^2.$$

Le système obtenu est hautement non linéaire.

On procède de la même façon, avec des coefficients inconnus α_2 et β_2 à estimer, pour la catégorie des revenus « supérieurs ».

3. Imputation et calage simultanés

Application au calage ultérieur.

Une fois estimés des revenus vrais $X_i^* (= \hat{\alpha} + \hat{\beta} Y_i)$, on pourra introduire dans les équations de calage de l'enquête un certain nombre de conditions relatives à ces revenus (ainsi, naturellement, que des conditions sur diverses autres variables socio-démographiques) :

- calage sur le revenu total : $\sum_{i \in s} \omega_i X_i^* = R_T$ (les nouveaux poids issus du calage étant notés ω_i) . **Cette équation est déjà satisfaite avec les poids de sondage $\frac{1}{\pi_i}$, par construction des X_i,** (du fait de la condition **C1**), ce qui devrait renforcer, mécaniquement, la proximité entre les nouveaux poids de calage et les poids initiaux de sondage.

- calage sur les quantiles de la vraie distribution : $\boxed{\sum_{i \in s} \omega_i 1_{X_i^* < q_\tau} = \tau \sum_{i \in s} \omega_i}$

  ou encore : $\boxed{\sum_{i \in s} \omega_i 1_{X_i^* < q_\tau} = \tau N}$ ($N =$ nombre total de ménages), si l'on impose par ailleurs la condition de calage sur l'effectif total : $\sum_{i \in s} \omega_i = N$.

Là encore, ces équations sont déjà satisfaites avec les poids de sondage initiaux, pour les quantiles sur lesquels on a ajusté la construction des estimations des revenus vrais X_i^*, mais on peut caler aussi l'enquête sur d'autres quantiles si on le souhaite.

4. Modélisation économétrique

Dans ce qui précède, on a en fait imposé une relation linéaire exacte entre revenu déclaré et revenu vrai. Il paraît plus naturel et plus efficace de postuler un modèle linéaire aléatoire reliant ces deux grandeurs.

Dans la nouvelle approche développée maintenant, on va montrer qu'on peut imputer à chaque ménage un vrai revenu ou plus exactement un prédicteur optimal du vrai revenu, en connaissant seulement un revenu déclaratif, exprimé soit en *valeur*, soit en *tranches*, mais en postulant un modèle économétrique très simple reliant les deux variables ***et en supposant connue la distribution des vrais revenus dans l'univers.***

► En particulier, on montre ainsi qu'on peut estimer les coefficients d'un modèle linéaire sans même disposer d'observations de la variable explicative !

L'approche est principalement stochastique, on verra ensuite comment la mettre en œuvre dans le cadre d'un modèle d'échantillonnage.

4.1. Cas où on connaît le revenu déclaratif

4.1.1. Préliminaires probabilistes

- Soit X une variable aléatoire réelle, de fonction de répartition F_0, supposée **continue**, de densité f_0 et **de carré intégrable**.

  On considère le modèle linéaire définissant une nouvelle variable aléatoire Y :

  $$Y = a + bX + U,$$

  où X et U sont **indépendantes** et $U \rightsquigarrow \mathcal{N}(0, \sigma^2)$, avec $b > 0$ et $\sigma > 0$.

- Soit *g* la densité de Y. Y est la somme de la variable aléatoire bX, de densité $\frac{1}{b} f_0\left(\frac{x}{b}\right)$ et de la variable $a + U$, de loi $\mathcal{N}(a, \sigma^2)$, les deux étant indépendantes.

  La densité *g* s'obtient donc par convolution des densités des lois de ces variables :

$$g(y) = \int_{-\infty}^{+\infty} \frac{1}{b} f_0\left(\frac{y-u}{b}\right) \frac{1}{\sigma} h\left(\frac{u-a}{\sigma}\right) du,$$

  en notant *h* la densité de $\mathcal{N}(0, 1)$.

  Le changement de variables $z = \frac{y-u}{b}$ conduit à :

$$\boxed{g(y) = \frac{1}{\sigma} \int_{-\infty}^{+\infty} f_0(z)\, h\left(\frac{y-a-b}{\sigma}\right) dz}.$$

- On considère maintenant une suite de couples $(X_i, U_i)_{i \in \mathbb{N}^*}$ mutuellement indépendants, où X_i et U_i sont indépendants entre eux et de même loi que le couple (X, U) défini précédemment, et l'on pose : $Y_i = a + bX_i + U_i$.

  Dans la pratique, les Y_i observés représenteront les revenus déclarés et les X_i inconnus, les revenus vrais.

  Si l'on dispose de *n* observations des Y_i, la vraisemblance du modèle s'écrira :

$$L(y_1, \ldots, y_n) = \prod_{i=1}^{n} g(y_i) = \frac{1}{\sigma^n} \prod_{i=1}^{n} \left[\int_{-\infty}^{+\infty} f_0(z)\, h\left(\frac{y_i - a - bz}{\sigma}\right) dz \right]$$

  et la log-vraisemblance :

$$\boxed{ln\, L(y_1, \ldots, y_n) = -n \ln \sigma + \sum_{i=1}^{n} ln\, g(y_i) = -n \ln \sigma + \sum_{i=1}^{n} ln \left[\int_{-\infty}^{+\infty} f_0(z)\, h\left(\frac{y_i - a - bz}{\sigma}\right) dz \right]}$$

4.1.2. Estimations des paramètres

- Il sera alors possible, théoriquement, d'estimer les paramètres a, b, σ par la méthode du maximum de vraisemblance, en maximisant la fonction $ln\, L$, **et ce sans connaître les observations des X_i mais seulement leur distribution, par l'intermédiaire de la fonction de densité f_0** (*sous réserve cependant de l'existence de solutions identifiables*).

- Équations du maximum de vraisemblance :

  Si l'on suppose qu'il est possible de dériver sous le signe intégrale, on a :

$$\frac{\partial}{\partial a} \left[\int_{-\infty}^{+\infty} f_0(z)\, h\left(\frac{y_i - a - bz}{\sigma}\right) dz \right] = \int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i - a - bz}{\sigma}\right) \left(-\frac{1}{\sigma}\right) dz$$

$$\frac{\partial}{\partial b} \left[\int_{-\infty}^{+\infty} f_0(z)\, h\left(\frac{y_i - a - bz}{\sigma}\right) dz \right] = \int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i - a - bz}{\sigma}\right) \left(-\frac{z}{\sigma}\right) dz$$

$$\frac{\partial}{\partial\sigma}\left[\int_{-\infty}^{+\infty} f_0(z)\, h\left(\frac{y_i-a-bz}{\sigma}\right)dz\right] = \int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i-a-bz}{\sigma}\right)\left(-\frac{y_i-a-bz}{\sigma^2}\right)dz.$$

D'où :

$$\frac{\partial \ln L}{\partial a} = -\frac{1}{\sigma}\sum_{i=1}^{n}\frac{1}{g(y_i)}\left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i-a-bz}{\sigma}\right)dz\right]$$

$$\frac{\partial \ln L}{\partial b} = -\frac{1}{\sigma}\sum_{i=1}^{n}\frac{1}{g(y_i)}\left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i-a-bz}{\sigma}\right)z\,dz\right]$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{1}{g(y_i)}\left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i-a-bz}{\sigma}\right)\left(-\frac{y_i-a-bz}{\sigma^2}\right)dz\right].$$

On en déduit les équations du maximum de vraisemblance :

$$\sum_{i=1}^{n}\frac{1}{g(y_i)}\left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i-a-bz}{\sigma}\right)dz\right] = 0$$

$$\sum_{i=1}^{n}\frac{1}{g(y_i)}\left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i-a-bz}{\sigma}\right)z\,dz\right] = 0$$

$$\sigma = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{g(y_i)}\left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i-a-bz}{\sigma}\right)(y_i-a-bz)\,dz\right]$$

En développant la dernière équation, et en réutilisant les deux premières, on obtient :

$$\boxed{\sigma = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{g(y_i)}\left[\int_{-\infty}^{+\infty} f_0(z)\, h'\left(\frac{y_i-a-bz}{\sigma}\right)dz\right].}$$

- On peut aussi, plus simplement, estimer les paramètres au moyen de la *méthode des moments* (sous réserve de l'existence des moments de X à un ordre adéquat).

  En effet, on a :

$$EY = a + bEX, \text{ d'où : } Y - EY = b(X - EX) + U$$
$$VY = b^2VX + \sigma^2$$
$$(Y-EY)^3 = b^3(X-EX)^3 + 3b^2(X-EX)^2U + 3b(X-EX)U^2 + U^3,$$

  d'où, en tenant compte de l'indépendance de X et U et de la loi normale centrée réduite de U :

$$E(Y-EY)^3 = b^3E(X-EX)^3.$$

  Si le moment centré d'ordre 3 de X, soit μ_3, n'est pas nul, la méthode des moments permettra d'écrire (pour les estimateurs des paramètres):

$$b^3 = \frac{1}{\mu_3}\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{Y})^3$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2 - b^2 VX$$

$$a = \overline{Y} - b\,EX = \frac{1}{n}\sum_{i=1}^{n} y_i - b\,EX.$$

On ne pourra donc estimer ainsi les paramètres que si $\boldsymbol{\mu_3 \neq 0}$.

Intuitivement, on ne peut capter l'effet de X sur Y que si la loi de X est *asymétrique*. Dans le cas où X suivrait une loi normale, les effets conjoints des deux variables inobservables X et U seraient « noyés » dans une même loi normale et les paramètres ne seraient pas identifiables (à moins d'introduire des contraintes *ex-ante*). Dans ce cas, les équations du maximum de vraisemblance ne permettent pas d'estimer les paramètres du modèle.

On supposera dans la suite que la loi de X a une forme qui permette l'identifiabilité des paramètres.

4.1.3. Prédiction de X à partir de de Y

- On va chercher l'expression de la meilleure approximation de X par une variable aléatoire dépendant de Y, de la forme $Z = u(Y)$, où u est une fonction à déterminer, au sens de la norme dans L_2, c'est-à-dire minimisant l'écart quadratique moyen $E(X-Z)^2$ [expression licite dès lors que X et Z sont dans L_2].

  La solution est, classiquement, l'espérance conditionnelle : $Z^* = u^*(Y) = E(X/Y)$.

  Elle s'exprime comme intégrale de la loi conditionnelle de X sachant Y = *y*, soit :

$$E(X/Y = y) = \int x \frac{f_{(X,Y)}(x,y)}{f_Y(y)}\,dx,$$

  en notant $f_{(X,Y)}(x,y)$ (resp. $f_X(y)$) la densité du couple (X, Y) (resp. la densité marginale de Y).

- Celles-ci se calculent à partir de la densité du couple (X, U) :

$$f_{(X,U)}(x,u) = f_0(x)\ \frac{1}{\sigma}\ h\left(\frac{u}{\sigma}\right), \text{ d'où : } f_{(X,Y)}(x,y) = f_0(x)\ \frac{1}{\sigma}\ h\left(\frac{y-a-b}{\sigma}\right)$$

$$f_Y(y) = \int f_{(X,Y)}(x,y)dx.$$

  Par suite :

$$E(X/\,Y = y) = \frac{\int x\, f_{(X,Y)}(x,y)dx}{\int f_{(X,Y)}(x,y)dx} = \frac{\int x\, f_0(x)\ h\left(\frac{y-a-bx}{\sigma}\right)dx}{\int f_0(x)\ h\left(\frac{y-a-bx}{\sigma}\right)dx}.$$

- Pour chaque observation y_0 de Y, on peut calculer le prédicteur optimal correspondent de X, soit :

$$\boxed{X^*(y_0) = E(X/\,Y = y_0) = \frac{\int x\, f_0(x)\;\; h\left(\frac{y_0 - a - bx}{\sigma}\right) dx}{\int f_0(x)\;\; h\left(\frac{y_0 - a - bx}{\sigma}\right)\, dx}.}$$

4.2. Cas où l'on ne connaît le revenu déclaratif qu'en tranches

On reprend le même cadre conceptuel que précédemment.

On n'observe plus maintenant les réalisations des Y_i mais, pour $K \geq 2$, on considère K réels :

$$s_1 < s_2 < \ldots < s_K.$$

On note N_1(resp. N_j, resp. N_{K+1}) le nombre de variables Y_i, parmi les *n* premières, qui prennent une valeur < s_1(resp. dans $[s_{j-1}, s_j[$, resp. $\geq s_K$). Naturellement, on a :

$$\sum_{j=1}^{K+1} N_j = n.$$

Les seules observations agrégées disponibles sont les N_j.

4.2.1. Loi des observations

- On pose : $p_1 = P\{Y < s_1\}\,; p_j = P\{s_{j-1} \leq \mathrm{Y} < s_j\}\,; p_{K+1} = P\{\mathrm{Y} \geq s_K\}.$

  Le K+1-uplet $(N_1, \ldots, N_{K+1})$ a pour loi la ***loi multinomiale*** définie par :

  $$L(n_1, \ldots, n_{K+1}) = P\{N_1 = n_1, \ldots, N_{K+1} = n_{K+1}\} = \frac{n!}{n_1! \ldots n_{K+1}!} p_1^{n_1} \ldots p_{K+1}^{n_{K+1}},$$

  où L est la "*vraisemblance*" associée aux observations $(n_1, \ldots, n_{K+1})$.

- Pour exprimer les p_j, on va utiliser la fonction de répartition G de Y :

  $$G(y) = P\{Y < y\} = P\{a + bX + U < y\} = P\left\{ \underbrace{X + \frac{U}{b}}_{=S} < \frac{y - a}{b} \right\}$$

  [car $b > 0$]

  Or :

  $\frac{U}{b} \rightsquigarrow \mathcal{N}(0, \frac{\sigma^2}{b^2})$, de densité $\frac{b}{\sigma} h\left(\frac{bu}{\sigma}\right)$, en notant *h* la densité de $\mathcal{N}(0, 1)$.

  Par convolution (pour l'expression d'une fonction de répartition de la somme S de deux v.a. indépendantes) :

  $$P\{S < s\} = \int_{-\infty}^{+\infty} F_0(s - u) \frac{b}{\sigma} h\left(\frac{bu}{\sigma}\right) du.$$

D'où :

$$G(y) = \int_{-\infty}^{+\infty} F_0\left(\frac{y-a}{b} - u\right)\frac{b}{\sigma}h\left(\frac{bu}{\sigma}\right)du.$$

Le changement de variables : $v = \frac{bu}{\sigma}, u = \frac{\sigma v}{b}, du = \frac{\sigma}{b}dv$ conduit à :

$$G(y) = \int_{-\infty}^{+\infty} F_0\left(\frac{y-a-\sigma v}{b}\right)h(v)\,dv. \textbf{ (1)}$$

- On a alors les relations définissant les p_j :

$$\begin{cases} p_1 = G(s_1) \\ p_j = G(s_j) - G(s_{j-1}) \; pour \; 2 \leq j \leq K \\ p_{K+1} = 1 - G(s_K). \end{cases}$$

4.2.2. Estimation des paramètres

- On peut estimer les paramètres a, b, σ par la méthode du maximum de vraisemblance, en maximisant la fonction L, d'où on tirera des estimateurs des p_i.

  On a : $ln\ L(n_1, \ldots, n_{K+1}) = ln\ \left(\frac{n!}{n_1! \ldots n_{K+1}!}\right) + \sum_{j=1}^{K+1} n_j\ ln\ p_j$.

  Les équations du 1er ordre s'écrivent :

$$\frac{\partial\ ln\ L}{\partial \phi} = 0, \text{ où } \phi = (a, b, \sigma), \text{ soit :}$$
$$\sum_{j=1}^{K+1} \frac{n_j}{p_j} \frac{\partial p_j}{\partial \phi} = 0.$$
(2)

- Les $\frac{\partial p_j}{\partial \phi}$ s'obtiennent à partir de la dérivation de G sous le signe intégrale (*si celle-ci est licite*). En supposant F_0 dérivable, de dérivée f_0 (= densité de X), on obtient :

$$\frac{\partial G(y)}{\partial a} = -\frac{1}{b} \int_{-\infty}^{+\infty} f_0\left(\frac{y - a - \sigma v}{b}\right) h(v) dv.$$
$$\frac{\partial G(y)}{\partial \sigma} = -\frac{1}{b} \int_{-\infty}^{+\infty} f_0\left(\frac{y - a - \sigma v}{b}\right) v\ \ h(v) dv.$$
$$\frac{\partial G(y)}{\partial b} = -\frac{1}{b^2} \int_{-\infty}^{+\infty} f_0\left(\frac{y - a - \sigma v}{b}\right) (y - a - \sigma v)\ h(v)\ dv.$$
(3)

  D'où :

$$\textbf{(4)} \begin{cases} \frac{\partial p_1}{\partial \phi} = \frac{\partial G(s_1)}{\partial \phi} \\ \frac{\partial p_j}{\partial \phi} = \frac{\partial G(s_j)}{\partial \phi} - \frac{\partial G(s_{j-1})}{\partial \phi} \\ \frac{\partial p_{K+1}}{\partial \phi} = -\frac{\partial G(s_K)}{\partial \phi} \end{cases},$$

  et l'on remplace ensuite les $\frac{\partial G(s_j)}{\partial \phi}$ par leurs expressions intégrales ci-dessus.

  En recombinant les différentes formules **(1)**, **(2)**, **(3)**, **(4)**, on obtient les équations de vraisemblance pour les paramètres a, b, σ.

► Ainsi, là encore, il est possible d'estimer (au moins numériquement) les paramètres du modèle $Y_i = a + bX_i + U_i$, ***sans connaître les observations des X_i, mais seulement la fonction de répartition théorique de leur distribution***, et en utilisant les observations des N_j qui résultent des observations ***tronquées*** des Y_i en tranches.

4.2.3. Prédiction optimale de X à partir des variables en tranches issues de Y

- Supposons dans un premier temps qu'on ne puisse observer que deux tranches : $Y < y$ ou $Y \geq y$ où y est un reel donné.

  On va chercher l'expression de la meilleure approximation de X, au sens de la norme dans L_2, par une variable aléatoire de la forme $Z = u(1_{Y<y})$, où *u* est une fonction à déterminer, qui minimiser l'écart quadratique moyen $E(X-Z)^2$ [expression licite dès lors que X et Z sont dans L_2].

  On remarque que *Z* ne prend que deux valeurs, selon les valeurs de $1_{Y<y}$, soit u_1 si cette indicatrice prend la valeur 1, u_0 sinon. On peut donc écrire : $Z = u_1 1_{Y<y} + u_0 1_{Y\geq y}$.

  Donc :

$$(X-Z)^2 = (X - u_1 1_{Y<y} - u_0 1_{Y\geq y})^2$$
$$= X^2 + u_1^2 1_{Y<y} + u_0^2 1_{Y\geq y} - 2u_1 X 1_{Y<y} - 2u_0 X 1_{Y\geq y} + u_1 u_0 \underbrace{1_{Y<y} 1_{Y\geq y}}_{=0}.$$

  Par suite :

$$E(X-Z)^2 = EX^2 + u_1^2 E(1_{Y<y}) + u_0^2 E(1_{Y\geq y}) - 2u_1 E(X 1_{Y<y}) - 2u_0 E(X 1_{Y\geq y})$$
$$= EX^2 + u_1^2 G(y) + u_0^2 (1-G(y)) - 2u_1 E(X 1_{Y<y}) - 2u_0 E(X 1_{Y\geq y}).$$

  ► **On cherche u_1 et u_0 minimisant cette expression.**

  Comme il s'agit d'une fonction du 2nd degré en chacun des paramètres u_i, on obtient aisément :

$$\boxed{u_1 = \frac{E(X 1_{Y<y})}{G(y)}, u_0 = \frac{E(X 1_{Y\geq y})}{1-G(y)}.}$$

  Les espérances $E(X 1_{Y<y})$ et $E(X 1_{Y\geq y})$ s'expriment au moyen de la loi conjointe de (X, Y).

  En effet, on a par exemple :

$$E(X 1_{Y<y}) = E(X 1_{U<y-a-bX}) = \int_{\mathbb{R}^2} x\, 1_{u<y-a-bx}\, f_0(x)\, \frac{1}{\sigma}\, h\left(\frac{u}{\sigma}\right)\, dx\, du$$

  [du fait de l'indépendance de X et U],

  soit :

$$E(X 1_{Y<y}) = \int_{\mathbb{R}^2} x\, 1_{v<\frac{y-a-b}{\sigma}}\; f_0(x)\; h(v)\, dx\, dv = \int_{\mathbb{R}} x\; f_0(x)\; H\left(\frac{y-a-bx}{\sigma}\right) dx$$

  en notant H la fonction de répartition de $\mathcal{N}(0, 1)$.

  En intervertissant l'ordre des intégrations, on peut écrire aussi :

$$E(X 1_{Y<y}) = \int_{\mathbb{R}^2} x\, 1_{x<\frac{y-a-\sigma}{b}}\; f_0(x)\; h(v)\, dx\, dv$$
$$= \int_{\mathbb{R}} \left[\int_{-\infty}^{\frac{y-a-\sigma}{b}} x\, f_0(x)\; dx\right] h(v)\, dv$$
$$= \frac{b}{\sigma} \int_{\mathbb{R}} \left[\int_{-\infty}^{z} x\, f_0(x)\; dx\right] h\left(\frac{y-a-bz}{\sigma}\right) dz.$$

Or :

$$\int_{-\infty}^{z} x\, f_0(x)\ dx = [x\, F_0(x)]_{-\infty}^{z} - \int_{-\infty}^{z} F_0(x)\, dx = z\, F_0(z) - \int_{-\infty}^{z} F_0(x)\, dx.$$

l'intégrabilité de X entraînant que : $\lim_{x \to -\infty} x\, F_0(x) = 0$
[cf. cours de théorie des probabilités de Marc CHRISTINE, chapitre 1, section 2.4.4].

Par suite :

$$E\big(X 1_{Y<y}\big) = \frac{b}{\sigma}\int_{\mathbb{R}} \left[z\, F_0(z) - \int_{-\infty}^{z} F_0(x)\, dx\right] h\left(\frac{y-a-bz}{\sigma}\right) dz$$

$$= \frac{b}{\sigma\sqrt{2\pi}}\int_{\mathbb{R}} \left[z\, F_0(z) - \int_{-\infty}^{z} F_0(x)\, dx\right] e^{-\frac{1}{2}\left(\frac{y-a-bz}{\sigma}\right)^2} dz.$$

Cette dernière expression se prête mieux à une discrétisation de F_0 lorsqu'on la calcule à partir d'une population finie.

On pourrait généraliser ces types de calcul en cherchant la meilleure approximation de X par une variable aléatoire de la forme $Z = u(1_{\alpha \le Y < \beta})$,

- Dans le cas qui nous intéresse, où l'on dispose de plusieurs tranches de revenus, on va chercher spécifiquement Z sous la forme :

$$Z = \sum_{k=1}^{K+1} u_k\, 1_{Y \in I_k}$$

avec : I_1=]-∞, s_1[, I_k=[s_{k-1}, s_k[$pour\ k = 2, \dots, K, I_{K+1} = [s_K, +\infty[$.

On a alors :

$$(X - Z)^2 = \left(X - \sum_{k=1}^{K+1} u_k\, 1_{Y \in I_k}\right)^2$$

$$= X^2 - 2\sum_{k=1}^{K+1} u_k\, X 1_{Y \in I_k} + \left(\sum_{j=1}^{K+1} u_k\, 1_{Y \in I_k}\right)^2$$

$$= X^2 - 2\sum_{k=1}^{K+1} u_k\, X 1_{Y \in I_k} + \sum_{k=1}^{K+1} u_k^2\, 1_{Y \in I_k} + \sum_{k \neq l} u_k\, u_l \underbrace{1_{Y \in I_k} 1_{Y \in I_l}}_{=0}$$

$$= X^2 + \sum_{k=1}^{K+1} (u_k^2\, 1_{Y \in I_k} - 2u_k\, X 1_{Y \in I_k}).$$

Par suite :

$$E(X - Z)^2 = EX^2 + \sum_{k=1}^{K+1} E(u_k^2\, 1_{Y \in I_k} - 2u_k\, X 1_{Y \in I_k})$$

$$= EX^2 + \sum_{k=1}^{K+1} [u_k^2\, E(1_{Y \in I_k}) - 2u_k\, E(X 1_{Y \in I_k})\}.$$

Les valeurs optimales u_k^* minimisant cette fonction sont, selon des formules analogues aux précédentes :

$$\boxed{u_k^* = \frac{E(X 1_{Y \in I_k})}{E(1_{Y \in I_k})} = \frac{E(X 1_{Y \in I_k})}{P\{Y \in I_k\})}.}$$

Elles s'expriment par des formules analogues à celles du cas initial simple où l'on n'avait que deux tranches, au moyen de la fonction de répartition de X et des paramètres du modèle linéaire.

Dans la pratique, on remplacera ces paramètres par les valeurs de leurs estimations par maximum de vraisemblance.

4.3. Application au problème

Dans les enquêtes auprès des ménages, on observe un **revenu déclaratif *Y* en tranches**:

$$Y \in I_k$$

qui n'est pas le vrai revenu *X* du ménage.

Mais on va postuler un modèle linéaire reliant *Y* à *X* avec une perturbation aléatoire normale et, par ailleurs, *sans observer les vrais revenus individuellement*, on suppose connue la vraie distribution de X (au moyen des **statistiques** fiscales exhaustives).

On peut alors à la fois estimer les paramètres du modèle linéaire à partir de l'ensemble des observations des tranches de revenus déclarés mais sans observer les vrais revenus individuels X_i et inférer, pour tout ménage *j*, une valeur approximante du vrai revenu, soit Z_j^*, lorsque l'on connaît la tranche à laquelle appartient le revenu déclaré Y_j.

On aura en effet :

$$Z_j^* = \sum_{k=1}^{K+1} u_k^* \, 1_{Y_j \in I_k}$$

avec les valeurs des u_k^* indiquées plus haut, remplacées par leurs estimateurs $\hat{u}_k^*$.

Résumé de la démarche :

- On observe les tranches de revenus déclarées.
- On estime les paramètres du modèle linéaire reliant le revenu déclaré au vrai revenu, par maximum de vraisemblance, en supposant connue la distribution théorique des vrais revenus, à partir de l'ensemble des observations disponibles du nombre de ménages dans chaque tranche de revenu déclaré. Ceci revient à ajuster la distribution empirique des tranches de revenu observées à la distribution théorique des vrais revenus.
- On impute le meilleur prédicteur du vrai revenu à chaque ménage à partir de la connaissance de sa tranche de revenu déclarée, au moyen d'une expression mettant en œuvre les paramètres du modèle linéaire (remplacés par leurs estimateurs) et la fonction de répartition théorique des vrais revenus.

- Les procédures de calage peuvent alors s'appliquer sans problème en utilisant ces valeurs de revenu imputées (plus celles d'autres variables éventuelles) pour caler sur un vrai revenu total et d'autres totaux connus.

4.3.1. Discussion

Un inconvénient de la procédure est qu'à une tranche de revenu déclaré donnée correspond une seule valeur possible de revenu vrai imputé (ceci étant, fondamentalement, les tranches de revenu déclaré constituent les modalités ***en nombre fini*** de la variable observée).

On pourrait avec un modèle linéaire comportant d'autres variables explicatives effectivement observées (et pas seulement par l'intermédiaire de leur distribution) remédier à cet inconvénient.

4.3.2. Modélisation lognormale

La plupart du temps, lorsqu'on traite de revenus (strictement positifs), on préfère un modèle en logarithmes sous la forme :

$$ln\, Y_i = a' + b'\, ln\, X_i + U_i,$$

avec l'hypothèse de normalité sur les U_i .

Ainsi, si la distribution des vrais revenus X_i est lognormale, celle des revenus déclarés Y_i le sera également.

Cette formulation ne modifie pas la logique de la méthode.

On trouvera en annexe une résolution théorique du problème dans le cas où la vraie distribution des revenus est exactement lognormale.

4.4. Mise en œuvre dans le cadre d'un processus d'échantillonnage

4.4.1. Cadre général

On suppose qu'on a ici une population finie de taille N. ***Mais les variables mesurées sur les individus de cette population sont supposées résulter d'une génération aléatoire***. Plus précisément, on suppose donnée une suite de couples $(X_i, U_i)_{i \in \{1,\dots,\mathrm{N}\}}$ mutuellement indépendants, où X_i et U_i sont indépendants entre eux et de même loi, et l'on pose : $Y_i = a + bX_i + U_i$.

Les observations dont on dispose sont les réalisations y_i des Y_i ou la seule appartenance des Y_i à une tranche donnée, sous la forme d'une indicatrice $1_{y_i \in I_k}$. **On n'observe ni les réalisations des U_i ni celles des X_i mais on suppose connue la loi des X_i, de densité f_0.**

Un échantillon *s* est tiré à l'aide d'un plan de sondage, indépendant du processus de génération des données sur la population. Ce plan de sondage est connu par l'intermédiaire des probabilités d'inclusion π_i affectées à chacun des individus *i* de la population.

Rappel : estimation du paramètre espérance dans ce cadre

Si les T_i sont des variables aléatoires indépendantes et de même loi, définies sur la population, d'espérance ET, on peut estimer sans biais l'espérance ET au moyen de l'estimateur :

$$\hat{T}_s = \frac{1}{N}\sum_{i \in s}\frac{T_i}{\pi_i}.$$

En effet :

$$\hat{T}_s = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}(T_i 1_{i \in s})$$

et : $E\hat{T}_s = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}E(T_i 1_{i \in s}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}(ET_i)\,E(1_{i \in s})$

du fait de l'indépendance entre le plan de sondage et les observations aléatoires sur la population,

d'où :

$$E\hat{T}_s = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}(ET)\,\pi_i = ET.$$

La démarche est alors la suivante :

4.4.2. Cas où l'on connaît les observations y_i des Y_i.

La vraie log-vraisemblance sur l'ensemble de la population :

$$ln\,L(y_1,\ldots,y_N) = -N\,ln\,\sigma + \sum_{i=1}^{N} ln\left[\int_{-\infty}^{+\infty} f_0(z)\,h\left(\frac{y_i - a - bz}{\sigma}\right)dz\right]$$

est estimée sans biais, conditionnellement aux observations y_i, par :

$$(ln\,\hat{L})_s = -N\,ln\,\sigma + \sum_{i \in s}\frac{1}{\pi_i} ln\left[\int_{-\infty}^{+\infty} f_0(z)\,h\left(\frac{y_i - a - bz}{\sigma}\right)dz\right].$$

On estime alors les paramètres a, b, σ en maximisant cette quantité $(ln\,\hat{L})_s$, d'où l'on obtient : $\hat{a}_s, \hat{b}_s, \hat{\sigma}_s$.

Puis on calcule le prédicteur optimal de X_j pour chaque individu j de l'échantillon pour lequel on dispose de l'observation y_j de Y_j :

$$\boxed{X^*(y_j) = E(X/\,Y = y_j) = \frac{\int x\,f_0(x)\;h\left(\frac{y_j - \hat{a}_s - \hat{b}_s x}{\hat{\sigma}_s}\right)dx}{\int f_0(x)\;h\left(\frac{y_j - \hat{a}_s - \hat{b}_s x}{\hat{\sigma}_s}\right)dx}}.$$

4.4.3. Cas où l'on ne connaît que les observations des y_i en tranches

La vraie log-vraisemblance est :

$$ln\,L(n_1,\ldots,n_{K+1}) = ln\left(\frac{n!}{n_1!\ldots n_{K+1}!}\right) + \sum_{k=1}^{K+1} n_k\,ln\,p_k,$$

où n_k est le nombre d'individus *i* tels que $Y_i \in I_k$, soit : $n_k = \sum_{i=1}^{N} 1_{Y_i \in I_k}$.

Les n_k sont estimés sans biais, conditionnellement aux observations y_i, par :

$$(\hat{n}_k)_s = \sum_{i \in s} \frac{1_{Y_i \in I_k}}{\pi_i}$$

On maximise alors la log-vraisemblance estimée :

$$(ln\, \hat{L})_s = ln\, \left(\frac{n!}{\hat{n}_1! \dots \hat{n}_{K+1}!}\right) + \sum_{k=1}^{K+1} (\hat{n}_k)_s\, ln\, p_k,$$

où les p_k sont fonctions des paramètres du modèle linéaire sous-jacent comme vu plus haut.

Puis on calcule le prédicteur optimal du vrai revenu à partir de l'observation du revenu déclaré en tranches en reprenant les formules du § 4.3 en remplaçant les différents paramètres par leurs estimateurs.

5. Perspectives futures

- Tester cette méthode sur données réelles et avec des algorithmes de résolution numérique des équations du maximum de vraisemblance : tester les différentes procédures proposées de reconstruction des revenus « vrais », à partir des données collectées dans les TCM de différentes enquêtes. , ainsi que de regarder la stabilité des modèles quand on raisonne par sous-catégorie de ménages ou de revenus.

- Faire tourner ces procédures par sous-population (catégories de ménages ou de revenus), pour chacune desquelles un modèle particulier peut jouer. Tester d'ailleurs si les paramètres des modèles diffèrent de manière significative.

- Enrichir le modèle linéaire en introduisant d'autres variables explicatives.

- Modifier la modélisation : modèle probit, modélisant la probabilité de répondre dans une tranche de revenu donnée, en fonction de différentes variables explicatives, dont évidemment le revenu vrai.

6. Conclusion

Comme chez l'auteur de la pièce[7], ces petits papiers cherchent un (co-)auteur pour écrire la suite de l'histoire, mettre en application les méthodes exposées sur des données réelles, tester leur pertinence et apporter tout complément…. Une discussion avec les experts métiers sera donc très bienvenue, s'ils la jugent utile…

[7] Luigi PIRANDELLO : « *Sei personnagi in cerca d'autore* » (1921)

Bibliographie

[1] Christine M., rapport du Groupe Marges, version révisée novembre 2013, note interne UMS, *unpublished paper*
[2] Vincent L., Faivre S., "Le projet Nautile (Nouvelle Application Utilisée pour le Tirage des Individus et des Logements des Enquêtes)", Actes des $13^{\text{ès}}$ Journées de méthodologie statistique de l'Insee sur http://jms-insee.fr
[3] Tronc commun des ménages (TCM), version 2022, document Insee du 29 octobre 2021
[4] BECK S., Réflexions sur l'allègement du module Revenus de l'enquête Patrimoine 2017-18, note n°622/DG75-F350/SB/ML du 9 avril 2015

ANNEXE 2 : Tronc Commun des Ménages (TCM) – version 2022

Bloc I. Revenus

Nous allons maintenant parler des ressources de votre ménage.
Y a-t-il actuellement, dans votre ménage, une ou plusieurs personnes qui perçoivent les ressources suivantes :

Salaires, traitements et primes ? 8	**Revenus d 'un e activité professionnelle non salariée (indépendant, profession libérale…) ?**	**Allocations de chômage ? 9**
1. Oui 2. Non	1. Oui 2. Non	1. Oui 2. Non
Préretraites, retraites ? 10	**Prestations liées à la maladie ou l'invalidité ? 11**	**Prestations familiales ? 12**
1. Oui 2. Non	1. Oui 2. Non	1. Oui 2. Non
Bourses scolaires ou bourses d'étudiants ? 13	**Allocations logement, aides au logement ?**	**RSA, prime d'activité ?**
1. Oui 2. Non	1. Oui 2. Non	1. Oui 2. Non
Loyers (y compris fermages) ? 14	**Intérêts, revenus d'épargne, dividendes, que peuvent vous procurer vos livrets d'épargne comme le livret A, PEL, PEP, LDD par exemple ?**	**Pensions alimentaires, aides financières régulières des parents, de la famille ou des amis, y compris paiement du loyer ?**
1. Oui 2. Non	1. Oui 2. Non	1. Oui 2. Non **RTRA**

Si RTRA = 1

TYPTRA **De quels types d'aides s'agit-il ? [15]**

1. le paiement d'un loyer
2. une pension alimentaire
3. une autre aide financière régulière

8 Y compris 13e mois, congés payés, heures supplémentaires, indemnités journalières pour un arrêt maladie de moins de 6 mois, rémunération des emplois temporaires, des activités secondaires, salaires des dirigeants salariés de leur entreprise, intéressements et participation

9 Allocation d'aide au retour à l'emploi (ARE), allocation de solidarité spécifique (ASS), rémunération de formation Pôle emploi (RFPE), rémunération de fin de formation (RFF), etc.

10 Y compris minimum vieillesse, pension d'ancien combattant, pension de réversion

11 AAH, pension invalidité, allocations liées à la dépendance, indemnités » journalières pour un arrêt maladie de 6 mois ou plus,...

12 Allocations familiales, complément familial, allocation pour jeune enfant, aides à la garde d'enfants, allocation de soutien familial, allocation de rentrée scolaire,...

13 Bourses pour les personnes âgées de 16 ans ou plus

14 Si vous avez des maisons, des appartements ou des terres que vous louez

15 Plusieurs réponses possibles

En prenant en compte tous les types de revenus que vous venez de mentionner, même s'il manque les revenus de certaines personnes, quel est actuellement le montant mensuel des ressources de l'ensemble de votre ménage ? [16]

[0 à 99 999]

ITOTREV **Ce montant prend-il en compte les revenus de tous les membres du ménage ?**

1. Oui
2. Non

Si ITOTREV = 2 **Si vous ne <u>pouvez</u> pas donner un montant précis des ressources de tous les membres du ménage, à combien environ les estimez-vous pour un mois ordinaire ? 17**	Sinon **Si vous ne <u>souhaitez</u> pas donner un montant précis des ressources de tous les membres du ménage, à combien environ les estimez-vous pour un mois ordinaire ?**
1. à moins de 400 € 2. de 400 € à moins de 600 € 3. de 600 € à moins de 800 € 4. de 800 € à moins de 1 000 € 5. de 1 000 € à moins de 1 200 € 6. de 1 200 € à moins de 1 500 € 7. de 1 500 € à moins de 1 800 € 8. de 1 800 € à moins de 2 000 € 9. de 2 000 € à moins de 2 500 € 10. de 2 500 € à moins de 3 000 € 11. de 3 000 € à moins de 4 000 € 12. de 4 000 € à moins de 6 000 € 13. de 6 000 € à moins de 10 000 € 14. à 10 000 € ou plus 98. refuse de répondre 99. ne sait pas	

Fin du TCM

16 Il s'agit du revenu net (de cotisations sociales et de CSG) avant impôts (donc en incluant le montant des impôts prélevés, indiqué sur les bulletins de paye). Si les revenus ont fluctuants, prendre une moyenne.

17 Il s'agit du revenu net (de cotisations sociales et de CSG) avant impôts (donc en incluant le montant des impôts prélevés, indiqué sur les bulletins de paye).

ANNEXE 2
Éléments de comparaison entre les revenus déclarés et les vrais revenus

Extrait de la note citée en [4]

Concernant les données de montants des revenus perçus, l'enquête Patrimoine 2009-10 permet de comparer les montants déclarés et les montants appariés. On y voit ainsi que les déclarations des enquêtés sont peu fiables. Le tableau ci-dessous montre en effet les différences entre revenu déclaré et revenu issu des sources fiscales, par quintiles, sur la base de l'enquête Patrimoine 2009-10.

		Revenu mensuel total (source fiscale), par quintiles				
		1	**2**	**3**	**4**	**5**
Revenu mensuel total (déclaré), par quintiles	**1**	**54,1**	20,1	9,0	7,9	8,9
	2	20,5	**54,5**	21,0	3,0	1,1
	3	7,4	17,8	**54,5**	17,1	3,2
	4	1,7	4,0	18,5	**57,5**	18,4
	5	2,9	4,0	5,2	15,4	**72,5**

On voit ainsi par exemple que 9 % des ménages se trouvant dans le 1er quintile de revenus déclarés se trouvent en fait dans le 3ème quintile de revenus fiscaux. De manière générale, sauf pour le dernier quintile, près de la moitié des ménages estiment mal leur place dans la distribution des revenus.

ANNEXE 3
Étude théorique du modèle lognormal (adapté quand X et Y sont des revenus)

Soit le modèle :

$$ln\, Y = a + b\, ln\, X + U$$

avec : $ln\, X$ et U indépendants, de lois normales respectives $\mathcal{N}(m, \tau^2)$ et $\mathcal{N}(0, \sigma^2)$.

On pose : $Z = \frac{ln\, X - m}{\tau}$, soit : $ln\, X = m + \tau Z$.

Le modèle s'écrit alors :

$$\boxed{ln\, Y = a + b(m + \tau Z) + \sigma V = a + bm + b\tau Z + \sigma V},$$

avec : $\begin{pmatrix} Z \\ V \end{pmatrix} \rightsquigarrow \mathcal{N}(0, I_2)$.

On pose : $\begin{cases} A = a + bm \\ B = b\tau \end{cases}$.

Donc : $\boxed{ln\, Y \rightsquigarrow \mathcal{N}(A, B^2 + \sigma^2).}$

La variable aléatoire Y a donc pour densité :

$$\frac{1}{\sqrt{B^2 + \sigma^2}}\, h\left(\frac{ln\, y - A}{\sqrt{B^2 + \sigma^2}}\right) \frac{1}{y}\, 1_{\mathbb{R}^{+*}}(y),$$

où h est la densité de la loi $\mathcal{N}(0,1)$.

La loi de X étant supposée connue, les paramètres m et τ le sont aussi. **On suppose aussi connus ici les paramètres du modèle : $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\sigma}$.**

<u>Prédiction de X à partir de Y</u>

► La meilleure approximation de X connaissant Y (au sens L_2) est :

$$E^*(X/Y) = E^*(X/T), \text{ avec : } T = ln\, Y$$

[du fait de la bijectivité de ln].

Or : $X = e^m\, e^{\tau Z}$, d'où : $E^*(X/T) = e^m E^*(e^{\tau Z}/T)$.

Cette espérance conditionnelle se calcule à partir de la loi conditionnelle de Z sachant $T = t$:

$$E(e^{\tau Z}/T = t) = \int_{\mathbb{R}} e^{\tau z} dP^{Z/T=t}\,(z).$$

<u>Calcul de la loi conditionnelle</u> $P^{Z/T=t}$.

Par transformation linéaire de $\begin{pmatrix} Z \\ V \end{pmatrix}$, le couple $\begin{pmatrix} T = ln\, Y \\ V \end{pmatrix}$ suit une loi normale :

- d'espérance $\begin{pmatrix} A \\ 0 \end{pmatrix}$
- de matrice de variance-covariance :

$$\begin{pmatrix} VT & Cov(T,Z) \\ Cov(T,Z) & VZ \end{pmatrix} = \begin{pmatrix} B^2 + \sigma^2 & B \\ B & 1 \end{pmatrix}.$$

La loi conditionnelle de $Z/T = t$ est donc la loi normale $\mathcal{N}(E^*(Z/T), V^*(Z/T))$,
avec :

$$E^*(Z/T) = EZ + \frac{Cov(Z,T)}{VT}(T - ET) = \frac{B}{B^2+\sigma^2}(T - A)$$
$$V^*(Z/T) = VZ - \frac{Cov^2(Z,T)}{VT} = 1 - \frac{B^2}{B^2+\sigma^2} = \frac{\sigma^2}{B^2+\sigma^2}.$$

Elle admet pour densité : $\frac{1}{\frac{\sigma}{\sqrt{(B^2+\sigma^2)}}} h\left(\frac{z-\frac{B}{B^2+\sigma^2}(t-A)}{\frac{\sigma}{\sqrt{(B^2+\sigma^2)}}}\right)$, où h est la densité de la loi $\mathcal{N}(0,1)$.

► On est donc amené à calculer une intégrale de la forme : $\int_{\mathbb{R}} e^{\tau z}\frac{1}{\theta}h(\frac{z-\mu}{\theta})dz$, avec :

$$\theta = \frac{\sigma}{\sqrt{(B^2+\sigma^2)}}$$
$$\mu = \frac{B}{B^2+\sigma^2}(t - A).$$

- $e^{\tau z}\frac{1}{\theta}h(\frac{z-\mu}{\theta}) = \frac{1}{\theta\sqrt{2\pi}}e^{\tau z - \frac{1}{2\theta^2}(z-\mu)^2}$.
-

$$\tau z - \frac{1}{2\theta^2}(z-\mu)^2 = -\frac{1}{2\theta^2}[(z-\mu)^2 - 2\theta^2\tau z] = -\frac{1}{2\theta^2}[z^2 - 2(\mu + \theta^2\tau)z + \mu^2]$$

$$= -\frac{1}{2\theta^2}[(z - (\mu + \theta^2\tau))^2 + \mu^2 - (\mu + \theta^2\tau)^2]$$

$$= -\frac{1}{2\theta^2}[(z - \mu - \theta^2\tau)^2 - \theta^2\tau(2\mu + \theta^2\tau)].$$

- $e^{\tau z}\frac{1}{\theta}h(\frac{z-\mu}{\theta}) = \frac{1}{\theta\sqrt{2\pi}}\exp -\frac{1}{2\theta^2}[(z - \mu - \theta^2\tau)^2 - \theta^2\tau(2\mu + \theta^2\tau)]$
  $= \frac{1}{\theta\sqrt{2\pi}}\exp \tau(\mu + \frac{\theta^2\tau}{2})\exp -\frac{1}{2\theta^2}(z - \mu - \theta^2\tau)^2.$

- Comme $\int_{\mathbb{R}}\frac{1}{\theta\sqrt{2\pi}}\exp -\frac{1}{2\theta^2}(z - \mu - \theta^2\tau)^2 = 1$, puisque c'est l'intégrale de la densité de la loi $\mathcal{N}(\mu + \theta^2\tau, \theta^2)$, on en déduit :

$$\int_{\mathbb{R}} e^{\tau z}\frac{1}{\theta}h(\frac{z-\mu}{\theta})dz = \exp[\tau(\mu + \frac{\theta^2\tau}{2})].$$

- Il ne reste plus alors qu'à remplacer μ et θ par leurs valeurs :

$$E(e^{\tau Z}/T = t) = \int_{\mathbb{R}} e^{\tau z}\frac{1}{\theta}h\left(\frac{z-\mu}{\theta}\right)dz = \exp\tau\left[\frac{B}{B^2+\sigma^2}(t - A) + \frac{\frac{\sigma^2}{B^2+\sigma^2}\tau}{2}\right]$$

$$= exp\frac{\tau}{B^2+\sigma^2}[B(t - A) + \frac{\sigma^2\tau}{2}].$$

► **Au final :**

$$E(X/T=t) = e^m exp\frac{\tau}{B^2+\sigma^2}[B(t-A)+\frac{\sigma^2\tau}{2}]$$

$$= e^m exp\frac{\tau}{b^2\tau^2+\sigma^2}[b\tau(t-a-bm)+\frac{\sigma^2\tau}{2}].$$

On remplace ensuite *t* par *ln y* pour obtenir $\boldsymbol{E(X/\,Y=y)}$.