UNIVERSITAS FRIBURGENSIS
FACULTÉ DES SCIENCES ÉCONOMIQUES & SOCIALES / DÉPARTEMENT D'ÉCONOMIE QUANTITATIVE

Modélisation des primes d'assurance et stratégies d'estimation : une étude empirique sur la base d'un panel d'assurances suisses

Travail de Master

en Statistique

sous la direction du Prof. Dr L. Donzé

par

Marie-Justine Leis

Fribourg, le 1er février 2008

Résumé

Le système de santé est une thématique répandue dans les médias. Les sujets de discussions portent sur les hausses de primes d'assurances, tout comme sur l'évolution des coûts de santé. L'Office fédéral de la santé publique (OFSP) a la tâche de surveiller les activités des assureurs maladie et accident. Il le fait à l'aide de données de surveillance qu'il récolte auprès des assureurs. Or, celles-ci ne sont pas toutes livrées dans les délais. Dès lors l'OFSP n'est pas en possession des données à temps pour pouvoir tirer des conclusions fiables sur l'évolution des principales variables. Le présent travail s'inscrit dans l'optique de pallier ce problème en trouvant des modèles robustes et des méthodes "approriées", permettant des estimations et des prévisions fiables. On se limitera aux primes d'assurance.

Mots clef : Données de panel, précision, imputation multiple, stratification optimale, allocation optimale, assurance maladie obligatoire

JEL-CODE : C23, C53, I11

Table des matières

Liste des tableaux

Table des figures

Liste des scripts

Liste des abréviations

ABB	Approximate Baysian Bootstrap
DFI	Département fédéral de l'intérieur
GHR	Groupe homogène de réponse
IID	Indépendant et identiquement distribué
LAMA	Loi fédérale sur l'assurance maladie et l'assurance accident
LAMal	Loi fédérale sur l'assurance maladie
LSDV	Least square dummy-variable
MCG	Moindres carrés généralisés
MCO	Moindres carrés ordinaires
OAMal	Ordonnance sur l'assurance maladie
OECD	Organisation de coopération et de développement économiques
OFAP	Office fédéral des assurances privées
OFAS	Office fédéral des assurances sociales
OFSP	Office fédéral de la santé publique

Introduction

Le système de santé est une thématique qui est omniprésente dans les médias. Les sujets de discussion portent sur les hausses de primes d'assurances, tout comme sur l'évolution des coûts de la santé. C'est également à cause des coûts élevés du système de santé que la loi fédérale sur l'assurance maladie (LAMal) a été instituée, d'une part, pour rendre l'assurance obligatoire plus solidaire et, d'autre part, afin de diminuer les coûts. Il n'est aujourd'hui pas possible de dire si la LAMal a eu un effet réducteur sur l'évolution des coûts. Cependant, les coûts de la santé sont susceptibles d'augmenter dans le futur, d'une part, à cause de la situation démographique et, d'autre part, à cause d'une évolution technologique fulgurante.

L'Office fédéral de la santé publique (OFSP) a comme tâche de surveiller les assureurs maladie et accident. Les assureurs doivent accorder à l'OFSP le libre accès à toute information que l'OFSP juge pertinente dans le cadre de l'inspection. Pour pouvoir surveiller les assurances l'OFSP contraint les assureurs de remplir un formulaire, afin d'évaluer leurs activités. C'est dans cette optique que s'inscrit ce travail. Notre objectif est de construire un modèle de prévision afin de pouvoir donner, en fonction des informations que l'OFSP récolte des assureurs, plus rapidement une appréciation de l'évaluation des variables de surveillance.

Le présent travail contient trois parties. La première traite dans un premier temps de l'historique et de la situation politique de l'assurance obligatoire. Dans un deuxième temps, les besoins de l'OFSP sont documentés. Finalement une explication des données à disposition est mentionnée. La deuxième partie traite, d'une part, des méthodologies économétriques utilisées et, d'autre part, des méthodes d'imputation utilisées. La dernière partie est réservée aux discussions des données, à l'interprétation des modèles et à la discussion des résultats.

Je tiens à remercier le Professeur Laurent Donzé qui m'a soutenu académiquement durant toute la conception de ce travail de Master. D'autre part, j'aimerais adresser à Monsieur Nicolas Siffert et ses supérieurs de l'Office fédéral de la santé publique ma plus vive reconnaissance pour avoir

suscité cette étude, fourni des commentaires judicieux sur le fonctionnement du système de surveillance des assurances et naturellement pour la mise à disposition des données nécessaires à l'accomplissement de l'étude.

Première partie

Champ d'étude

Chapitre 1

L'assurance maladie obligatoire

1.1 Les débuts de l'assurance maladie

La première loi sur l'assurance maladie se basait sur la LAMA (Loi fédérale sur l'assurance maladie et l'assurance accident) acceptée en 1911. Cette loi n'a pas vu de modification jusqu'en 1964 ou une révision partielle consistant en un allégement des conditions d'entrée dans l'assurance maladie, une adaptation tarifaire, une amélioration des prestations et une augmentation des subsides fédéraux a été mise en place. Cependant, cette loi fédérale ne prévoyait pas l'introduction d'une assurance maladie obligatoire. On constate néanmoins, que les 99% de la population s'étaient assurés auprès d'un assureur et que plusieurs cantons ont obligé certaines catégories de la population de s'assurer. Il y a eu, toutefois, quelques tentatives de réformer l'assurance maladie. En 1974, on projetait une réforme complète du système, et en 1981 on a voulu réformer partiellement la loi en vigueur. Cependant, les deux projets ont échoué au vote (OECD, 1994).

La loi avait deux problèmes majeurs. Premièrement, la prime d'assurance était calculée sur la base du sexe et de l'âge de l'assuré, ce qui faisait qu'une femme payait en moyenne 10 pour-cent de plus qu'un homme de même âge. Non seulement l'âge, mais également la durée du contrat d'assurance, avait un impact sur la hauteur de la prime. On comprend que la solidarité était difficile sous ces conditions. Deuxièmement, les subsides de l'État versés aux assureurs maladies ne prenaient pas en compte le portfolio de risque des différents assureurs. Ce système impliquait que la relation, entre les assurés et l'assureur, d'une part, et la relation entre les assureurs maladie, d'autre part, était injuste (Colombo, 2001).

Le système a commencé à basculer avec l'entrée de nouvelles assurances agressives sur le marché suisse. Celle-ci offraient des primes basses pour les personnes jeunes, considérées comme de "bons risques" et des primes éle-

vées pour les personnes d'âge plus avancé, considérées comme de "mauvais risques". Ceci a entraîné une rupture du pacte de solidarité entre les générations au sein d'un assureur, ce qui a mis en danger les assureurs traditionnels qui se trouvaient avec un mauvais portfolio de risques. Cette situation a fait appel à une nouvelle solution. Le Conseil fédéral a réagi à cet appel en instituant un fond d'ajustement de risque, afin d'ajuster les risques associés à l'âge et au sexe des assurés. Ce fond a été mis en place, par un décret fédéral urgent, pour renforcer la solidarité entre les générations (Crivelli, 2005).

1.2 L'entrée en vigueur de la LAMal et ces objectifs

La loi fédérale sur l'assurance maladie actuelle a été approuvée par le Parlement fédéral et ratifiée par un référendum en 1994. Elle est entrée en vigueur en 1996. La nouvelle loi introduit sept grandes lignes de changements part rapport au régime précédent (Siffert, 2007) :

- Introduction de l'obligation de s'assurer auprès d'un assureur autorisé par la Confédération avec un catalogue exhaustif de prestations dans l'assurance obligatoire des soins ;
- Introduction de primes uniques pour les assurés par caisse et par canton. Il existe des primes réduites pour les enfants et les jeunes adultes.
- Garantie pour les assurés du libre choix de leur caisse maladie et libre passage d'une assurance à l'autre. Les assureurs ont l'obligation d'accepter l'assuré sans réserves ;
- Libre choix du produit d'assurance ;
- Réduction individuelle des primes. Les subventions sont accordées aux assurés selon leur situation économique, subventions qui étaient allouées aux assureurs auparavant ;
- Prolongation de la compensation des risques jusqu'en 2005. Ce mécanisme a été établi, afin d'éliminer les attraits pour les assureurs de poursuivre une stratégie de sélection de risques. La compensation des risques a été prolongée d'une durée de 5 ans jusqu'en 2010 ;
- Encouragement de la concurrence entre les fournisseurs de prestations et les caisses maladies.

On peut résumer cette nouvelle loi par trois objectifs qui sont, d'une part, le renforcement de la solidarité entre les assurés, en assurant l'équité entre les individus de revenus et de santé différentes et, d'autre part, le plafonnement des dépenses du secteur de santé et, la garantie d'une qualité adéquate et de haut niveau de l'assurance sociale. Ces objectifs ont également été sujets du message concernant la révision de la loi fédérale sur l'assurance maladie du 6 novembre 1991 (Colombo, 2001).

1.3 Les organes de surveillance

Selon l'art. 21 de la LAMal, le Conseil fédéral surveille la mise en oeuvre de l'assurance maladie. Cependant, c'est l'OFSP à qui le Conseil fédéral confie l'essentiel de la surveillance de l'assurance maladie et accidents (Nyffeler, 2006). Les assureurs doivent accorder à l'OFSP le libre accès à toutes les informations que cet office juge pertinentes dans le cadre de l'inspection. Ils doivent lui communiquer leurs rapports et leurs comptes annuels qu'ils rédigent selon l'art. 60 LAMal. L'OFSP a également comme mission, de maintenir et de développer les volets "maladie" et "accidents" de l'assurance sociale. Il est dans sa compétence de définir les prestations qui sont prises en charges par les assurances maladies. L'OFSP est sous l'auspice du Département fédéral de l'intérieur (DFI) et rend régulièrement compte de ses activités.

L'Office fédéral de l'assurance privée (OFAP) complète l'OFSP en surveillant les caisses maladies reconnues dans le domaine de l'assurance complémentaire (art. 24 OAMal). Les cantons ont également une tâche de surveillance. Leurs champs de responsabilité est de veiller à ce que tous les individus qui résident sur leur territoire soient affilés à une caisse maladie (art. 3 LAMal). Il est dans leur compétence d'assurer automatiquement les individus qui n'ont pas répondu à cette obligation dans les délais (art.10 OAMal).

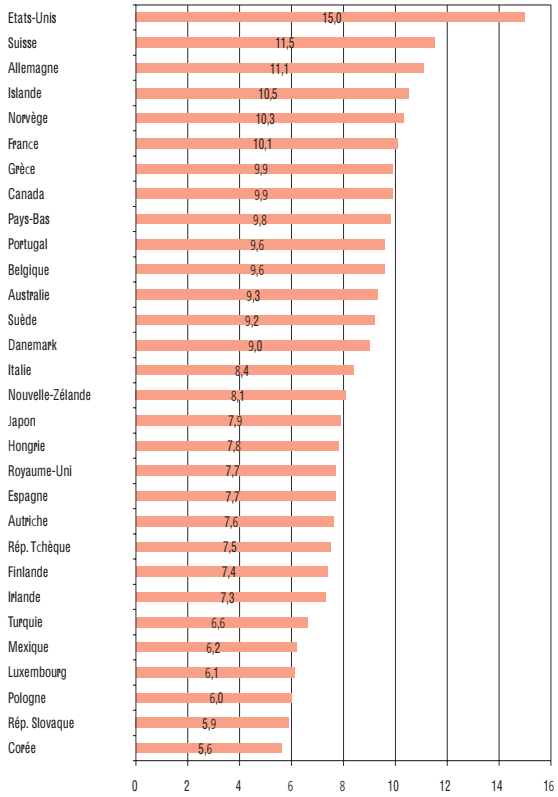
1.4 Structure de coût et de financement de l'assurance maladie

La LAMal a été introduite afin de renforcer la solidarité, d'augmenter la qualité du réseau de santé et de diminuer les coûts. L'analyse du nouveau régime a montré que les deux premiers objectifs ont été atteints. Toutefois, il n'est pas certain, si la loi a réussi de diminuer les coûts (OFAS, 2001).

On peut se poser la question de savoir quelle est la définition des coûts de la santé. L'Office fédéral de la statistique définit les coûts liés à la santé comme (Siffert, 2002) :

> "Les 'coûts du système de santé' comprennent toutes les dépenses des établissements et des personnes exerçant des activités médicales et paramédicales, la vente de médicaments et appareils médicaux ainsi que les frais de gestion du système de santé et de prévention. En sont par contre exclus les coûts de formation des professions de la santé, les travaux de recherche médicale et les prestations en espèces des assurances qui ne servent pas directement à la guérison ou au maintien de la santé comme par exemple les indemnités journalières (IJ) pour pertes de gain.[1]"

[1] Nicolas Siffert (2002), p.2



Source: R. Rossel, Sécurité sociale 1/2006

FIG. 1.1 – Dépenses totales de santé en pour-cent du produit intérieur brut en 2003

1.4.1 L'évolution des coûts

Le produit intérieur brut (PIB) est souvent utilisé comme mesure pour apprécier l'activité de production d'une économie. Le PIB constitue un indicateur qui met en évidence la part des ressources économiques que le pays consacre à des tâches spécifiques (Rossel et Siffert, 2003).

Si on observe les dépenses totales pour la santé en pourcentage du PIB sur le plan international en 2003 du graphique 1.1, on s'aperçoit que la Suisse est parmi les pays en tête du peloton avec les États-Unis, l'Allemagne, l'Islande, la Norvège et la France. Cependant, les positions de certains pays comme les pays nordiques et le Royaume-Uni peuvent surprendre. Ceci n'est pas seulement dû aux efforts entrepris afin de maîtriser les coûts de la santé. En effet, ces pays ne rapportent pas ou peu les dépenses de soins de longue durée dans leurs comptes de santé (Rossel, 2006).

On peut se poser la question de savoir quelle est l'évolution de cet indi-

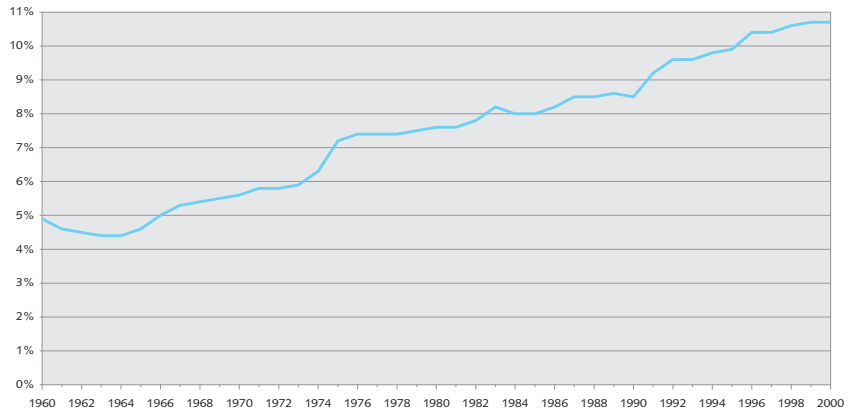


FIG. 1.2 – Coûts du système suisse en pour-cent du produit intérieur brut (PIB)

cateur en Suisse. On déduit de la figure 1.2 que la part du PIB consacrée à la santé a augmenté de façon assez régulière, en passant d'environ 4.9 pour-cent en 1960 à 10.7 pour-cent en 2000. C'est-à-dire que les ressources allouées au système de santé ont évolué en moyenne de 2 pour-cent annuellement entre 1960-2000. En observant la figure on aperçoit trois périodes d'accélération des dépenses de santé. Celles-ci ont eu lieu entre 1970 et 1976, 1990 et 1993 et entre 1995 et 1996. L'explication se trouve surtout dans une croissance économique faible, voire négative. Ceci implique que la croissance des dépenses de la santé ne peut pas être financée par la croissance du PIB. Des transferts supplémentaires entre les agents de financement qui sont les ménages, les assurances sociales et les collectivités publiques sont nécessaires afin d'absorber les coûts (Rossel et Siffert, 2003). En ce qui concerne les années 2000-2004, l'indicateur reste stable (Gerber, 2006).

Néanmoins, l'augmentation des coûts en terme absolu croît de façon constante, comme on peut l'observer à la figure 1.3. La première tendance nous indique les coûts du système de santé réels corrigés de toute inflation. Le montant des dépenses de 1,9 milliards de 1960 est corrigé au niveau des prix de 2000 à un montant de 7,4 milliard, soit 3.82 fois de plus qu'en 1960 (Rossel et Siffert, 2003). La deuxième tendance montre l'évolution des coûts à leur valeur nominale.

Les mêmes observations peuvent être faites lorsqu'on analyse les années 2000-2005. De façon plus précise, l'évolution des coûts progresse de 43,4 milliards en 2000 à 51.6 milliards en 2004 et à 53.8 milliards de francs en 2005 (Gerber, 2006, Rossel et Siffert, 2003, Siffert, 2002).

Les facteurs coupables de l'augmentation des coûts de la santé sont,

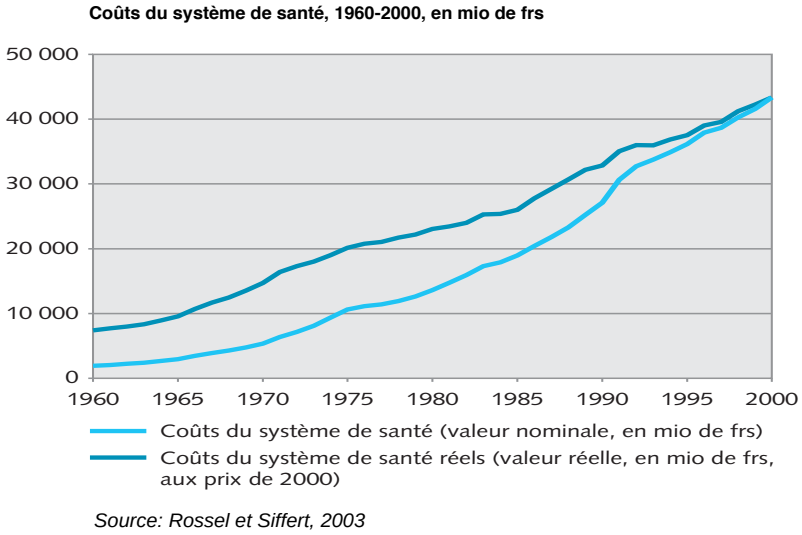


FIG. 1.3 – Comparaison des coûts réels et nominaux

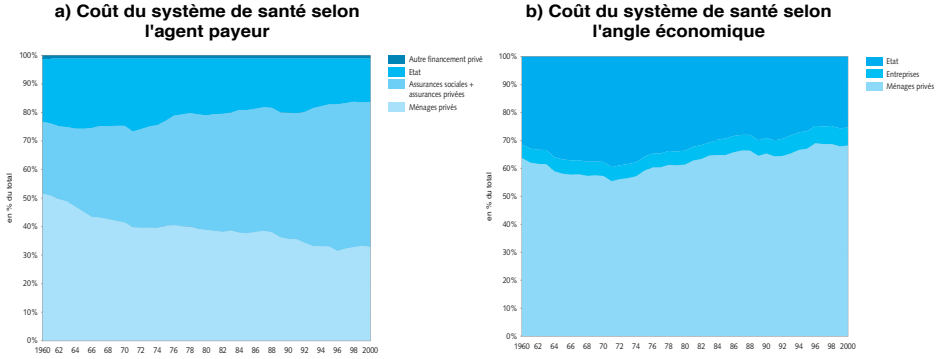
d'une part, le vieillissement démographique, d'autre part, la progression de la quantité des prestations, mais surtout l'amélioration de la qualité des soins (Rossel et Siffert, 2003). Il faut ajouter que, vu l'évolution démographique et le développement des nouvelles technologies, il est clair que les dépenses vont encore augmenter. L'enjeu est ici la préparation de notre système de santé, afin qu'il puisse affronter l'augmentation des dépenses à venir (Sordat Fornerod, 2007).

1.4.2 Le financement du système

Le financement du système de santé suisse est porté par trois piliers qui sont l'État, les assurances sociales et les ménages (Greppi *et al.*, 1998).

L'État finance en matière de santé publique le subventionnement des hôpitaux, des centres socio-médicaux et des soins à domicile. En matière de sécurité sociale, il finance la réduction des primes de l'assurance maladie de base, l'aide sociale, les prestations complémentaires de l'AVS/AI et les oeuvres diverses en faveur des invalides. Ensuite, il intervient par des services de prévention et d'administration. Les assurances sociales, les assurances accidents, l'AVS/AI et l'assurance militaire sont considérées comme des agents de financement séparés de l'État, même si l'État intervient dans leur financement.

Les ménages participent au financement par le payement de primes aux assureurs maladie, aux frais de l'assurance maladie (franchises, quotes-parts)



Source: Rossel et Siffert, 2003

FIG. 1.4 – Agent payeur et agent économique

et par des paiements directs de prestations non couvertes par les assurances maladies. Finalement, il existe également "les autres financements privés", parmi lesquels on considère le financement propre par dons et légations d'institutions à but non-lucratif, comme des homes pour personnes âgées, des services de soins à domicile et des institutions pour handicapés (Siffert, 2002). Les autres financements privés correspondent seulement à une part marginale du financement.

L'évolution du financement du système de santé selon l'agent payeur est illustrée par le graphique 1.4, a). Les évolutions suivantes peuvent être observées. La part du financement de l'État a peu diminué passant de 22.2 pour-cent en 1960 à 15.2 pour-cent en 2000. La charge augmente surtout pour les assurances sociales et privées qui passent d'un financement de 25.2 pour-cent à 50.9 pour-cent. La charge des ménages diminue continuellement de 51.5 pour-cent à un niveau de 32.9 pour-cent en 2000. Cependant, il faut nuancer ce résultat, car les assurances sociales et privées sont financées pour la plus grande partie par les ménages et l'État. Ceci se remarque dans la partie b) du graphique 1.4. Sous l'angle économique, on s'aperçoit, que les ménages subissent 68.2 pour-cent des coûts de santé (Rossel et Siffert, 2003).

1.4.3 L'influence de la LAMal sur les coûts

On ne peut pas encore analyser exactement les effets de la LAMal sur les coûts. On peut seulement se limiter à constater que la LAMal n'a pas occasionné de coûts supplémentaires au système de santé. Néanmoins, il faut également ajouter que les effets attendus de la diminution des coûts ne sont pas encore mesurables. Ceci est dû à l'échelonnement des dispositions faites afin de diminuer les coûts comme la planification hospitalière ou les tarifs

unifiés. D'autre part, les données statistiques à la base des estimations des coûts du système de santé ont connu des changements, ce qui rend difficile la comparaison (Rossel, 2000).

1.5 Le système de la compensation des risques en Suisse

Sur le marché suisse d'assurances, les assureurs sont contraints d'appliquer des primes d'assurances uniques par classes d'âge, même si le risque pour chaque individu n'est pas le même. Ceci force les assureurs d'élaborer des stratégies de sélection de risques, pour attirer les bons risques et dissuader les mauvais risques, afin de persister sur le marché (Beck *et al.*, 2006). Pour lutter contre ce comportement, l'État a mis en place la compensation des risques, visant à garantir à toute la population des soins de base appropriés, avec le but d'atteindre à moyen terme que les assureurs aient une structure de risques plus homogène (Bandi, 1999, Beck *et al.*, 2003, Spycher, 1999). Cependant, il faut ajouter que ce mécanisme n'est pas encore suffisant pour retenir les assureurs de poursuivre un comportement stratégique de sélection de risques (van de Ven, 2004).

1.5.1 Calcul de la compensation des risques

La représentation graphique 1.5 donne une idée schématique du fonctionnement du système d'assurance social en Suisse. On s'aperçoit que les consommateurs, ou les ménages, interviennent par le payement de primes. Le fond central de l'État règle la compensation des risques directement avec les caisses maladies selon leur portfolio de risques.

Pour calculer la compensation des risques les assurés sont répartis en 15 groupes de risques séparés, cependant, comme le sexe est également un critère de classification, le calcul de la compensation des risques se fait finalement avec 30 classes de risques. On ne prend pas en compte dans le calcul les enfants et les jeunes en dessous de 18 ans (Spycher, 1999, Bandi, 1999). Afin de déterminer, si un assureur est un contributeur net ou un receveur net du fond central, l'assureur calcule d'abord la moyenne des coûts par assuré pour l'ensemble des classes. Il calcule ensuite la moyenne des coûts par groupe de risque. Finalement, le calcul de la différence entre les moyennes de l'ensemble des groupes indique si l'assureur doit contribuer à la compensation des risques ou s'il reçoit des contributions de cette dernière (Spycher, 1999, van de Ven, 2004). Un contributeur net a typiquement un portfolio avec des jeunes et des individus masculins, tandis qu'un receveur net a tendance d'assurer des personnes âgées et des femmes (Beck *et al.*,

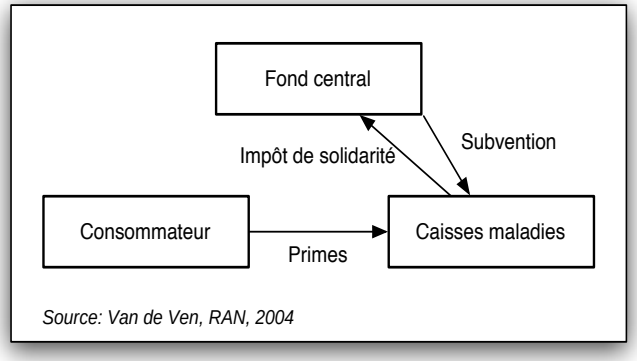


FIG. 1.5 – Le système de compensation des risques en Suisse

2003). L'objectif de la compensation des risques est de prévenir toute distorsion de la concurrence, afin qu'elle ne joue pas sur la sélection des risques (Bandi, 1999).

1.5.2 Les désavantages de la compensation des risques

Les assureurs peuvent choisir entre deux stratégies pour influencer leurs coûts. La première s'effectue par la gestion et le contrôle des coûts. La deuxième utilise la sélection de risques par laquelle les assureurs maladies peuvent systématiquement refuser les individus qui sont susceptibles de coûter plus que la moyenne de leurs assurés (Bandi, 1999). Dans l'actuelle compensation des risques, les assureurs continuent à pratiquer une sélection intensive de risques, ce qui produit une distorsion de la concurrence entre les caisses (Spycher, 1999). Il est vrai que l'assureur peut par le biais de produits d'assurances spécifiques attirer de bons risques (Beck, 2000, Becker et Zweifel, 2005).

Une grande partie des économistes de santé suisses partagent l'opinion que la compensation des risques est nécessaire aussi longtemps qu'il existe une prime d'assurance unitaire. Ensuite, la Suisse est le pays avec la forme de compensation des risques la plus rudimentaire d'Europe si on la compare avec les pays qui ont un système similaire (Beck, 2004, Spycher, 2004b, Holly *et al.*, 2004a, Beck *et al.*, 2006). La conséquence d'une formule de compensation des risques insuffisante qui se base sur l'âge, comme en Suisse, est que les politiciens sont confrontés à un conflit difficile entre la solidarité, l'efficacité et les effets négatifs de la compensation des risques. D'autre part, on observe également une forte segmentation du marché où les bons risques payent une prime basse et les mauvais risques payent une prime élevée, ce qui

nous renvoie à la situation initiale en ce qui concerne le critère de solidarité (van de Ven, 2004).

Beck (2000) et Spycher (1999), conseillent une amélioration de la compensation des risques. Mais il s'agit d'analyser minutieusement le régime actuel et les propositions de réforme en profondeur, si on prend en considération qu'il n'existe pas de compensation des risques optimale (Spycher, 1999).

1.5.3 Projets de réformes

L'actuelle compensation des risques part du principe que les coûts dépendent fortement du sexe et de l'âge d'un individu. Il est cependant clair que ceci n'est pas forcément vrai, car il y a des personnes âgées qui jouissent d'une excellente santé et des jeunes qui souffrent de maladies chroniques et qui engendrent d'importants coûts de santé. Les propositions de réformes partent du principe que l'état de santé influence directement les coûts (Bandi, 1999). Pour la Suisse, elles peuvent se résumer de la façon suivante :

- L'idée sous-jacente des coûts par diagnostique (DCG), est la possibilité de prédire les coûts de santé sur la base des diagnostiques déjà faits. On obtient non seulement des indications sur l'état de santé du patient, mais également sur l'évolution de la santé du patient en général (Holly *et al.*, 2004a).
- Si les coûts par diagnostique ne sont pas disponibles, il existe d'autres indicateurs, moins exacts, pour évaluer l'état de santé d'une personne. Un indicateur souvent mentionné est "les séjours hospitaliers de l'année précédente". Où on peut distinguer, d'une part, entre des séjours hospitaliers induits par des maladies chroniques et, d'autre part, des séjours hospitaliers qui n'ont qu'un impact temporaire sur la santé de l'individu (Holly *et al.*, 2004a, van de Ven et van Vliet, 1992). Beck (2000) comme Holly *et al.* (2004b) recommandent l'introduction de cet indicateur en Suisse. Les deux articles concluent que cet indicateur améliore la formule actuelle en force en Suisse. Cette proposition de réforme a été récemment approuvée par le Conseil national.
- La réassurance est une autre alternative. L'assureur se réassurerait contre le risque de voir son coût moyen ultérieur dépasser d'un pourcentage donné (par exemple 10 pour-cent) ceux de tous les assureurs du canton. Toutefois, ceci implique une régulation du marché de la réassurance. Les réassureurs devraient appliquer "la clause du plus favorisé". Chaque contrat que le réassureur tient avec un assureur doit être proposé à chaque autre assureur qui l'exige (Bandi, 1999, Oggier, 2004, Spycher, 2004a,b).
- Une autre possibilité est l'établissement d'un "pool pour risques élevés". Néanmoins, il faut faire attention de ne pas choisir un seuil top

élevé, afin que celui-ci diminue de manière significative le comportement sélectif des assureurs, comme ceci a été le cas en Allemagne (Spycher, 2004b).

Cependant, Bandi (1999) remarque que l'introduction de facteurs supplémentaires peut avoir comme conséquence de rendre la compensation des risques compliquée et induire une perte de transparence.

Chapitre 2

Les besoins de l'OFSP

L'OFSP rassemble par le biais du formulaire EF123[1] un grand nombre d'informations sur les assurances qui sont contraintes de communiquer leurs rapports et leurs comptes annuels. Ces informations sont importantes afin de surveiller les assureurs et leur prestations. Malheureusement, la plupart des assureurs ne rendent pas à temps le formulaire, ce qui pose problème à l'OFSP. En effet, en pratique les assureurs rendent leur questionnaire pour la fin mars, les derniers questionnaires arrivant à la fin mai. Environ la moitié des questionnaires remplis, arrivent en début mai. Or, sans ces informations, l'OFSP n'est pas en mesure de donner des informations précises sur les variables de surveillance. Celles-ci qui, comme on l'a noté au premier chapitre, donnent des informations sur l'état des assurances et de l'évolution des coûts de santé. Informations qui sont réclamées d'urgence par différents groupes d'intérêt et de politiciens. Il s'en suit, que l'OFSP n'a pas la possibilité de répondre rapidement aux pressions politiques, en fournissant des informations précises en matière d'évolution des coûts.

Le but du travail est de concevoir des modèles avec lesquels on puisse faire une prévision sur les variables de surveillance[2]. Dans la présente étude, nous nous limitons à une variable de surveillance. Il s'agit ici de trouver pour les recettes totales, un modèle robuste et des méthodes "appropriées", permettant des estimations et des prévisions fiables pour l'année en cours.

[1] Ce formulaire se trouve sur le site de l'OFSP ; http ://www.bag.admin.ch/themen/krankenversicherung/01156/01157/index.html ?lang=fr

[2] Un aperçu des variables de surveillance de l'OFSP se trouvent aux pages 120 à 122 de la Statistique de l'assurance maladie obligatoire de l'OFSP de 2005 (Siffert, 2007)

Chapitre 3

Données à disposition

3.1 Provenance des données

Les données proviennent de l'OFSP. Ce sont les données que l'OFSP rassemble, afin d'assurer la tâche de surveillance auprès des assurances sociales, que lui a confiée le DFI. Les données se basent sur le formulaire EF123 qui se trouve sur le site Internet de l'OFSP et que chaque assurance a l'obligation de remplir, afin de posséder l'autorisation d'offrir des prestations[1]. Ce formulaire consiste en trois parties. La première partie donne des informations sur les comptes de l'assurance de façon générale. La deuxième, traite le compte d'exploitation en détail, en analysant le compte d'exploitation pour chaque produit offert par l'assureur en différenciant les postes "maladie" et "accident". Et la dernière partie informe sur l'effectif des assureurs, cela donne une image de la structure de risque du portfolio de l'assurance en question. Un aperçu des variables issues de ce formulaire se trouve en annexe (tableau A.1).

Les données ont été fournies selon la logique du formulaire. Chaque tableau reçu contient des informations sur un produit d'assurance distinct. Ceci vaut également pour les différents effectifs des assureurs, chaque effectif étant fourni dans un tableau séparé. Afin de pouvoir travailler avec les données, il a fallu procéder à une fusion des fichiers, avec le but de faciliter les calculs effectués avec notre logiciel statistique (voir en annexe le job B.1).

3.2 Description des données

De manière générale, on observe que les données sont des données de panel, avec des observations allant de 1996 à 2006. Certains assureurs exercent leurs activités dans plusieurs cantons, et d'autres au niveau régional. Ceci a

[1] http ://www.bag.admin.ch/themen/krankenversicherung/01156/01157/index.html ?lang=fr

également un impact sur leur grandeur : une assurance avec un cercle d'activités nationale a forcément un effectif plus important qu'une assurance régionale. Malheureusement, les données sont incomplètes (*Unbalanced*) : il n'existe pas d'observations pour chaque assureur pour chaque année. Ce phénomène peut être illustré à l'aide d'un exemple. Le tableau 3.1 contient quelques variables. La variable $YEAR$ est l'aspect chronologique, tandis que $NAME\_INS$ énonce l'aspect individuel. C'est-à-dire que l'on observe chaque assureur sur une période prédéterminée qui est dans notre cas de 1996 à 2006. On s'aperçoit qu'on possède pour l'assurance $assureur1$ des observations pour toute la période, tandis que tous les autres assureurs n'ont pas d'observation pour toutes les années. Ceci est dû, d'un côté, au fait qu'il y a des assurances qui cessent leur activité en cours de période et, de l'autre, qu'il existe des assurances qui commencent leur activité en cours de période, comme l'assurance $assureur3$.

Dans le cas d'un panel incomplet, il faut contrôler de près, si l'hypothèse d'exogénéité est remplie dans le cas d'effets fixes. Inversement, il est important de vérifier l'indépendance de la constante par rapport aux variables explicatives, dans le cas d'effets aléatoires. Si ces conditions sont remplies, les estimateurs pour effets fixes ou effets aléatoires peuvent êtres utilisés avec peu d'ajustements, comme nous allons le voir au prochain chapitre (Cameron et Trivedi, 2005, Baltagi, 2005).

TAB. 3.1 – Exemple

YEAR	NAME_INS	rec.tot1	EF32_T_BE	EF33_EF_O_T	EF22_M_6
1996	assureur1	1441198696	NA	200434	322113914
1997	assureur1	1497829002	361139	189174	189686870
1998	assureur1	1628099892	350298	243558	327459028
1999	assureur1	1172492841	272612	151510	232372506
2000	assureur1	1094295054	257483	136225	215044708
2001	assureur1	1089142106	244673	131203	223388090
2002	assureur1	1149091981	227027	117768	227712107
2003	assureur1	1143779736	215385	109880	230672102
2004	assureur1	1147217295	206923	124617	253076119
2005	assureur1	1147877188	201247	122573	332894742
2006	assureur1	1053036275	201283	106256	313858121
2003	assureur2	110913	4	20	13036
2004	assureur2	2389279	14	365	1050761
2005	assureur2	18951584	66	1443	4149788
2006	assureur2	40500343	2803	5969	14163040
2003	assureur2	21265237	8440	15055	8655391
2004	assureur2	69407722	9839	24100	28123728
2005	assureur2	148585880	11786	41489	58120225
2006	assureur2	211520953	13324	51170	84339148
2004	assureur3	0	0	0	0
2005	assureur3	3600290	561	1170	2379332
2006	assureur3	5820989	968	1137	2364715
2005	assureur4	31719239	618	4712	10374944
2006	assureur4	50571482	1754	7576	19284960
2006	assureur5	84317	7	7	6426
2006	assureur6	3908069	139	475	1085558

Source : Office fédéral de la santé publique (OFSP), 2007.
Les noms des assureurs ont été changés pour des raisons de confidencialité.

Deuxième partie

Outils d'estimation et de prévision

Chapitre 4

Modèle de régression et économétrie de panel

4.1 Modèle classique de régression multiple

Le modéle classique de régression multiple est

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \tag{4.1}$$

où $\boldsymbol{y}$ est la variable expliquée ; $\boldsymbol{X}$ est la matrice des régresseurs plus une constante ; $\boldsymbol{u}$ est le terme d'erreur et où $\boldsymbol{\beta}$ est un vecteur qui contient les paramètres inconnus qui doivent être estimés. On peut estimer le vecteur $\boldsymbol{\beta}$ par la méthodes des moindres carrés ordinaires (MCO) :

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{4.2}$$

Cependant, cette formulation n'est pas appropiée, car nous avons une structure spéciale des données. Cela nous conduit à'utiliser les méthodes de l'économétrie de panel.

4.2 Intérêts de l'économétrie de panel

Les données de panel sont caractérisées, d'une part, par un aspect transversal ("cross-section data") et, d'autre part, par un aspect chronologique ("longitudinal data"). L'observation d'agents économiques sur une certaine période, fournit des données de panel. On peut se poser la question de savoir quel est l'interêt d'utiliser des données de panel.

Premièrement, comme on l'a vu à l'exemple 1 au tableau 3.1, les données de panel nous fournissent des observations séquentielles d'un nombre d'individus. Ceci nous permet de distinguer des différences intra-individuelles et inter-individuelles. Il s'en suit, qu'on a la possibilité de construire et de tester des comportements plus complexes. Un autre aspect intéressant est

l'augmentation de la précision des estimations, due à un nombre plus important d'observations. Ce fait nous donne également la possibilité d'estimer des "effets fixes" tenant compte de l'hétérogénéité entre les individus qui n'est pas prise en considération dans les variables explicatives de la régression. On peut dire de façon générale, que les observations de données de panel contiennent plus d'informations, plus de variabilité, plus de degrés de liberté, plus d'efficacité et moins de colinéarité qui est un grand problème dans les études de séries temporelles. Un troisième aspect est la possibilité de modéliser le comportement dynamique des agents économiques. Dans le cas d'un panel de ménages, on observe dans quelle mesure le niveau de vie d'un individu augmente ou s'abaisse dans le temps. De plus, on peut également observer de quelle façon le niveau de vie se modifie. En effet, les données de panel peuvent mettre en relation le comportement d'un individu d'une période à l'autre. Ceci est surtout intéressant pour l'évaluation de politiques économiques sur les agents économiques.

Les difficultés liées aux données de panel sont surtout dues à des enquêtes mal conçues. Il s'agit d'erreurs de mesure, plus précisément de questions mal formulées, de réponses délibérément fausses ou encore de non-compréhension de questions du questionnaire. Un autre aspect est le problème de sélectivité, qui apparait sous forme de non-réponse ou "d'usure", où certains individus ne répondent tout à coup plus aux questionnaires. Les raisons peuvent être diveres, le répondant est peut-être décédé, a déménagé ou n'as simplement plus envie d'y répondre. Néanmoins, il ne faut pas sous-estimer les difficultés liées au choix du modèle. Il s'agit ici d'utiliser la "bonne méthode" d'estimation pour obtenir une estimation convergente des paramètres des modèles postulés (Baltagi, 2005, Cameron et Trivedi, 2005, Hsiao, 2003, 2001, Klevmarken, 1989). Notons que le chapitre 4 se base essentiellement sur le livre de Cameron et Trivedi (2005).

4.3 Le modèle "pooled"

4.3.1 Le modèle linéaire général

Le modèle linéaire général pour les données de panel se distingue d'un modèle linéaire classique par une double indexation. Ceci permet à la constante et aux paramètres de varier en fonction de l'individu et du temps. Le modèle s'écrit comme

$$y_{it} = \alpha_{it} + \boldsymbol{x}'_{it}\boldsymbol{\beta}_{it} + u_{it}; \qquad t = 1, ..., T; \qquad i = 1, ..., N, \tag{4.3}$$

où i désigne l'agent économique et t le temps. Dans ce cas l'indice i indique la composante transversale, tandis que t désigne la période, c'est-à-dire la coupe transversale ; y_{it} est la variable dépendante ; $\boldsymbol{\beta}_{it}$ est un vecteur d'ordre $K \times 1$ de paramètres ; $\boldsymbol{x}_{it}$ est un vecteur $K \times 1$ de variables indépendantes ;

u_{it} est le terme d'erreur.

Sous cette forme, ce modèle n'est pas estimable, car il contient trop de paramètres. Des contraintes doivent êtres imposées sur les α_{it}, $\boldsymbol{\beta}_{it}$ et les u_{it}. Il faut surtout être vigilant aux hypothèses qu'on spécifie pour les u_{it}.

4.3.2 Le modèle “pooled”

Le modèle le plus restrictif est le modèle “pooled” qui suppose que les coefficients sont constants. Le modèle s'écrit comme

$$y_{it} = \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + u_{it}. \tag{4.4}$$

C'est le modèle usuel utilisé dans l'analyse de type transversale. Comme on empile les observations i sur la composante temps t on estime les paramètres de l'équation de régression avec les NT observations. On peut pour cela utiliser la méthode des MCO. Si $Cov[u_{it}, \boldsymbol{x}_{it}] = \mathbf{0}$ alors $N \longrightarrow \infty$ ou $T \longrightarrow \infty$ est suffisant pour la convergence des paramètres.

Neanmoins, la matrice de variances-covariances par MCO est généralement fausse. Il est probable que les erreurs ne sont pas indépentantes et identiquements distriuées (IID) et il est en général pas adéquat de supposer l'indépendance entre les i. Toutefois, il est important de noter, que les erreurs peuvent êtres potentiellement corrélées temporellement pour chaque i et/ou être hétéroscédastique (Hsiao, 2003). Ceci est surtout vrai pour une régression “pooled”. En effet, dans ce cas, l'effet individuel n'est pas pris en compte, car les coefficients sont supposés constants. Il s'en suit, que la $Cov[u_{it}, u_{is}] > 0$ pour tout $t \neq s$. Si on ignore la corrélation temporelle, on obtiendra des erreurs standard sous-estimées, ce qui conduira à des statistiques de tests surestimées.

Pour conclure, on peut affirmer que l'estimation par MCO est convergente et asymptotiquement normale. Seul problème est que la matrice de variances-covariances $\Omega \neq \sigma^2 \boldsymbol{I}_{NT}$. La possibilité d'y remédier est d'utiliser les moindres carrés généralisés (MCG). Les estimations sont convergentes et efficaces et une spécification de Ω correcte est possible.

4.3.3 Le modèle à variables muettes

Le modèle à variables muettes est une variante du modèle “pooled” qui permet à la constante de varier pour chaque individu en chaque période. On peut donc écrire (4.4) comme

$$y_{it} = \sum_{j=1}^{N} \alpha_j d_{j,it} + \sum_{s=2}^{T} \gamma_s d_{s,it} + \boldsymbol{x}'_{it}\boldsymbol{\beta} + u_{it}, \tag{4.5}$$

où les N variables muettes pour l'individu $d_{j,it}$ sont égales à 1 si $i = j$ et zéro sinon et les $(T-1)$ variables muettes pour le temps $d_{s,it}$ sont égales à

1 si $t = s$ et zéro sinon. On suppose que $\boldsymbol{x}'_{it}$ ne contient pas de constante. Si on inclut une constante, alors il faut supprimer une des variables muettes.

Ce modèle a $N + (T-1) + dim[\boldsymbol{x}]$ paramètres qui peuvent êtres estimés de façon convergente si $N \longrightarrow \infty$ et $T \longrightarrow \infty$.

4.3.4 Limitation du modèle "pooled"

Une des limitations de l'estimation "pooled" est caractérisée par le fait qu'on n'utilise pas tout le potentiel des données. En effet, l'estimation par le modèle "pooled" cause une surévaluation du contenu d'information, parce qu'on traite la composante t comme une information supplémentaire. Dès lors, ceci n'est pas justifié car on peut admettre que la composante individuelle est fortement corrélée dans le temps, et que ceci implique une redondance d'informations des t. Ce fait a également un impact sur la matrice de variances-covariances, qui n'est plus égale à $\sigma^2 \boldsymbol{I}_{NT}$.

4.4 Le modèle à effets fixes

4.4.1 Hypothèses générales pour les modèles à composantes d'erreur

Soit le modèle

$$y_{it} = \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + u_{it}, \tag{4.6}$$

où les u_{it} sont définis par

$$u_{it} = \mu_i + \varepsilon_{it}, \tag{4.7}$$

et où μ_i désigne l'effet individuel inobservable (l'hétérogénéité) ; α est un scalaire et où ε_{it} indique le terme d'erreur. Notons que μ_i ne varie pas dans le temps et qu'il capte l'effet individuel qui n'est pas inclus dans la régression. En exprimant (4.6) sous forme matricielle on comprend pourquoi ce modèle est une simplification du modèle à variables muettes. On obtient

$$\boldsymbol{y} = \alpha\boldsymbol{\iota}_{NT} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} = \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{u}, \tag{4.8}$$

où $\boldsymbol{y}$ est d'ordre $NT \times 1$, $\boldsymbol{X}$ de $NT \times K$, $\boldsymbol{Z} = [\boldsymbol{\iota}_{NT}, \boldsymbol{X}]$, $\boldsymbol{\delta}' = (\alpha', \boldsymbol{\beta}')$ et $\boldsymbol{\iota}_{NT}$ est un vecteur de 1 de dimension NT. Dans ce cas (4.7) peut être écrit comme

$$\boldsymbol{u} = \boldsymbol{Z}_\mu\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{4.9}$$

où $\boldsymbol{u}' = (u_{11}, ..., u_{1T}, u_{21}, ..., u_{2T}, ..., u_{N1}, ..., u_{NT})$ avec les observations empilées de façon a ce que les i soient au-dessus des t ; $\boldsymbol{Z}_\mu = \boldsymbol{I}_N \otimes \boldsymbol{\iota}_T$, où $\boldsymbol{I}_N$ est une matrice identité de dimension N, $\boldsymbol{\iota}_T$ est un vecteur de dimension T

et $\otimes$ dénote le produit de Kronecker. Et où $\boldsymbol{Z}_\mu$ est une matrice de sélection de un et de zéros d'où la relation avec le modèle à variables muettes. Et où $\boldsymbol{\mu}' = (\mu_1, ..., \mu_N)$ et $\boldsymbol{\varepsilon}' = (\varepsilon_{11}, ..., \varepsilon_{1T}, ..., \varepsilon_{N1}, ..., \varepsilon_{NT})$. Les μ_i sont des variables aléatoires qui détectent l'hétérogénéité non observée dans le modèle (Baltagi, 2005). Par conséquent, l'estimateur MCO est aussi nommé "least squares dummy-variable" (LSDV) (Hsiao, 2003).

Notons également que l'on fait en général l'hypothèse d'une exogénéité forte ou stricte pour les erreurs, i.e.

$$E[\varepsilon_{it}|\mu_i, \boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}] = 0; \qquad t = 1, ..., T, \tag{4.10}$$

ce qui implique que le terme d'erreur a une espérance conditionnelle nulle pour la valeur antérieure, présente et future des régresseures (Arellano, 2001).

4.4.2 Les effets fixes

Pour les modèles à effets fixes on admet que μ_i est un paramètre fixe à estimer. Les termes d'erreur ε_{it} sont $IID(0, \sigma_\varepsilon^2)$, i.e. indépendants et identiquement distribués, d'espérance 0 et de variance σ_ε^2, et indépendants des régresseurs $\boldsymbol{x}_{it}$ pour tout i et t. Ce modèle est approprié par exemple pour un grand nombre N d'entreprises ou de pays. Dans ce cas, l'inférence s'effectue conditionnellement à ces N entreprises ou pays (Baltagi, 2005). En effet, μ_i est potentiellement corrélé avec les $\boldsymbol{x}_{it}$. Dans ce cas, une estimation par MCO n'est pas appropriée. Des méthodes alternatives d'estimation doivent être utilisées afin d'assurer une estimation convergente de $\boldsymbol{\beta}$. Parmi celles-ci on peut considérer l'estimateur "within" et l'estimateur des "différences premières". Le principe de ces méthodes est d'abord de procéder à une transformation et puis d'une estimation par MCO.

4.4.3 L'estimateur "within"

L'estimateur "within" ou estimateur à effets fixes est construit de la manière suivante. Soit

$$y_{it} = \alpha + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \mu_i + \varepsilon_{it}. \tag{4.11}$$

En prennant la moyenne sur la dimension temps, on obtient

$$\overline{y}_i = \alpha + \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + \mu_i + \overline{\varepsilon}_i. \tag{4.12}$$

Considérons la différence de ces deux équations

$$y_{it} - \overline{y}_i = (\boldsymbol{x}_{it} - \overline{\boldsymbol{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \overline{\varepsilon}_i); \qquad i = 1, ..., N; \quad t = 1, ..., T. \tag{4.13}$$

On constate que les μ_i et α ont été éliminés. L'estimateur "within" est l'estimateur par MCO de ce dernier modèle qui est donné par

$$\hat{\boldsymbol{\beta}}_W = \Big[\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x}_{it}-\overline{\boldsymbol{x}}_i)(\boldsymbol{x}_{it}-\overline{\boldsymbol{x}}_i)'\Big]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{x}_{it}-\overline{\boldsymbol{x}}_i)(y_{it}-\overline{y}_i) \quad (4.14)$$

pour $\boldsymbol{\beta}$. Si le modèle à effets fixes est vrai, l'estimateur "within" conduit à des estimateurs de $\boldsymbol{\beta}$ convergents. Ce qui est le cas lorsque

$$plim(NT)^{-1}\sum_i\sum_t(\boldsymbol{x}_{it}-\overline{\boldsymbol{x}}_i)(\varepsilon_{it}-\overline{\varepsilon}_i)=0, \quad (4.15)$$

qui est vrai si soit $N\longrightarrow\infty$ ou si $T\longrightarrow\infty$ et seulement si la condition d'exogénéité forte ou stricte est verifiée, de façon à ce que $E[\varepsilon_{it}-\overline{\varepsilon}_i|\boldsymbol{x}_{i1},...,\boldsymbol{x}_{iT}]=0$.

Dans ce modèle, μ_i est traité comme un "paramètre de nuisance" ou "noise parameter". Cela a peu d'importance si on s'intéresse seulement à l'estimation de $\boldsymbol{\beta}$. En revanche, si on s'intéresse également à μ_i on peut l'estimer par

$$\hat{\mu}_i=\overline{y}_i+\overline{\boldsymbol{x}}_i'\hat{\boldsymbol{\beta}}_W; \qquad i=1,...,N. \quad (4.16)$$

Il faut noter que dans des panels courts (T est fixe et $N\longrightarrow\infty$) l'estimation de μ_i peut être non convergente (Baltagi, 2005).

La distribution de $\hat{\boldsymbol{\beta}}_W$ semble être compliquée parce que l'erreur $(\varepsilon_{it}-\overline{\varepsilon}_i)$ du modèle within est corrélée dans le temps pour i. Cependant, sous l'hypothèse forte que les ε_{it} sont IID on peut écrire

$$V(\hat{\boldsymbol{\beta}}_W)=\sigma_\varepsilon^2\Big[\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{\boldsymbol{x}}_{it}\tilde{\boldsymbol{x}}_{it}')\Big]^{-1}, \quad (4.17)$$

où $\tilde{\boldsymbol{x}}_{it}=\boldsymbol{x}_{it}-\overline{\boldsymbol{x}}_i$ et où σ_ε^2 est estimé par $\hat{\sigma}_\varepsilon^2=[N(T-1)-K]^{-1}\sum_i\sum_t\hat{\varepsilon}_{it}^2$ et est non biaisé. Et où $\hat{\varepsilon}_{it}$ est le terme d'erreur estimé.

4.4.4 L'estimateur des "différences premières"

Définissons le modèle suivant

$$y_{it}=\alpha+\boldsymbol{x}_{it}'\boldsymbol{\beta}+\mu_i+\varepsilon_{it}. \quad (4.18)$$

Retardons-le d'une période

$$y_{i,t-1}=\alpha+\boldsymbol{x}_{i,t-1}'\boldsymbol{\beta}+\mu_i+\varepsilon_{i,t-1}, \quad (4.19)$$

et prenons la différence de ces deux équations

$$y_{it} - y_{i,t-1} = (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1}); \quad t = 2, ..., T. \tag{4.20}$$

On constate que μ_i et α sont également éliminées pour cet estimateur. L'estimateur des "différences premières" est l'estimateur par MCO de ce dernier modèle, i.e.

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{DP} = \Big[\sum_{i=1}^{N}\sum_{t=2}^{T}(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})'\Big]^{-1} \\ \sum_{i=1}^{N}\sum_{t=2}^{T}(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})(y_{it} - y_{i,t-1}).\end{aligned} \tag{4.21}$$

Cet estimateur conduit, sous un modèle à effets fixes, à des estimations convergentes de $\boldsymbol{\beta}$. Notons qu'il n'y a seulement que $N(T-1)$ observations dans cette régression. Afin que l'estimateur des "différences premières" soit convergent il faut que l'hypothèse d'exogénéité $E[\varepsilon_{it} - \varepsilon_{i,t-1} | \boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1}]$ soit remplie. Ceci est une hypothèse moins contraignante que celle de l'estimateur "within".

En ce qui concerne la variance, il faut prendre en compte qu'on pourrait avoir une corrélation dans le temps du terme d'erreur $\varepsilon_{it} - \varepsilon_{i,t-1}$. Afin d'obtenir la variance asymptotique de $\hat{\boldsymbol{\beta}}_{DP}$, il faut "empiler" le modèle de la manière suivante :

$$\Delta \boldsymbol{y}_i = \Delta \boldsymbol{X}_i \boldsymbol{\beta} + \Delta \boldsymbol{\varepsilon}_i, \tag{4.22}$$

où $\Delta \boldsymbol{y}_i$ est un vecteur de dimension $(T-1) \times 1$ avec les valeurs $(y_{i2} - y_{i1}), ..., (y_{iT} - y_{i,T-1})$; $\Delta \boldsymbol{\varepsilon}_i$ est un vecteur de dimension $(T-1) \times 1$ avec les valeurs $(\varepsilon_{i2} - \varepsilon_{i1}), ..., (\varepsilon_{iT} - \varepsilon_{i,T-1})$ et $\Delta \boldsymbol{X}_i$ est une matrice d'ordre $(T-1) \times K$ avec les valeurs $(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})', ..., (\boldsymbol{x}_{iT} - \boldsymbol{x}_{i,T-1})'$. Alors on peut réécrire $\hat{\boldsymbol{\beta}}_{DP}$ comme

$$\hat{\boldsymbol{\beta}}_{DP} = \Big[\sum_{i=1}^{N}(\Delta \boldsymbol{X}_i)'(\Delta \boldsymbol{X}_i)\Big] \sum_{i=1}^{N}(\Delta \boldsymbol{X}_i)'(\Delta \boldsymbol{y}_i) \tag{4.23}$$

qui a une variance

$$\begin{aligned}V(\hat{\boldsymbol{\beta}}_{DP}) = \Big[\sum_{i=1}^{N}(\Delta \boldsymbol{X}_i)'(\Delta \boldsymbol{X}_i)\Big]^{-1}\Big[\sum_{i=1}^{N}(\Delta \boldsymbol{X}_i)'V[\Delta \boldsymbol{\varepsilon}_i | \Delta \boldsymbol{X}_i](\Delta \boldsymbol{X}_i)\Big] \\ \Big[\sum_{i=1}^{N}(\Delta \boldsymbol{X}_i)'(\Delta \boldsymbol{X}_i)\Big]^{-1}.\end{aligned} \tag{4.24}$$

On suppose, en général, que les ε_{it} sont corrélées dans le temps pour un i donné, de façon à ce que la $Cov[\varepsilon_{it}, \varepsilon_{is}] \neq 0$ pour tout $t \neq s$.

4.4.5 Limitation du modèle à effets fixes

Les estimateurs à effets fixes souffrent d'une perte de degrés de libertés importants lorsque $N \longrightarrow \infty$. Par conséquent, une forte augmentation des paramètres peut aggraver le problème de multicolinéarité entre les régresseurs. D'autre part, les estimateurs à effets fixes ne peuvent pas estimer les régresseurs invariants par rapport au temps. Ce qui est un problème, car en pratique ce type de régresseurs apparaît très souvent (Baltagi, 2005).

Notons que lorsque $T \longrightarrow \infty$, l'estimateur à effets fixes est convergent. En revanche, ceci n'est pas le cas lorsque $N \longrightarrow \infty$. On a, en effet, convergence pour l'estimateur de $\boldsymbol{\beta}$; cependant, l'estimateur pour l'effet individuel (α, μ_i) ne l'est pas, parce que le nombre de paramètres augmente lorsque N augmente. Ceci est connu dans la littérature sous le nom de "incidental parameter problem" examiné par Neyman et Scott (1948) et récemment revu par Lancaster (2000).

4.5 Le modèle à effets aléatoires

Le modèle à effets aléatoires repose également sur le modèle de référence discuté au chapitre 4.4.1. Toutefois, cette fois les hypothèses sur μ_i et ε_{it} sont différentes. Tout d'abord, on pose les hypothèses suivantes

$$\mu_i \sim [0, \sigma_\mu^2]; \qquad \varepsilon_{it} \sim [0, \sigma_\varepsilon^2], \tag{4.25}$$

où l'indépendance vaut entre μ_i et ε_{it}. Tout $\boldsymbol{x}_{it}$ est également indépendant de μ_i et ε_{it}. Ceci est un modèle approprié, si on tire un échantillon aléatoire d'une grande population. L'effet individuel est aléatoire et l'inférence concerne donc la population de laquelle on a tiré aléatoirement l'échantillon. Le modèle à effets aléatoires peut être considéré comme un cas spécifique du modèle "pooled", parce que μ_i peut être ajouté au terme d'erreur (4.9). La matrice de variances-covariances des erreurs s'écrit comme

$$\begin{aligned} \Omega = E(\boldsymbol{u}\boldsymbol{u}') &= \boldsymbol{Z}_\mu E(\boldsymbol{\mu}\boldsymbol{\mu}')\boldsymbol{Z}_\mu' + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\ &= \sigma_\mu^2(\boldsymbol{I}_N \otimes \boldsymbol{J}_T) + \sigma_\varepsilon^2(\boldsymbol{I}_N \otimes \boldsymbol{I}_T) \end{aligned} \tag{4.26}$$

où $\boldsymbol{J}_T$ est une matrice de 1. Cette équation implique une variance homoscédastique $Var(u_{it}) = \sigma_\mu^2 + \sigma_\varepsilon^2$ pour tout i et t et une matrice de variances-covariances diagonale en bloc équicorrélée qui montre la corrélation dans le temps entre le terme d'erreur du même individu. En effet,

$$Cov[u_{it}, u_{js}] = \begin{cases} \sigma_\mu^2 + \sigma_\varepsilon^2 & \text{si } i = j, t = s; \\ \sigma_\mu^2 & \text{si } i = j, t \neq s \end{cases} \tag{4.27}$$

et zéro ailleurs. Ceci implique que le coefficient de corrélation entre u_{it} et u_{js} est

$$\rho = Cor(u_{it}, u_{js}) = \begin{cases} 1 & \text{si } i = j, t = s; \\ \sigma_\mu^2/(\sigma_\mu^2 + \sigma_\varepsilon^2) & \text{si } i = j, t \neq s \end{cases} \tag{4.28}$$

et zéro ailleurs. Une difficulté du modèle aléatoire est l'ampleur de la matrice de variances-covariances des erreurs, qui est de dimension $NT \times NT$. Ce qui rend son inversion très difficile. Il existe différentes techniques afin d'estimer cette matrice, citons ici notamment Amemiyia (1971), Nerlove (1971), Wallace et Hussain (1969) et Swamy et Aurora (1972). Ou la méthode de Swamy et Aurora (1972) est la méthode par défaut du logiciel R. Cette technique consiste de faire deux régressions, une de type "within" et l'autre de type "between", afin de calculer les composantes des l'erreurs de ces régressions. Avec l'aide des variances estimées par les modèles "within" et "between" il est possible d'estimer la matrice de variances-covariances d'un modèle à effets aléatoires. Ce qu'on verra plus en détail ci-dessous.

4.5.1 L'estimateur "between"

L'estimateur "between" prend en compte la variation entre individus. Considérons le modèle tenant compte d'effets spécifiques individuels

$$y_{it} = \mu_i + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \tag{4.29}$$

et prenons la moyenne sur les périodes t. On a :

$$\overline{y}_i = \alpha + \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + (-\alpha + \mu_i + \overline{\varepsilon}_i); \quad i = 1, ..., N, \tag{4.30}$$

où $\overline{y}_i = T^{-1}\sum_t y_{it}$, $\overline{\varepsilon}_i = T^{-1}\sum_t \varepsilon_{it}$ et $\overline{\boldsymbol{x}}_i = T^{-1}\sum_t \boldsymbol{x}_{it}$ sont les moyennes des observations dans le temps et où α est un scalaire.

L'estimateur "between" est l'estimation par MCO de la régression $\overline{y}_i$ sur une constante et $\overline{\boldsymbol{x}}_i$. Pour $T = 1$, c'est l'estimateur classique sur des données transversales. L'estimateur "between" est convergent si les régresseurs sont indépendants du terme d'erreur ε_i. Par conséquent, ceci ne sera pas le cas, lorsque le vrai modèle est une modèle à effets fixes.

4.5.2 L'estimateur à effets aléatoires

Posons le modèle

$$y_{it} = \alpha + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \mu_i + \varepsilon_{it}. \tag{4.31}$$

Supposons qu'il s'agisse d'un modèle à effets aléatoires. L'estimateur "pooled" est convergent, mais sera moins efficace que l'estimateur "pooled" par

MCG. L'estimateur MCG ("feasible GLS estimator") est l'estimateur à effets aléatoires. Celui-ci peut être résolu en appliquant les MCO au modèle transformé suivant

$$y_{it} - \hat{\lambda}\overline{y}_i = (1-\hat{\lambda})\alpha + (\boldsymbol{x}_{it} - \hat{\lambda}\overline{\boldsymbol{x}}_i)'\boldsymbol{\beta} + \psi_{it} \tag{4.32}$$

où les $\psi_{it} = (1-\hat{\lambda})\mu_i + (\varepsilon_{it} - \hat{\lambda}\overline{\varepsilon}_i)$ sont asymtpotiquement IID et $\hat{\lambda}$ est convergent pour

$$\lambda = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_\mu^2}}. \tag{4.33}$$

Cette estimation s'obtient en deux étapes. D'abord on estime λ, puis on estime le modèle transformé. Il s'en suit que l'estimateur à effets aléatoires $\hat{\boldsymbol{\delta}}_{EA}$ se calcule par

$$\begin{aligned}\hat{\boldsymbol{\delta}}_{EA} = \begin{bmatrix} \hat{\alpha}_{EA} \\ \hat{\boldsymbol{\beta}}_{EA} \end{bmatrix} = \Big[\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{w}_{it} - \hat{\lambda}\overline{\boldsymbol{w}}_i)(\boldsymbol{w}_{it} - \hat{\lambda}\overline{\boldsymbol{w}}_i)'\Big]^{-1} \\ \sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{w}_{it} - \hat{\lambda}\overline{\boldsymbol{w}}_i)(y_{it} - \hat{\lambda}\overline{y}_i),\end{aligned} \tag{4.34}$$

où $\boldsymbol{w}_{it}' = [1 \quad \boldsymbol{x}_{it}']$ et $\overline{\boldsymbol{w}}_i' = [1 \quad \overline{\boldsymbol{x}}_i']$. Afin que l'estimateur soit convergent, $NT \longrightarrow \infty$, i.e. soit $N \longrightarrow \infty$ ou $T \longrightarrow \infty$.

Si on part du principe que les ε_{it} et les μ_i sont IID, alors on peut utiliser la formule suivante pour calculer la variance

$$V\begin{bmatrix} \hat{\alpha}_{EA} \\ \hat{\boldsymbol{\beta}}_{EA} \end{bmatrix} = \sigma_\varepsilon^2\Big[\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{w}_{it} - \hat{\lambda}\overline{\boldsymbol{w}}_i)(\boldsymbol{w}_{it} - \hat{\lambda}\overline{\boldsymbol{w}}_i)'\Big]^{-1}. \tag{4.35}$$

Cette estimation demande une estimation convergente de σ_ε^2 et de σ_μ^2. De la régression par effets fixes de $(y_{it} - \overline{y}_i)$ sur $(\boldsymbol{x}_{it} - \overline{\boldsymbol{x}}_i)$, on obtient

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N(T-1) - K}\sum_i\sum_t((y_{it} - \overline{y}_i) - (\boldsymbol{x}_{it} - \overline{\boldsymbol{x}}_i)'\hat{\boldsymbol{\beta}}_W)^2, \tag{4.36}$$

et de la régression "between" de $\overline{y}_i$ avec un terme d'erreur égal à $\sigma_\mu^2 + \sigma_\varepsilon^2/T$, on obtient

$$\hat{\sigma}_\mu^2 = \frac{1}{N - (K+1)}\sum_i(\overline{y}_i - \hat{\mu}_B - \overline{\boldsymbol{x}}_i'\hat{\boldsymbol{\beta}}_B)^2 - \frac{1}{T}\hat{\sigma}_\varepsilon^2. \tag{4.37}$$

Il est vrai qu'il existe des estimateurs plus efficaces pour les composantes de la variance σ_ε^2 et σ_μ^2, toutefois, ceux-ci ne vont pas augmenter l'efficacité

de l'estimateur, $\hat{\boldsymbol{\beta}}_{EA}$. Pour une preuve plus rigoureuse de l'estimation de la variance et de ces composantes, nous conseillons les ouvrages de Baltagi (2005) et de Hsiao (2003).

Remarquons que si $\hat{\lambda} = 0$, on obtient l'estimateur MCO "pooled". Si $\hat{\lambda} = 1$, on obtient l'estimateur "within". Finalement, $\hat{\lambda} \longrightarrow 1$ lorsque $T \longrightarrow \infty$. Le modèle n'est pas convergent si le vrai modèle est un modèle à effets fixes.

4.5.3 Limitations du modèle à effets aléatoires

L'estimateur à effets aléatoires utilise également le potentiel des données de panel en utilisant la composante individuelle et temporelle. Cependant, les gains d'efficacité par rapport à une estimation "pooled" ne sont pas forcément importants.

4.6 Problèmes de modélisation

4.6.1 Les effets fixes contre les effets aléatoires

Il n'est souvent pas simple de décider quel type de modèle est le plus adéquat, étant donné les jeux de données. Souvent le choix du modèle est lié à la question d'exogénéite ou d'endogénéité. Car les modèles à effets aléatoires supposent une exogénéité forte ou stricte de tous les régresseurs, tandis que le modèle à effets fixes permet aux régresseurs d'être endogènes. Il s'agit ici, de choisir entre les deux extrêmes (Baltagi, 2005).

Si on est certain de la causalité, alors le modèle à effets aléatoires est approprié. Cependant, il n'est pas dans l'habitude des économistes d'utiliser les modèles à effets aléatoires. Parce que souvent les économistes cherchent exactement à trouver cette causalité. Neanmoins, puisque le modèle à effets fixes est souvent lié à une perte de degrés de liberté et qu'il est impossible d'estimer des variables invariantes au temps, les économistes peuvent également avoir recours à un modèle aléatoire.

Le test de Hausmann offre un soutien lors du choix du bon modèle. De façon générale, on peut affirmer que si le modèle vrai est un modèle à effets fixes, les estimateurs du modèle à effets aléatoires ne seront pas convergents. C'est pourquoi il est important de savoir s'il y a des effets fixes en testant ces effets. Le test de spécification de Hausman permet de faire cela. Une valeur élevée de la statistique du test de Hausman, conduit à rejeter l'hypothèse nulle, que les effets spécifiques individuels ne sont pas corrélés avec les régresseurs et donc qu'il y a des effets fixes.

Hausman (1978) montre qu'une variante asymptotiquement équivalente de son test est d'effectuer un test de Wald sur $\mu = 0$ où μ est estimé par MCO dans l'équation de régression

$$y_{it} - \hat{\lambda}\overline{y}_i = (1 - \hat{\lambda})\alpha + (\boldsymbol{x}_{1it} - \hat{\lambda}\overline{\boldsymbol{x}}_{1i})'\boldsymbol{\beta}_1 + (\boldsymbol{x}_{1it} - \overline{\boldsymbol{x}}_{1i})'\mu + \varepsilon_{it}, \qquad (4.38)$$

où $\boldsymbol{x}_{1it}$ est le vecteur des régresseurs qui varie dans le temps et ou $\hat{\lambda}$ est défini à l'équation (4.32).

4.6.2 Données de panel incomplètes

Jusqu'à présent, on est parti du principe que les données sont complètes ("balanced"). C'est-à-dire qu'on a des données pour chaque individu i dans chaque période t. Malheureusement ceci n'est pas toujours le cas. Souvent on a des jeux de données qui ne sont pas complets ("unbalanced"). Soit d_{it} une variable indicatrice égale à 1, lorsque la i-ème valeur est observée et zéro sinon. Alors, l'hypothèse d'exogénéité forte ou stricte pour le modèle à effets fixes est

$$E[u_{it}|\mu_i, \boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT}, d_{i1}, ..., d_{iT}] = 0. \tag{4.39}$$

L'estimateur à effets aléatoires est convergent si μ_i est indépendant des autres variables de conditionnement. Sous cette condition, les estimateurs à effets fixes et à effets aléatoires devraient être applicables sans trop d'ajustements. Baltagi (2005) traite longuement les modèles à données de panel incomplètes au chapitre 9 et 10.

4.6.3 Le modèle à deux composantes d'erreur

Le modèle à "deux composantes d'erreur" (tow-way effects model) introduit la composante temporelle au modèle qu'on a traité jusqu'ici. Ce modèle est défini par la modélisation suivante

$$u_{it} = \mu_i + \gamma_t + \varepsilon_{it}; \qquad t = 1, ..., T, \tag{4.40}$$

où μ_i est l'effet individuel non observable discuté à l'équation (4.7) ; γ_t indique l'effet non observable temporel et où ε_{it} est le terme d'erreur. Notons que γ_t est invariant par rapport à l'individu et capte seulement les effets qui ne sont pas inclus dans les régresseurs. De façon équivalente, on peut exprimer (4.40) sous forme vectorielle.

$$\boldsymbol{u} = \boldsymbol{Z}_\mu \boldsymbol{\mu} + \boldsymbol{Z}_\gamma \boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{4.41}$$

où $\boldsymbol{Z}_\mu$, $\boldsymbol{\mu}$ et $\boldsymbol{\varepsilon}$ sont définis au chapitre 4.4.1. $\boldsymbol{Z}_\gamma = \boldsymbol{\iota}_N \otimes \boldsymbol{I}_T$ est la matrice des variables muettes qu'on introduit dans la régression afin d'estimer γ_t dans le cas d'effets fixes et où $\boldsymbol{\gamma}' = (\gamma_1, ..., \gamma_T)$. L'estimation considère μ_i et γ_t comme des effets fixes et $\varepsilon_{it} \sim (0, \sigma_\varepsilon^2)$. $\boldsymbol{x}_{it}$ est considérée indépendante de ε_{it} pour tout i et t.

Pour le modèle à effets aléatoires on admet que les $\mu_i \sim (0, \sigma_\mu^2)$, $\gamma_t \sim (0, \sigma_\gamma^2)$ et $\varepsilon_{it} \sim (0, \sigma_\varepsilon^2)$ sont mutuellement indépendants. D'autre part, $\boldsymbol{x}_{it}$ est indépendant de μ_i, γ_t et de ε_{it} pour tout i et t.

Pour plus d'informations sur les modèles et estimateurs du modèle à "deux composantes d'erreur", les livres de Baltagi (2005) et Hsiao (2003) sont recommandés.

Chapitre 5

Imputation des données

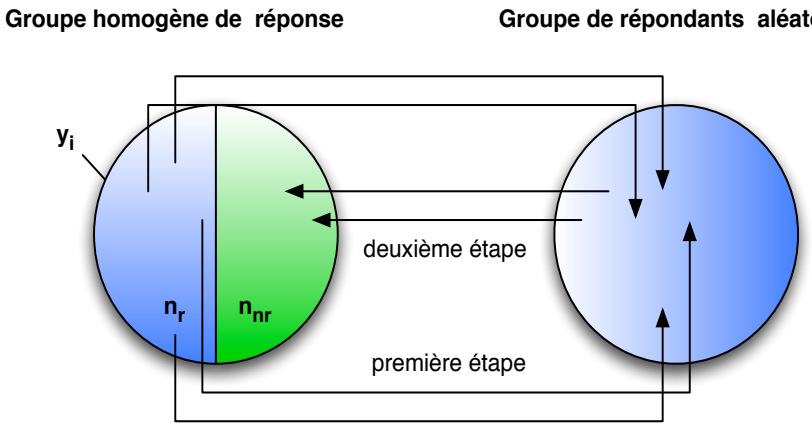
5.1 Méthode d'imputation simple

Nous allons utiliser les méthodes d'imputation des données pour générer les valeurs qui manquent à notre estimation. On parle d'imputation de données, lorsque l'on assigne à chaque valeur manquante une valeur, qui peut être trouvée par exemple auprès d'un donneur (répondant) "proche", en termes de caractéristiques du non-répondant. Le modèle implicite de cette façon de procéder est celui des groupes homogènes de réponses (GHR). Le processus d'imputation conduit à une matrice de données complète.

Il existe un grand nombre de méthodes d'imputation qu'on peut également combiner entre elles (Rubin, 1987). Dans ce travail, nous avons utilisé deux procédures. La procédure ABB ("Approximate Bayesian Bootstrap") et une procédure basée sur un modèle de régression linéaire. Toutes deux sont des méthodes d'imputation "appropriées" qui garantissent des inférences valides.

5.1.1 Méthode d'imputation ABB

La procédure d'imputation ABB est illustrée au graphique 5.1. Considérons un GHR. Pour la variable y_i on enregistre n_r répondants et $n_{nr} = n - n_r$ non-répondants. Dans une première étape, on tire pour chacun des K ensembles d'imputation, aléatoirement avec remise, dans l'ensemble des répondants, n valeurs possibles de y_i. Dans une deuxième étape, on impute les n_{nr} valeurs manquantes de y_i, en tirant aléatoirement avec remise les valeurs tirées à l'étape ultérieure. Ce processus conduit à une base de données complète.



Source: Elaboration personnelle

FIG. 5.1 – Procédure d'imputation ABB

5.1.2 Modèle d'imputation sur la base d'un modèle de régression linéaire avec y_i univarié

Une autre manière de procéder à une imputation est par le biais d'une régression linéaire (Rubin, 1987). Sous l'hypothèse que y_i est univarié remplissant la condition

$$y_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2), \tag{5.1}$$

où $\boldsymbol{\beta}$ est un vecteur de q composantes et σ un scalaire. Nous admettons que $n_1 > q$; où n_1 est le nombre de répondants (valeurs non manquantes). On peut démontrer en suivant un raisonnement de type bayésien que à posteriori, σ^2 peut être exprimé comme $\hat{\sigma}_1^2(n_1 - q)$ divisé par une $\chi^2_{n_1-q}$ générée aléatoirement. Et que $\boldsymbol{\beta}$, étant donné σ^2, soit distribué normalement avec une moyenne de $\hat{\boldsymbol{\beta}}_1$ et une matrice de variances-covariances de $\sigma^2 V$. Connaissant la procédure par MCO on sait que

$$\sigma_1^2 = \sum_{obs}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_i)^2/(n_1 - q), \tag{5.2}$$

où $\boldsymbol{x}_i$ est le vecteur des régresseurs et $(n_1 - q)$ sont les degrés de libertés d'une χ^2 générée aléatoirement. Les $\hat{\boldsymbol{\beta}}_1$ sont calculés par

$$\hat{\boldsymbol{\beta}}_1 = V\Big[\sum_{obs}\boldsymbol{x}_i y_i\Big], \tag{5.3}$$

où

$$V = \Big[\sum_{obs} \boldsymbol{x}_i \boldsymbol{x}_i' \Big]. \tag{5.4}$$

Cette argumentation nous donne la démarche d'imputation suivante. Premièrement, on génère aléatoirement une observation g d'une loi $\chi^2_{n_1-q}$, afin de calculer

$$\sigma_*^2 = \sigma_1^2 (n_1 - q)/g. \tag{5.5}$$

Deuxièmement, nous générons un nombre q d'une normale centrée réduite afin de construire le vecteur $\boldsymbol{z}$. Nous pouvons alors écrire

$$\boldsymbol{\beta}_* = \hat{\boldsymbol{\beta}}_1 + \sigma_* [V]^{1/2} \boldsymbol{z}, \tag{5.6}$$

où $[V]^{1/2}$ est l'une des matrices obtenues par une factorisation de Cholesky de V.
Finalement, on impute chaque valeur, n_0, des $\boldsymbol{y}_{mis}$ observations manquantes par

$$y_{i*} = \boldsymbol{x}_i' \boldsymbol{\beta}_* + z_i \sigma_*, \tag{5.7}$$

où les z_i sont des observations indépendantes issues d'une distribution $N(0,1)$.

Une nouvelle valeur imputée pour $\boldsymbol{y}_{mis}$ est initiée lorsqu'on génére une nouvelle valeur du parametre σ_*^2. Cette procédure est répétée le nombre de fois qu'on désire imputer la valeur manquante.

5.2 Méthode d'imputation multiple

Sans procédure additionelle, le traitement de données ne considère ni la variabilité supplémentaire apportée par l'imputation, ni la variance additionnelle due à la non-réponse elle-même. Il s'en suit que les procédures statistiques standard, appliquées aux données, notamment le calcul de certains estimateurs, sont biaisées. Généralement la variance est sous estimée. C'est pourquoi, on ne devrait pas traiter les données imputées de la même manière que les données réellement observées dans l'enquête.

L'imputation multiple peut pallier ces désavantages (Rubin, 1987). L'idée est de constituer par imputation plusieurs bases de données complètes. C'est-à-dire qu'on n'impute pas une seule fois la base de données mais K-fois. Le calcul des estimateurs combine les estimateurs obtenus de façon standard de chaque base de données complète. L'avantage de cette méthode est qu'on peut prendre en compte la variabilité supplémentaire introduite par l'imputation. D'autre part, elle nous permet de trouver des estimateurs ponctuels plus efficaces. De façon générale, on constate que plus le nombre de bases de

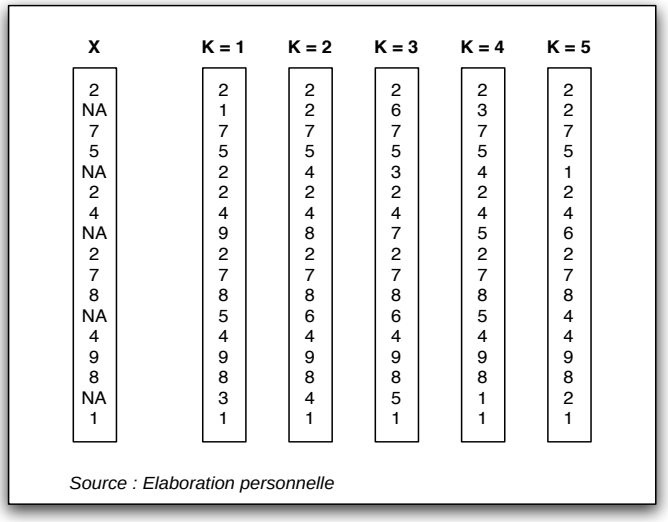


FIG. 5.2 – Procédure d'imputation multiple

données imputées est grand, plus les estimateurs seront précis. En pratique on remarque qu'on obtient déjà de bons résultats avec $K = 5$.

Les méthodes d'imputation "appropriées" ABB et sur la base d'un modèle de régression linéaire sont également applicables pour l'imputation multiple.

Une application de l'imputation multiple est montrée au graphique 5.2. On a une base de données X avec un certain nombre de valeurs manquantes. A l'aide d'une méthode d'imputation simple "appropriée" on impute les données manquantes et on répéte le procéssus $K = 5$. Ceci nous fournira 5 bases de données complètes. Les valeurs imputées différeront d'une base de données à l'autre.

Pour le calcul d'un estimateur $\hat{\theta}$ d'un paramètre θ, on prend la moyenne des estimateurs $\hat{\theta}_k$, $k = 1, ..., K$, obtenus à partir de chaque une des K bases de données.

5.3 Stratification optimale

Un plan d'échantillonnage aléatoire simple stratifié divise une population U en H strates distinctes. La population est divisée en H strates mutuellement exclusives U_i et un échantillon aléatoire est tiré pour chaque strate. On a

$$U = \bigcup_{i=1}^{H} U_i, \qquad U_i \cap U_j = \emptyset, \quad i \neq j \in \{1, 2, ..., H\}. \tag{5.8}$$

L'objectif de toute stratégie de stratification est de maximiser la précision de l'estimateur. Le plan d'échantillonnage stratifié diffère selon (i) le nombre de strates utilisées, (ii) l'allocation de l'échantillon dans la strate et (iii) de la construction des bornes de chaque strate (Horgan, 2006).

Gunning et Horgan (2004), à la suite des travaux de Dalenius (1950), Ekman (1959), Lavallée et Hidiroglou (1988), ont présenté un algorithme simple permettant de trouver les bornes tout en évitant les problèmes discutés dans d'autres méthodes. En se basant sur l'expérience, on va faire les hypothèses suivantes :

- La distribution de la population est supposée asymétrique ;
- Dans chaque strate, la distribution des données (de la variable X) est approximativement uniforme ;
- Les coefficients de variation de chaque strate sont plus ou moins égaux.

La combinaison de ces hypothèses mène à une approximation plus simple vers la solution optimale, qui n'implique aucune itération. Par une solution optimale, on comprend une variance minimale pour la moyenne de chaque strate. Notons qu'on prend une variable auxiliaire X et non pas la variable d'enquête Y afin de calculer les strates, puisque la variable Y est inconnue. Pour calculer les bornes des strates, on va admettre que les coefficients de variation de chaque strate sont égaux, i.e.

$$\frac{\sigma_1}{\mu_1} = \frac{\sigma_2}{\mu_2} = ... = \frac{\sigma_H}{\mu_H}. \tag{5.9}$$

Sous l'hypothèse que chaque strate soit uniformément distribuée, on a

$$\mu_h \approx \frac{k_h + k_{h-1}}{2} \tag{5.10}$$

et

$$\sigma_h \approx \frac{1}{\sqrt{12}} k_h - k_{h-1} \tag{5.11}$$

où μ_h et σ_h sont respectivement la moyenne et l'écart type de la strate h, $1 \leq h \leq H$, et k_h est la borne supérieur de la strate h. Ce qui signifie que le coefficient de variation, $c_h = \sigma_h/\mu_h$, est approximativement

$$c_h \approx \frac{(k_h - k_{h-1})/\sqrt{12}}{(k_h + k_{h-1})/2} \tag{5.12}$$

pour une strate h. Avec l'égalité des c_h on obtient

$$\frac{k_{h+1} - k_h}{k_{h+1} + k_h} = \frac{k_h - k_{h-1}}{k_h + k_{h-1}} \tag{5.13}$$

qui peut être réduit à

$$k_h^2 = k_{h+1} k_{h-1}. \tag{5.14}$$

Ainsi, on peut écrire les bornes des strates comme les termes d'une série géométrique

$$k_h = ar^h; \qquad h = 0, 1, ..., H. \tag{5.15}$$

Pour une population finie, si $a = x_1$ est la valeur minimale de la variable et $ar^H = x_N$ est la valeur maximale de la variable, on fixera $r = x_N/{x_1}^{1/H}$ et l'on obtient les bornes

$$k_0 = x_1 = a, ar, ar^2, ..., ar^H = k_H = x_N. \tag{5.16}$$

5.4 Allocation optimale

Pour une population dont les strates sont fixées, on aimerait déterminer de manière optimale la taille des échantillons n_h à tirer par strates. Si n est la taille de l'échantillon de façon que

$$n = \sum_{h=1}^{H} n_h, \tag{5.17}$$

alors, une allocation optimale de l'échantillon par strate peut s'obtenir par

$$n_h = n \frac{N_h S_{y_{U_h}}}{\sum_{h=1}^{H} N_h S_{y_{U_h}}}. \tag{5.18}$$

où $S_{y_{U_h}}$ est l'écart-type par strate de la variable d'étude (ou d'une variable auxilière suffisamment proche de la variable d'étude) et N_k est le nombre d'unités d'une population appartennant à la strate h. Ce résultat est nommé "*the Neyman allocation*"[1]. L'optimum est atteint lorsque la variance par strate est minimale. En pratique nous ne trouverons pas une allocation optimale. Cependant, il est possible de calculer avec cette méthode une allocation presque optimale.

[1] c.f Särndal *et al.* (1997)

Troisième partie

Stratégies d'estimation et résultats

Chapitre 6

Les données

6.1 Analyse descriptive

Comme on le voit au tableau 6.1, il est évident, que les variables[1] diffèrent l'une de l'autre, ce qui peut s'expliquer, d'une part, par la différence de caractéristiques des variables et, d'autre part, par la différence de taille des assurances. Ensuite, les données montrent également un nombre considérable de valeurs nulles et de variables manquantes, certaines variables n'ont pas moins de 1111 observations nulles sur un total de 1221 observations. Ce qui nous a forcé d'éliminer un certain nombre d'observations au cours de l'analyse. Ceci est également dû à la structure des données. En effet, une structure de données de panel incomplète implique un nombre de données manquantes élevées.

6.2 Choix des variables

Notre variable d'étude est les recettes totales (RT). On l'obtient de la manière suivante. On soustrait les produits neutres (PN) aux produits d'assurance (PA) :

$$RT = PA - PN. \tag{6.1}$$

On observe qu'il y a un nombre élevé de valeurs nulles et manquantes dans les variables composant la variable expliquée, comme on le voit très clairement au tableau 6.1. On s'aperçoit que les observations nulles et manquantes peuvent être rattachées à certains produits d'assurance. On va donc se limiter aux quantités principales de la recette totale. On utilise uniquement les variables des produits, franchise ordinaire ($EF22\_A\_6$, $EF22\_M\_6$, $EF22\_A\_7$, $EF22\_M\_7$) et franchise avec option ($EF23\_A\_6$, $EF23\_M\_6$, $EF23\_A\_7$, $EF23$

[1] Une déscription de l'ensemble de nos données figure en annexe A

_M_7), afin de calculer les recettes totales. On posera

$$PA = EF22\_A\_6 + EF22\_M\_6 + EF23\_A\_6 + EF23\_M\_6, \tag{6.2}$$

$$PN = EF22\_A\_7 + EF22\_M\_7 + EF23\_A\_7 + EF23\_M\_7. \tag{6.3}$$

La RT sera modélisée par une équation de régression. Ce sera la variable expliquée du modèle. Les variables explicatives seront essentiellement les effectifs des produits d'assurance. Cependant, la décision a été prise, d'éliminer les produits d'assurance avec bonus ($EF33\_EF\_BO\_T$), produits d'assurance avec choix limités ($EF33\_EF\_CL\_T$) et l'effectif des assureurs qui bénéficient d'une réduction de primes ($EF33\_EF\_ARP\_T$), ceux-ci sont caractérisés par un grand nombre de valeurs nulles et manquantes. En ce qui concerne la variable $EF33\_EF\_ARP\_T$, on peut également la justifier économiquement. Les assurés ne sont pas suffisamment informés, s'ils peuvent bénéficier d'une réduction de primes lorsqu'ils sont assujettis à une autre assurance, comme l'assurance accident payée par l'employeur. Tous les produits de primes selon le modèle d'assurance ($EF34$...) n'ont pas été pris en considération pour cause d'un risque d'identité avec la variable explicative. Ensuite, les effectifs des assurés selon l'âge ($EF31$...) n'ont pas été considérés pour cause d'une colinéarité possible avec les assurés selon le modèle d'assurance. On peut justifier ceci par le fait que le total de l'effectif d'une assurance selon le modèle d'assurance est également donné par le total de l'effectif des assurés selon l'âge. Nous avons en outre construit la variable *taille*.

La variable *taille* est une variable catégorielle se basant sur le volume de primes qu'une assurance encaisse. Nous avons divisé les assurances en trois catégories, les petites, les moyennes et les grandes. On considère les assurances comme petites lorsque le volume de primes est entre zéro et $4'470'000$ francs suisses. Les assurances moyennes encaissent un volume de primes de $4'470'000$ à $197'000'000$ francs suisses. Les grandes assurances sont celles qui encaissent plus que $197'000'000$ francs suisses de primes par années. Cette variable a été construite afin de capter les différences entre les petites et grandes assurances. En ce qui concerne la variable $EF\_310\_X\_CH$, il s'agit d'une variable binaire. Cette variable prend la valeur 1 lorsque l'assureur à un rayon d'activité s'étendant sur toute la Suisse et la valeur 0 lorsque l'assurance offre des prestations à un niveau régional. Cette variable capte également la différence de grandeur entre les assureurs.

Enfin, les dépenses de publicité ($EF\_16\_X\_450$), nous indique le volume de dépenses qu'une assurance dépense en matière de publicité. Il est clair que cette variable nous donne d'une part, des informations sur la taille d'une assurance et d'autre part si l'assureur a un cercle d'activité régional ou national.

TAB. 6.1 – Statistiques descriptives des variables utilisées

	nbr.val	nbr.null	nbr.na	min	max	range	sum
EF33_EF_O_T	1221	71	30	0	718641	718641	42730516
EF33_EF_FO_T	1221	209	30	0	782085	782085	30915759
EF33_EF_BO_T	1221	1111	30	0	20671	20671	120741
EF33_EF_CL_T	1221	736	30	0	303003	303003	6885625
EF33_EF_ARP_T	1075	874	176	0	108581	108581	551106
EF16_X_450	1249	296	2	-8573	18952889	18961462	1129073930
EF310_X_CH	1248	623	3	0	1	1	625
EF31_T_A_MOY	1075	127	176	0	126	126	35892
EF34_PRI_BO_T	1221	1111	30	0	28783256	28783256	199768580
EF34_PRI_CL_T	1221	737	30	0	598976680	598976680	12592558326
EF34_PRI_FO_T	1221	208	30	0	1383607046	1383607046	68374512535
EF34_PRI_O_T	1221	71	30	0	1329760717	1329760717	83201529531
EF22_A_6	1207	245	44	-21510	70016751	70038261	3840201100
EF23_A_6	1160	263	91	-95114	49950467	50045581	2430290517
EF24_A_6	988	884	263	0	780016	780016	6133858
EF25_A_6	989	857	262	0	10597523	10597523	86959557
EF26_A_6	1052	624	199	-65	15841428	15841493	323267728
EF27_A_6	599	563	652	0	11175735	11175735	45681060
EF22_M_6	1207	58	44	-516233	1263850106	1264366339	79485072838
EF23_M_6	1160	146	91	-2282740	1331553971	1333836711	65745911445
EF24_M_6	988	878	263	0	27982125	27982125	192398593
EF25_M_6	989	843	262	0	97420761	97420761	2255053023
EF26_M_6	1052	598	199	-7196	476594603	476601799	8531918433
EF27_M_6	599	563	652	0	271039805	271039805	1358615009
EF22_A_7	1207	378	44	-1246857	8749175	9996031	76276376
EF23_A_7	1160	381	91	-1456954	2464185	3921139	37833487
EF24_A_7	988	889	263	-5382	12278	17659	94636
EF25_A_7	989	869	262	-95604	234651	330255	1304175
EF26_A_7	1052	672	199	-209070	1550206	1759276	6006059
EF27_A_7	599	569	652	-49731	158674	208405	383100
EF22_M_7	1207	62	44	-20164914	55907943	76072856	1383114102
EF23_M_7	1160	249	91	-15973403	67034442	83007845	944183823
EF24_M_7	988	882	263	-138895	401376	540271	2910409
EF25_M_7	989	853	262	-2145229	4439583	6584812	32410587
EF26_M_7	1052	636	199	-4110559	8976667	13087226	114908531
EF27_M_7	599	569	652	-1206110	4533450	5739560	11926265
rec.tot1	1160	53	91	-4513193	2256684514	2261197707	148884531918

Source : Office fédéral de la santé publique (OFSP), 2007.

Chapitre 7

Estimation du modèle

7.1 Estimation par modèle de régression classique

Nous avons testé plusieurs possibilités de modéliser les recettes totales. Dans ce chapitre, nous avons retenu les modèles avec le R^2 le plus élevé et les critères d'informations les plus bas[1]. Notre période de référence est de 1996 à 2006.

7.1.1 Modèle général

La forme structurelle du modèle de base, nommé modèle1, est la suivante. On régresse les recettes totales en fonction des effectifs par produits d'assurances. On ajoute encore la variable binaire $EF310\_X\_CH$, les dépenses de publicité ($EF16\_X\_450$), et la variable muette de taille. Les coefficients de ce premier modèle se trouvent dans le tableau 7.1. Les effectifs des assurés selon le modèle d'assurance sont des variables très sinificatives. La variable binaire de "régionalité" n'est pas significative et la variable catégorielle du volume des primes n'est seulement significative que pour la troisième catégorie par rapport à la première ($taille^b$). D'autre part, le coefficient des dépenses de publicité est négatif, ce qui contredit notre intuition économique. Le modèle a un R^2 très élevé de 0.952, mais le niveau des critères d'informations (AIC et BIC) est extrêmement élevé.

7.1.2 Modèle semi-logarithmique

Ce modèle utilise le logarithme des recettes totales comme variable dépendante. Les variables explicatives restent les mêmes. Ce modèle est nommé

[1]Les modèles négligés sont notamment des modèles avec des effectifs moyens pour les variables explicatives et la variable expliquée, et des modèles semi-logarithmiques avec des effectifs moyens. Ceci parce que les coefficients n'avaient ni le signe attendu, ni un niveau de signification satisfaisant.

Tab. 7.1 – Coefficients du modèle de régression classique

	modèle1	modèle2	modèle3
Coefficients			
(Intercept)	−907571.812	14.185***	8.384***
	(5257494.601)	(0.051)	(0.084)
EF33_EF_FO_T	2026.537***[c]	0.000***	
	(51.865)	(0.000)	
EF33_EF_O_T	1691.501***	0.000***	
	(42.174)	(0.000)	
EF16_X_450	2.625	0.000***	0.000
	(1.382)	(0.000)	(0.000)
EF310_X_CH : Suisse-Régional	−450791.278	0.693***	0.171***
	(6412079.798)	(0.062)	(0.025)
taille[a]	1833307.116	2.405***	0.302***
	(6861736.938)	(0.067)	(0.037)
taille[b]	66341138.596***	4.167***	0.593***
	(12451081.341)	(0.121)	(0.067)
log(EF33_EF_FO_T)			0.312***
			(0.009)
log(EF33_EF_O_T)			0.625***
			(0.013)
Summaries			
R-squared	0.952	0.875	0.982
adj. R-squared	0.952	0.874	0.982
sigma	80295693.434	0.780	0.292
F	2890.7	1011.4	8122.1
p	0.000	0.000	0.000
Log-likelihood	−17183.8	−1021.5	−159.9
Deviance	5602789195796274176.0	528.3	73.9
AIC	34383.5	2058.9	335.7
BIC	34421.7	2097.1	373.9
N	876	876	876

Source : Elaboration personnelle

a La catégorie(4'470'000,197'000'000]par rapport à la catégorie [0,4'470'000].

b La catégorie >197'000'000 par rapport à la catégorie[0,4'470'000]

c Les étoiles donnent une indication sur le niveau de significativité des coefficients. Où *** indique un niveau de significativité à plus que 0.001%,** indique un niveau de significativité de 0.001%, et où * indique un niveau de significativité de 0.05, les coefficients sans étoiles ont une significativité en dessous de 0.1.

modèle2. Les coefficients sont tous significatifs et positifs. Les critères d'informations s'améliorent par rapport au modèle1. Néanmoins, le R^2 diminue un peu. Il faut ajouter que ce modèle n'est pas robuste. Un changement de variable diminue le R^2 et les coefficients sont moins significatifs.

7.1.3 Modèle logarithmique

Le modèle logarithmique (modèle3) montre de bonnes propriétés. L'avantage d'un modèle logarithmique est le fait qu'on puisse interpréter les coefficients en termes d'élasticités. Les coefficients sont tous significatifs et positifs avec un R^2 très élevé. Le modèle est robuste : La valeur des coefficients change seulement de peu, lorsqu'on enlève ou ajoute des variables. Le modèle a également des critères d'informations faibles. Ceci est la raison pourquoi nous allons retenir ce modèle pour l'estimation par les modèles de panel.

7.2 Estimation par modèle de panel

Nous allons considérer uniquement la forme logarithmique des modèles de panel. Nos estimations sont résumées dans le tableau 7.2.

7.2.1 Estimations sous l'hypotèse d'effets fixes

Le modèle4 considère l'estimation "within" qui est l'estimation par effets fixes. Pour les modèles à effets fixes on admet que les effets individuels et temporels sont des paramètres fixes à estimer. Les effets fixes sont potentiellement corrélés avec les régresseurs. Avec cette modèlisation, nous n'avons pas de constante. En effet, la procédure "within" fait en sorte d'éliminer les effets fixes et la constante avant de procéder à une estimation par MCO.

Les effectifs par le produit d'assurance sont très significatifs ce qui vaut également pour la variable catégorielle qui est dans les deux cas significative. D'autre part, les dépenses de publicité tout comme la variable binaire de "régionalité" ne sont pas significatives. Le R^2 est extrêmement élevé.

7.2.2 Estimations sous l'hypotèse d'effets aléatoires

Le modèle5 considère l'estimation "random" qui est l'estimation par effets aléatoires. L'hypothèse sous-jacente de ce modèle est que les effets temporels et individuels ne sont ni corréles avec les régresseurs, ni mutuellement dépendants ou encore dépendants du terme d'erreur. Malheureusement, l'estimation n'as pas pu être effectué avec les deux effets, car le logiciel R ne soutient pas l'estimation "random" à deux effets pour les jeux de données non complètes. Les résultats affichés au tableau 7.2 prennent seulement en compte l'effet individuel. On a également ici un R^2 élevé. Les effectifs selon le modèle d'assurance sont significatifs et également la variable *taille*. De surcroît, l'estimation "random" calcule en plus la constante qui est hautement significative.

TAB. 7.2 – Coefficients du modèle de régression de panel

	modèle4	modèle5	modèle6
Coefficients			
(Intercept)		8.118***	8.362***
		(0.108)	(0.086)
log(EF33_EF_FO_T)	0.260***[c]	0.300***	0.307***
	(0.010)	(0.009)	(0.009)
log(EF33_EF_O_T)	0.724***	0.678***	0.630***
	(0.019)	(0.016)	(0.013)
EF16_X_450	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
EF310_X_CH : Suisse-Régional	−0.011	0.002	0.172***
	(0.022)	(0.024)	(0.026)
taille[a]	0.182***	0.280***	0.300***
	(0.030)	(0.035)	(0.038)
taille[b]	0.279***	0.462***	0.604***
	(0.053)	(0.062)	(0.069)
Summaries			
Multiple R-squared	0.932	0.976	0.983
Residual Sum of Squares	15.7	29.5	68.8
Total Sum of Squares	232	1250	4090
F	1559.43	5632.76	7971.11
p	0.000	0.000	0.000
N	824	824	824

Source : Elaboration personnelle

a La catégorie(4'470'000,197'000'000]par rapport à la catégorie [0,4'470'000].

b La catégorie >197'000'000 par rapport à la catégorie[0,4'470'000]

c Les étoiles donnent une indication sur le niveau de significativité des coefficients. Où *** indique un niveau de significativité à plus que 0.001%,** indique un niveau de significativité de 0.001%, et où * indique un niveau de significativité de 0.05, les coefficients sans étoiles ont une significativité en dessous de 0.1.

7.2.3 Estimations “pooled”

Le dernier modèle est une estimation “pooled”. Qui est l'estimation MCO des observations empilées. C'est le même modèle que le modèle3 sauf que le calcul de la régression se fait avec un nombre d'observation N plus petit que dans la régression classique. Ce qui explique également pourquoi les coefficients et le R^2 ne sont pas exactement la même chose. Ceci est dû au processus de calcul en R de la fonction plm (panel linear model).

Chapitre 8

Stratégies de prévisions

La situation de base est la suivante. Nous sommes en possession des données complètes de 1996 à 2006. Afin de pouvoir procéder à des stratégies de prévision, nous allons utiliser la période 1996 à 2006 comme base, pour l'estimation des modèles où l'on espère pouvoir tirer des avantages de la structure de panel. L'année 2006 nous sert, d'une part, de contrôle pour juger la qualité des prévisions par les modèles et, d'autre part, à tirer aléatoirement un certain nombre d'assureurs, pour simuler les assureurs qui rendent à temps le formulaire.

Nous avons procédé de la façon suivante. Nous avons tiré aléatoirement de l'année 2006 un nombre n d'assureurs. Afin de simuler les assureurs qui rendent à temps le formulaire EF123. Le reste des assureurs pour l'année 2006 sont des valeurs manquantes. Celles-ci seront imputées par deux méthodes d'imputation "appropriées", la méthode d'imputation ABB et une méthode d'imputation basée sur un modèle linéaire.

Les stratégies utilisées sont, d'un côté, l'utilisation des valeurs estimées issues de modèles et, de l'autre, l'utilisation des vraies valeurs de 2006. Les stratégies de prévision, que nous allons voir plus en détail plus bas, donnent de mauvais résultats si on utilise la méthode d'imputation sans stratification.

Comme on l'a constaté en pratique il est important de pouvoir procéder à l'imputation par GHR. Dans le cas présent, une procédure de stratification dite optimale nous permettera de constituer ces groupes. Malheureusement, la stratification optimale à elle seule n'est, dans notre cas, pas suffisante pour produire des résultats stables de prévision. Il est judicieux de prendre également en compte la distribution des assureurs. Il se peut qu'on ait par exemple une densité plus importante de grandes assurances que de petites assurances. C'est pourquoi on ne peut pas donner le même poids à chaque strate. En procédant à une allocation optimale du nombre d'unités à tirer par strates, on prend également en compte la distribution des assureurs. Ce qui nous aide de concevoir des résultats stables de prévision.

8.1 Procédure

La procédure adoptée est la suivante. Dans un premier temps, il a fallu procéder à la construction des strates. Basé sur l'expérience, il est possible de dire que 5 strates sont suffisantes pour distinguer les assureurs selon leur grandeur. L'allocation des différentes assurances dans leurs strates respectives a été calculée par la stratification optimale selon Gunning et Horgan (2004). La construction des strates a été faite avec les données réelles de 2005 (recettes totales 2005). Sous l'hypothèse que la grandeur des assurances reste stable dans le temps, la stratification de 2005 a été appliquée aux assurances de 2006. Une fois que les strates sont définies, on peut procéder au tirage aléatoire de l'échantillon de n assureurs de l'année 2006 qui simuleront les assureurs qui ont rendu à temps le formulaire. Comme nous l'avons vu plus haut, la stratification n'est pas suffisante afin d'arriver à de bons résultats de prévision. Nous avons procédé à une allocation optimale dite de "*Neyman*", afin d'allouer de façon optimale l'échantillon n dans les strates.

Le deuxième volet traite la modélisation. Différents modèles ont été traités, notamment une estimation "pooled" à variables muettes, une estimation "random" et une estimation "within" qui ont été estimées avec le jeux de données complet de 1996 à 2006. Une fois les valeurs estimées, on tire un échantillon de valeurs estimées de 2006. C'est avec cet échantillon qu'on va procéder à une imputation multiple "appropriée". Il faut noter qu'heuristiquement on a remarqué qu'un nombre de $K = 5$ imputations est suffisant pour de bons résultats. La recette totale estimées pour 2006 est obtennue en prenant la moyenne des recettes totales estimées à partir des K bases de données imputées.

Finalement, nous allons calculer la précision de la prévision. La différence entre les recettes totales estimées ($rec.tot1.fitted$) et les vraies recettes totales 2006 ($rec.tot.2006$) nous donne une idée sur la qualité de la précision de la prévision. Exprimée en pour-cent, nous avons la mesure suivante :

$$Precision = \frac{\sum rec.tot1.fitted - \sum rec.tot.2006}{\sum rec.tot.2006}. \tag{8.1}$$

Afin de pouvoir tirer des conclusions sur les résultats de la prévision, on procède à une simulation de la procédure. Nous répetons 100 fois la démarche expliquée. En tirant la moyenne arithmétique sur les résultats de l'itération, nous obtenons la précision moyenne de toutes les simulations. Les précisions de prévision moyennes se trouvent au tableau 8.1 et sont discutées au chapitre suivant.

8.2 Stratégies

8.2.1 Stratégies avec des valeurs estimées

Premièrement, nous allons discuter les résultats qui ont été générés par la méthode d'imputation "appropriée" ABB. Sur la base des données estimées par les modèles développés au chapitre précédent, nous avons pu produire, pour différentes tailles d'échantillon, les résultats du tableau 8.1.

On remarque, que les résultats pour les estimations "random" et "within" sont extrêmement proches. On peut en déduire que l'effet temporel n'est peut-être pas aussi important qu'on le croyait. On voit que pour l'estimation "within", on a une précision de la prévision de 21% pour un échantillon de taille $n = 15$; on augmente la précision à un niveau de 12% pour un échantillon de taille $n = 25$. Avec un peu près un tiers des assureurs ($n = 36$) nous obtenons une précision de 4,1%. Avec l'estimation "within" nous avons une sous estimation systématique des recettes totales. Les résultats sont similaires pour l'estimation "random". En effet, nous obtenons dans les simulations exactement les mêmes résultats qu'avec l'estimation "within".

Lorsqu'on procède à une prévision avec une estimation par MCO nous avons une amélioration de la précision surprenante. En effet, nous avons une précision de 0.48% pour un échantillon de $n = 15$, de 0.3% pour un échantillon de $n = 25$ et de 0.038% pour un échantillon de $n = 36$. Cependant, nous avons pour une estimation par MCO une légère surestimation de recettes totales, car les signes sont négatifs.

Deuxièmement, nous allons aborder le modèle d'imputation sur la base d'un modèle linéaire. Comme on le voit au tableau 8.1, les résultats de cette méthode d'imputation sont excessivement mauvais. C'est pourquoi nous n'avons représenté que l'estimation par MCO dans le tableau. Ce résultat déçoit par le fait qu'on impute les valeurs estimées sur la base d'un modèle de régression, duquel on attend en général de meilleurs résultats.

La dernière colonne du tableau 8.1, indique le niveau de précision lorsqu'on procéde à une allocation non optimale de l'échantillon par strates. Les strates ont été allouées de façon similaire, chaque strate comptant le même nombre d'assureurs. Si on a un échantillon de par exemple $n = 15$, les strates seront allouées de manière à avoir 3 assurances par strates. On s'aperçoit que la précision est moins bonne que la précision des résultats obtenus par une allocation optimale.

8.2.2 Stratégies avec les valeurs réelles

Pour la prévision avec les valeurs réelles nous avons procédé, comme pour les modèles, à une stratification optimale et une allocation optimale des strates. Cependant, nous avons pris les échantillons sur la base des données réelles et non pas sur celle des valeurs estimées. Nous obtenons ainsi des prévisions très exactes. Sur la base d'un échantillon de $n = 15$, nous

TAB. 8.1 – Précision de la prévision

i=100[a]	estimation rec.tot	rec.tot 2006	précision[b]	précision (ano)[bc]
Données estimées				
Impuatation par ABB				
WITHIN				
n=15	12745532113.94	16175236711.93	0.21	0.75
n=25	14165335018.38	16175236711.93	0.12	0.59
n=36	15509097504.22	16175236711.93	0.041	0.52
RANDOM				
n=15	12721150373.70	16175236711.93	0.21	0.76
n=25	14190037714.16	16175236711.93	0.12	0.6
n=36	15505953705.73	16175236711.93	0.041	0.52
MCO				
n=15	16253117089.37	16175236711.93	-0.0048	0.028
n=25	16223562711.47	16175236711.93	-0.003	0.026
n=36	16180454032.65	16175236711.93	-0.00032	0.012
Imputation par modèle linéaire				
MCO				
n=15	163228802520059072	16175236711.93	-10091277	-17761679
n=25	506304109663426880	16175236711.93	31301186	-5419891
n=36	47192979365244640	16175236711.93	-2917606	-5419891
Données réelles[d]				
n=15	16286687432.91	16300874623.81	0.00087	-0.0017
n=25	16314705884.35	16300874623.81	-0.00084	-0.0010
n=36	16300024596.03	16300874623.81	0.000052	0.0047

Source : Elaboration personnelle.
a i est le nombre d'itérations de la simulation.
b En pourcent des recettes totales 2006 estimées, ou par exemple 0.5 est égal à 50%.
c Précision lorsque l'allocation des strates est non-optimale.
d Les ret.tot 2006 des données réelles différent par rapport au données estimées, parce que un plus grand nombre d'assureurs a été pris en compte dans le calcul.

obtenons déjà une précision de 0.03%. En augmentant l'échantillon, la prévision devient encore plus précise. Avec $n = 36$ nous avons calculé une précision de 0.0053%, ce qui est meilleur que tout résultat produit sur la base d'un modèle. Ces résultats oscillent avec une faible variance autour de zéro. On est très précis, mais on ne sait pas si on surestime ou sous-estime légèrement la vraie valeur. L'allocation des strates est également importante pour la prévision avec les valeurs réelles, car elle augmente la précision des résultats.

8.3 Recommandations

Nous avons vu que les résultats basés sur l'estimation des valeurs de 2006 donnent de moins bons résultats que les valeurs réelles. Ce qui peut surprendre, parce que avec un modèle on utilise également les informations disponibles des années antérieurs, et l'on pourrait suspecter d'obtenir de meilleurs résultats. Il n'en demeure pas moins que, vu les résultats, de recommander l'utilisation de la méthode d'imputation "appropriée" ABB avec les données réelles. Nous recommandons de procéder aux étapes suivantes :

- En début d'année, stratifier la variable d'intérêt à partir de l'année précédente ;
- Calculer l'allocation optimale pour différentes tailles d'échantillon, par exemple $n = 15$;
- Lors de premières estimations, répartir par strate selon la stratification optimale les assurances qui sont arrivées et constater dans quelle mesure cette allocation diverge de l'allocation optimale ;
- D'effectuer la prévision par la méthode d'imputation multiple ABB sur la base des données réelles de l'année en cours ;
- Dans l'appréciation du résultat de bien noter la différence qui existe entre la répartition des assurances par strate optimale et allocation optimale.

Conclusion

Dans le présent travail, il s'agit de trouver un modèle robuste et des méthodes appropriées, permettant des estimations et des prévisions fiables pour l'année en cours. On a pu montrer que pour les recettes totales, les modèles de panel s'appliquaient bien, car nous avons pu trouver des modèles robustes. Cependant, les modèles se sont avérés peu effectifs pour la prévision sous-estimant les recettes totales de façon conséquente.

L'apport du travail montre la puissance de précisison de la méthode d'imputation multiple ABB. Néanmoins, il est important de procéder à une stratification optimale de la variable expliquée, où l'on prend en considération la grandeur des assurances. D'autre part, il faut procéder à une allocation optimale des strates. En effet, la stratification optimale est seulement effective lorsqu'on prend en compte la distribution des assureurs. En ce qui concerne la puissance de précision, celle-ci est surtout surprenante lors de l'application de la méthode d'imputation multiple ABB aux valeurs réelles et à une application par MCO.

Pour calculer les algorithmes nous avons utilisé le logiciel R. Ce logiciel est à disposition de tout le monde. En effet, il peut être télécharché gratuitement sur le site Internet suivant : http ://www.r-project.org/. De surcroît, une vaste documentation sur son utilisation se trouve en-ligne.

Données

Tab. A.1: Les données de l'OFSP et leur description

	Variable	Sinification
1	EF22_A_3	Franchise ordinaire - Total des charges d'assurance - accident
2	EF22_A_30.33	Franchise ordinaire - Prestations payées - accident
3	EF22_A_30.35	Franchise ordinaire - Prestation brutes - accident
4	EF22_A_31	Franchise ordinaire - Prestations - accident
5	EF22_A_32	Franchise ordinaire - Participation des assurés aux frais - accident
6	EF22_A_3.4	Franchise ordinaire - Charges d'assurance et d'exploitation - accident
7	EF22_A_34	Franchise ordinaire - Autres charges d'assurance - accident
8	EF22_A_35	Franchise ordinaire - Provisions pour cas d'assurance non liquidés - accident
9	EF22_A_36	Franchise ordinaire - Part des prestations remboursées par les réassureurs - accident
10	EF22_A_37	Franchise ordinaire - Compensation des risques - accident
11	EF22_A_4	Franchise ordinaire - Total des charges d'exploitation - accident
12	EF22_A_40.47	Franchise ordinaire - Charges d'administration - accident
13	EF22_A_48	Franchise ordinaire - Amortissements - accident
14	EF22_A_49	Franchise ordinaire - Autres charges d'exploitation - accident
15	EF22_A_6	Franchise ordinaire - Produits d'assurance - accident
16	EF22_A_60.65	Franchise ordinaire - Primes brutes - accident
17	EF22_A_60.66	Franchise ordinaire - Primes propres - accident
18	EF22_A_61	Franchise ordinaire - Primes selon coûts effectifs - accident
19	EF22_A_64	Franchise ordinaire - Déductions accordées sur primes - accident
20	EF22_A_65	Franchise ordinaire - Autres primes - accident
21	EF22_A_66	Franchise ordinaire - Part de primes des réassureurs - accident
22	EF22_A_67	Franchise ordinaire - Réduction des primes et autres contribution - accident
23	EF22_A_68	Franchise ordinaire - Déduction des parts de primes des assurés - accident
24	EF22_A_69	Franchise ordinaire - Autres produits d'exploitation - accident
25	EF22_A_7	Franchise ordinaire - Charges et produits neutres - accident
26	EF22_A_RES	Franchise ordinaire - Résultat d'exploitation de l'assurance - accident
27	EF22_A_R.GEN	Franchise ordinaire - Résultat du compte d'exploitation général - accident
28	EF22_M_3	Franchise ordinaire - Total des charges d'assurance - maladie
29	EF22_M_30.33	Franchise ordinaire - Prestations payées - maladie
30	EF22_M_30.35	Franchise ordinaire - Prestation brutes - maladie
31	EF22_M_31	Franchise ordinaire - Prestations - maladie
32	EF22_M_32	Franchise ordinaire - Participation des assurés aux frais - maladie
33	EF22_M_3.4	Franchise ordinaire - Charges d'assurance et d'exploitation - maladie
34	EF22_M_34	Franchise ordinaire - Autres charges d'assurance - maladie
35	EF22_M_35	Franchise ordinaire - Provisions pour cas d'assurance non liquidés - maladie
36	EF22_M_36	Franchise ordinaire - Part des prestations remboursées par les réassureurs - maladie
37	EF22_M_37	Franchise ordinaire - Compensation des risques - maladie
38	EF22_M_4	Franchise ordinaire - Total des charges d'exploitation - maladie
39	EF22_M_40.47	Franchise ordinaire - Charges d'administration - maladie
40	EF22_M_48	Franchise ordinaire - Amortissements - maladie
41	EF22_M_49	Franchise ordinaire - Autres charges d'exploitation - maladie
42	EF22_M_6	Franchise ordinaire - Produits d'assurance - maladie
43	EF22_M_60.65	Franchise ordinaire - Primes brutes - maladie
44	EF22_M_60.66	Franchise ordinaire - Primes propres - maladie
45	EF22_M_61	Franchise ordinaire - Primes selon coûts effectifs - maladie
46	EF22_M_64	Franchise ordinaire - Déductions accordées sur primes - maladie
47	EF22_M_65	Franchise ordinaire - Autres primes - maladie
48	EF22_M_66	Franchise ordinaire - Part de primes des réassureurs - maladie
49	EF22_M_67	Franchise ordinaire - Réduction des primes et autres contribution - maladie
50	EF22_M_68	Franchise ordinaire - Déduction des parts de primes des assurés - maladie
51	EF22_M_69	Franchise ordinaire - Autres produits d'exploitation - maladie
52	EF22_M_7	Franchise ordinaire - Charges et produits neutres - maladie
53	EF22_M_RES	Franchise ordinaire - Résultat d'exploitation de l'assurance - maladie

Suite à la page suivante

Suite

	Variable	Sinification
54	EF22_M_R.GEN	Franchise ordinaire - Résultat du compte d'exploitation général - maladie
55	EF23_A_3	Assurance avec franchise à option - Total des charges d'assurance - accident
56	EF23_A_30.33	Assurance avec franchise à option - Prestations payées - accident
57	EF23_A_30.35	Assurance avec franchise à option - Prestation brutes - accident
58	EF23_A_31	Assurance avec franchise à option - Prestations - accident
59	EF23_A_32	Assurance avec franchise à option - Participation des assurés aux frais - accident
60	EF23_A_3.4	Assurance avec franchise à option - Charges d'assurance et d'exploitation - accident
61	EF23_A_34	Assurance avec franchise à option - Autres charges d'assurance - accident
62	EF23_A_35	Assurance avec franchise à option - Provisions pour cas d'assurance non liquidés - accident
63	EF23_A_36	Assurance avec franchise à option - Part des prestations remboursées par les réassureurs - accident
64	EF23_A_37	Assurance avec franchise à option - Compensation des risques - accident
65	EF23_A_4	Assurance avec franchise à option - Total des charges d'exploitation - accident
66	EF23_A_40.47	Assurance avec franchise à option - Charges d'administration - accident
67	EF23_A_48	Assurance avec franchise à option - Amortissements - accident
68	EF23_A_49	Assurance avec franchise à option - Autres charges d'exploitation - accident
69	EF23_A_6	Assurance avec franchise à option - Produits d'assurance - accident
70	EF23_A_60.65	Assurance avec franchise à option - Primes brutes - accident
71	EF23_A_60.66	Assurance avec franchise à option - Primes propres - accident
72	EF23_A_61	Assurance avec franchise à option - Primes selon coûts effectifs - accident
73	EF23_A_64	Assurance avec franchise à option - Déductions accordées sur primes - accident
74	EF23_A_65	Assurance avec franchise à option - Autres primes - accident
75	EF23_A_66	Assurance avec franchise à option - Part de primes des réassureurs - accident
76	EF23_A_67	Assurance avec franchise à option - Réduction des primes et autres contribution - accident
77	EF23_A_68	Assurance avec franchise à option - Déduction des parts de primes des assurés - accident
78	EF23_A_69	Assurance avec franchise à option - Autres produits d'exploitation - accident
79	EF23_A_7	Assurance avec franchise à option - Charges et produits neutres - accident
80	EF23_A_RES	Assurance avec franchise à option - Résultat d'exploitation de l'assurance - accident
81	EF23_A_R.GEN	Assurance avec franchise à option - Résultat du compte d'exploitation général - accident
82	EF23_M_3	Assurance avec franchise à option - Total des charges d'assurance - maladie
83	EF23_M_30.33	Assurance avec franchise à option - Prestations payées - maladie
84	EF23_M_30.35	Assurance avec franchise à option - Prestation brutes - maladie
85	EF23_M_31	Assurance avec franchise à option - Prestations - maladie
86	EF23_M_32	Assurance avec franchise à option - Participation des assurés aux frais - maladie
87	EF23_M_3.4	Assurance avec franchise à option - Charges d'assurance et d'exploitation - maladie
88	EF23_M_34	Assurance avec franchise à option - Autres charges d'assurance - maladie
89	EF23_M_35	Assurance avec franchise à option - Provisions pour cas d'assurance non liquidés - maladie
90	EF23_M_36	Assurance avec franchise à option - Part des prestations remboursées par les réassureurs - maladie
91	EF23_M_37	Assurance avec franchise à option - Compensation des risques - maladie
92	EF23_M_4	Assurance avec franchise à option - Total des charges d'exploitation - maladie
93	EF23_M_40.47	Assurance avec franchise à option - Charges d'administration - maladie
94	EF23_M_48	Assurance avec franchise à option - Amortissements - maladie
95	EF23_M_49	Assurance avec franchise à option - Autres charges d'exploitation - maladie
96	EF23_M_6	Assurance avec franchise à option - Produits d'assurance - maladie

Suite à la page suivante

Suite

	Variable	Sinification
97	EF23_M_60.65	Assurance avec franchise à option - Primes brutes - maladie
98	EF23_M_60.66	Assurance avec franchise à option - Primes propres - maladie
99	EF23_M_61	Assurance avec franchise à option - Primes selon coûts effectifs - maladie
100	EF23_M_64	Assurance avec franchise à option - Déductions accordées sur primes - maladie
101	EF23_M_65	Assurance avec franchise à option - Autres primes - maladie
102	EF23_M_66	Assurance avec franchise à option - Part de primes des réassureurs - maladie
103	EF23_M_67	Assurance avec franchise à option - Réduction des primes et autres contribution - maladie
104	EF23_M_68	Assurance avec franchise à option - Déduction des parts de primes des assurés - maladie
105	EF23_M_69	Assurance avec franchise à option - Autres produits d'exploitation - maladie
106	EF23_M_7	Assurance avec franchise à option - Charges et produits neutres - maladie
107	EF23_M_RES	Assurance avec franchise à option - Résultat d'exploitation de l'assurance - maladie
108	EF23_M_R.GEN	Assurance avec franchise à option - Résultat du compte d'exploitation général - maladie
109	EF24_A_3	Assurance avec bonus - Total des charges d'assurance - accident
110	EF24_A_30.33	Assurance avec bonus - Prestations payées - accident
111	EF24_A_30.35	Assurance avec bonus - Prestation brutes - accident
112	EF24_A_31	Assurance avec bonus - Prestations - accident
113	EF24_A_32	Assurance avec bonus - Participation des assurés aux frais - accident
114	EF24_A_3.4	Assurance avec bonus - Charges d'assurance et d'exploitation - accident
115	EF24_A_34	Assurance avec bonus - Autres charges d'assurance - accident
116	EF24_A_35	Assurance avec bonus - Provisions pour cas d'assurance non liquidés - accident
117	EF24_A_36	Assurance avec bonus - Part des prestations remboursées par les réassureurs - accident
118	EF24_A_37	Assurance avec bonus - Compensation des risques - accident
119	EF24_A_4	Assurance avec bonus - Total des charges d'exploitation - accident
120	EF24_A_40.47	Assurance avec bonus - Charges d'administration - accident
121	EF24_A_48	Assurance avec bonus - Amortissements - accident
122	EF24_A_49	Assurance avec bonus - Autres charges d'exploitation - accident
123	EF24_A_6	Assurance avec bonus - Produits d'assurance - accident
124	EF24_A_60.65	Assurance avec bonus - Primes brutes - accident
125	EF24_A_60.66	Assurance avec bonus - Primes propres - accident
126	EF24_A_61	Assurance avec bonus - Primes selon coûts effectifs - accident
127	EF24_A_64	Assurance avec bonus - Déductions accordées sur primes - accident
128	EF24_A_65	Assurance avec bonus - Autres primes - accident
129	EF24_A_66	Assurance avec bonus - Part de primes des réassureurs - accident
130	EF24_A_67	Assurance avec bonus - Réduction des primes et autres contribution - accident
131	EF24_A_68	Assurance avec bonus - Déduction des parts de primes des assurés - accident
132	EF24_A_69	Assurance avec bonus - Autres produits d'exploitation - accident
133	EF24_A_7	Assurance avec bonus - Charges et produits neutres - accident
134	EF24_A_RES	Assurance avec bonus - Résultat d'exploitation de l'assurance - accident
135	EF24_A_R.GEN	Assurance avec bonus - Résultat du compte d'exploitation général - accident
136	EF24_M_3	Assurance avec bonus - Total des charges d'assurance - maladie
137	EF24_M_30.33	Assurance avec bonus - Prestations payées - maladie
138	EF24_M_30.35	Assurance avec bonus - Prestation brutes - maladie
139	EF24_M_31	Assurance avec bonus - Prestations - maladie
140	EF24_M_32	Assurance avec bonus - Participation des assurés aux frais - maladie
141	EF24_M_3.4	Assurance avec bonus - Charges d'assurance et d'exploitation - maladie
142	EF24_M_34	Assurance avec bonus - Autres charges d'assurance - maladie
143	EF24_M_35	Assurance avec bonus - Provisions pour cas d'assurance non liquidés - maladie

Suite à la page suivante

Suite

	Variable	Sinification
144	EF24_M_36	Assurance avec bonus - Part des prestations remboursées par les réassureurs - maladie
145	EF24_M_37	Assurance avec bonus - Compensation des risques - maladie
146	EF24_M_4	Assurance avec bonus - Total des charges d'exploitation - maladie
147	EF24_M_40.47	Assurance avec bonus - Charges d'administration - maladie
148	EF24_M_48	Assurance avec bonus - Amortissements - maladie
149	EF24_M_49	Assurance avec bonus - Autres charges d'exploitation - maladie
150	EF24_M_6	Assurance avec bonus - Produits d'assurance - maladie
151	EF24_M_60.65	Assurance avec bonus - Primes brutes - maladie
152	EF24_M_60.66	Assurance avec bonus - Primes propres - maladie
153	EF24_M_61	Assurance avec bonus - Primes selon coûts effectifs - maladie
154	EF24_M_64	Assurance avec bonus - Déductions accordées sur primes - maladie
155	EF24_M_65	Assurance avec bonus - Autres primes - maladie
156	EF24_M_66	Assurance avec bonus - Part de primes des réassureurs - maladie
157	EF24_M_67	Assurance avec bonus - Réduction des primes et autres contribution - maladie
158	EF24_M_68	Assurance avec bonus - Déduction des parts de primes des assurés - maladie
159	EF24_M_69	Assurance avec bonus - Autres produits d'exploitation - maladie
160	EF24_M_7	Assurance avec bonus - Charges et produits neutres - maladie
161	EF24_M_RES	Assurance avec bonus - Résultat d'exploitation de l'assurance - maladie
162	EF24_M_R.GEN	Assurance avec bonus - Résultat du compte d'exploitation général - maladie
163	EF25_A_3	HMO - Total des charges d'assurance - accident
164	EF25_A_30.33	HMO - Prestations payées - accident
165	EF25_A_30.35	HMO - Prestation brutes - accident
166	EF25_A_31	HMO - Prestations - accident
167	EF25_A_32	HMO - Participation des assurés aux frais - accident
168	EF25_A_3.4	HMO - Charges d'assurance et d'exploitation - accident
169	EF25_A_34	HMO - Autres charges d'assurance - accident
170	EF25_A_35	HMO - Provisions pour cas d'assurance non liquidés - accident
171	EF25_A_36	HMO - Part des prestations remboursées par les réassureurs - accident
172	EF25_A_37	HMO - Compensation des risques - accident
173	EF25_A_4	HMO - Total des charges d'exploitation - accident
174	EF25_A_40.47	HMO - Charges d'administration - accident
175	EF25_A_48	HMO - Amortissements - accident
176	EF25_A_49	HMO - Autres charges d'exploitation - accident
177	EF25_A_6	HMO - Produits d'assurance - accident
178	EF25_A_60.65	HMO - Primes brutes - accident
179	EF25_A_60.66	HMO - Primes propres - accident
180	EF25_A_61	HMO - Primes selon coûts effectifs - accident
181	EF25_A_64	HMO - Déductions accordées sur primes - accident
182	EF25_A_65	HMO - Autres primes - accident
183	EF25_A_66	HMO - Part de primes des réassureurs - accident
184	EF25_A_67	HMO - Réduction des primes et autres contribution - accident
185	EF25_A_68	HMO - Déduction des parts de primes des assurés - accident
186	EF25_A_69	HMO - Autres produits d'exploitation - accident
187	EF25_A_7	HMO - Charges et produits neutres - accident
188	EF25_A_RES	HMO - Résultat d'exploitation de l'assurance - accident
189	EF25_A_R.GEN	HMO - Résultat du compte d'exploitation général - accident
190	EF25_M_3	HMO - Total des charges d'assurance - maladie
191	EF25_M_30.33	HMO - Prestations payées - maladie
192	EF25_M_30.35	HMO - Prestation brutes - maladie
193	EF25_M_31	HMO - Prestations - maladie
194	EF25_M_32	HMO - Participation des assurés aux frais - maladie
195	EF25_M_3.4	HMO - Charges d'assurance et d'exploitation - maladie
196	EF25_M_34	HMO - Autres charges d'assurance - maladie
197	EF25_M_35	HMO - Provisions pour cas d'assurance non liquidés - maladie

Suite à la page suivante

Suite

	Variable	Sinification
198	EF25_M_36	HMO - Part des prestations remboursées par les réassureurs - maladie
199	EF25_M_37	HMO - Compensation des risques - maladie
200	EF25_M_4	HMO - Total des charges d'exploitation - maladie
201	EF25_M_40.47	HMO - Charges d'administration - maladie
202	EF25_M_48	HMO - Amortissements - maladie
203	EF25_M_49	HMO - Autres charges d'exploitation - maladie
204	EF25_M_6	HMO - Produits d'assurance - maladie
205	EF25_M_60.65	HMO - Primes brutes - maladie
206	EF25_M_60.66	HMO - Primes propres - maladie
207	EF25_M_61	HMO - Primes selon coûts effectifs - maladie
208	EF25_M_64	HMO - Déductions accordées sur primes - maladie
209	EF25_M_65	HMO - Autres primes - maladie
210	EF25_M_66	HMO - Part de primes des réassureurs - maladie
211	EF25_M_67	HMO - Réduction des primes et autres contribution - maladie
212	EF25_M_68	HMO - Déduction des parts de primes des assurés - maladie
213	EF25_M_69	HMO - Autres produits d'exploitation - maladie
214	EF25_M_7	HMO - Charges et produits neutres - maladie
215	EF25_M_RES	HMO - Résultat d'exploitation de l'assurance - maladie
216	EF25_M_R.GEN	HMO - Résultat du compte d'exploitation général - maladie
217	EF26_A_3	Réseau santé - Total des charges d'assurance - accident
218	EF26_A_30.33	Réseau santé - Prestations payées - accident
219	EF26_A_30.35	Réseau santé - Prestation brutes - accident
220	EF26_A_31	Réseau santé - Prestations - accident
221	EF26_A_32	Réseau santé - Participation des assurés aux frais - accident
222	EF26_A_3.4	Réseau santé - Charges d'assurance et d'exploitation - accident
223	EF26_A_34	Réseau santé - Autres charges d'assurance - accident
224	EF26_A_35	Réseau santé - Provisions pour cas d'assurance non liquidés - accident
225	EF26_A_36	Réseau santé - Part des prestations remboursées par les réassureurs - accident
226	EF26_A_37	Réseau santé - Compensation des risques - accident
227	EF26_A_4	Réseau santé - Total des charges d'exploitation - accident
228	EF26_A_40.47	Réseau santé - Charges d'administration - accident
229	EF26_A_48	Réseau santé - Amortissements - accident
230	EF26_A_49	Réseau santé - Autres charges d'exploitation - accident
231	EF26_A_6	Réseau santé - Produits d'assurance - accident
232	EF26_A_60.65	Réseau santé - Primes brutes - accident
233	EF26_A_60.66	Réseau santé - Primes propres - accident
234	EF26_A_61	Réseau santé - Primes selon coûts effectifs - accident
235	EF26_A_64	Réseau santé - Déductions accordées sur primes - accident
236	EF26_A_65	Réseau santé - Autres primes - accident
237	EF26_A_66	Réseau santé - Part de primes des réassureurs - accident
238	EF26_A_67	Réseau santé - Réduction des primes et autres contribution - accident
239	EF26_A_68	Réseau santé - Déduction des parts de primes des assurés - accident
240	EF26_A_69	Réseau santé - Autres produits d'exploitation - accident
241	EF26_A_7	Réseau santé - Charges et produits neutres - accident
242	EF26_A_RES	Réseau santé - Résultat d'exploitation de l'assurance - accident
243	EF26_A_R.GEN	Réseau santé - Résultat du compte d'exploitation général - accident
244	EF26_M_3	Réseau santé - Total des charges d'assurance - maladie
245	EF26_M_30.33	Réseau santé - Prestations payées - maladie
246	EF26_M_30.35	Réseau santé - Prestation brutes - maladie
247	EF26_M_31	Réseau santé - Prestations - maladie
248	EF26_M_32	Réseau santé - Participation des assurés aux frais - maladie
249	EF26_M_3.4	Réseau santé - Charges d'assurance et d'exploitation - maladie
250	EF26_M_34	Réseau santé - Autres charges d'assurance - maladie
251	EF26_M_35	Réseau santé - Provisions pour cas d'assurance non liquidés - maladie
252	EF26_M_36	Réseau santé - Part des prestations remboursées par les réassureurs - maladie
253	EF26_M_37	Réseau santé - Compensation des risques - maladie

Suite à la page suivante

Suite

	Variable	Sinification
254	EF26_M_4	Réseau santé - Total des charges d'exploitation - maladie
255	EF26_M_40.47	Réseau santé - Charges d'administration - maladie
256	EF26_M_48	Réseau santé - Amortissements - maladie
257	EF26_M_49	Réseau santé - Autres charges d'exploitation - maladie
258	EF26_M_6	Réseau santé - Produits d'assurance - maladie
259	EF26_M_60.65	Réseau santé - Primes brutes - maladie
260	EF26_M_60.66	Réseau santé - Primes propres - maladie
261	EF26_M_61	Réseau santé - Primes selon coûts effectifs - maladie
262	EF26_M_64	Réseau santé - Déductions accordées sur primes - maladie
263	EF26_M_65	Réseau santé - Autres primes - maladie
264	EF26_M_66	Réseau santé - Part de primes des réassureurs - maladie
265	EF26_M_67	Réseau santé - Réduction des primes et autres contribution - maladie
266	EF26_M_68	Réseau santé - Déduction des parts de primes des assurés - maladie
267	EF26_M_69	Réseau santé - Autres produits d'exploitation - maladie
268	EF26_M_7	Réseau santé - Charges et produits neutres - maladie
269	EF26_M_RES	Réseau santé - Résultat d'exploitation de l'assurance - maladie
270	EF26_M_R.GEN	Réseau santé - Résultat du compte d'exploitation général - maladie
271	EF27_A_3	Autres produits d'assurances - Total des charges d'assurance - accident
272	EF27_A_30.33	Autres produits d'assurances - Prestations payées - accident
273	EF27_A_30.35	Autres produits d'assurances - Prestation brutes - accident
274	EF27_A_31	Autres produits d'assurances - Prestations - accident
275	EF27_A_32	Autres produits d'assurances - Participation des assurés aux frais - accident
276	EF27_A_3.4	Autres produits d'assurances - Charges d'assurance et d'exploitation - accident
277	EF27_A_34	Autres produits d'assurances - Autres charges d'assurance - accident
278	EF27_A_35	Autres produits d'assurances - Provisions pour cas d'assurance non liquidés - accident
279	EF27_A_36	Autres produits d'assurances - Part des prestations remboursées par les réassureurs - accident
280	EF27_A_37	Autres produits d'assurances - Compensation des risques - accident
281	EF27_A_4	Autres produits d'assurances - Total des charges d'exploitation - accident
282	EF27_A_40.47	Autres produits d'assurances - Charges d'administration - accident
283	EF27_A_48	Autres produits d'assurances - Amortissements - accident
284	EF27_A_49	Autres produits d'assurances - Autres charges d'exploitation - accident
285	EF27_A_6	Autres produits d'assurances - Produits d'assurance - accident
286	EF27_A_60.65	Autres produits d'assurances - Primes brutes - accident
287	EF27_A_60.66	Autres produits d'assurances - Primes propres - accident
288	EF27_A_61	Autres produits d'assurances - Primes selon coûts effectifs - accident
289	EF27_A_64	Autres produits d'assurances - Déductions accordées sur primes - accident
290	EF27_A_65	Autres produits d'assurances - Autres primes - accident
291	EF27_A_66	Autres produits d'assurances - Part de primes des réassureurs - accident
292	EF27_A_67	Autres produits d'assurances - Réduction des primes et autres contribution - accident
293	EF27_A_68	Autres produits d'assurances - Déduction des parts de primes des assurés - accident
294	EF27_A_69	Autres produits d'assurances - Autres produits d'exploitation - accident
295	EF27_A_7	Autres produits d'assurances - Charges et produits neutres - accident
296	EF27_A_RES	Autres produits d'assurances - Résultat d'exploitation de l'assurance - accident
297	EF27_A_R.GEN	Autres produits d'assurances - Résultat du compte d'exploitation général - accident
298	EF27_M_3	Autres produits d'assurances - Total des charges d'assurance - maladie
299	EF27_M_30.33	Autres produits d'assurances - Prestations payées - maladie
300	EF27_M_30.35	Autres produits d'assurances - Prestation brutes - maladie
301	EF27_M_31	Autres produits d'assurances - Prestations - maladie
302	EF27_M_32	Autres produits d'assurances - Participation des assurés aux frais - maladie
303	EF27_M_3.4	Autres produits d'assurances - Charges d'assurance et d'exploitation - maladie
304	EF27_M_34	Autres produits d'assurances - Autres charges d'assurance - maladie

Suite à la page suivante

Suite

	Variable	Sinification
305	EF27_M_35	Autres produits d'assurances - Provisions pour cas d'assurance non liquidés - maladie
306	EF27_M_36	Autres produits d'assurances - Part des prestations remboursées par les réassureurs - maladie
307	EF27_M_37	Autres produits d'assurances - Compensation des risques - maladie
308	EF27_M_4	Autres produits d'assurances - Total des charges d'exploitation - maladie
309	EF27_M_40.47	Autres produits d'assurances - Charges d'administration - maladie
310	EF27_M_48	Autres produits d'assurances - Amortissements - maladie
311	EF27_M_49	Autres produits d'assurances - Autres charges d'exploitation - maladie
312	EF27_M_6	Autres produits d'assurances - Produits d'assurance - maladie
313	EF27_M_60.65	Autres produits d'assurances - Primes brutes - maladie
314	EF27_M_60.66	Autres produits d'assurances - Primes propres - maladie
315	EF27_M_61	Autres produits d'assurances - Primes selon coûts effectifs - maladie
316	EF27_M_64	Autres produits d'assurances - Déductions accordées sur primes - maladie
317	EF27_M_65	Autres produits d'assurances - Autres primes - maladie
318	EF27_M_66	Autres produits d'assurances - Part de primes des réassureurs - maladie
319	EF27_M_67	Autres produits d'assurances - Réduction des primes et autres contribution - maladie
320	EF27_M_68	Autres produits d'assurances - Déduction des parts de primes des assurés - maladie
321	EF27_M_69	Autres produits d'assurances - Autres produits d'exploitation - maladie
322	EF27_M_7	Autres produits d'assurances - Charges et produits neutres - maladie
323	EF27_M_RES	Autres produits d'assurances - Résultat d'exploitation de l'assurance - maladie
324	EF27_M_R.GEN	Autres produits d'assurances - Résultat du compte d'exploitation général - maladie
325	EF31_T_0_5	enfants 0-5
326	EF31_T_6_10	enfants 6-10
327	EF31_T_11_15	enfants 11-15
328	EF31_T_16_18	enfants 16-18
329	EF31_T_19_20	adultes 19-20
330	EF31_T_21_25	adultes 21-25
331	EF31_T_26_30	adultes 26-30
332	EF31_T_31_35	adultes 31-35
333	EF31_T_36_40	adultes 36-40
334	EF31_T_41_45	adultes 41-45
335	EF31_T_46_50	adultes 46-50
336	EF31_T_51_55	adultes 51-55
337	EF31_T_56_60	adultes 56-60
338	EF31_T_61_65	adultes 61-65
339	EF31_T_66_70	adultes 66-70
340	EF31_T_71_75	adultes 71-75
341	EF31_T_76_80	adultes 76-80
342	EF31_T_81_85	adultes 81-85
343	EF31_T_86_90	adultes 86-90
344	EF31_T_91_95	adultes 91-95
345	EF31_T_96_100	adultes 96-100
346	EF31_T_100_	adultes plus de 100
347	EF31_T_A_MOY	agemoyen
348	EF31_T_A_TOT	age total
349	EF31_T_E_A_TOT	enfants et adultes total
350	EF31_T_E_TOT	enfants total
351	EF31_T_INCO	age inconnu
352	EF31_T_JA_TOT	jeunes adultes total
353	EF32_T_AG	Canton de domicil de l'assuré - Agrovie
354	EF32_T_AI	Canton de domicil de l'assuré - Appenzell Rhodes-Intérieures
355	EF32_T_AR	Canton de domicil de l'assuré - Appenzell Rhodes-Extérieures
356	EF32_T_BE	Canton de domicil de l'assuré - Berne
357	EF32_T_BL	Canton de domicil de l'assuré - Bâle-Ville

Suite à la page suivante

Suite

	Variable	Sinification
358	EF32_T_BS	Canton de domicil de l'assuré - Bâle-Campagne
359	EF32_T_ETR	Canton de domicil de l'assuré - étranger
360	EF32_T_FR	Canton de domicil de l'assuré - Fribourg
361	EF32_T_GE	Canton de domicil de l'assuré - Genève
362	EF32_T_GL	Canton de domicil de l'assuré - Glaris
363	EF32_T_GR	Canton de domicil de l'assuré - Grisons
364	EF32_T_INC	Canton de domicil de l'assuré - inconnu
365	EF32_T_JU	Canton de domicil de l'assuré - Jura
366	EF32_T_LU	Canton de domicil de l'assuré - Lucerne
367	EF32_T_NE	Canton de domicil de l'assuré - Neuchâtel
368	EF32_T_NW	Canton de domicil de l'assuré - Nidwald
369	EF32_T_OW	Canton de domicil de l'assuré - Obwald
370	EF32_T_SG	Canton de domicil de l'assuré - St.Gall
371	EF32_T_SH	Canton de domicil de l'assuré - Shaffhouse
372	EF32_T_SO	Canton de domicil de l'assuré - Soleure
373	EF32_T_SZ	Canton de domicil de l'assuré - Schwytz
374	EF32_T_TG	Canton de domicil de l'assuré - Thurgovie
375	EF32_T_TI	Canton de domicil de l'assuré - Tessin
376	EF32_T_TOT	Canton de domicil de l'assuré - Total cantons
377	EF32_T_UR	Canton de domicil de l'assuré - Uri
378	EF32_T_VD	Canton de domicil de l'assuré - Vaud
379	EF32_T_VS	Canton de domicil de l'assuré - Vallais
380	EF32_T_ZG	Canton de domicil de l'assuré - Zoug
381	EF32_T_ZH	Canton de domicil de l'assuré - Zurich
382	EF33_EF_ARP_T	Effectif des assurés selon le modèle d'assurance - dont assurés au bénéfice d'une réduction de prime
383	EF33_EF_BO_T	Effectif des assurés selon le modèle d'assurance - Assurance avec bonus
384	EF33_EF_CL_T	Effectif des assurés selon le modèle d'assurance - Assurance avec choix limité des fournisseurs de prestations
385	EF33_EF_FO_T	Effectif des assurés selon le modèle d'assurance - Assurance avec franchise à option
386	EF33_EF_O_T	Effectif des assurés selon le modèle d'assurance - Assurance ordinaire
387	EF33_EF_T_T	Effectif des assurés selon le modèle d'assurance - Total
388	EF34_PRI_BO_T	Produits des primes selon le modèle d'assurance - Assurance avecc bonus
389	EF34_PRI_CL_T	Produits des primes selon le modèle d'assurance - Assurance avec choix limité des fournisseurs de prestations
390	EF34_PRI_FO_T	Produits des primes selon le modèle d'assurance - Assurance avec franchise à option
391	EF34_PRI_O_T	Produits des primes selon le modèle d'assurance - Assurance ordinaire
392	EF34_PRI_T_T	Produits des primes selon le modèle d'assurance - Total
393	EF35_PRE_BO_T	Prestations selon le modèle d'assurance - Assurance avecc bonus
394	EF35_PRE_CL_T	Prestations selon le modèle d'assurance - Assurance avec choix limité des fournisseurs de prestations
395	EF35_PRE_FO_T	Prestations selon le modèle d'assurance - Assurance avec franchise à option
396	EF35_PRE_O_T	Prestations selon le modèle d'assurance - Assurance ordinaire
397	EF35_PRE_T_T	Prestations selon le modèle d'assurance - Total
398	EF36_T_A_AMB	Prestation par groupes de frais - autres prestations ambulatoires
399	EF36_T_A_STA	Prestation par groupes de frais - autres prestations stationaire
400	EF36_T_CHIRO	Prestation par groupes de frais - Chiropracticiens
401	EF36_T_EMS	Prestation par groupes de frais - Etablissemnets médicosociaux
402	EF36_T_HMO	Prestation par groupes de frais - Contributions aux HMO's
403	EF36_T_HO_AMB	Prestation par groupes de frais - Hôpital ambulatoire
404	EF36_T_HO_STA	Prestation par groupes de frais - Hôpital stationnaire
405	EF36_T_LABO	Prestation par groupes de frais - Laboratoires
406	EF36_T_MED_AMB	Prestation par groupes de frais - Médecin ambulatoire
407	EF36_T_MEDIC_HO_AMB	Prestation par groupes de frais - Médicaments hôpital ambulatoire
408	EF36_T_MEDIC_MED	Prestation par groupes de frais - Médicaments médecin
409	EF36_T_MEDIC_HA	Prestation par groupes de frais - Médicaments pharmacie
410	EF36_T_MOY_A	Prestation par groupes de frais - Moyens et appareilles

Suite à la page suivante

Suite

	Variable	Sinification
411	EF36_T_HYSIO	Prestation par groupes de frais - Physiothérapeutes
412	EF36_T_SITEX	Prestation par groupes de frais - Spitex
413	EF36_T_TOT	Prestation par groupes de frais - Total des prestations
414	EF371_T_MAL	Nombre de malades
415	EF372_T_JO_HOSP	Nombre de séjours hospitaliers
416	EF372_T_JO_MAT	dont séjours pour maternité
417	EF372_T_SEJ_HOP	Nombre de jours d'hospitalisation
418	EF372_T_SEJ_MAT	dont jours pour maternité
419	EF383_T_A	Admission d'assureés - adultes
420	EF383_T_E	Admission d'assureés - enfants
421	EF383_T_JA	Admission d'assureés - jeunes adultes
422	EF383_T_NN	Admission d'assureés - dont nouveaux nés
423	EF383_T_TOT	Admission d'assureés - total
424	EF384_T_A	Démission d'assurés - adulte
425	EF384_T_DEC	Démission d'assurés - dont décès
426	EF384_T_E	Démission d'assurés - enfants
427	EF384_T_JA	Démission d'assurés - jeunes adultes
428	EF384_T_TOT	Démission d'assurés - total
429	EF310_X_CH	Rayon d'activité - Suisse
430	EF310_X_REG	Rayon d'activité - régional
431	ID_INS	Nummero d'identité d'une assurance
432	YEAR	Année de référence
433	rec.tot1	Recettes totales
434	taille	Variable catégorielle selon le niveau de primes

Source : Office fédéral de la santé publique (OFSP), 2007

Scipts R

```
#-------------------------------------------------------------------------------
#
#                -/                                                    -/-/
#              -/     -/-/        -/-/-/     -/-/-/  -/     -/     -/
#            -/   -/-/-/-/   -/-/        -/     -/  -/     -/   -/-/-/-/
#          -/   -/                  -/-/   -/     -/   -/     -/     -/
#        -/     -/-/-/   -/-/-/        -/-/-/     -/-/-/     -/
#
#
#                    Laboratoire d'économétrie et de statistique
#                          de l'Université de Fribourg
#
#  AUTEUR:
#
#     Marie-Justine Leis
#     Département d'économie quantitative
#     Université de Fribourg (Suisse)
#     Bd de Pérolles 90
#     1700 Fribourg
#
#  COURRIEL:
#
#     Marie-Justine.Leis@UniFr.ch
#
#  PROJET:
#
#     Travail de Master
#
#  DATE:
#
#     29 Juin 2007
#
#  PROGRAMME:
#
#     Merge02.r
#
#  BUT:
#
#     Fusioner les données de differents ficher csv
#
#  DONNEES:
#
#     Type            : Panel
#
#     N. obs.         : 1221
#
#     Fichiers        : Tableau_22A.csv, Tableau_22M.csv,
#                       Tableau_23A.csv, Tableau_23M.csv, Tableau_24A.csv,
#                       Tableau_24M.csv, Tableau_25A.csv, Tableau_25M.csv,
#                       Tableau_26A.csv, Tableau_26M.csv, Tableau_27A.csv,
#                       Tableau_27M.csv, Tableau_31.csv, Tableau_32.csv,
#                       Tableau_34.csv, Tableau_35.csv, Tableau_36.csv,
#                       Tableau_112A.csv, Tableau_371-372.csv,
#                       Tableau_383-384.csv,Tableau_16_450.csv
#
#     Source          : Office fédéral de la santé publique, données
#                       d'observation des assurances maladiés du formulaire
#                       EF123.
#
#     Description     : Observations des assureurs depuis 1996
#
#
#
#-------------------------------------------------------------------------------

#
# Initialisation
#---------------

# Chemins d'accès aux dossiers de données
ddpath <- "/Users/marie-justineleis/Masterarbeit/Master R/données/"


# Chemins d'accès aux dossiers de travail
wdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/work/"


# fixe le dossier de travail et de données
setwd(wdpath)



#===============================================================================
#
```

```
#                                   MERGE
#
#-------------------------------------------------------------------------------


# 1. Données de base
#-------------------

# Lecture des données de base
Tableau_22A <-read.csv2(file(paste(ddpath,"Tableau_22A.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_22M <-read.csv2(file(paste(ddpath,"Tableau_22M.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_23A <-read.csv2(file(paste(ddpath,"Tableau_23A.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_23M <-read.csv2(file(paste(ddpath,"Tableau_23M.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_24A <-read.csv2(file(paste(ddpath,"Tableau_24A.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_24M <-read.csv2(file(paste(ddpath,"Tableau_24M.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_25A <-read.csv2(file(paste(ddpath,"Tableau_25A.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_25M <-read.csv2(file(paste(ddpath,"Tableau_25M.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_26A <-read.csv2(file(paste(ddpath,"Tableau_26A.csv",sep=""),encoding="latin1"),
                  header=T,sep=";", dec=".")
Tableau_26M <-read.csv2(file(paste(ddpath,"Tableau_26M.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_27A <-read.csv2(file(paste(ddpath,"Tableau_27A.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_27M <-read.csv2(file(paste(ddpath,"Tableau_27M.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_31 <-read.csv2(file(paste(ddpath,"Tableau_31.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_32 <-read.csv2(file(paste(ddpath,"Tableau_32.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_33 <-read.csv2(file(paste(ddpath,"Tableau_33.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_34 <-read.csv2(file(paste(ddpath,"Tableau_34.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_35 <-read.csv2(file(paste(ddpath,"Tableau_35.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_36 <-read.csv2(file(paste(ddpath,"Tableau_36.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_112A <-read.csv2(file(paste(ddpath,"Tableau_112A.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_371 <-read.csv2(file(paste(ddpath,"Tableau_371.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_383 <-read.csv2(file(paste(ddpath,"Tableau_383.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")
Tableau_16_450 <-read.csv2(file(paste(ddpath,"Tableau_16_450.csv",sep=""),encoding="latin1")
    ,
                    header=T,sep=";", dec=".")

Tableau_3102 <-read.csv2(file(paste(ddpath,"Tableau_3102.csv",sep=""),encoding="latin1"),
                    header=T,sep=";", dec=".")

# 2. Fusion des ficher
#---------------------

ofsp1 <- merge (Tableau_22A, Tableau_22M, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp2 <- merge (ofsp1, Tableau_23A, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp3 <- merge (ofsp2, Tableau_23M, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp4 <- merge (ofsp3, Tableau_24A, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp5 <- merge (ofsp4, Tableau_24M, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp6 <- merge (ofsp5, Tableau_25A, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp7 <- merge (ofsp6, Tableau_25M, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp8 <- merge (ofsp7, Tableau_26A, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp9 <- merge (ofsp8, Tableau_26M, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp10 <- merge (ofsp9, Tableau_27A, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp11 <- merge (ofsp10, Tableau_27M, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp12 <- merge (ofsp11, Tableau_31, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp13 <- merge (ofsp12, Tableau_32, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp14 <- merge (ofsp13, Tableau_33, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp15 <- merge (ofsp14, Tableau_34, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp16 <- merge (ofsp15, Tableau_35, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp17 <- merge (ofsp16, Tableau_36, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp18 <- merge (ofsp17, Tableau_112A, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp19 <- merge (ofsp18, Tableau_371, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp20 <- merge (ofsp19, Tableau_383, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp21 <- merge (ofsp20, Tableau_16_450, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
ofsp <- merge (ofsp21, Tableau_3102, by = c("ID_INS", "NAME_INS", "YEAR"), all = T)
```

```
# 3. Sauvegarde du ficher de données en format Binaire
#-----------------------------------------------------

#Fichier complet:
save (list=("ofsp"), file=("ofsp.RData"))


# 4. Chargement du ficher Binaire en R
#-------------------------------------

load(file="ofsp.RData")
head(ofsp)
```

Script B.1 – Fusion des fichiers par la commande “merge”

```
#-------------------------------------------------------------------------------
#
#                -/                                                      -/-/
#              -/    -/-/      -/-/-/    -/-/-/  -/    -/    -/
#             -/  -/-/-/-/  -/-/       -/    -/  -/    -/  -/-/-/-/
#           -/  -/                -/-/  -/    -/  -/    -/    -/
#          -/    -/-/-/  -/-/-/      -/-/-/    -/-/-/    -/
#
#
#                   Laboratoire d'économétrie et de statistique
#                         de l'Université de Fribourg
#
#  AUTEUR:
#
#     Marie-Justine Leis
#     Département d'économie quantitative
#     Université de Fribourg (Suisse)
#     Bd de Pérolles 90
#     1700 Fribourg
#
#  COURRIEL:
#
#     Marie-Justine.Leis@UniFr.ch
#
#  PROJET:
#
#     Travail de Master
#
#  DATE:
#
#     Novembre 2007 update Janvier 2008
#
#  PROGRAMME:
#
#     job04.r
#
#  BUT:
#
#     Estimation par MCO - Imputation ABB - Prévision
#
#  DONNEES:
#
#     Type            : Panel
#
#     N. obs.         : 1221
#
#     Fichiers        : "ofsp.Rdata"
#     Titre           :  ofsp
#
#     Source          : Office fédéral de la santé publique, données
#                       d'observation des assurances maladiés du formulaire
#                       EF123.
#
#     Description     : Observations des assureurs depuis 1996 à 2006
#
#
#-------------------------------------------------------------------------------

# Initialisation
#---------------

# Chemins d'accès aux dossiers de données
  ddpath <- "/Users/marie-justineleis/Masterarbeit/Master R/données/"

# Chemins d'accès aux dossiers de travail
  wdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/work/"

# Chemins d'accès aux jobs
  jdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/Jobs J-edit/librairies/"

# fixe le dossier de travail et de données
  setwd(wdpath)

# Librairies utilisées
#---------------------

library(MASS)       # vcov.lm()
library(Hmisc)      # procédures: contents, label
library(plm)        # pdata.frame
library(lmtest)     # wdtest

# Mes librairies

source(file(paste(jdpath, "lib01.r", sep="")))
```

```
#=============================================================================
#
#                              Estimation par MCO
#
#-----------------------------------------------------------------------------

# 1. Lecture des données de base
#-------------------------------

load(file="ofsp.RData")

# 2. Création du data frame de travail (WDF)
#-------------------------------------------

# Variables retenues

vname <- c("ID_INS","YEAR","NAME_INS", "EF33_EF_FO_T", "EF33_EF_O_T",
           "EF16_X_450", "EF310_X_CH",
           "EF22_A_6", "EF23_A_6", "EF22_A_7", "EF23_A_7",
           "EF22_M_6", "EF23_M_6", "EF22_M_7", "EF23_M_7")

WDF <- ofsp[vname]
rm(ofsp)

# 3. Création de nouvelles variables
#-----------------------------------

# Création d'une nouvelle variable (recettes totales)

proda.A <- WDF$EF22_A_6 + WDF$EF23_A_6
proda.M <- WDF$EF22_M_6 + WDF$EF23_M_6
prodn.A <- WDF$EF22_A_7 + WDF$EF23_A_7
prodn.M <- WDF$EF22_M_7 + WDF$EF23_M_7

rec.tot1 <- proda.A + proda.M - prodn.A - prodn.M

# Création d'un facteur "taille"

taille <- cut(rec.tot1,
              c(min(rec.tot1, na.rm=T), 4470000, 197000000, max(rec.tot1, na.rm=T)),
              include.lowest=T)

label(taille) <- "Taille de l'assurance selon les primes"

# Création d'un facteur régional vs suisse

WDF$EF310_X_CH <- as.factor(WDF$EF310_X_CH);
levels(WDF$EF310_X_CH) <- c("Régional", "Suisse")

# Mise à jour de la base de données

WDF <- data.frame(WDF, rec.tot1, taille)

# 4. Analyse de régression
#-------------------------

# Traitement des valeurs 0 pour les variables
# rec.tot1, EF33_EF_FO_T, EF33_EF_O_T, EF16_X_450

obs <- rep(1, nrow(WDF))
obs[WDF$rec.tot1 <= 0] <- 0
obs[WDF$EF33_EF_FO_T <= 0] <- 0
obs[WDF$EF33_EF_O_T <= 0] <- 0
obs[WDF$EF16_X_450 <= 0] <- 0

# 4.1 Estimation à partir d'un modèle:
#     données de base: 1996-2005 + un échantillon de 2006
#-----------------------------------------------------------

# Prévision sur la base d'un échantillon aléatoire d'observations de 2006
# Un modèle est estimé sur la base des données de 1996-2005 et
# d'un échantillon de données de 2006
# Les données estimées pour 2006 sont générées par imputation (multiple)
# Méthode d'imputation ABB
# Taille de l'échantillon: nsam

# Déclaration des variables du loop:
fitimp <- numeric()
tot2006 <-numeric()
pct <- numeric()

for (loop in 1:100){
```

```
# Paramètres à initialiser

nsam ← 30 # Nbr. obs. de l'échantillon
nH ← 5    # Nbr. strates
nimp ← 10  # Nbr. d'imputations


# Stratification optimale selon rec.tot1 2005
# (cf. Gunning & Horgan (2004))

rec.tot1.2005 ←  with(subset(WDF, c(YEAR == 2005)), rec.tot1)
rec.tot1.2005[rec.tot1.2005 <= 0] ← NA

strates ← stratopt(rec.tot1.2005, nH)

# Nos d'assurances 2005 avec nos des strates

ID_INS.2005 ←  with(subset(WDF, c(YEAR == 2005)), ID_INS)

INS.2005 ← data.frame(ID_INS.2005, strates)

# Nos d'assurances 2006 avec nos des strates

INS.2006 ← data.frame(with(subset(WDF, c((YEAR == 2006)&(obs==1))), ID_INS))
names(INS.2006) ← c("ID_INS.2006")

INS.2006 ← merge(INS.2006, INS.2005, by.x = c("ID_INS.2006"),
                 by.y = c("ID_INS.2005"), all.x=T)

# Allocation optimale de l'échantillon

(alloc ← allocopt(rec.tot1.2005, strates, nsam))

# Ajout de deux unités dans les strates vides ou égales à 1

alloc[alloc <= 1] ← 2

# Tirage de l'échantillon

ID_INS.sam ← vector("numeric")

for (i in 1:nH){
    pop.i ← INS.2006$ID_INS.2006[INS.2006$strates == i]
    pop.i ← pop.i[!is.na(pop.i)]
    n.i ← alloc[i]
    sam.i ← sample(pop.i, n.i)
    ID_INS.sam ← c(ID_INS.sam, sam.i)
}

# Indices des observations sélectionnées

obs.new ← obs
obs.new[(WDF$YEAR == 2006)&(!is.element(WDF$ID_INS, unique(ID_INS.sam)))] ← 0

# 4.2 Analyse de régression
# ------------------------

# Estimation du modèle

yfit ← lm(log(rec.tot1) ~ log(EF33_EF_FO_T) + log(EF33_EF_O_T) +
                          log(EF16_X_450) + EF310_X_CH + taille + YEAR,
          data=WDF,
          subset=c(obs.new == 1),
          na.action=na.exclude)
summary(yfit)

# Valeurs estimées des observations correspondant à
# l'échantillon aléatoire d'observations de 2006.
# En utilisant le modèle estimé.

p.yfit ← predict(yfit, WDF[((WDF$YEAR==2006)&(obs.new==1)),],
                 interval = c("prediction"))

# 4.3 Prévision sur la base d'un modèle estimé
# --------------------------------------------

# Recettes totales 2006 estimées

rec.tot1.2006 ← exp(p.yfit[,1])

fit ← data.frame(WDF$ID_INS[((WDF$YEAR==2006)&(obs.new==1))], rec.tot1.2006)
names(fit) ← c("ID_INS.2006", "rec.tot1")
```

```
# Constitution pour 2006 d'une base de données comprenant
# les nos d'assurances, de strates et les valeurs prédites de l'échantillon

datafit <- merge(INS.2006, fit, by.x = c("ID_INS.2006"),
                 by.y = c("ID_INS.2006"), all.x = T)

# Imputation multiple par ABB

imp.fit <- multimpute(datafit$rec.tot1, "fit", datafit$strates, nimp)


datafit <- data.frame(datafit[,-3], imp.fit[-1])

# Ajout de la série observée pour 2006

data.2006 <- with(subset(WDF, c(YEAR == 2006)),
                  data.frame(ID_INS, rec.tot1))

datafit <- merge(datafit, data.2006, by.x = c("ID_INS.2006"),
                 by.y = c("ID_INS"), all.x =T)

# Remplacement des valeurs estimées de l'échantillon
# par les véritables valeurs observées

ind <- is.element(datafit$ID_INS.2006,unique(ID_INS.sam))
datafit[ind,(3:(nimp+2))] <- datafit[ind, (nimp+3)]

# Estimation du total

ind <- (datafit$rec.tot1 > 0) & (!is.na(datafit$strates))

fitimp[loop] <- mean(colSums(datafit[ind, 3:(nimp+2)], na.rm=T))
tot2006[loop] <- sum(datafit[ind, (nimp+3)], na.rm=T)
pct[loop] <- (tot2006[loop]-fitimp[loop])/tot2006[loop]

# fin de la boucle
}

# Résumé des résultats

Result <- numeric()
Result[1] <- mean(fitimp)
Result[2] <- mean(tot2006)
Result[3] <- mean(pct)
Result[4] <- sd(fitimp)
Result
```

Script B.2 – Estimation par MCO - Imputation ABB - Prévision

```
#-------------------------------------------------------------------------------
#
#                -/                                                      -/-/
#              -/    -/-/      -/-/-/    -/-/-/  -/    -/    -/
#             -/  -/-/-/-/  -/-/       -/    -/  -/    -/  -/-/-/-/
#           -/  -/                -/-/  -/    -/  -/    -/    -/
#          -/    -/-/-/  -/-/-/      -/-/-/    -/-/-/    -/
#
#
#                   Laboratoire d'économétrie et de statistique
#                         de l'Université de Fribourg
#
#  AUTEUR:
#
#     Marie-Justine Leis
#     Département d'économie quantitative
#     Université de Fribourg (Suisse)
#     Bd de Pérolles 90
#     1700 Fribourg
#
#  COURRIEL:
#
#     Marie-Justine.Leis@UniFr.ch
#
#  PROJET:
#
#     Travail de Master
#
#  DATE:
#
#     Novembre 2007 update Janvier 2008
#
#  PROGRAMME:
#
#     job03.r
#
#  BUT:
#
#     Estimation Within - Imputation ABB - Prévision
#
#  DONNEES:
#
#     Type            : Panel
#
#     N. obs.         : 1221
#
#     Fichiers        : "ofsp.Rdata"
#     Titre           :  ofsp
#
#     Source          : Office fédéral de la santé publique, données
#                       d'observation des assurances maladiés du formulaire
#                       EF123.
#
#     Description     : Observations des assureurs depuis 1996 à 2006
#
#
#-------------------------------------------------------------------------------

# Initialisation
#---------------

# Chemins d'accès aux dossiers de données
  ddpath <- "/Users/marie-justineleis/Masterarbeit/Master R/données/"

# Chemins d'accès aux dossiers de travail
  wdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/work/"

# Chemins d'accès aux jobs
  jdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/Jobs J-edit/librairies/"

# fixe le dossier de travail et de données
  setwd(wdpath)

# Librairies utilisées
#---------------------

library(MASS)       # vcov.lm()
library(Hmisc)      # procédures: contents, label
library(plm)        # pdata.frame
library(lmtest)     # wdtest

# Mes librairies

source(file(paste(jdpath, "lib01.r", sep="")))
```

```
#=============================================================================
#
#                              Estimation Within
#
#-----------------------------------------------------------------------------

# 1. Lecture des données de base
#-------------------------------

load(file="ofsp.RData")

# 2. Création du data frame de travail (WDF)
#-------------------------------------------

# Variables retenues

vname <- c("ID_INS","YEAR","NAME_INS", "EF33_EF_FO_T", "EF33_EF_O_T",
           "EF16_X_450", "EF310_X_CH",
           "EF22_A_6", "EF23_A_6", "EF22_A_7", "EF23_A_7",
           "EF22_M_6", "EF23_M_6", "EF22_M_7", "EF23_M_7")

WDF <- ofsp[vname]
rm(ofsp)

# 3. Création de nouvelles variables
#-----------------------------------

# Création d'une nouvelle variable (recettes totales)

proda.A <- WDF$EF22_A_6 + WDF$EF23_A_6
proda.M <- WDF$EF22_M_6 + WDF$EF23_M_6
prodn.A <- WDF$EF22_A_7 + WDF$EF23_A_7
prodn.M <- WDF$EF22_M_7 + WDF$EF23_M_7

rec.tot1 <- proda.A + proda.M - prodn.A - prodn.M

# Création d'un facteur "taille"

taille <- cut(rec.tot1,
              c(min(rec.tot1, na.rm=T), 4470000, 197000000, max(rec.tot1, na.rm=T)),
              include.lowest=T)

label(taille) <- "Taille de l'assurance selon les primes"

# Création d'un facteur régional vs suisse

WDF$EF310_X_CH <- as.factor(WDF$EF310_X_CH);
levels(WDF$EF310_X_CH) <- c("Régional", "Suisse")

# Mise à jour de la base de données

WDF <- data.frame(WDF, rec.tot1, taille)

# Nettoyage de l'espace de travail

rm(proda.A, proda.M, prodn.A, prodn.M, rec.tot1, taille)

# 4. Estimation à partir d'un modèle
#-----------------------------------

# Prévision sur la base d'un échantillon aléatoire de valeur espérée de 2006
# Un modèle est estimé sur la base des données de 1996-2005 et
# d'un échantillon de données de 2006
# Les données estimées pour 2006 sont générées par imputation (multiple)
# Méthode d'imputation ABB

# Déclaration des variables du loop:
fitimp <- numeric()
tot2006 <-numeric()
pct <- numeric()

for (loop in 1:100){



# 4.1 Données de base: 1996-2005 + un échantillon de 2006
#--------------------------------------------------------

# Paramètres à initialiser

nsam <- 15 # Nbr. obs. de l'échantillon
nH <- 5    # Nbr. strates
nimp <- 5  # Nbr. d'imputations
```

```
# Traitement des valeurs 0 pour les variables
# rec.tot1, EF33_EF_FO_T, EF33_EF_O_T, EF16_X_450

obs ← rep(1, nrow(WDF))
obs[WDF$rec.tot1 <= 0] ← 0
obs[WDF$EF33_EF_FO_T <= 0] ← 0
obs[WDF$EF33_EF_O_T <= 0] ← 0
obs[WDF$EF16_X_450 <= 0] ← 0

# Stratification optimale selon rec.tot1 2005
# (cf. Gunning & Horgan (2004))

rec.tot1.2005 ← with(subset(WDF, c(YEAR == 2005)), rec.tot1)
rec.tot1.2005[rec.tot1.2005 <= 0] ← NA

strates ← stratopt(rec.tot1.2005, nH)

# Nos d'assurances 2005 avec nos des strates

ID_INS.2005 ← with(subset(WDF, c(YEAR == 2005)), ID_INS)

INS.2005 ← data.frame(ID_INS.2005, strates)

# Nos d'assurances 2006 avec nos des strates

INS.2006 ← data.frame(with(subset(WDF, c((YEAR == 2006)&(obs==1))), ID_INS))
names(INS.2006) ← c("ID_INS.2006")

INS.2006 ← merge(INS.2006, INS.2005, by.x = c("ID_INS.2006"),
                 by.y = c("ID_INS.2005"), all.x=T)

# Allocation optimale de l'échantillon

(alloc ← allocopt(rec.tot1.2005, strates, nsam))

# Ajout de deux unités dans les strates vides ou égales à 1

alloc[alloc <= 1] ← 2

# Tirage de l'échantillon

ID_INS.sam ← vector("numeric")

for (i in 1:nH){
    pop.i ← INS.2006$ID_INS.2006[INS.2006$strates == i]
    pop.i ← pop.i[!is.na(pop.i)]
    n.i ← alloc[i]
    sam.i ← sample(pop.i, n.i)
    ID_INS.sam ← c(ID_INS.sam, sam.i)
}

# 4.2 Analyse de régression
# ————————————————————

# Mise à jour des indices des observations sélectionnées

obs.new ← obs
obs.new[(WDF$YEAR == 2006)&(!is.element(WDF$ID_INS, unique(ID_INS.sam)))] ← 0

obs.compcase ← complete.cases(WDF$rec.tot1,
                              WDF$EF33_EF_FO_T,
                              WDF$EF33_EF_O_T,
                              WDF$EF16_X_450,
                              WDF$EF310_X_CH,
                              WDF$taille)

obs.new[!obs.compcase] ← 0

# Données de l'analyse de régression

WDF.subsam ← subset(WDF, obs.new==1)

# Variables retenues et formule

vname ← c("ID_INS","YEAR", "EF33_EF_FO_T", "EF33_EF_O_T",
          "EF16_X_450", "EF310_X_CH", "rec.tot1", "taille")

form ← log(rec.tot1) ~ log(EF33_EF_FO_T) + log(EF33_EF_O_T) +
                       EF16_X_450 + EF310_X_CH + taille

# Mise à jour de la base de données

WDF.subsam ← WDF.subsam[vname]
```

```
rownames(WDF.subsam) ← 1:nrow(WDF.subsam)

# Conversion en base de données de panel

pWDF ← plm.data(WDF.subsam, index=c("ID_INS", "YEAR"))
pdim(pWDF)
pvar(pWDF)

# Estimation du modèle par PLM

yfit.plm ← plm(form, data=pWDF, model="within", effect = "twoways")
summary(yfit.plm)

# 4.3 Prévision sur la base du modèle estimé
#-----------------------------------------

# Recettes totales 2006 estimées (espérée)

rec.tot1.2006 ← exp(yfit.plm$fitted.values)[pWDF$YEAR==2006]

fit ← data.frame(pWDF$ID_INS[pWDF$YEAR==2006], rec.tot1.2006)
names(fit) ← c("ID_INS.2006", "rec.tot1")

# Constitution pour 2006 d'une base de données comprenant
# les nos d'assurances, de strates et les valeurs prédites de l'échantillon

datafit ← merge(INS.2006, fit, by.x = c("ID_INS.2006"),
                by.y = c("ID_INS.2006"), all.x = T)

# Imputation multiple par ABB

imp.fit ← multimpute(datafit$rec.tot1, "fit", datafit$strates, nimp)

datafit ← data.frame(datafit[,-3], imp.fit[-1])

# Ajout de la série observée pour 2006

data.2006 ← with(subset(WDF, c(YEAR == 2006)),
                 data.frame(ID_INS, rec.tot1))

datafit ← merge(datafit, data.2006, by.x = c("ID_INS.2006"),
                by.y = c("ID_INS"), all.x =T)

# Remplacement des valeurs estimées de l'échantillon
# par les véritables valeurs observées

ind ← is.element(datafit$ID_INS.2006,unique(ID_INS.sam))
datafit[ind,(3:(nimp+2))] ← datafit[ind, (nimp+3)]

# Estimation du total

ind ← (datafit$rec.tot1 > 0) & (!is.na(datafit$strates))

fitimp[loop] ← mean(colSums(datafit[ind, 3:(nimp+2)], na.rm=T))
tot2006[loop] ← sum(datafit[ind, (nimp+3)], na.rm=T)
pct[loop] ← (tot2006[loop]-fitimp[loop])/tot2006[loop]

# fin de la boucle
}

# Résumé des résultats

Result ← numeric()
Result[1] ← mean(fitimp)
Result[2] ← mean(tot2006)
Result[3] ← mean(pct)
Result[4] ← sd(fitimp)
Result
```

Script B.3 – Estimation within - Imputation ABB - Prévision

```
#-------------------------------------------------------------------------------
#
#                -/                                                      -/-/
#              -/    -/-/      -/-/-/    -/-/-/  -/    -/    -/
#             -/  -/-/-/-/  -/-/       -/    -/  -/    -/  -/-/-/-/
#           -/  -/                -/-/  -/    -/  -/    -/    -/
#          -/    -/-/-/  -/-/-/      -/-/-/    -/-/-/    -/
#
#
#                    Laboratoire d'économétrie et de statistique
#                          de l'Université de Fribourg
#
#  AUTEUR:
#
#     Marie-Justine Leis
#     Département d'économie quantitative
#     Université de Fribourg (Suisse)
#     Bd de Pérolles 90
#     1700 Fribourg
#
#  COURRIEL:
#
#     Marie-Justine.Leis@UniFr.ch
#
#  PROJET:
#
#     Travail de Master
#
#  DATE:
#
#     Novembre 2007 update Janvier 2008
#
#  PROGRAMME:
#
#     job02.r
#
#  BUT:
#
#     Estimation Random - Imputation ABB - Prévision
#
#  DONNEES:
#
#     Type            : Panel
#
#     N. obs.         : 1221
#
#     Fichiers        : "ofsp.Rdata"
#     Titre           :  ofsp
#
#     Source          : Office fédéral de la santé publique, données
#                       d'observation des assurances maladiés du formulaire
#                       EF123.
#
#     Description     : Observations des assureurs depuis 1996 à 2006
#
#
#-------------------------------------------------------------------------------

# Initialisation
#---------------

# Chemins d'accès aux dossiers de données
  ddpath <- "/Users/marie-justineleis/Masterarbeit/Master R/données/"

# Chemins d'accès aux dossiers de travail
  wdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/work/"

# Chemins d'accès aux jobs
  jdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/Jobs J-edit/librairies/"

# fixe le dossier de travail et de données
  setwd(wdpath)

# Librairies utilisées
#---------------------

library(MASS)       # vcov.lm()
library(Hmisc)      # procédures: contents, label
library(plm)        # pdata.frame
library(lmtest)     # wdtest

# Mes librairies

source(file(paste(jdpath, "lib01.r", sep="")))
```

```
#=============================================================================
#
#                              Estimation Random
#
#-----------------------------------------------------------------------------

# 1. Lecture des données de base
#-------------------------------

load(file="ofsp.RData")

# 2. Création du data frame de travail (WDF)
#-------------------------------------------

# Variables retenues

vname <- c("ID_INS","YEAR","NAME_INS", "EF33_EF_FO_T", "EF33_EF_O_T",
           "EF16_X_450", "EF310_X_CH",
           "EF22_A_6", "EF23_A_6", "EF22_A_7", "EF23_A_7",
           "EF22_M_6", "EF23_M_6", "EF22_M_7", "EF23_M_7")

WDF <- ofsp[vname]
rm(ofsp)

# 3. Création de nouvelles variables
#-----------------------------------

# Création d'une nouvelle variable (recettes totales)

proda.A <- WDF$EF22_A_6 + WDF$EF23_A_6
proda.M <- WDF$EF22_M_6 + WDF$EF23_M_6
prodn.A <- WDF$EF22_A_7 + WDF$EF23_A_7
prodn.M <- WDF$EF22_M_7 + WDF$EF23_M_7

rec.tot1 <- proda.A + proda.M - prodn.A - prodn.M

# Création d'un facteur "taille"

taille <- cut(rec.tot1,
              c(min(rec.tot1, na.rm=T), 4470000, 197000000, max(rec.tot1, na.rm=T)),
              include.lowest=T)

label(taille) <- "Taille de l'assurance selon les primes"

# Création d'un facteur régional vs suisse

WDF$EF310_X_CH <- as.factor(WDF$EF310_X_CH);
levels(WDF$EF310_X_CH) <- c("Régional", "Suisse")

# Mise à jour de la base de données

WDF <- data.frame(WDF, rec.tot1, taille)

# Nettoyage de l'espace de travail

rm(proda.A, proda.M, prodn.A, prodn.M, rec.tot1, taille)

# 4. Estimation à partir d'un modèle
#-----------------------------------

# Prévision sur la base d'un échantillon aléatoire de valeur espérée de 2006
# Un modèle est estimé sur la base des données de 1996-2005 et
# d'un échantillon de données de 2006
# Les données estimées pour 2006 sont générées par imputation (multiple)
# Méthode d'imputation ABB

# Déclaration des variables du loop:
fitimp <- numeric()
tot2006 <-numeric()
pct <- numeric()

for (loop in 1:100){



# 4.1 Données de base: 1996-2005 + un échantillon de 2006
#--------------------------------------------------------

# Paramètres à initialiser

nsam <- 15 # Nbr. obs. de l'échantillon
nH <- 5    # Nbr. strates
nimp <- 10  # Nbr. d'imputations
```

```
# Traitement des valeurs 0 pour les variables
# rec.tot1, EF33_EF_FO_T, EF33_EF_O_T, EF16_X_450

obs ← rep(1, nrow(WDF))
obs[WDF$rec.tot1 <= 0] ← 0
obs[WDF$EF33_EF_FO_T <= 0] ← 0
obs[WDF$EF33_EF_O_T <= 0] ← 0
obs[WDF$EF16_X_450 <= 0] ← 0

# Stratification optimale selon rec.tot1 2005
# (cf. Gunning & Horgan (2004))

rec.tot1.2005 ← with(subset(WDF, c(YEAR == 2005)), rec.tot1)
rec.tot1.2005[rec.tot1.2005 <= 0] ← NA

strates ← stratopt(rec.tot1.2005, nH)

# Nos d'assurances 2005 avec nos des strates

ID_INS.2005 ← with(subset(WDF, c(YEAR == 2005)), ID_INS)

INS.2005 ← data.frame(ID_INS.2005, strates)

# Nos d'assurances 2006 avec nos des strates

INS.2006 ← data.frame(with(subset(WDF, c((YEAR == 2006)&(obs==1))), ID_INS))
names(INS.2006) ← c("ID_INS.2006")

INS.2006 ← merge(INS.2006, INS.2005, by.x = c("ID_INS.2006"),
                 by.y = c("ID_INS.2005"), all.x=T)

# Allocation optimale de l'échantillon

(alloc ← allocopt(rec.tot1.2005, strates, nsam))

# Ajout de deux unités dans les strates vides ou égales à 1

alloc[alloc <= 1] ← 2

# Tirage de l'échantillon

ID_INS.sam ← vector("numeric")

for (i in 1:nH){
    pop.i ← INS.2006$ID_INS.2006[INS.2006$strates == i]
    pop.i ← pop.i[!is.na(pop.i)]
    n.i ← alloc[i]
    sam.i ← sample(pop.i, n.i)
    ID_INS.sam ← c(ID_INS.sam, sam.i)
}

# 4.2 Analyse de régression
# ————————————————————

# Mise à jour des indices des observations sélectionnées

obs.new ← obs
obs.new[(WDF$YEAR == 2006)&(!is.element(WDF$ID_INS, unique(ID_INS.sam)))] ← 0

obs.compcase ← complete.cases(WDF$rec.tot1,
                              WDF$EF33_EF_FO_T,
                              WDF$EF33_EF_O_T,
                              WDF$EF16_X_450,
                              WDF$EF310_X_CH,
                              WDF$taille)

obs.new[!obs.compcase] ← 0

# Données de l'analyse de régression

WDF.subsam ← subset(WDF, obs.new==1)

# Variables retenues et formule

vname ← c("ID_INS","YEAR", "EF33_EF_FO_T", "EF33_EF_O_T",
          "EF16_X_450", "EF310_X_CH", "rec.tot1", "taille")

form ← log(rec.tot1) ~ log(EF33_EF_FO_T) + log(EF33_EF_O_T) +
                       EF16_X_450 + EF310_X_CH + taille

# Mise à jour de la base de données

WDF.subsam ← WDF.subsam[vname]
```

```
rownames(WDF.subsam) ← 1:nrow(WDF.subsam)

# Conversion en base de données de panel

pWDF ← plm.data(WDF.subsam, index=c("ID_INS", "YEAR"))
pdim(pWDF)
pvar(pWDF)

# Estimation du modèle par PLM

yfit.plm ← plm(form, data=pWDF, model="random")
summary(yfit.plm)

# 4.3 Prévision sur la base du modèle estimé
#-----------------------------------------

# Recettes totales 2006 estimées (espérée)

rec.tot1.2006 ← exp(yfit.plm$fitted.values)[pWDF$YEAR==2006]

fit ← data.frame(pWDF$ID_INS[pWDF$YEAR==2006], rec.tot1.2006)
names(fit) ← c("ID_INS.2006", "rec.tot1")

# Constitution pour 2006 d'une base de données comprenant
# les nos d'assurances, de strates et les valeurs prédites de l'échantillon

datafit ← merge(INS.2006, fit, by.x = c("ID_INS.2006"),
                by.y = c("ID_INS.2006"), all.x = T)

# Imputation multiple par ABB

imp.fit ← multimpute(datafit$rec.tot1, "fit", datafit$strates, nimp)

datafit ← data.frame(datafit[,-3], imp.fit[-1])

# Ajout de la série observée pour 2006

data.2006 ← with(subset(WDF, c(YEAR == 2006)),
                 data.frame(ID_INS, rec.tot1))

datafit ← merge(datafit, data.2006, by.x = c("ID_INS.2006"),
                by.y = c("ID_INS"), all.x =T)

# Remplacement des valeurs estimées de l'échantillon
# par les véritables valeurs observées

ind ← is.element(datafit$ID_INS.2006,unique(ID_INS.sam))
datafit[ind,(3:(nimp+2))] ← datafit[ind, (nimp+3)]

# Estimation du total

ind ← (datafit$rec.tot1 > 0) & (!is.na(datafit$strates))

fitimp[loop] ← mean(colSums(datafit[ind, 3:(nimp+2)], na.rm=T))
tot2006[loop] ← sum(datafit[ind, (nimp+3)], na.rm=T)
pct[loop] ← (tot2006[loop]-fitimp[loop])/tot2006[loop]

# fin de la boucle
}

# Résumé des résultats

Result ← numeric()
Result[1] ← mean(fitimp)
Result[2] ← mean(tot2006)
Result[3] ← mean(pct)
Result[4] ← sd(fitimp)
Result
```

Script B.4 – Estimation random - Imputation ABB - Prévision

```
#-------------------------------------------------------------------------------
#
#                -/                                                      -/-/
#              -/    -/-/       -/-/-/    -/-/-/  -/     -/     -/
#             -/  -/-/-/-/  -/-/        -/     -/  -/     -/  -/-/-/-/
#           -/  -/                 -/-/  -/     -/  -/     -/     -/
#          -/     -/-/-/  -/-/-/       -/-/-/     -/-/-/     -/
#
#
#                    Laboratoire d'économétrie et de statistique
#                          de l'Université de Fribourg
#
#  AUTEUR:
#
#     Marie-Justine Leis
#     Département d'économie quantitative
#     Université de Fribourg (Suisse)
#     Bd de Pérolles 90
#     1700 Fribourg
#
#  COURRIEL:
#
#     Marie-Justine.Leis@UniFr.ch
#
#  PROJET:
#
#     Travail de Master
#
#  DATE:
#
#     Novembre 2007 update Janvier 2008
#
#  PROGRAMME:
#
#     job06.r
#
#  BUT:
#
#     Estimation par MCO - Imputation sur la base d'un modèle linéaire -
#     Prévision
#
#  DONNEES:
#
#     Type            : Panel
#
#     N. obs.         : 1221
#
#     Fichiers        : "ofsp.Rdata"
#     Titre           :  ofsp
#
#     Source          : Office fédéral de la santé publique, données
#                       d'observation des assurances maladiés du formulaire
#                       EF123.
#
#     Description     : Observations des assureurs depuis 1996 à 2006
#
#
#-------------------------------------------------------------------------------

# Initialisation
#---------------

# Chemins d'accès aux dossiers de données
  ddpath <- "/Users/marie-justineleis/Masterarbeit/Master R/données/"

# Chemins d'accès aux dossiers de travail
  wdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/work/"

# Chemins d'accès aux jobs
  jdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/Jobs J-edit/librairies/"

# fixe le dossier de travail et de données
  setwd(wdpath)

# Librairies utilisées
#---------------------

library(MASS)       # vcov.lm()
library(Hmisc)      # procédures: contents, label
library(plm)        # pdata.frame
library(lmtest)     # wdtest

# Mes librairies
```

```
source(file(paste(jdpath, "lib01.r", sep="")))

#=============================================================================
#
#                              Estimation par MCO
#
#-----------------------------------------------------------------------------

# 1. Lecture des données de base
#-------------------------------

load(file="ofsp.RData")

# 2. Création du data frame de travail (WDF)
#-------------------------------------------

# Variables retenues

vname <- c("ID_INS","YEAR","NAME_INS", "EF33_EF_FO_T", "EF33_EF_O_T",
           "EF16_X_450", "EF310_X_CH",
           "EF22_A_6", "EF23_A_6", "EF22_A_7", "EF23_A_7",
           "EF22_M_6", "EF23_M_6", "EF22_M_7", "EF23_M_7")

WDF <- ofsp[vname]
rm(ofsp)

# 3. Création de nouvelles variables
#-----------------------------------

# Création d'une nouvelle variable (recettes totales)

proda.A <- WDF$EF22_A_6 + WDF$EF23_A_6
proda.M <- WDF$EF22_M_6 + WDF$EF23_M_6
prodn.A <- WDF$EF22_A_7 + WDF$EF23_A_7
prodn.M <- WDF$EF22_M_7 + WDF$EF23_M_7

rec.tot1 <- proda.A + proda.M - prodn.A - prodn.M

# Création d'un facteur "taille"

taille <- cut(rec.tot1,
              c(min(rec.tot1, na.rm=T), 4470000, 197000000, max(rec.tot1, na.rm=T)),
              include.lowest=T)

label(taille) <- "Taille de l'assurance selon les primes"

# Création d'un facteur régional vs suisse

WDF$EF310_X_CH <- as.factor(WDF$EF310_X_CH);
levels(WDF$EF310_X_CH) <- c("Régional", "Suisse")

# Mise à jour de la base de données

WDF <- data.frame(WDF, rec.tot1, taille)

# Nettoyage de l'espace de travail

rm(proda.A, proda.M, prodn.A, prodn.M, rec.tot1, taille)

# 4. Estimation à partir d'un modèle
#-----------------------------------

# Prévision sur la base d'un échantillon aléatoire de valeur espérée de 2006
# Un modèle est estimé sur la base des données de 1996-2005 et
# d'un échantillon de données de 2006
# Les données estimées pour 2006 sont générées par imputation (multiple)
# Méthode d'imputation ABB

# Déclaration des variables du loop:
fitimp <- numeric()
tot2006 <-numeric()
pct <- numeric()

for (loop in 1:100){

# 4.1 données de base: 1996-2005 + un échantillon de 2006
#--------------------------------------------------------

# Paramètres à initialiser

nsam <- 30 # Nbr. obs. de l'échantillon
nH <- 5    # Nbr. strates
nimp <- 5  # Nbr. d'imputations
```

```
# Traitement des valeurs 0 pour les variables
# rec.tot1, EF33_EF_FO_T, EF33_EF_O_T, EF16_X_450

obs <- rep(1, nrow(WDF))
obs[WDF$rec.tot1 <= 0] <- 0
obs[WDF$EF33_EF_FO_T <= 0] <- 0
obs[WDF$EF33_EF_O_T <= 0] <- 0
obs[WDF$EF16_X_450 <= 0] <- 0

# Stratification optimale selon rec.tot1 2005
# (cf. Gunning & Horgan (2004))

rec.tot1.2005 <- with(subset(WDF, c(YEAR == 2005)), rec.tot1)
rec.tot1.2005[rec.tot1.2005 <= 0] <- NA

strates <- stratopt(rec.tot1.2005, nH)

# Nos d'assurances 2005 avec nos des strates

ID_INS.2005 <- with(subset(WDF, c(YEAR == 2005)), ID_INS)

INS.2005 <- data.frame(ID_INS.2005, strates)

# Nos d'assurances 2006 avec nos des strates

INS.2006 <- data.frame(with(subset(WDF, c((YEAR == 2006)&(obs==1))), ID_INS))
names(INS.2006) <- c("ID_INS.2006")

INS.2006 <- merge(INS.2006, INS.2005, by.x = c("ID_INS.2006"),
                  by.y = c("ID_INS.2005"), all.x=T)

# Allocation optimale de l'échantillon

(alloc <- allocopt(rec.tot1.2005, strates, nsam))

# Ajout de deux unités dans les strates vides ou égales à 1

alloc[alloc <= 1] <- 2

# Tirage de l'échantillon

ID_INS.sam <- vector("numeric")

for (i in 1:nH){
    pop.i <- INS.2006$ID_INS.2006[INS.2006$strates == i]
    pop.i <- pop.i[!is.na(pop.i)]
    n.i <- alloc[i]
    sam.i <- sample(pop.i, n.i)
    ID_INS.sam <- c(ID_INS.sam, sam.i)
}

# 4.2 Analyse de régression
# -------------------------

# Mise à jour des indices des observations sélectionnées

obs.new <- obs
obs.new[(WDF$YEAR == 2006)&(!is.element(WDF$ID_INS, unique(ID_INS.sam)))] <- 0

# Estimation du modèle

yfit <- lm(log(rec.tot1) ~ log(EF33_EF_FO_T) + log(EF33_EF_O_T) +
                           log(EF16_X_450) + EF310_X_CH + taille + YEAR,
           data=WDF,
           subset=c(obs.new == 1),
           na.action=na.exclude,
           x=T, qr=T)
(yfits <- summary(yfit))

# 4.3 Prévision sur la base du modèle estimé
# ------------------------------------------

# Recettes totales 2006 estimées (espérée)

# Valeurs estimées des observations correspondant à
# l'échantillon aléatoire d'observations de 2006.
# En utilisant le modèle estimé.

p.yfit <- yfit$fitted.values
ind.rnames <- is.element(names(p.yfit),
                         rownames(WDF)[((WDF$YEAR==2006)&(obs.new==1))])
p.yfit <- p.yfit[ind.rnames]
X.p.yfit <- yfit$x[ind.rnames,]
```

```
# Log des recettes totales 2006 estimées

rec.tot1.2006 ← p.yfit

fit ← data.frame(WDF$ID_INS[((WDF$YEAR==2006)&(obs.new==1))], rec.tot1.2006,
                 names(p.yfit))
names(fit) ← c("ID_INS.2006", "rec.tot1", "ind.rnames")

# Constitution pour 2006 d'une base de données comprenant
# les nos d'assurances, de strates et les valeurs prédites de l'échantillon

datafit ← merge(INS.2006, fit, by.x = c("ID_INS.2006"),
                by.y = c("ID_INS.2006"), all.x = T)

# Imputation multiple par regression

sigma_squared_1 ← yfits$sigma^2
V ← yfits$cov.unscaled
df ← yfit$df.residual

imp.fit ← reg.multimpute(datafit$rec.tot1, X.p.yfit, datafit$ind.rnames,
                         yfit$coefficients, sigma_squared_1, V, df, "fit",
                         datafit$strates, nimp)

# Recettes totales estimées

imp.fit ← exp(imp.fit)

datafit ← data.frame(datafit[,-c(3, 4)], imp.fit[-1])

# Ajout de la série observée pour 2006

data.2006 ← with(subset(WDF, c(YEAR == 2006)),
                 data.frame(ID_INS, rec.tot1))

datafit ← merge(datafit, data.2006, by.x = c("ID_INS.2006"),
                by.y = c("ID_INS"), all.x =T)

# Remplacement des valeurs estimées de l'échantillon
# par les véritables valeurs observées

ind ← is.element(datafit$ID_INS.2006,unique(ID_INS.sam))
datafit[ind,(3:(nimp+2))] ← datafit[ind, (nimp+3)]

# Estimation du total

ind ← (datafit$rec.tot1 > 0) & (!is.na(datafit$strates))

fitimp[loop] ← mean(colSums(datafit[ind, 3:(nimp+2)], na.rm=T))
tot2006[loop] ← sum(datafit[ind, (nimp+3)], na.rm=T)
pct[loop] ← (tot2006[loop]-fitimp[loop])/tot2006[loop]

# fin de la boucle
}

# Résumé des résultats

Result ← numeric()
Result[1] ← mean(fitimp)
Result[2] ← mean(tot2006)
Result[3] ← mean(pct)
Result[4] ← sd(fitimp)
Result
```

Script B.5 – Estimation par MCO - Imputation sur un modèle linéaire - Prévision

```
#-------------------------------------------------------------------------------
#
#                -/                                                      -/-/
#              -/    -/-/       -/-/-/    -/-/-/  -/     -/     -/
#             -/  -/-/-/-/  -/-/        -/     -/  -/     -/  -/-/-/-/
#           -/  -/                 -/-/  -/     -/  -/     -/     -/
#          -/     -/-/-/  -/-/-/       -/-/-/     -/-/-/     -/
#
#
#                    Laboratoire d'économétrie et de statistique
#                          de l'Université de Fribourg
#
#  AUTEUR:
#
#     Marie-Justine Leis
#     Département d'économie quantitative
#     Université de Fribourg (Suisse)
#     Bd de Pérolles 90
#     1700 Fribourg
#
#  COURRIEL:
#
#     Marie-Justine.Leis@UniFr.ch
#
#  PROJET:
#
#     Travail de Master
#
#  DATE:
#
#     Novembre 2007 update Janvier 2008
#
#  PROGRAMME:
#
#     job05.r
#
#  BUT:
#
#     Données réelles - Prévision
#
#  DONNEES:
#
#     Type            : Panel
#
#     N. obs.         : 1221
#
#     Fichiers        : "ofsp.Rdata"
#     Titre           :  ofsp
#
#     Source          : Office fédéral de la santé publique, données
#                       d'observation des assurances maladiés du formulaire
#                       EF123.
#
#     Description     : Observations des assureurs depuis 1996 à 2006
#
#
#-------------------------------------------------------------------------------

# Initialisation
#---------------

# Chemins d'accès aux dossiers de données
  ddpath <- "/Users/marie-justineleis/Masterarbeit/Master R/données/"

# Chemins d'accès aux dossiers de travail
  wdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/work/"

# Chemins d'accès aux jobs
  jdpath <- "/Users/marie-justineleis/Masterarbeit/Master R/Jobs J-edit/librairies/"

# fixe le dossier de travail et de données
  setwd(wdpath)

# Librairies utilisées
#---------------------

library(MASS)       # vcov.lm()
library(Hmisc)      # procédures: contents, label
library(plm)        # pdata.frame
library(lmtest)     # wdtest

# Mes librairies

source(file(paste(jdpath, "lib01.r", sep="")))
```

```
#=============================================================================
#
#                              Données réelles
#
#-----------------------------------------------------------------------------

# 1. Lecture des données de base
#-------------------------------

load(file="ofsp.RData")

# 2. Création du data frame de travail (WDF)
#-------------------------------------------

# Variables retenues

vname <- c("ID_INS","YEAR","NAME_INS", "EF33_EF_FO_T", "EF33_EF_O_T",
           "EF16_X_450", "EF310_X_CH",
           "EF22_A_6", "EF23_A_6", "EF22_A_7", "EF23_A_7",
           "EF22_M_6", "EF23_M_6", "EF22_M_7", "EF23_M_7")

WDF <- ofsp[vname]
rm(ofsp)

# 3. Création de nouvelles variables
#-----------------------------------

# Création d'une nouvelle variable (recettes totales)

proda.A <- WDF$EF22_A_6 + WDF$EF23_A_6
proda.M <- WDF$EF22_M_6 + WDF$EF23_M_6
prodn.A <- WDF$EF22_A_7 + WDF$EF23_A_7
prodn.M <- WDF$EF22_M_7 + WDF$EF23_M_7

rec.tot1 <- proda.A + proda.M - prodn.A - prodn.M

# Création d'un facteur "taille"

taille <- cut(rec.tot1,
              c(min(rec.tot1, na.rm=T), 4470000, 197000000, max(rec.tot1, na.rm=T)),
              include.lowest=T)

label(taille) <- "Taille de l'assurance selon les primes"

# Création d'un facteur régional vs suisse

WDF$EF310_X_CH <- as.factor(WDF$EF310_X_CH);
levels(WDF$EF310_X_CH) <- c("Régional", "Suisse")

# Mise à jour de la base de données

WDF <- data.frame(WDF, rec.tot1, taille)

# Traitement des valeurs 0 pour les variables
# rec.tot1, EF33_EF_FO_T, EF33_EF_O_T, EF16_X_450

obs <- rep(1, nrow(WDF))
obs[WDF$rec.tot1 <= 0] <- 0
obs[WDF$EF33_EF_FO_T <= 0] <- 0
obs[WDF$EF33_EF_O_T <= 0] <- 0
obs[WDF$EF16_X_450 <= 0] <- 0


# 4. Estimation à partir des données observées sur un échantillon
#----------------------------------------------------------------

# Prévision sur la base d'un échantillon aléatoire d'observations de 2006
# Sans utilisation de modèle
# Taille de l'échantillon: nsam

# Déclaration des variables du loop:
fitimp <- numeric()
tot2006 <- numeric()
pct <- numeric()

for (loop in 1:1000){


# Paramètres à initialiser

nsam <- 35 # Nbr. obs. de l'échantillon
nH <- 5    # Nbr. strates
nimp <- 10  # Nbr. d'imputations
```

```
# Stratification optimale selon rec.tot1 2005
# (cf. Gunning & Horgan (2004))

rec.tot1.2005 <- with(subset(WDF, c(YEAR == 2005)), rec.tot1)
rec.tot1.2005[rec.tot1.2005 <= 0] <- NA

strates <- stratopt(rec.tot1.2005, nH)

# Nos d'assurances 2005 avec nos des strates

ID_INS.2005 <- with(subset(WDF, c(YEAR == 2005)), ID_INS)

INS.2005 <- data.frame(ID_INS.2005, strates)

# Nos d'assurances 2006 avec nos des strates

INS.2006 <- data.frame(with(subset(WDF, c(YEAR == 2006)), ID_INS))
names(INS.2006) <- c("ID_INS.2006")

INS.2006 <- merge(INS.2006, INS.2005, by.x = c("ID_INS.2006"),
                  by.y = c("ID_INS.2005"), all.x=T)

# Allocation optimale de l'échantillon

(alloc <- allocopt(rec.tot1.2005, strates, nsam))

# Ajout de deux unités dans les strates vides ou égales à 1

alloc[alloc <= 1] <- 2

# Tirage de l'échantillon

ID_INS.sam <- vector("numeric")

for (i in 1:nH){
    pop.i <- INS.2006$ID_INS.2006[INS.2006$strates == i]
    pop.i <- pop.i[!is.na(pop.i)]
    n.i <- alloc[i]
    sam.i <- sample(pop.i, n.i)
    ID_INS.sam <- c(ID_INS.sam, sam.i)
}

# Recettes totales 2006 pour l'échantillon

data.2006 <- with(subset(WDF, c(YEAR == 2006)),
                  data.frame(ID_INS, rec.tot1))

fit <- data.2006[is.element(data.2006$ID_INS, unique(ID_INS.sam)),]

# Constitution pour 2006 d'une base de données comprenant
# les nos d'assurances, de strates et les valeurs prédites de l'échantillon

datafit <- merge(INS.2006, fit, by.x = c("ID_INS.2006"),
                 by.y = c("ID_INS"), all.x = T)

# Imputation multiple par ABB

imp.fit <- multimpute(datafit$rec.tot1, "fit", datafit$strates, nimp)

datafit <- data.frame(datafit[,-3], imp.fit[-1])

# Ajout de la série observée pour 2006

datafit <- merge(datafit, data.2006, by.x = c("ID_INS.2006"),
                 by.y = c("ID_INS"), all.x =T)

# Estimation du total

ind <- (datafit$rec.tot1 > 0) & (!is.na(datafit$strates))

fitimp[loop] <- mean(colSums(datafit[ind, 3:(nimp+2)], na.rm=T))
tot2006[loop] <- sum(datafit[ind, (nimp+3)], na.rm=T)
pct[loop] <- (tot2006[loop]-fitimp[loop])/tot2006[loop]

# fin de la boucle
}

# Résumé des résultats

Result <- numeric()
Result[1] <- mean(fitimp)
Result[2] <- mean(tot2006)
Result[3] <- mean(pct)
```

```
Result[4] ← sd(fitimp)
Result


Vnimp ← function(nimp, datafit, fitimp){
    nimp ← as.numeric()
    V ← (1/nimp(nimp−1))∗(colSums(datafit[ind, 3:(nimp+2)], na.rm=T)+ fitimp)
    return(V)
}
```

Script B.6 – Données réelles - Imputation ABB - Prévision

```
#-------------------------------------------------------------------------------
#
#                -/                                                      -/-/
#              -/     -/-/       -/-/-/     -/-/-/  -/     -/     -/
#             -/  -/-/-/-/  -/-/        -/     -/  -/     -/  -/-/-/-/
#           -/  -/                 -/-/  -/     -/  -/     -/     -/
#          -/     -/-/-/  -/-/-/       -/-/-/     -/-/-/     -/
#
#
#                    Laboratoire d'économétrie et de statistique
#                          de l'Université de Fribourg
#
#  AUTEUR:
#
#     Marie-Justine Leis
#     Département d'économie quantitative
#     Université de Fribourg (Suisse)
#     Bd de Pérolles 90
#     1700 Fribourg
#
#  COURRIEL:
#
#     Marie-Justine.Leis@UniFr.ch
#
#  PROJET:
#
#     Travail de Master
#
#  DATE:
#
#     Novembre 2007 update Janvier 2008
#
#  LIBRAIRIE:
#
#     lib01.r
#
#  BUT:
#
#     Diverses fonctions ad hoc
#
#-------------------------------------------------------------------------------

#===============================================================================
#
#                                IMPUTATION MULTIPLE
#
#-------------------------------------------------------------------------------

# 1. Méthode ABB
#---------------

abb1 <- function(x){
    n <- length(x)
    yrep <- x[!is.na(x)]
    (sample(yrep, n, replace=T))
}

abb2 <- function(x, abb1){
    n_nr <- sum(is.na(x))
    (sample(abb1, n_nr, replace=T))
}

abb <- function(x){
    ynrep <- abb2(x, abb1(x))
    yimp <- x
    yimp[is.na(yimp)] <- ynrep
    return(yimp)
}

impute <- function(x, CI){
    CI <- as.numeric(CI)
    yimp <- x
    for (i in 1:max(CI, na.rm=T)){
        yimp[(CI==i)&(!is.na(CI))] <- abb(yimp[(CI==i)&(!is.na(CI))])
    }
    return(yimp)
}

multimpute <- function(x, varname, CI, K){
    dataimp <- as.vector(x)
    for (k in 1:K){
        yimp <- x
        dataimp <- cbind(dataimp,as.vector(impute(yimp, CI)))
    }
```

```
    dataimp ← as.data.frame(dataimp)
    names(dataimp) ← c(varname, paste(varname, "I", 1:K, sep=""))
    return(dataimp)
}

# 2. Méthode "Modèle de régression linéaire"
#-------------------------------------------

sigma_squared_star ← function(sigma_squared_1, df){
    sigma_squared_1*df/rchisq(1,df)
}

beta_star ← function(beta_1, sigma_squared_star, V){
    sigma_star ← sqrt(sigma_squared_star)
    q ← length(beta_1)
    Z←rnorm(q)
    V05 ← chol(V)
    b ←beta_1 + sigma_star*V05 %*% as.vector(Z)
    return(b)
}

ystar ← function(xi, beta_1, sigma_squared_1, V, df){
    s2 ← sigma_squared_star(sigma_squared_1, df)
    b ← beta_star(beta_1, s2, V)
    yi ← xi %*% b + rnorm(1) * sqrt(s2)
    return(yi)
}

reg.impute ← function(y, X, rnames, beta_1, sigma_squared_1, V, df, CI){
    names(y) ← rnames
    ind ← 1:length(y)
    CI ← as.numeric(CI)
    yimp ← y
    for (i in 1:max(CI, na.rm=T)){
        ind.miss ← ind[((is.na(yimp))&(CI==i))]
        ind.miss ← ind.miss[!is.na(ind.miss)]
        nmiss ← length(ind.miss)
        if(nmiss!=0){
            ind.nomiss ← ind[((!is.na(yimp))&(CI==i))]
            ind.nomiss ← ind.nomiss[!is.na(ind.nomiss)]
            ind.nomiss ← sample(ind.nomiss, nmiss, replace=T)
            for (j in 1:nmiss){
                xi ← X[rnames[ind.nomiss[j]] == rownames(X),]
                yimp[ind.miss[j]] ← ystar(xi, beta_1, sigma_squared_1, V, df)
            }
        }
    }
    return(yimp)
}

reg.multimpute ← function(y, X, rnames, beta_1, sigma_squared_1,
                          V, df, varname, CI, K){
    dataimp ← as.vector(y)
    for (k in 1:K){
        yimp ← reg.impute(y, X, rnames, beta_1, sigma_squared_1, V, df, CI)
        dataimp ← cbind(dataimp,as.vector(yimp))
    }
    dataimp ← as.data.frame(dataimp)
    names(dataimp) ← c(varname, paste(varname, "I", 1:K, sep=""))
    return(dataimp)
}


#=============================================================================
#
#                              ALLOCATION OPTIMALE
#
#-----------------------------------------------------------------------------


allocopt ← function(x, strates, nsam){
    Nh.vec ← tapply(x, strates, length)
    sd.vec ← tapply(x, strates, sd)
    alloc0 ← Nh.vec
    alloc1 ← rep(0, length(Nh.vec))
    i ← 1
    while(i > 0){
        alloc ← round(nsam * (Nh.vec*sd.vec)/sum(Nh.vec*sd.vec))
        dif ← ((alloc0-alloc) < 0)
        i ← ifelse(sum(dif)>0, 1, 0)
        if(i > 0){
            alloc1[dif] ← Nh.vec[dif]
            alloc0 ← alloc0[!dif]
            Nh.vec ← Nh.vec[!dif]
```

```
            sd.vec ← sd.vec[!dif]
            nsam ← nsam − sum(alloc1[dif])
        }
        else{
            alloc1[alloc1==0] ← alloc
            }
    }
    return(alloc1)
}


#=============================================================================
#
#                          STRATIFICATION OPTIMALE
#
#-----------------------------------------------------------------------------


stratopt ← function(x, L){
    H ← L + 1 # Nbr de strates désiré + 1
    hbreaks ← vector("numeric", H)
    hbreaks[1]← min(x, na.rm=T)
    hbreaks[H]← max(x, na.rm=T)
    for (h in 2:(H−1)){
        hbreaks[h] ← hbreaks[1]∗((hbreaks[H]/hbreaks[1])^(1/H))^h
        }
    strates ← cut(x, hbreaks, include.lowest=T, labels=1:(H−1))
    return(strates)
}
```

Script B.7 – Librairie avec les fonctions d'imputation multiple et les fonctions de stratification et d'allocation optimale

Bibliographie

T. Amemiyia : The estimation of the variance-components model. *International Economic Review*, 12:1–13, 1971.

M. Arellano : Panel Data Models : Some Recent Developments. *In Handbook of Econometrics*, vol. 5, chap. 53, p. 3229–3296. Elesevier Sience B.V., 2001.

B. H. Baltagi : *Econometric Analysis of Panel Data.* John Wiley & Sons Ltd, 3 édn, 2005.

T. Bandi : Compensation des risques dans l'assurance-maladie : Une amélioration par l'extension ? *Sécurité sociale*, p. 94–98, 1999.

K. Beck, M. Trottmann, U.Käser, B. Keller, S. V. Rotz et P. Zweifel : *Nachhaltige Gestaltung des Risikoausgleichs in der Schweizer Krankenversicherung.* hep, Ott Verlag, Bern, 1. édn, 2006.

K. Beck : Growing importance of capitation in Switzerland. *Health Care Manangement Sience*, p. 111–119, 2000.

K. Beck : Risikoausgleich in der Krankenversicherung-Wozu ? *In Reformstau beim Risikoausgleich ?*, p. 4–5. Risk Adjustment Network, Luzern, 2004.

K. Beck, S. Spycher, A. Holly et L. Gardiol : Risk adjustment in Switzerland. *Health Policy*, 65:63–74, 2003.

K. Becker et P. Zweifel : Cost Sharing in Health Insurance : An Instrument for Risk Selection ? Sozialökonomisches Institut, Universität Zürich, 2005.

A. Cameron et P. K. Trivedi : *Microeconometrics : Methods and Applications.* Cambridge University Press, New York, 2005.

F. Colombo : Towards more Choice in Social Protection ? Individual Choice of Insurer in Basic Mandatory Health Insurance in Switzerland. *OECD Labour Market and Social Policy Occasional Papers*, 2001.

L. Crivelli : Improving the risk adjustment formula. *Health Policy Monitor*, 2005.

T. Dalenius : The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, p. 203–213, 1950.

G. Ekman : An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30:219–229, 1959.

Y.-A. Gerber : *Kosten und Finanzierung des Gesundheitswesens : Detaillierte Ergebnisse 2004 und jüngste Entwicklung.* Bundesamt für Statistik, 2006.

S. Greppi, R. Rossel et W. Strüwe : Effets de la LAMal dans le financement de la santé. *Sécurité sociale*, p. 95–99, 1998.

P. Gunning et J. M. Horgan : Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête*, 30(2):177–185, 2004.

J. Hausman : Specification tests in econometrics. *Econometrica*, 46:1251–1271, 1978.

A. Holly, L. Gardiol, Y. Eggli et T. Yalcin : Gesundheitsbasierter Risikoausgleich in der Schweiz : Eine Untersuchung mit Hilfe medizinischer Informationen aus vorhergehneden Spitalaufenthalten. *In Reformstau beim Risikoausgleich ?*, p. 25–29. Risk Adjustment Network, Luzern, 2004a.

A. Holly, L. Gardiol, Y. Eggli et T. Yalcin : Health-Based Risk Adjustment in Switzerland : an Exploration Using Medical Information from Prior Hospitalisation. Rap. tech., Institut d'économie et management de santé (IEMS), 2004b. Financed by the Swiss National Fund.

J. M. Horgan : Stratification of skewed populations : A review. *International Statistical Review*, 74(1):67–76, 2006.

C. Hsiao : Panel data models. *In A companion to theorectical econometrics*, chap. 16, p. 349–365. Blackwell Publishers, 2001.

C. Hsiao : *Analysis of Panel Data.* Economic Society Monographs. Cambridge University Press, 2003.

N. A. Klevmarken : Introduction - Panel Studies : What Can We Learn From Them ? *European Economic Review*, 33:523–529, 1989.

T. Lancaster : The incidental parameter problem since 1948. *Journal of Econometrics*, 95:391–413, 2000.

P. Lavallée et M. Hidiroglou : On the stratification of skewed populations. *Survey Methodology*, 14:33–43, 1988.

M. Nerlove : Further evidence on the estimation of dynamic economic relations from a time-series of cross-sections. *Econometrica*, 39:359–382, 1971.

J. Neyman et E. L. Scott : Consistent estimation from partially consistent observations. *Econometrica*, 16:1–32, 1948.

R. Nyffeler : Surveillance de l'assurance-maladie sociale. *Sécurité sociale*, 2006.

OECD : *La réfome des systèmes de santé : Etude de dix-sept pays de l'OCDE*. OECD, Paris, 1994.

OECD : *OECD-Berichte ; Über Gesundheitssysteme : Schweiz*. OECD, Paris, 2006.

OFAS : *Analyse des effets de la LAMal : Rapport de Santé*. Office fédéral des assurances sociales, 2001.

W. Oggier : Risikoausgleich oder Risikoselektion ? : Einige gesundheitsökonomische Gedanken zu den aktuellen Reformvorhaben auf Bundesebene. *Schweizerische Ärztezeitung*, Nr 31:1626–1629, 2004.

J. C. Pinheiro et D. M. Bates : *Mixed-Effects Models in S and S-Plus*. Springer, 2000.

R Development Core Team : *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. URL `http://www.R-project.org`.

R. Rossel : Effets de la LAMal dans le coût et le financement de la santé. *Sécurité sociale*, p. 153–157, 2000.

R. Rossel : Coût des systèmes de santé. *Sécurité sociale*, p. 48–53, 2006.

R. Rossel et N. Siffert : *Coûts de la santé en suisse : Evolution de 1960 à 2000*. StatSanté : Résultats des statistiques suisses de la santé. Office fédéral de la statistique, 2003.

D. B. Rubin : *Multiple Imputation for Nonresponse in Surveys*. Wiley series in probability and mathematical statistics. John Wiley & Sons, 1987. ISBN 0-387-98957-0.

N. Siffert : *Flux financiers dans le système suisse de santé*. StatSanté : Résultats des statistiques suisses de la santé. Office fédéral de la statistique, 2002.

N. Siffert : *Statistique de l'assurance-maladie obligatoire 2005.* Office fédéral de la santé publique, 2007.

D. Sordat Fornerod : Das schweizerische Gesundeheitswesen : Analyse und Empfehlungen der OECD und der WHO. *Soziale Sicherheit*, p. 38–43, 2007.

S. Spycher : Assurance-maladie : les effets sous-évalués de la compensation des risques. *Sécurité sociale*, p. 202–205, 1999.

S. Spycher : Compensation des risques dans la LAMal - et la suite ? *Sécurité sociale*, p. 109–112, 2004a.

S. Spycher : Die politische und wissenschaftliche Diskussion in der Schweiz. *In Reformstau beim Risikoausgleich ?*, p. 21–24. Risk Adjustment Network, Luzern, 2004b.

P. Swamy et S. Aurora : The exacte finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica*, 40:261–275, 1972.
wWr97

C. E. Särndal, B. Swensson et J. Wretman : *Model Assisted Survey Sampling.* Springer Series in Statistics. Springer-Verlag, 1997.

W. P. van de Ven : Risikoselektion im Krankenversicherungsmarkt. *In Reformstau beim Risikoausgleich ?*, p. 6–10. Risk Adjustment Network, Luzern, 2004.

W. P. van de Ven et R. P. Ellis : Risk adjustment in competitive health plan markets. *In Handbook of Health Economics*, chap. 17. North Holland Publishing Company, 1999.

W. P. van de Ven et R. C. van Vliet : How can we prevent cream sikimming in a competitive health insurance market ? The great challange of de 90's. *In Health Economics Worldwide*, p. 23–46. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.

T. Wallace et A. Hussain : The use of error components models in combining cross section with time series data. *International Economic Review*, 37(1):55–72, 1969.

J. Wasem, F. Buchner et C. Behrend : Der Risikostrukturausgleich in der deutschen gesetzlichen Krankenversicherung. *In Reformstau beim Risikoausgleich ?*, p. 17–20. Risk Adjustment Network, Luzern, 2004.

Déclaration

Je soussignée, déclare sur mon honneur, que j'ai personnellement préparé le travail qui précède et que celui-ci est conforme aux normes de l'honnêteté scientifique.

J'ai pris connaissance de la décision du Conseil de Faculté du 9 novembre 2004 l'autorisant à me retirer le titre conféré sur la base du présent travail dans le cas où ma déclaration ne correspondrait pas à la vérité.

Fribourg, le 3 juin 2008 Marie-Justine Leis