Chapitre 2

Coût et qualité de l'offre de soins, quelle(s) rémunération(s) ? Une application au Québec[1]

> *«The re-engineering of health care will certainly require a reform in the way that medical providers are paid.[. . .] Fee-for-service payment promotes productivity but also encourages over-use ; conversely, paying doctors a straight salary in heavily regulated markets may lead to under performance, because it fails to reward productivity. Successful reform will involve remoulding payments systems so that they reward quality and performance. »*
>
> *"Keep taking the medicine"* , *The Economist* (15 Juillet 2004, p.17)

Dans le dernier classement qu'elle a consacré à la performance des systèmes de santé, en 2000, l'OMS plaçait la France au premier rang mondial.[2] La même année, la France

[1] Ce chapitre est inspiré d'un travail réalisé en collaboration avec Bernard Fortin et Bruce Shearer.

[2] Organisation Mondiale de la Santé (2000). *Rapport sur la santé dans le monde 2000*, Genève.

se situait parmi les systèmes de santé les plus chers de l'OCDE, se situant au 4° rang avec 9.4% de son PIB consacré aux dépenses de santé (à égalité avec le Canada, contre 10.3% pour l'Allemagne et 12.9% pour les Etats-Unis).[3] Le rapprochement de ces deux faits illustre parfaitement les termes contemporains dans lesquels se pose la question de la gestion des soins de santé. Après des décennies d'efforts dirigés presque exclusivement vers la maîtrise des coûts, c'est aujourd'hui vers la rentabilité de ces dépenses, en termes d'amélioration de la santé, que se déplace le débat (McGlynn, Asch, Adams & *al.*, 2003). L'accroissement tendanciel des ressources consacrées à la santé apparaît ainsi comme l'orientation naturelle d'une société développée[4] et l'objectif assigné à la politique de santé est dès lors formulé selon une exigence d'efficience plutôt que d'économies. Cette réorientation des objectifs nécessite un renouvellement profond des réflexions quant à l'efficacité des instruments de la politique de santé (Dudley, Miller, Korenbrot & *al.*, 1998 ; Malin & Keating, 2005), tant les spécificités du secteur médical et le souci de contrôle des coûts s'opposent à l'amélioration de la qualité (McNeil, 2001).

En raison de l'importance de son expertise, le médecin est au cœur du fonctionnement du système de santé (Arrow, 2001b). En termes de coûts, il représente à la fois un poste de dépenses important[5] et l'origine indirecte de la grande majorité des dépenses de santé, à travers son rôle de prescripteur. Cette préoccupation d'amélioration de la qualité s'est donc, en particulier, traduite par d'importantes innovations destinées à mieux contrôler la qualité de l'offre de soins, telle que l'instauration d'objectifs-cibles (Kiefe, Allison, Williams & *al.*, 2001).[6] En accord avec la majorité de la profession

[3]Sources : Direction de la Recherche, des Etudes, de l'Evaluation et des statistiques (Drees). Comparaison internationale des dépenses de santé, *Etudes et Résultats* n°175, 2002 ; Institut Canadien d'Information sur la santé (ICIS). Les soins de santé au Canada, 2000.

[4]Voir, par exemple, Le cercle des Economistes (2004). Economie de la santé : Une réforme ? Non, une révolution, *Cahier n°6.*

[5]Dans le cas du Canada, Deber, Narine, Baranek & *al.* (1998) évaluent à 15% la part des dépenses de santé inhérente à la rémunération des médecins.

[6]Voir également J. Carroll "Quality Counts : So Why Not Offer Physicians Bonuses ?", *California Health Consensus* (1 Janvier 2003), pour un survol d'initiatives prises récemment en ce sens aux Etats-unis et Rosenthal, Fernandopulle, Song & *al.* (2004) pour une revue critique.

(McKenna, 2002), le gouvernement britannique a récemment franchi un pas important dans cette direction, instituant, en 2002, un système de rémunération fondé sur un score de qualité (Shekelle, 2003) et susceptible d'affecter jusqu'à 18% de la rémunération totale des médecins généralistes (Smith & York, 2004). Le nouveau système s'appuie sur une évaluation large de la qualité de la pratique, puisque ce score est calculé à partir d'un tableau de bord composé de plus de 70 indicateurs. Le revers de cette innovation ambitieuse est d'avoir considérablement accru la complexité du système. Ce dernier effet semble l'avoir emporté, et explique en grande partie l'échec de cette réforme (Smith, 2003). Cette expérience milite donc en faveur de l'exploration des possibilités d'amélioration de la qualité offertes par les sytèmes de rémunération traditionnels, présentant l'avantage de la simplicité («*it may be more appropriate to pursue quality-oriented refinements of traditional payment approaches, rather than radical transformation.*». Cunningham, 2004, p.36). C'est la voie suivie dans ce chapitre, qui se propose d'évaluer la capacité des systèmes de rémunération fondés sur les mesures de performance traditionnelles à résoudre la contradiction entre amélioration de la qualité et contrôle des coûts.

La capacité des incitations financières traditionnelles à orienter la pratique professionnelle est pourtant moins évidente dans le cas des médecins que pour la plupart des autres professions. L'arbitrage entre consommation et loisir des médecins, en particulier, est fortement influencé par leur appartenance aux catégories de revenus les plus élevées.[7] Pour les individus appartenant à cette couche de la population, en effet, le niveau de rémunération est tel que l'effet revenu est susceptible de prendre l'avantage sur l'effet substitution (Feldstein, 1995). Dans ce cas, les médecins prendraient alors leurs décisions le long d'une fonction d'offre de travail à rebroussement, si bien que les variations de rémunération auraient un effet ambigu sur le nombre d'heures de travail ; positif ou négatif selon la position initiale le long de la fonction d'offre. Les premiers travaux consacrés à cette question ont confirmé cette crainte (Sloan, 1975). Feldstein

[7]Dans le cas des Etats-Unis, Showalter & Thurston (1997) indiquent ainsi que les médecins représentent 15% de la moitié la plus riche de la population.

(1970) estime ainsi à 60% la probabilité que l'offre de travail des médecins américains inclus dans son échantillon présente un rebroussement au salaire courant. Bien que l'effet revenu estimé reste important (Rizzo & Blumenthal, 1994), les travaux plus récents tendent cependant à renverser ces premières conclusions et obtiennent une élasticité positive de l'offre de travail au niveau de salaire offert (Showalter & Thurston, 1997). En raison de l'importance de l'effet revenu, cette élasticité tend à être inférieure au niveau généralement observé, oscillant entre 10% (Sæther 2003) et 30% (Baltagi, Bratberg & Holmas, 2005) en fonction des données et méthodes utilisées.

S'ils apparaissent comme un instrument utile de pilotage du comportement des médecins, les systèmes de rémunération doivent encore répondre à la double exigence de maîtrise des coûts et de promotion de la qualité des soins. Outre le volume de rémunération, et le volume d'heures de soins qui en résulte, c'est ainsi la nature de ces rémunérations et le type de pratiques qu'ils encouragent qui constituent la question centrale de la gestion des soins de santé.[8] En accord avec la préoccupation de Sloan (1975, p.554), pour qui «*no single variable may appropriately be used as an indication of physician supply behavior*», résoudre cette question implique en particulier de prendre en compte l'ensemble des dimensions de l'activité médicale. A cet égard, la réforme du mode de rémunération des médecins réalisée au Québec en 1999 constitue une innovation importante, par le type de rémunération instauré comme par les objectifs poursuivis.

Cette réforme consiste à offrir aux médecins spécialistes d'opter librement pour un mode de *rémunération mixte* alternatif aux modes de rémunération existants. En termes de politique de rémunération, ce nouveau système offre la particularité de combiner plusieurs instruments, puisqu'il associe un salaire forfaitaire, appelé *per diem*, à une diminution du taux de rémunération proportionnelle au nombre d'actes réalisés. L'adoption du nouveau mode de rémunération est de plus laissé à la discrétion des médecins,

[8] «*Empirical research as demonstrated financing methods as important tools in the management of health service. Knowledge of possible health effects for the patients as a consequence of financing methods seems limited.*»Aas (1995, p.205).

qui peuvent choisir de conserver le système antérieur. Pour reprendre les termes de ses concepteurs, la rémunération mixte a été introduite de façon à créer «*un mode équitable qui incite [les médecins] à avoir des comportements permettant à la fois d'améliorer les services à la population et d'être plus efficients.*»[9] Ces principes généraux se déclinent en trois objectifs. Les autorités souhaitent d'abord encourager la diversification des activités des médecins, en accordant le versement du *per diem* à toute activité médicale, du temps passé avec les patients aux heures de travail consacrées à l'administration des établissements ou à l'enseignement. La rémunération de ces activités constitue une nouveauté importante en comparaison du système le plus largement répandu jusqu'alors, la *rémunération à l'acte*, qui consiste uniquement en une rémunération proportionnelle aux actes délivrés et ignore le temps de travail qui n'est pas consacré aux patients. Cet élargissement des activités rémunérées devrait donc, ensuite, promouvoir l'équité des rémunérations, en offrant une compensation financière aux médecins qui acceptent de les exercer. En réduisant la rémunération proportionnelle aux actes réalisés, enfin, la rémunération mixte est également destinée à encourager les médecins à accroître le temps qu'ils consacrent à chaque patient et à améliorer ainsi la qualité des soins prodigués.

Par les instruments de rémunération qu'elle mobilise comme par ses résultats attendus, la rémunération mixte constitue donc une expérience originale, qui recouvre assez largement les préoccupations actuelles quant aux modalités de gestion de l'offre de soins de santé. Afin de mieux comprendre l'influence des incitations sur le coût et la qualité de l'offre de soins, ce chapitre propose une analyse des effets attendus et des résultats effectifs de l'introduction de la rémunération mixte. Dans ce but, l'analyse théorique et le modèle économétrique proposés intègrent explicitement les déterminants de l'arbitrage entre marges extensives (nombre d'actes, heures et semaines de travail) et marge intensive (temps consacré aux actes) de la pratique médicale. Dans ce cadre, la contrainte budgétaire qui gouverne les choix de pratique sous la rémunération mixte présente d'importantes non-linéarités. La méthode adoptée est destinée à résoudre les

[9] Conseil Médical du Québec (1997, p.13). Avis pour un mode mixte de rémunération des médecins de 2^{e} et 3^{e} lignes lié à leurs responsabilités, *Avis n° 97-03*.

problèmes analytiques posés par cette propriété.

Une première non-linéarité est due à l'endogénéité des prix lorsque le choix porte simultanément sur le nombre d'actes et le temps qui leur est consacré. Ainsi, par exemple, le prix qui rémunère une heure de travail consacrée aux patients dépend du nombre d'actes réalisés pendant ces heures, qui constitue lui-même une variable de choix. Cette première non-linéarité est analogue à celle que rencontrent les modèles d'arbitrage entre quantités et qualité dans les choix de consommation (Becker & Lewis, 1973). Le modèle théorique proposé emprunte par conséquent à cette littérature, en définissant des prix virtuels qui permettent une linéarisation locale de la contrainte budgétaire. La statique comparative du modèle permet de prédire les effets attendus du passage à la rémunération mixte. La combinaison des instruments – variation simultanée du salaire fixe et du taux de rémunération des actes – et l'arbitrage entre marges intensive et extensive rendent très ambigu l'impact des incitations sur les choix de pratique. Dans le cas général, rien ne garantit en particulier que la relation négative traditionnellement attendue entre le volume de soins (nombre d'actes et heures de travail) et l'affaiblissement des incitations financières soit respectée. Le modèle permet cependant d'isoler des conditions sur les préférences des médecins suffisantes à ce que ces résultats apparaissent.

L'analyse théorique est considérablement simplifiée par le caractère volontaire du passage à la rémunération mixte, qui permet d'interpréter son adoption comme une préférence révélée. En termes économétriques, en revanche, cet aspect se traduit par le risque que les estimations soient sujettes à un biais de sélection. Si les médecins qui choisissent la rémunération mixte se distinguent de ceux qui la refusent par des caractéristiques individuelles inobservables et corrélées avec les choix de pratique, les méthodes classiques de régression multiple échouent en effet à identifier l'effet des incitations sur ces choix. Ce problème est résolu en spécifiant un modèle économétrique structurel, dans lequel les paramètres estimés gouvernent les choix optimaux le long de la contrainte budgétaire.

L'endogénéité du schéma de rémunération affecte également la forme de cette contrainte budgétaire elle-même. En raison du libre choix entre les modes de rémunération alternatifs, la contrainte budgétaire est générée pour un choix de pratique donné par le mode de rémunération qui maximise le revenu. La contrainte budgétaire est donc linéaire par morceau, façonnée dans chaque espace de choix de pratique par le mode de rémunération optimal. Cette co-existence entre le mode de rémunération mixte et la rémunération à l'acte s'étend également aux choix de pratique conditionnels à son adoption. En raison des dispositions de sa mise en œuvre, les médecins qui ont choisi la rémunération mixte voient une partie de leurs activités rémunérées selon le mode de rémunération à l'acte. La contrainte budgétaire correspond alors à l'un ou l'autre des systèmes de rémunération, en fonction de l'adéquation de la pratique aux dispositions qui ouvrent droit à la rémunération mixte. Par définition, le passage de la rémunération à l'acte à la rémunération mixte provoque un changement simultané du salaire fixe reçu de la pratique – ordonnée à l'origine de la contrainte budgétaire – et du taux de rémunération des actes – pente dans le plan des actes. Les contraintes budgétaires associées à chaque mode de rémunération sont donc sécantes ; et chacune de ces deux propriétés est à l'origine de non-linéarités supplémentaires de la contrainte budgétaire le long de laquelle les médecins prennent leurs décisions.

Le choix de la méthode d'estimation du modèle répond au souci d'intégrer l'ensemble de ces non-linéarités. La solution la plus couramment adoptée consiste à utiliser l'algorithme d'Hausman, fondé sur un balayage de l'ensemble des segments linéaires de la contrainte (Burtless & Hausman, 1978 ; Hausman , 1979 ; 1980 ; 1985). Des travaux récents ont cependant montré que cette méthode restreint de façon importante les paramètres estimés. La convexité des préférences, en particulier, est imposée *a priori* par la méthode d'estimation (MaCurdy, Green & Paarsch, 1990), ce qui peut conduire à rejeter à tort les conditions de Slutsky (Meyer & Heim, 2003). Afin de laisser les comportements observés définir librement les préférences estimées, nous optons par conséquent pour une estimation par discrétisation de la contrainte budgétaire (Zabalza, Pissarides & Barton, 1980). Cette stratégie d'estimation n'impose aucune contrainte sur les para-

mètres estimés. La cohérence de la méthode requiert uniquement que l'utilité marginale du revenu soit positive (van Sœst, 1995).[10]

A travers l'évaluation de la rémunération mixte, ces outils permettent d'approfondir l'analyse théorique et empirique de la réponse optimale des choix de pratique aux variations des incitations. Un élément clé de cette approche, fondée sur la maximisation contrainte d'utilité, est la précision avec laquelle le niveau de consommation engendré par les choix de pratique est décrit. Dans notre cas, la contrainte budgétaire résulte des modalités institutionnelles qui gouvernent la rémunération des médecins du Québec avant et après la réforme (Section 2.1). Les traits essentiels de la rémunération mixte et les objectifs qui ont présidé à son instauration constituent les principaux ingrédients de l'analyse théorique (Section 2.2). Afin de lever les ambiguïtés qui en découlent, l'analyse économétrique intègre les dispositions précises de la rémunération mixte. Le modèle permet d'estimer les paramètres de préférences des médecins de l'échantillon, sous l'hypothèse que les choix observés maximisent l'utilité sous une contrainte de revenu discrétisée (Section 2.3).

L'identification empirique du modèle repose sur le comportement de pratique observé de l'ensemble des médecins du Québec entre 1996 et 2002. Les données trimestrielles utilisées couvrent donc, sous forme de panel, une période de 6 ans centrée sur l'année de la réforme. Surtout, ces données combinent des résultats d'enquête sur les semaines et les heures travaillées (ventilées selon le type d'activités) et des données administratives sur le volume d'actes délivrés et le revenu tiré de la pratique. Elles permettent donc

[10] Cette capacité à offrir un traitement adéquat des non-convexités de l'ensemble budgétaire a motivé de nombreuses applications de cette méthode en économétrie de l'offre de travail, au nombre desquels s'inscrivent, notamment, Hoynes (1996), Colombino (1998), Keane & Moffitt (1998), Euwals & van Soest (1999), Blundell, Duncan, McCrae & *al.* (2000) et van Sœst, Das & Gong (2002). Nyffeler (2005) propose une synthèse critique de l'adéquation de cette méthode à l'analyse de l'offre de travail. A notre connaissance, Sæther (2005) est le seul exemple d'application de cette méthode à l'économie de la santé, consacrée aux choix des médecins en termes d'établissements de pratique et d'heures de travail clinique.

d'associer le niveau réel de consommation à la pratique observée, lacune qui a constitué jusqu'à présent une barrière importante à l'analyse des choix de pratique des médecins.[11] A partir de ces observations, l'estimation structurelle des préférences des médecins et la modélisation de la contrainte budgétaire permettent d'évaluer l'effet propre de la rémunération mixte sur les comportements de pratique et de simuler l'effet potentiel de réformes alternatives (Section 2.4). Ces résultats montrent en particulier l'importance que revêt la liberté de choix du mode de rémunération, et participe par là aux réflexions contemporaines sur la capacité des incitations à améliorer la qualité des soins à un coût maîtrisé (Section 2.5).

2.1 Institutions : la contrainte budgétaire des médecins du Québec

Le gouvernement fédéral du Canada conditionne le financement des soins de santé à la conformité à un standard national. La politique publique de santé reste cependant une prérogative assez largement provinciale. Cette autonomie se traduit notamment par une grande diversité des politiques de santé au Canada. Cette section se limite au cas Québecois, d'où proviennent les données utilisées dans la partie empirique de notre analyse. L'introduction d'un mode de rémunération mixte, en 1999, constitue un changement profond dans les rémunérations des médecins du Québec, puisque la rémunération à l'acte était, jusqu'alors, le système le plus largement répandu.[12] Ces disposi-

[11] « *The greatest impediment to understanding physician behavior in Canada is the lack of data linking details of physician practice setting with individual and household physician income.* », Ferrall, Gregory & Tholl (1998, p.24).

[12] Emery, Auld & Lu (1999) proposent une synthèse très complète des différences institutionnelles entre les Etats de l'ensemble du Canada. On y trouvera également une discussion des motifs historiques et politiques qui ont conduit à la prédominance du mode de rémunération à l'acte, qui concerne 84% des médecins canadiens (Ferrall, Gregory & Tholl, 1998). Pour le cas Européen, voir, par exemple, la synthèse théorique d'Abel-Smith & Mossialos (1994) et la description des institutions du système de

tions institutionnelles gouvernent la contrainte budgétaire des médecins du Québec, le long de laquelle sont choisis les comportements de pratique optimaux. A titre d'analyse théorique préliminaire, la prochaine section propose un aperçu des effets attendus des instruments d'incitation sur ces choix.

2.1.1 Modes de rémunération et comportements de pratique : un survol

Comme nous l'avons souligné plus haut, une importante littérature empirique atteste de la sensibilité de la pratique des médecins au volume de rémunération. Ces travaux, qui empruntent pour la plupart aux méthodes d'analyse de l'offre de travail, reposent sur une simulation du salaire horaire basée sur le rapport entre le revenu total et le nombre d'heures travaillées. Les schémas de rémunération qui sont à l'origine de ce revenu total présentent cependant une très grande diversité. Ils sont traversés, en particulier, par la distinction importante établie par la littérature d'économie du travail (voir, par exemple, la synthèse de Lazear, 1995) entre rémunérations fixes et variables. Au-delà du montant du revenu, ces modes de rémunération influencent considérablement les choix de pratique. Cette section présente une courte synthèse des propriétés les plus connues de chacun d'entre eux.

Une rémunération variable consiste à verser un paiement proportionnel à une mesure de performance vérifiable. En matière médicale, les unités de mesure les plus couramment utilisées sont soit l'acte délivré – *rémunération à l'acte* – soit le nombre de patients traités – *capitation*. Bien que ces mesures de performances introduisent des différences importantes, ces modes de rémunération présentent donc tous deux les propriétés essentielles d'une rémunération à la pièce.[13] Ils en partagent par conséquent les qualités,

santé français fournie par M. Duriez (2000), *Le système de santé en France*, Rapport du Haut Comité de Santé Publique.

[13] A notre connaissance, le modèle de Selden (1990) est la seule analyse théorique explicitement

comme les défauts.

Les méthodes de rémunération à la pièce sont réputées pouvoir réconcilier les intérêts du principal (les autorités qui l'administrent dans le cas de la politique de santé) et ceux de l'agent (le médecin) par la dépendance de la rémunération sur la production (Prendergast, 1999). Lorsque la mesure de performance constitue une mesure adéquate des intérêts du principal et de l'activité de l'agent, ces schémas de rémunération permettent alors une amélioration de la performance (Lazear, 2000a). Dans le cas de la rémunération des médecins, l'étude de Hemenway, Killen, Cashman & *al.* (1990) établit ainsi que le passage à une rémunération à la performance, fondée en l'occurrence sur le revenu généré par l'activité du médecin pour son hôpital d'appartenance, permet d'augmenter significativement (12% ici) le nombre de patients traités. Au-delà des ces principes généraux, les spécificités de la pratique médicale peuvent en partie contrarier l'efficacité de ces modes de rémunération. Ces difficultés proviennent des asymétries d'information qu'engendrent les activités du médecin – avec l'organisme qui le rémunère comme avec le patient – et du caractère multi-dimensionnel de son activité.

Une première source d'asymétrie d'information provient de l'information cachée dont dispose le médecin dans sa relation avec le patient (Arrow, 1963). Le médecin est en effet le seul capable de juger à la fois de l'adéquation des soins aux affections dont souffre le patient et du diagnostic de ces affections elles-mêmes. Le médecin se trouve donc en position de manipuler la demande de soins en multipliant les prescriptions au-delà de ce qu'exige la préservation de la santé. En raison de cette "demande induite", la demande de soins qui s'adresse aux médecins est donc endogène (Evans, Parish & Sully, 1973). Dans ce cas, la mesure de performance sur laquelle est fondée la rémunération devient pour les médecins un instrument de maîtrise de leur propre revenu. Bien que ce mécanisme soit désormais théoriquement bien connu (voir, par exemple, De Jaegher

consacrée à la rémunération par capitation. Hutchison, Birch, Hurley & *al.* (1996) proposent une comparaison empirique entre ces deux types de rémunération à la pièce ; Gosden , Forland , Kristiansen & *al.* (2001) comparent l'efficacité de la capitation à celles de rémunérations fixes telles que le salaire.

& Jegers (2000) pour un modèle récent fondé sur le comportement du médecin), sa pertinence empirique a suscité d'intenses débats.[14]

Il semble cependant qu'un consensus se dégage pour admettre l'existence d'une demande induite,[15] compte tenu du faisceau convergent de confirmations empiriques qu'elle a reçues dans le cas du Canada (Schaafsma, 1994), de la France (Delattre & Dormont, 2003) et des Etats-Unis. Pour ce dernier cas, Gruber & Owings (1996) utilisent par exemple l'expérience naturelle offerte par le déclin de la fertilité pour évaluer la réaction des obstétriciens à une baisse exogène de la demande de soins. Les résultats confirment une induction de la demande de la part des médecins, puisque la baisse de la demande s'accompagne d'un déplacement des prescriptions vers des soins plus coûteux – donc plus rémunérateurs sous un régime de rémunération à la pièce – tels que les césariennes. La Norvège constitue à cet égard une exception persistante, puisque les travaux qui lui sont consacrés concluent systématiquement à l'absence d'induction de la demande (Carlsen & Grytten, 1998 ; 2000 ; Sørensen & Jostein, 1999 ; Grytten & Sørensen, 2001). Quoi qu'il en soit de son universalité, la demande induite constitue au minimum une éventualité que les médecins ajustent l'intensité de l'activité médicale

[14]Pour ne citer que les plus intenses, on pourra consulter à ce sujet les doutes émis par Feldman & Sloan (1988) et la réponse de Rice & Labelle (1989) ainsi que les débats qui opposent Labelle, Stoddart & Rice (1994a, 1994b) et Culyer & Evans (1996) à Pauly (1994a, 1994b).

[15]D'un point de vue méthodologique, le scepticisme le plus fondé ne peut qu'être ébranlé par la réaction de Fuchs (1986) à ces critiques, qui s'attend à ce que les «*economists will react to the study and the critique as they have in the past on this issue, with the fervant hope that maybe there is no inducement. This reaction has always reminded me of the story of the Frenchman who suspected that his wife was unfaithful. When he told his friend that the uncertainty was ruining his life, the friend suggested hiring a private detective to resolve the matter once and for all. He did so, and a few days later the detective came and gave his report : "One evening when you were out of town I saw your wife get dressed in a slinky black dress, put on perfume, and go down to the local bar. She had several drinks with the piano player and when the bar was closed they came back to your house. They sat in the living room, had a few more drinks, danced, and kissed." The Frenchman listened intently as the detective went on : "Then they went upstairs to the bedroom, they playfully undressed one another, and got into bed. Then they put out the light and I could see no more." The Frenchman sighed "Always that doubt, always that doubt".*

de façon à manipuler le revenu tiré de la pratique. Cette manipulation ne saurait avoir lieu si la rémunération était indépendante de l'intensité de l'activité, et seules les rémunérations à la pièce présentent par conséquent le risque d'y être sujettes (Grytten & Sørensen, 2001).

Outre le volume de soins délivrés, les médecins peuvent également manipuler les taux de rémunération à la pièce qui s'appliquent à ces soins. Dans la plupart des systèmes de santé, les taux de rémunération à la pièce varient en effet selon la nature des actes de façon à refléter leur difficulté de réalisation ainsi que leur efficacité en termes de santé. A cet égard, les médecins sont les seuls à connaître la nature des soins prodigués aux patients. La qualification des actes délivrés, qui sert de base à leur rémunération, est donc en général laissée à leur discrétion. Les médecins peuvent alors exploiter cette seconde information privée, dans la relation avec le principal qui décide de leur rémunération (Etat, direction du service, de l'hôpital,...), en falsifiant dans leurs déclarations les soins effectivement délivrés. Ce risque de sur-facturation (*billing-creep*) limite la possibilité de différencier les tarifs en fonction de la nature des actes (Evans, 1983). Il constitue donc une contrainte importante sur la capacité du principal à transmettre à l'agent, par l'intermédiaire des taux de rémunération, la hiérarchie de ses intérêts quant à l'importance des actes.

Compte tenu de la complexité de l'activité médicale, cette difficulté rend particulièrement délicate le choix des taux de rémunération à la pièce. La pratique médicale s'assimile en effet assez largement à une situation multitâche dont le diagnostic, la quantité de soins, la qualité des soins, le coût des soins et leur adéquation ou encore la gestion des établissements de santé sont autant de dimensions. Pour que la rémunération à la pièce soit efficace il convient alors que les taux de rémunérations relatifs reflètent la structure de priorités du principal (Holmstrom & Milgrom, 1991). Si des contraintes s'imposent à la différenciation des taux de rémunération, la rémunération à la pièce peut biaiser les choix de pratique dans un sens opposé à ses souhaits. Surtout, le système d'incitation encourage dans ce cadre l'abandon des activités pour lesquelles

il n'existe pas de mesure de performance vérifiable. Il est ainsi fondé théoriquement (Stiglitz, 1975) comme empiriquement (Paarsch & Shearer, 1999 ; Shearer & Paarsch, 2000) que les systèmes de rémunération à la pièce tendent à encourager la quantité au détriment de la qualité, plus difficilement mesurable. Bien que peu d'études empiriques soient consacrées à cette question, la pratique médicale ne semble pas faire exception à cette propriété (Jencks, Cuerdon, Burwen & *al.*, 2000).

En raison des nombreuses difficultés que posent la conception de rémunérations à la pièce, un principe de rémunération fixe leur est parfois préféré. Ce mode de rémunération consiste en général en un salaire constant, obtenu quel que soit le comportement de pratique dès lors que les termes du contrats (temps de présence dans l'établissement, participation aux réunions, ...) sont respectés. A l'inverse des rémunérations à la performance, ce schéma tend donc à déconnecter les choix de pratique de la rémunération. Elle laisse en conséquence les médecins libres de diversifier leurs activités, et rend non coûteux l'investissement en qualité. A partir de données d'enquête sur les médecins canadiens, Ferrall, Gregory & Tholl (1998) montrent ainsi que les médecins salariés consacrent 5.5 heures de plus par semaines à leur travail que les médecins rémunérés à la pièce, alors même qu'ils consacrent 5.9 heures de moins aux soins des patients.

A cette diversification s'ajoute un accroissement de l'attention consacrée aux soins, classiquement interprétée comme une mesure de qualité (Glazer & McGuire, 1993). La synthèse des résultats empiriques opérée par Gosden, Pedersen & Torgerson (1999) établit par exemple que la rémunération par un salaire est associée à des consultations plus longues ainsi qu'un nombre d'actes par patient et de patients par médecin moindre. L'envers de cette amélioration de la qualité reste cependant l'inévitable accroissement des coûts associé à un système de rémunération qui néglige la performance. Malgré les résultats de Gosden, Sibbald, Williams & *al.* (2003), qui montrent que le passage à une rémunération fixe a été sans effet sur la productivité des médecins généralistes en Angleterre, la très grande majorité des travaux empiriques soulignent la diminution dans le volume de soins associée à l'instauration d'un salaire fixe (Gosden, Forland,

Kristiansen & *al.*, 2001 ; Gaynor & Gertler, 1995).

Le choix entre rémunérations fixes et variables est donc gouverné par un arbitrage entre le volume de soins délivrés – éventuellement au-delà de ce que requiert l'amélioration de la santé – et la qualité de la pratique. Comme le soulignent Ma & McGuire (1997), cette tension entre les différents modes de rémunération peut s'interpréter comme une insuffisance du nombre d'instruments utilisés au regard du nombre d'objectifs poursuivis. Si l'efficience de la pratique médicale dépend à la fois de la quantité de soins et de leur qualité, il convient en effet que ce double objectif soit servi par au moins deux instruments. Un certain nombre de travaux se sont en conséquence tournés vers les modes de rémunération qui combinent des rémunérations fixe et variable.

Lorsque la demande de soins est excédentaire, Ma (1994) et Rogerson (1994) montrent ainsi que l'efficience, définie selon ces deux dimensions, nécessite que la rémunération soit une combinaison linéaire entre remboursement prospectif – enveloppe prévisionnelle, indépendante des soins effectifs – et remboursement des coûts. Bien qu'ils se dotent d'un instrument supplémentaire, ces modes de rémunération n'échappent pas aux difficultés liées aux asymétries d'information inhérentes à la pratique médicale. Même dans les cas où l'efficience l'exigerait, il est ainsi impossible d'utiliser une rémunération variable négative, au risque que le volume d'actes délivrés soit falsifié par le médecin – pour qui la rémunération variable devient un coût – en accord avec le patient – qui reste redevable de la partie des soins qui n'est pas couverte (Ma & McGuire, 1997). L'existence d'une partie variable maintient en outre la dépendance de la rémunération sur le volume de soins délivrés. Cette propriété perpétue par conséquent le risque de demande induite, bien que l'association à une rémunération fixe permette d'en diminuer l'importance (Levaggi & Rochaix, 2003).

Outre la combinaison de plusieurs instruments, offrir un menu de modes de rémunération alternatifs peut également permettre de renforcer l'efficacité des incitations par un effet de sélection (Encinosa, Gaynor & Rebitzer, 1997 ; Barro & Beaulieu, 2003).

La réponse des comportements de pratique aux incitations offertes dépend en effet de façon importante de caractéristiques individuelles inobservables, telles que la compétence (Dranove, 1988). Sous cette hypothèse d'hétérogénéité, offrir un menu de modes de rémunérations permet alors d'instaurer une auto-sélection des médecins (Demange & Geoffard, 2003) par laquelle le choix du mode de rémunération révèle ces caractéristiques inobservables.

La rémunération des médecins du Québec mobilise chacun des instruments décrits dans cette section. Les résultats théoriques présentés offrent donc un premier aperçu de leurs effets attendus et permettent, en particulier, d'évaluer l'adéquation du dispositif de rémunération mixte aux objectifs qui ont motivé son instauration.

2.1.2 Le règne de la rémunération à l'Acte

L'impulsion initiale de la rémunération mixte est née de la volonté de rééquilibrer les effets pervers de la rémunération à l'acte. La répartition des sources de revenu des médecins du Québec lève toute ambiguïté sur sa prédominance. La rémunération à l'acte représente en effet 80.57% des revenus des praticiens exerçant au Québec en 1996 et cette proportion reste stable jusqu'en 1999. Les 19.43% restants se répartissent entre les autres types de rémunération, très largement minoritaires, que sont les salaires (0.6%), les vacations (9.23%) – qui rémunérent des heures de travail ponctuelles dans un établissement – et les rémunérations provenant d'activités en laboratoire (9.6%).[16]

Contrairement au système américain où la rémunération des actes est un prix de marché qui varie selon les praticiens, celle-ci résulte, au Québec, de négociations entre le gouvernement provincial et les organisations professionnelles. Les prix sont donc exogènes du point de vue des médecins. Outre l'administration du mode de rémunération, le revenu des médecins du Québec a également fait l'objet de nombreuses mesures vi-

[16]La capitation n'est pas utilisée au Québec.

sant tant la réduction des coûts que l'amélioration des soins offerts à la population. Contrairement aux autres provinces, les ressortissants du Québec rencontrent en effet des barrières linguistiques et culturelles importantes à la mobilité. Cette caractéristique a permis la mise en oeuvre de mesures drastiques de maîtrise du revenu, qui expliquent en partie la faiblesse du revenu moyen des médecins Québecois en comparaison des autres provinces (Ferrall, Gregory & Tholl, 1998).

Le souci de maîtrise des coûts a en particulier conduit, dès 1976 et pour la première fois au Canada, à imposer un système de plafonnement des rémunérations. Une fois le montant mensuel du plafond atteint, ce système consiste à amputer les revenus de pratique de 75% de leur valeur. Le niveau des plafonds a fait l'objet d'ajustements constants, en réponse à l'évolution du pouvoir d'achat et aux spécificités des spécialités de pratique. Ainsi, les revenus de pratique en cabinet privé sont réduits de 35% (75% pour la radiologie diagnostique) avant application du plafond, afin de prendre en compte les charges liées aux frais professionnels. Pendant la période couverte par notre étude, les plafonds étaient fixés à 128 750$[17] par semestre pour toutes les spécialités à l'exception de la neurologie (142 000$), de l'endrocrinologie (103 500$), et de la pédiatrie (105 000$) jusqu'au premier trimestre 2001. Leur montant a ensuite été porté à 140 000$ pour toutes les spécialités à l'exception de la pédiatrie (115 000$). Pour l'ensemble des spécialités, les revenus provenant des services d'urgence étaient, jusqu'au premier trimestre 2000, exclus du revenu admissible au plafond. Cette mesure est étendue, depuis cette date, à l'ensemble de la pratique exercée en hôpital. Quoiqu'elles s'adaptent aux spécifités de pratique, ces dispositions peuvent être considérées comme très contraignantes. A titre de comparaison, les mesures de plafonnement instaurées en Ontario en 1993 réduisent seulement d'un tiers les revenus de pratique annuels qui dépassent 400 000$ (soit un plafond près de trois fois supérieur à celui qui s'applique à la plupart des spécialités).[18]

[17]Tous les montants monétaires sont exprimés en Dollars Canadiens.

[18]Ces dispositions de plafonnement des rémunérations sont peu utilisées ailleurs dans le monde. Les rares études qui leur sont consacrées tendent à montrer que l'effet des plafonds sur les choix des médecins s'apparente à celui qu'a une taxe sur le revenu des contribuables (Kralj, Kantarevic &

Au-delà de la maîtrise des coûts, les autorités québécoises ont également mené une politique active de réduction des inégalités régionales en termes de densité médicale. Dans ce but, un taux de rémunération différenciée déforme, depuis 1982, le prix payé pour les actes en fonction de différentes caractéristiques de pratique telles que la spécialité, la région administrative et la ville d'exercice. Cette mesure créé d'importants écarts de rémunération, puisque, s'il pénalise l'exercice dans les régions à forte densité médicale, le taux de rémunération différenciée accroît le prix payé dans les zones défavorisées. Ces distortions importantes dans les prix des services se sont avérées efficaces pour encourager l'installation dans les zones à faible densité médicale (Bolduc, Fortin & Fournier, 1996).

L'ensemble de ces mesures crée une déconnexion importante entre le revenu qui devrait résulter des choix de pratique – appelé *consommation potentielle* – et le revenu effectivement touché par les médecins, qui constitue leur *consommation effective*. Pour en tenir compte, notre approche consiste à modéliser l'ensemble de ces mesures. A ce titre, il faut noter que les mesures de plafonnement s'appliquent à l'ensemble du revenu après application du taux de rémunération différenciée. En notant $\widetilde{X}_i$ le revenu potentiel du médecin i[19] ; τ_i le taux de rémunération différenciée induit par ses caractéristiques individuelles ($\tau_i > 1$ dans les zones subventionnées, $\tau_i < 1$ sinon) et C_i le niveau du plafond au-delà duquel le revenu est diminué de 75%, la consommation effective, X_i, d'un médecin du Québec est donc :

$$X_i = min\left[\widetilde{X}_i, C_i\right] + max\left[0.25\ (\widetilde{X}_i - C_i), 0\right] + \tau_i \widetilde{X}_i \tag{2.1}$$

Ces mesures s'appliquent quelles que soient les dispositions qui président à la rémunération de la pratique. En particulier, elles s'appliquent dans les mêmes termes aux médecins rémunérés selon le mode de rémunération mixte, introduit en 1999.

Weinkauf, 2005). S'y ajoute cependant un effet de demande induite lié à la réduction du taux au-delà du plafond (Nassiri & Rochaix-Ranson, 2000).

[19] Afin de simplifier les expressions algébriques, nous omettons la dépendance des variables sur le temps aussi souvent que cela ne crée pas d'ambiguïté.

2.1.3 Le mode de Rémunération Mixte

Contrairement au système de rémunération à l'acte, qui se limite à rémunérer les soins délivrés selon un prix indexé sur la nature et la difficulté des actes, la rémunération mixte repose sur deux composantes : un taux partiel de rémunération à l'acte, qui rémunère les soins délivrés à un taux réduit (en comparaison du taux sous la rémunération à l'acte "pure", présentée ci-dessus) ; mais aussi un *per diem*, rémunération fixe assise sur un large éventail d'activités.

a) Objectifs et dispositions

L'instauration de la rémunération mixte, à compter du quatrième trimestre 1999, a fait l'objet d'une étroite collaboration entre les autorités provinciales et les organisations professionnelles. Les objectifs qui ont présidé à son instauration reflètent donc à la fois le souci de maîtrise des coûts et d'amélioration des soins et des exigences issues de l'expérience de pratique.

Cette réforme visait d'abord à donner aux médecins les moyens de consacrer une partie de leur temps aux charges administratives et à l'enseignement. Bien qu'elles constituent un élément essentiel du fonctionnement du système de santé – en termes de circulation de l'information sur les patients, de gestion des établissements et de transmission des connaissances aux nouvelles générations de médecins – ces activités sont en effet exercées à titre exclusivement bénévole sous la rémunération à l'acte. Un objectif connexe était donc de rétablir une certaine équité entre les médecins, qu'ils consacrent ou non une partie importante de leur temps à ces activités.

Pour tenir compte de cette hétérogénéité dans les choix de pratique, les autorités ont choisi d'assortir l'introduction de la rémunération mixte d'une assez grande flexibilité.

Plutôt qu'un nouveau mode de rémunération universel et obligatoire, la rémunération mixte est en effet une alternative à laquelle les praticiens peuvent librement adhérer. Ainsi, après 1999, les médecins dont l'activité est très largement consacrée aux activités cliniques peuvent choisir de conserver la rémunération à l'acte ; tandis que les praticiens qui privilégient les activités non cliniques peuvent librement opter pour la rémunération mixte.[20] Sous la rémunération mixte, ces activités sont en effet rémunérées par un salaire fixe, appelé *per diem*. Lorsque leurs heures de travail sont couvertes par un *per diem*, les médecins peuvent exercer diverses activités *admissibles* incluant l'enseignement, les activités administratives et les activités cliniques.[21]

Le *per diem* rémunère sans distinction toutes les activités admissibles, y compris les activités cliniques. Les activités cliniques présentent pourtant une très grande diversité, qui se manifeste, notamment, dans la difficulté des actes et dans le temps qui leur est consacré. Les organisations professionnelles ont donc fait valoir la nécessité d'une rémunération spécifique afin de garantir la continuité des soins. Dans ce but, une rémunération à l'acte s'ajoute au *per diem*.

Plutôt qu'une rémunération à taux plein, les autorités ont cependant choisit une rémunération à l'acte partielle, qui rémunère les actes pratiqués à un taux réduit en comparaison du taux en vigueur sous la rémunération à l'acte. Cette réduction du

[20]Plus précisément, l'adhésion à la rémunération mixte requiert l'unanimité des médecins appartenant à une unité médicale (en général un service). Nous sommes contraints d'ignorer cet aspect faute d'information sur l'établissement d'appartenance. Dans ce qui suit, nous faisons donc l'hypothèse que les médecins ont la possibilité de recourir au "vote par les pieds", et de changer de service en fonction de leurs préférences de rémunération. Cette hypothèse semble assez conforme à la pratique. Le vice-président du CMQ soulignait ainsi en Novembre 2000 que «*those specialties which depends on physicians spending large amounts of time with their patients and especially the university hospitals and pediatric hospitals are keen to adopt the new system*». (cité par S. Benady "Mixed payment a go in Quebec", *The Medical Post*, 7 Novembre 2000).

[21]La recherche, exclue des activités admissibles, constitue une exception importante. Elles sont considérées comme directement rémunérées par les établissements (en général les hôpitaux) où elles sont exercées.

taux des actes est destinée à améliorer la qualité des soins en accroissant le temps consacré à chacun. La combinaison du *per diem* à ce taux réduit devrait en outre participer à atténuer le risque qu'une induction de la demande accompagne cette baisse de rémunération.

En raison de ces dispositions, la rémunération mixte a attiré diversement les médecins en fonction de leur spécialité. Le Tableau 2.1 présente les taux d'adhésion à la rémunération mixte et les taux de rémunération des actes en fonction par spécialité en 2002, qui est la dernière année post-réforme de notre échantillon. Le taux de rémunération des actes est en moyenne diminué de 50% avec le passage à la rémunération mixte. La variété des types de pratique entre spécialités conduit cependant à une assez grande variabilité du taux de rémunération, oscillant entre 30% et 90%. L'adhésion à la rémunération mixte partage la population en deux sous-ensembles de tailles sensiblement égales (taux d'adhésion supérieur à 40% dans l'ensemble de la population), variant entre les spécialités de moins de 3% à près de 90%.

b) Consommation potentielle sous la rémunération mixte

Bien que la combinaison d'un *per diem* et d'un taux de rémunération à l'acte réduit soit la caractéristique essentielle de la rémunération mixte, ses modalités pratiques font intervenir de très nombreux critères qui complexifient considérablement son application. Cette section est consacrée à une description détaillée des dispositions adoptées, destinée à modéliser la consommation potentielle des médecins qui choisissent la rémunération mixte.

Lorsqu'un médecin choisit la rémunération mixte, un demi *per diem*, d'un montant D, lui est versé pour chaque tranche de $\overline{d} = 3.5$ heures de travail fournies. Le nombre maximum de demis *per diems* dont peut bénéficier un médecin est cependant limité à 28 par période de deux semaines, soit un *per diem* par jour ouvrable. Cette notion

Tableau 2.1 – Statistiques descriptives de la rémunération mixte

Spécialité	Taux d'adhésion (%)	Nombre total d'observations	Taux de rémunération moyen
Anesthésie-réanimation	55.0	122	0.5
Cardiologie	5.1	11	0.5
Chirurgie générale	36.5	121	0.6
Chirurgie orthopédique	13.4	84	0.7
Chirurgie plastique	3.2	21	0.6
Chirurgie thoracique	0.0	2	.
Dermatologie	28.6	44	0.5
Gastro-entérologie	5.9	34	0.8
Obstétrique-gynécologie	12.9	76	0.5
Pneumologie	20.3	24	0.8
Médecine interne	27.7	71	0.7
Physiatrie	64.6	16	0.5
Neuro-chirurgie	89.7	7	.
Neurologie	33.3	28	0.3
Ophtalmologie	6.9	40	0.6
Oto-rhino-laryngologie	21.3	37	0.6
Pédiatrie	65.2	120	0.3
Radiologie diagnostique	0.0	5	.
Radio-oncologie	79.3	13	0.8
Urologie	18.1	34	0.7
Chirurgie cardio-vasculaire	2.9	9	0.6
Néphrologie	18.7	19	0.3
Endocrinologie	42.7	28	0.3
Rhumatologie	62.8	19	0.3
Autres	69.6	235	0.4
Total	**40.9**	**1220**	**0.5**

Note. Le *taux d'adhésion* est mesuré par la proportion des individus d'une spécialité qui ont obtenu une partie de leur revenu sous la rémunération mixte. Le *taux de rémunération* correspond au rapport entre le taux de rémunération des actes (mesuré par l'indice de prix, voir Section 2.4.2) sous la rémunération mixte et le taux sous la rémunération à l'acte en 2002. La spécialité "Autres" regroupe des champs de pratique non reconnus par la Corporation professionnelle des médecins du Québec (CPMQ) telles que l'allérgie, l'immunologie clinique, l'anatomo-pathologie,... Les taux de rémunération sont manquants pour toute spécialité dont aucun professionnel n'a choisi la rémunération mixte.

est d'ailleurs au coeur du système, puisque seules les plages horaires d'une semaine traditionnellement ouvrée (du lundi au vendredi, de 7h à 12h et de 14h à 19h) sont admissibles au *per diem*.

En notant h le nombre d'heures de travail hebdomadaires d'un médecin qui choisit la rémunération mixte, le nombre de demis *per diems* versés par semaine est donc :[22]

$$N = \frac{\min\left\{floor\left(\frac{2.h}{d}\right), 28\right\}}{2} \tag{2.2}$$

Le montant versé dans le cadre d'un demi *per diem* est resté constant et fixé à $D = 300\$$ pendant l'ensemble de la période qui retient notre intérêt. Par la suite, il a été porté à $D = 308\$$ en Avril 2003, puis $335\$$ en Juillet 2003.

Outre le versement d'un salaire, la réclamation d'un *per diem* a par ailleurs des conséquences sur la rémunération des actes réalisés. Les actes sont en effet distingués selon qu'ils ont été pratiqués pendant des heures de travail couvertes par un *per diem*. Les actes réalisés en dehors d'un *per diem* (i.e. pendant des heures de travail non couvertes par un *per diem*) sont en effet rémunérés selon les conditions qui prévalent sous la rémunération à l'acte. Seuls les actes délivrés sous un *per diem* sont donc, à l'inverse, rémunérés à taux réduit. Ainsi, en notant P le prix versé pour un acte représentatif sous la rémunération à l'acte "pure", cet acte est rémunéré au prix P s'il est réalisé en dehors d'un *per diem* et au taux $(1-\alpha)P$, $\alpha < 1$, sinon. Le taux de réduction dans la rémunération des actes, α, dépend non seulement de la spécialité de pratique mais également de la nature des actes eux-même. Une nomenclature associe ainsi à chaque code d'acte un prix de rémunération à taux plein et un taux de réduction applicable si l'acte est réalisé sous un *per diem*.

[22] La fonction *floor* transforme un nombre décimal en sa partie entière. A titre d'illustration, on a : $floor(\frac{7}{2}) = floor(3.5) = 3$

Si la plupart des actes se voit attribuer un taux de réduction partiel, un certain nombre d'entre eux sont considérés comme étant rémunérés directement par le *per diem* et ne sont donc assortis d'aucune rémunération spécifique (on a donc $\alpha \in [0, 1]$). Cette caractéristique établit de fait une distinction importante entre les actes rémunérés sous la rémunération mixte – que nous appellerons *actes facturables*, notés AF – et les actes *non facturables*, notés ANF, pour lesquels les médecins n'ont aucune incitation sous la rémunération mixte.

Compte tenu de l'ensemble de ces mesures, la consommation potentielle d'un médecin varie donc considérablement selon le mode de rémunération choisi. On note d_i la variable binaire indiquant le mode de rémunération, $d_i = 1$ lorsque le médecin i a choisi la rémunération mixte. Pour tenir compte des différences de rémunération engendrées par le versement d'un *per diem*, nous distinguons les variables de pratique, $V = \{AF, ANF\}$, selon la période pendant laquelle elles ont été exercées. On note ainsi V^{RM} la variable de pratique V lorsqu'elle est réalisée sous un *per diem* et V^{RA} lorsqu'elle est réalisée en dehors d'un *per diem*. La consommation potentielle d'un médecin du Québec qui travaille W_i semaines par an s'écrit donc :

$$\begin{aligned}\widetilde{X}_i = d_i[W_i\ N_i\ D + (1-\alpha)P\ AF_i^{RM} + P\ (AF_i^{RA} + ANF_i^{RA})] \\ + (1-d_i)\ P\ (AF_i^{RA} + ANF_i^{RM})\end{aligned} \tag{2.3}$$

La consommation réelle d'un médecin – c'est à dire le revenu de pratique qui lui est effectivement versé – dépend des mesures de maîtrise des revenus qui s'appliquent à lui. La contrainte budgétaire des médecins du Québec correspond donc à l'ensemble des équations (2.1) à (2.3). Le Tableau 2.2 propose une synthèse des dispositions décrites dans cette section.

Tableau 2.2 – Rémunération des médecins du Québec considérés dans l'analyse

<table>
<tr><th>RA</th><th colspan="2">RM</th></tr>
<tr><td>Pas de rémunération fixe
Heures non cliniques (h^o)
non rémunérées</td><td>Demi
Per diem :</td><td>- Rémunère chaque tranche de 3.5 h en établissement
- Toutes les heures de pratique sont admissibles (h^c, h^o)
- Plafonné à 28 toutes les 2 semaines</td></tr>
<tr><td rowspan="4">Actes
rémunérés
au prix P</td><td>Actes
Facturables :</td><td>- Rémunérés à $(1-\alpha)P$ pendant les heures *per diem*
- Rémunérés au prix P en dehors des heures *per diem*</td></tr>
<tr><td>Actes
Non facturables :</td><td>- Non rémunérés pendant les heures *per diem*
- Rémunérés au prix P en dehors des heures *per diem*</td></tr>
<tr><td colspan="2">Rémunération différenciée sur critères géographiques</td></tr>
<tr><td colspan="2">Plafonnement des rémunérations†</td></tr>
</table>

†A l'exception des activités en urgence jusqu'en 2001, et de toutes les activités en hôpital depuis. Voir Section 2.4.2.

c) Réalisations : un premier aperçu

La rémunération mixte repose sur un salaire fixe et un taux réduit de rémunération des actes, dont la combinaison est destinée à encourager une augmentation des heures de travail non clinique et une diminution du nombre d'actes réalisés par heure. Dans la mesure où l'adoption de ce nouveau mode de rémunération est un choix volontaire des médecins, un effet de sélection peut cependant amplifier cet effet d'incitation. Si le passage à la rémunération mixte repose sur une différence systématique de préférence entre les médecins, on peut en effet s'attendre à ce que cette différence influence également leur réaction aux incitations fournies sous la rémunération mixte. Par exemple, il est raisonnable de penser que des médecins qui ont de fortes préférences pour les activités non cliniques réagissent à une baisse du taux de rémunération des actes par une forte diminution des activités cliniques.[23]

[23]Nous proposons une illustration graphique de cette intuition dans la Section 2.2.2.

TABLEAU 2.3 – STATISTIQUES DESCRIPTIVES DE L'EFFET DE LA RÉFORME

	Médecins ayant choisi la RM					
	Sous la RA		Sous la RM		Total	
	Moyenne	Ecart type	Moyenne	Ecart type	Moyenne	Ecart type
Heures hebdomadaires totales (h)	49.17	12.8	46.6	11.79	48.38	12.55
——— Cliniques (h^c)	41.39	13.5	40.03	12.73	40.98	13.28
——— Non-cliniques (h^{nc})	7.77	8.25	6.57	8.46	7.4	8.34
Semaines de travail (W)	45.55	4.62	45.43	4	45.52	4.44
Actes[a]	122.87	70.70	101.81	64.59	116.41	69.56
——— Non Facturables (ANF)	28.42	462.94	19.33	41.17	25.64	44.98
——— Facturales[b] (AF)	94.44	60.18	81.72	52.45	90.54	58.21
Revenu annuel[a] (X)	130.80	73.40	188.26	71.77	148.41	77.56

	Médecins constamment sous la RA					
	Avant la réforme		Après la réforme		Total	
	Moyenne	Ecart type	Moyenne	Ecart type	Moyenne	Ecart type
Heures hebdomadaires totales (h)	49.23	14.66	48.57	13.26	49.11	14.41
——— Cliniques (h^c)	41.97	15.22	43.33	14.03	42.22	15.01
——— Non-cliniques (h^{nc})	7.26	8.95	5.24	7.88	6.89	8.8
Semaines de travail (W)	45.15	5.29	45.18	3.88	45.16	5.06
Actes [a]	151.59	116.10	168.58	106.66	154.71	114.61
——— Non Facturables (ANF)	54.09	64.07	58.97	77.79	54.98	66.83
——— Facturables (AF)	97.50	104.23	110.67	92.74	99.92	102.34
Revenu annuel[a] (X)	165.42	96.05	224.02	117.71	176.18	102.91

[a]En milliers de Dollars. Les actes sont mesurés en Dollars constants (base 1996).

[b]Inobservables pour les médecins ayant choisit la RM lorsqu'ils sont pratiqués pendant des heures couvertes par un *per diem*. Voir la Section (2.3.2).

Note. Profil de pratique moyen des médecins du Québec sur la période 1996-1998 et 2002. *Moitié supérieure* : Médecins ayant obtenu une partie de leurs revenus sous la rémunération mixte pendant la période d'observations, avant (partie gauche) et après (partie droite) l'avoir adoptée. *Moitié inférieure* : Médecins dont 100% du revenu provient de la rémunéation à l'acte, avant (partie gauche) et après (partie droite) l'introdcution de la réforme.

Le Tableau 2.3 propose un premier aperçu de l'effet de la rémunération mixte sur les comportements de pratique.[24] Ces derniers sont décrits par le nombre d'heures hebdomadaires, consacrées respectivement aux activités cliniques (notées h^c) et non cliniques (h^{nc}), le nombre annuel de semaines de travail, le nombre d'actes pratiqués, distingués selon qu'ils sont facturables ou non sous la rémunération mixte, et, enfin, le revenu annuel. La moyenne et l'écart-type de chacune de ces variables sont calculés pour chacun des deux groupes de médecins créés par la réforme. La partie supérieure du tableau résume en effet le comportement de pratique des médecins qui sont passés à la rémunération mixte pendant notre période d'étude (1996-2002) ; la partie inférieure celui des médecins qui n'ont jamais abandonné la rémunération à l'acte pendant cette période. Avant l'introduction de la réforme, la comparaison entre les parties haute et basse du tableau permet donc d'apprécier l'ampleur de l'effet de sélection en comparant les choix des médecins selon leur groupe d'appartenance. A cet effet, nous nous intéressons, pour chaque groupe de médecins, au comportement de pratique sous la rémunération à l'acte (pendant la période qui précède l'introduction de la réforme pour les médecins qui sont restés à la rémunération à l'acte), sous la rémunération mixte (après la réforme pour ces mêmes médecins) et sur l'ensemble de la période. L'effet d'incitation émerge ainsi des comparaisons entre les deux premières colonnes du tableau.

En ce qui concerne l'effet de sélection, il semble que les médecins des deux groupes se distinguent moins par les heures de travail choisies que par la quantité d'actes pratiqués. Le nombre moyen d'heures hebdomadaires de travail (49.17 pour les médecins ayant choisit la rémunération mixte contre 49.23 pour ceux qui restent à la rémunération à l'acte) comme la répartition de ces heures entre les activités cliniques (41.39 contre 41.97) et non-cliniques (7.77 contre 7.26) sont en effet très similaires. A l'inverse, les médecins qui choisiront la rémunération mixte réalisent beaucoup moins d'actes que ceux qui restent à la rémunération à l'acte (122869 pour les premiers et 151591 pour les seconds). Alors que les actes facturables pratiqués sont très proches d'un groupe à l'autre (94440 contre 97503), cette différence se manifeste principalement dans les actes

[24] Les données utilisées pour construire ce tableau sont décrites dans la Section 2.4.1.

non facturables réalisés (28429 contre 54087). Il en résulte une importante différence de revenu, les médecins qui passeront à la rémunération mixte bénéficiant d'un revenu annuel significativement inférieur (130795$ contre 165422$).

Au total, les médecins qui choisiront la rémunération mixte se distinguent principalement par une faible quantités d'actes pratiqués, réalisant 18% d'actes de moins que les médecins qui conservent la rémunération à l'acte. Cette différence est presque entièrement imputable à un important écart de comportement en termes d'actes non-facturables. Sous les mêmes conditions d'incitation – la rémunération à l'acte – les médecins qui manifesteront leur préférence pour la rémunération mixte choisissent en effet une quantité d'actes facturables de 48% inférieure à celle des médecins qui la refusent. Ces résultats confirment l'existence d'un effet de sélection dans le passage à la rémunération mixte, fondé principalement sur des préférences divergentes vis-à-vis des actes pratiqués.

En utilisant les comportements de pratique moyens au sein de chaque groupe, un premier aperçu de l'effet d'incitation peut être obtenu par un estimateur de Différences en Différence (DD). Cet estimateur utilise le groupe de médecins qui ne sont pas affectés par la réforme comme un *groupe contrôle* des médecins qui passent à la rémunération mixte (*groupe traitement*). Sa validité repose donc sur l'hypothèse que le comportement des médecins du groupe contrôle reflète celui qu'auraient adopté les médecins du groupe traitement en l'absence de réforme (Heckman & Smith, 1995). L'effet de la réforme estimé correspond alors à la variation dans les différences de comportement, entre les médecins du groupe traitement et ceux du groupe contrôle, induite par la réforme. La dernière colonne du Tableau 2.4 présente l'estimateur DD et le t de Student associé pour chaque variable de pratique. Il correspond à la différence entre les deux premières colonnes du tableau, dans lesquelles sont calculées les différences entre les médecins des groupes contrôle et traitement avant et après la réforme.

La réforme semble rencontrer un succès mitigé en termes d'heures de travail. L'écart

Tableau 2.4 – Estimateurs de Différence en Différence

	Avant	**Après**	**DD**
Heures hebdomadaires totales	-.06 (-.231)	-1.97 (-4.838)	-1.9 (-3.66)
——— cliniques (h^c)	-.58 (-2.016)	-3.3 (-7.621)	-2.72 (-5.00)
——— non cliniques (h^{nc}	.51 (3.028)	1.33 (5.168)	.82 (2.53)
Semaines (W)	.4 (4.033)	.26 (2.071)	-.14 (-0.78)
Actes[a] totaux	-28721.63 (-13.796)	-66769.94 (-22.272)	-38048.32 (-9.72)
—— non facturables (ANF)	-25658.49 (-21.928)	-39631.22 (-18.485)	-13972.73 (-7.45)
——— facturables (AF)	-3063.13 (-1.646)	-28946.78 (-11.223)	-25883.65 (-6.04)
Revenu annuel[a] (X)	-34626.44 (-19.599)	-35756.15 (-10.796)	-1129.72 (-0.32)

[a]En milliers de Dollars. Les actes sont mesurés en Dollars constants (base 1996).

Note. *Deux premières colonnes* : Différence dans les profils de pratique moyens entre les médecins qui ont choisi la rémunération mixte (moitié supérieure du Tableau 2.3) et ceux qui sont restés à la rémunération à l'acte (moitié inférieure du Tableau 2.3), avant et après la réforme. *Dernière colonne* : Différence en Différence, *i.e.* différence entre les deux premières colonnes. Entre parenthèses, t de Student des différences de moyennes.

de comportement entre les médecins du groupe traitement et ceux du groupe contrôle se creuse après la réforme en termes d'heures de travail tant clinique que non-clinique. Pourtant, si cet écart se creuse positivement pour les heures non-cliniques, suggérant un effet positif de la réforme (qui incite les médecins à augmenter le temps consacré à ces activités de 0.8 heures), il est en effet négatif pour les heures de travail clinique. La réforme tend à engendrer une diminution de 2.7 heures du temps de travail consacré à cette activité. Il en résulte un effet net, sur les heures de travail totales, négatif, puisque les heures de travail clinique diminuent plus que les heures non-cliniques n'augmentent suite à la réforme.

Il n'est pas possible, à ce stade de l'analyse, de tirer des conclusions des observations

sur les choix d'actes non-facturables et, par conséquent, sur les choix d'actes totaux. Pendant les heures de travail couvertes par un *per diem*, les actes non-facturables sont en effet inobservables par définition. Les niveaux d'actes non-facturables présentés dans les tableaux proviennent donc, sous la RM, de la pratique exercée en dehors d'un *per diem* – sous la rémunération à l'acte avant la réforme, pendant des heures de travail rémunérées à l'acte après – et constituent, par conséquent, une borne inférieure des actes effectivement pratiqués. Au regard des actes facturables, observés en toutes circonstances, la réforme semble cependant avoir un fort impact négatif puisqu'elle provoque une diminution de 27% des actes pratiqués. Combinée à une diminution moins que proportionnelle des heures cliniques (6%), cette variation suggère un accroissement du temps consacré à chaque acte.

En résumé, le passage à la rémunération mixte s'appuie sur un effet de sélection important mais limité aux actes facturables. A heures de travail clinique identiques, les médecins qui choisissent la rémunération mixte pratiquent systématiquement moins de ces actes que ceux qui conservent la rémunération à l'acte. Si l'on interprète le temps consacré à chaque acte comme un critère de qualité des soins prodigués, la rémunération mixte tend donc à attirer des médecins qui valorisent relativement plus la qualité des soins. A cet effet de sélection s'ajoute un effet d'incitation, plus particulièrement marqué quant aux heures de travail et aux actes facturables. Les actes facturables connaissent une diminution importante suite au passage à la rémunération mixte. En termes d'heures de travail, enfin, la réforme influence à la fois la quantité d'heures travaillées et leur répartition entre les types d'activités. Si la rémunération mixte parvient à encourager une augmentation modérée des heures non-cliniques, elle se traduit simultanément par une diminution importante des heures de travail clinique. Il en résulte une diminution non négligeable des heures de travail totales.

Si les conditions qui assurent la validité de l'estimateur de Différence en Différence sont respectées,[25] ces résultats correspondent à l'effet de la réforme sur les médecins qui

[25]L'hypothése fondamentale est que seule la réforme est susceptible de modifier les différences de

ont choisi d'y adhérer, ou encore à l'effet de *traitement sur les traités* (Heckman, 1997). Si l'effet de la réforme est hétérogène – au sens où les caractéristiques inobservables des médecins influencent leur sensibilité à la réforme – ils offrent donc une compréhension assez restreinte de ses effets en se limitant à un sous-ensemble de la population. Pour y remédier, nous proposons une estimation structurelle des préférences des médecins, permettant de prédire la réaction de l'ensemble de la population aux variations des incitations. Ce modèle intègre en particulier les possibilités d'arbitrage entre la qualité et les quantités de soins délivrés. La prochaine section présente une version simplifiée du modèle (décrit en détail dans la Section 2.3) qui met en évidence les conséquences de cet arbitrage.

2.2 Analyse théorique du passage à la Rémunération Mixte

Comme l'a fait apparaître la synthèse proposée dans la Section 2.1.1, le choix entre une rémunération fixe et une rémunération variable se résume assez largement à un arbitrage entre l'intensité de l'activité médicale (qui assure une allocation efficace des ressources consacrées aux soins de santé) et la qualité des soins délivrés (qui détermine l'amélioration de la santé permise par un volume donné de moyens). Le recours à des modes de rémunération mixtes tente de tirer parti de chacun de ces avantages. Le modèle présenté dans cette section en évalue la possibilité en intégrant dans l'analyse l'arbitrage que réalisent les médecins entre marges intensive (qualité) et extensive (quantité) en réaction aux variations dans les incitations.

comportement entre les groupes contrôle et traitement. Voir Bertrand, Duflo & Mullainathan (2004) pour une présentation critique, qui montre en outre que les écart-types estimés par cette méthode sont non-convergents en présence d'auto-corrélation.

2.2.1 Modélisation du comportement des médecins

L'hypothèse retenue quant aux déterminants du comportement des médecins oriente de façon cruciale l'analyse théorique de la sensibilité des choix de pratique aux incitations offertes. Le choix de la fonction objectif peut en effet déterminer entièrement le changement dans les comportements de pratique induit par une variation des incitations. [26] Ainsi, comme le soulignent McGuire & Pauly (1991), l'hypothèse traditionnelle de maximisation du profit prédit de façon certaine une réduction des actes pratiqués en réponse à une diminution de leur taux de rémunération. A l'inverse, l'hypothèse de revenu-cible – selon laquelle les médecins ajustent leurs choix de pratique de façon à maintenir leur revenu à un niveau désiré – implique un accroissement du nombre d'actes suite à une réduction du taux de rémunération (Rice, 1983). Au regard de ces prédictions, la maximisation d'utilité constitue une hypothèse très générale puisqu'elle se réduit soit à la maximisation du profit, soit à la recherche d'un revenu-cible, en fonction de l'importance relative des effets revenu et substitution (McGuire & Pauly, 1991). Nous nous en remettons donc aux comportements observés pour discriminer entre ces hypothèses de comportement, et nous analysons les comportements de pratique qui résultent de la maximisation d'utilité des médecins.

Afin de simplifier l'analyse, nous nous limitons aux choix de pratique journaliers. Les activités cliniques sont décrites par le nombre d'heures qui leur sont consacrées par jour, h^c,[27] et le nombre d'actes, A, pratiqués.[28] Le nombre d'actes par heure, $e = A/h^c$, est donc une variable endogène du modèle, qui peut être interprétée comme une forme d'effort. A nombre d'heures cliniques donné, accroître le nombre d'actes

[26] McGuire (2000) propose une discussion très complète des résultats existants quant aux motivations des médecins.

[27] A la différence des autres sections, toutes les notations font référence, ici, aux choix de pratique journaliers.

[28] L'objectif de l'analyse est de mettre en évidence l'influence sur les choix de pratique de l'arbitrage entre marges intensive et extensive. Nous négligeons donc la distinction entre actes facturables et non-facturables, qui concerne les profils de substitution au sein de la marge intensive.

pratiqués requiert en effet une plus grande rapidité d'exécution et correspond donc à une augmentation de l'intensité des heures de travail. Simultanément, cette augmentation conduit également à une diminution du temps consacré à chaque acte ainsi qu'à une augmentation des soins prodigués à la population. Ces multiples effets sont résumés par la fonction de production du médecin, que nous supposons être le niveau de santé atteint par la population, s. Interprétant le temps consacré à chaque acte comme un critère de qualité des soins, nous supposons en effet que la santé est une fonction décroissante de l'effort. A effort donné, le nombre d'actes est un indicateur de la quantité de soins prodigués et est donc supposé accroître la santé. Au total, la fonction de production des médecins est donc : $s = s(\underset{+}{A}, \underset{-}{e})$.

Adoptant une hypothèse devenue classique dans la littérature (Dranove, 1988 ; Rochaix, 1989), nous supposons que les activités cliniques affectent le bien-être des médecins par l'intermédiaire de cette fonction de production. Cette hypothèse peut s'interpréter comme une norme éthique, par laquelle les médecins internalisent l'objectif d'amélioration de la santé dans la population (Arrow, 1963 ; Evans, 1974).

Le temps journalier laissé libre par les activités cliniques, $T - h^c$, est réparti entre le loisir pur, l, et les heure de travail non-clinique, h^{nc}. Ces dernières recouvrent l'ensemble des activités, telles que l'enseignement ou les tâches administratives, qui ne sont pas rémunérées sous la rémunération à l'acte. Malgré cette absence d'incitations, les médecins consacrent de fait une partie de leur temps aux activités non-cliniques sous la rémunération à l'acte (voir le Tableau 2.3). Cette observation révèle donc un goût pour les activités non-cliniques, au sens où celles-ci accroissent l'utilité (en diminuant la pression des pairs, en accroissant la satisfaction au travail, ...) alors même qu'elles laissent la consommation, X, inchangée. Cet aspect est pris en compte en modélisant les les heures de travail non-clinique comme une forme particulière de loisir.

Au total, la fonction d'utilité des médecins s'écrit donc : $U = U(X, l, h^{nc}, s)$. Au Québec – comme dans de nombreux pays industrialisés – les soins de santé sont très

largement pris en charge par les institutions de mutualisation des risques telles que l'assurance sociale, les mutuelles, etc... Dans ce contexte, la demande de soins de santé est déconnectée des variations de prix, et ne résulte que des besoins réels de la population. Nous supposons par conséquent que la sensibilité des choix de pratique aux variations de rémunération résultent exclusivement de la maximisation d'utilité sous contrainte budgétaire, et que les médecins peuvent donc, en particulier, allouer librement leur temps entre les différents types d'activités.

La contrainte budgétaire dépend à la fois des variables de pratique et du mode de rémunération adopté. On note ainsi y le revenu obtenu indépendamment du nombre d'actes pratiqués, correspondant donc au *per diem* sous la rémunération mixte[29], $y = D$, et fixé à $y = 0$ sous la rémunération à l'acte. Le revenu hors-travail – non affecté par la réforme – est supposé être nul. La contrainte budgétaire journalière d'un médecin peut donc être schématiquement résumée par $X = fA + y$, où f désigne le taux de rémunération des actes, égal à P sous la rémunération à l'acte et $(1-\alpha)P$ sous la rémunération mixte, $\alpha \in [0,1]$. Les variables décrivant les comportements de pratique – en ignorant les solutions de coin – sont donc les solutions du programme :

$$\begin{aligned} \underset{\{X,l,h^o,A,e\}}{Max} \quad & U = U\left(X, l, h^{nc}, s\left(A, e\right)\right) \\ s.c. \quad & (i) \quad T = h^{nc} + l + h^c \\ & (ii) \quad A = eh^c \\ & (iii) \quad X = fA + y \end{aligned}$$

En remplaçant les heures cliniques et le nombre d'actes par les valeurs imposées respectivement par les contraintes (i) et (ii), la fonction d'utilité des médecins s'écrit encore $U\left(X, l, h^o, s\left((T - h^o - l)e, e\right)\right)$ soit, en forme réduite : $\widetilde{U}\left(X, l, h^o, e\right)$. Les variables de pratique laissées libres par les contraintes techniques qui s'imposent aux choix sont alors

[29] Nous supposons que les heures de travail journalières respectent la condition $h^c + h^{nc} \geq \bar{h}$, où $\bar{h}$ est le nombre d'heures ouvrant droit à un *per diem* – 7h dans le dispositif actuel, voir Section 2.1.3.

solution de :

$$\begin{array}{ll} \underset{\{X,l,h^{nc},e\}}{Max} & \widetilde{U}\left(X,l,h^{nc},e\right) \\ s.c. & X - fe\left[T - h^{nc} - l\right] = y \end{array} \tag{2.4}$$

Compte tenu de la description des comportements de pratique retenue, l'arbitrage traditionnel entre consommation et loisir est donc généralisé pour inclure deux types de loisir (l, h^{nc}) ainsi que l'intensité des heures de travail (e). En raison de cette dernière propriété, la contrainte budgétaire que nous étudions est non-linéaire dans les variables endogènes. Plus précisément, les prix des variables de pratique sont eux-mêmes endogènes puisque, par exemple, le prix qui rémunère les heures de travail clinique (fe) dépend du niveau d'effort exercé pendant ces heures. Cette caractéristique nécessite de mettre en oeuvre des outils d'analyse spécifiques, qui permettent notamment d'étudier l'influence du mode de rémunération sur l'arbitrage entre qualité (effort) et quantités (heures et nombre d'actes). Cette approche sera présentée dans la Section 2.2.3. Nous nous limitons dans un premier temps à analyser l'ajustement des marges extensives aux modes de rémunération, en supposant un effort constant. Bien qu'il nécessite d'être réévalué pour intégrer l'endogénéité des choix d'effort, ce modèle apporte en effet une première compréhension des effets de la réforme et des mécanismes à l'oeuvre dans la réponse optimale aux changements de rémunération.

2.2.2 Quantités optimales : analyse du modèle à effort exogène

Si l'effort est supposé constant (et normalisé, $e = 1$) le programme d'optimisation (2.4) ne fait plus intervenir que les heures de travail :

$$\begin{array}{ll} \underset{\{X,l,h^{nc}\}}{Max} & \widetilde{U}\left(X,l,h^{nc}\right) \\ s.c. & X - f\left[T - h^{nc} - l\right] = y \end{array} \tag{2.5}$$

et les choix optimaux qui en résultent s'écrivent comme des fonctions des paramètres

du mode de rémunération : $h^{nc} = h^{nc}(P^c, y)$, $l = l(P^c, y)$.[30]

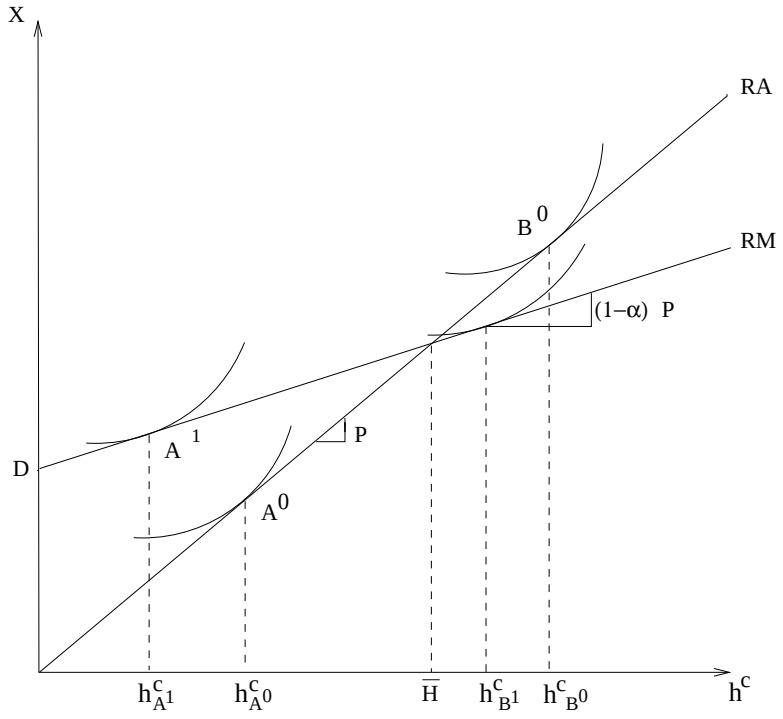
Cette hypothèse simplificatrice permet donc d'éliminer les problèmes de non-linéarité de la contrainte budgétaire liés à l'endogénéité des prix. Elle permet également d'obtenir une illustration graphique simplifiée des déterminants du passage à la rémunération mixte, fournie dans le Graphique 2.1. Lorsque seule la rémunération mixte existe, la contrainte budgétaire est la droite $\boldsymbol{RA}$, de pente P et passant par l'origine. La réforme introduit une seconde contrainte budgétaire – droite $\boldsymbol{RM}$ – dont l'ordonnée à l'origine correspond au *per diem*, D. En raison de la réduction de taux qui s'applique à la rémunération des actes (α), la pente de cette seconde contrainte est inférieure à la première. Les courbes se croisent donc pour un niveau donné d'heures de travail, noté $\overline{H}$.[31] Pour tout niveau des heures de travail en deçà de $\overline{H}$, le revenu sous la rémunération mixte domine donc strictement celui qui aurait résulté du mode de rémunération à l'acte. A l'inverse, le revenu est systématiquement supérieur sous la rémunération à l'acte pour tous les choix de pratique par lesquels les heures de travail excèdent $\overline{H}$. Comme l'adhésion à la rémunération mixte est un choix volontaire des médecins, ces contraintes budgétaires ne sont pas mutuellement exclusives. Les deux segments que nous venons de décrire – droite $\boldsymbol{RM}$ avant $\overline{H}$, $\boldsymbol{RA}$ au delà – constituent donc, ensemble, la contrainte budgétaire efficiente des médecins après la réforme. Cette combinaison de deux droites de pentes différentes est à l'origine d'une seconde non-linéarité dans la contrainte budgétaire.

Ces segments déterminent aussi, et surtout, la décision de passer à la rémunération mixte après la réforme. Les médecins qui décident d'adopter la rémunération mixte sont en effet ceux dont les choix optimaux, après la réforme, se situent sur le premier segment – droite $\boldsymbol{RM}$, avant $\overline{H}$ – de la contrainte budgétaire efficiente. Etant donné que les choix optimaux dépendent des préférences, puisqu'ils résultent d'un programme

[30] Les conditions du premier ordre (CPO) du programme à effort exogène sont omises dans cette section. Elles sont identiques aux CPO du programme à effort endogène, présentées en (2.10), après substitution de la normalisation $e = 1$.

[31] Formellement, ces heures de travail sont telles que le revenu sous la rémunération à l'acte et le revenu sous la rémunération mixte concordent, c'est à dire : $D = \alpha\ P\ \overline{H}$.

Graphique 2.1 – Passage à la rémunération mixte



de maximisation de l'utilité, le passage à la rémunération mixte repose donc sur une auto-sélection des médecins. A titre d'illustration, les préférences de deux médecins-types sont représentées sur le Graphique 2.1. Avant l'introduction de la réforme, le médecin ***A*** comme le médecin ***B*** choisissent les heures de travail qui maximisent leur utilité le long de la contrainte budgétaire ***RA***, matérialisées respectivement par les points $\mathbf{A_0}$ et $\mathbf{B_0}$. Leurs préférences sont telles, cependant, que le premier choisit des heures de travail inférieures à $\overline{H}$, tandis que le second se trouve au delà de $\overline{H}$. Après la réforme, les médecins choisissent leurs heures de travail le long de la contrainte budgétaire efficiente. Pour le médecin ***A***, la possibilité de choisir la rémunération mixte permet une amélioration de bien-être, qui le conduit du point $\mathbf{A_0}$ au point $\mathbf{A_1}$. Il fera donc partie des médecins qui décident d'adopter la rémunération mixte, et une réduction des heures de travail clinique accompagne cette adoption. A l'inverse, le passage du point $\mathbf{B_0}$ au point $\mathbf{B_1}$ correspondrait, pour le médecin ***B***, à une diminution d'utilité. On s'attend donc à ce que ses choix de pratique restent inchangés après la réforme.

Comme le suggère cet exemple graphique, l'adoption de la rémunération mixte par un médecin révèle donc en partie ses préférences à l'égard de la pratique médicale. En notant $\Delta V|_{RM}$ la variation de la variable V engendrée par le passage à la rémunération mixte, on a donc pour tous les médecins dont la pratique est affectée par la réforme : $\Delta\widetilde{U}\Big|_{RM} > 0$. Cette propriété simplifie considérablement l'analyse théorique de l'effet du passage à la rémunération mixte sur les choix de pratique. En notant $\widetilde{V}$ la demande Hicksienne de la variable V et $E(f, \widetilde{U})$ la fonction de dépense, les variations respectives des heures de travail totales et des heures de travail non cliniques peuvent en effet être déduites des expressions suivantes :

$$\Delta h|_{RM} \approx \frac{\partial \widetilde{h}}{\partial f}(-\alpha\ P) + \frac{\partial h}{\partial y} E_u\ \Delta\widetilde{U}\Big|_{RM} \tag{2.6}$$

$$\Delta h^{nc}|_{RM} \approx \frac{\partial \widetilde{h}^{nc}}{\partial f}(-\alpha\ P) + \frac{\partial h^{nc}}{\partial y} E_u\ \Delta\widetilde{U}\Big|_{RM} \tag{2.7}$$

Preuve Les demandes d'heures de travail non clinique et de loisir s'écrivent en fonction des paramètres de rémunération : $l(f, y), h^{nc}(f, y)$. En utilisant la contrainte d'allocation du temps (i), on a donc : $h^c = T - l(f, y) - h^{nc}(f, y) = h^c(f, y)$. Les heures de travail totales correspondent à la somme du temps consacré au travail clinique et non-clinique et on obtient de la même façon : $h = h(f, y)$. Par définition de la fonction de dépense, la demande d'heures de travail s'écrit encore : $h(f, y) = h(f, E(f, \widetilde{U}))$. Soit $\widetilde{U}_C$ l'utilité optimale atteinte sous le mode de rémunération C. La variation induite par le passage à la rémunération mixte correspond alors à la différence : $\Delta h|_{RM} = h\left[(1-\alpha)\ P\ ,\ E\left((1-\alpha)\ P\ ,\ \widetilde{U}_{RM}\right)\right] - h\left[P\ ,\ E\left(P\ ,\ \widetilde{U}_{RA}\right)\right]$. Autour de l'équilibre, cette quantité peut être approximée par l'expression :

$$\Delta h|_{RM} \approx \frac{\partial h}{\partial f}\Delta f + \frac{\partial h}{\partial y}\left[E_f \Delta f + E_{\widetilde{U}}\right] \Delta\ \widetilde{U}\Big|_{RM}$$

où E_i indique la dérivée première de la fonction de dépense par rapport à l'argument i. Par ailleurs, la décomposition de Slutsky met en évidence la combinaison des effets revenu et substitution selon l'expression : $\dfrac{\partial h}{\partial f} = \dfrac{\partial \widetilde{h}}{\partial f} - \dfrac{\partial h}{\partial y}\ f$. On sait en outre, par le lemme de Shephard, que : $E_f = f$. Par ailleurs, la variation du taux de rémunération des actes dans le passage à la rémunération mixte correspond au taux de réduction, soit : $\Delta f = (1-\alpha)\ P - P = -\alpha\ P$. Par substitution, l'expression de la variation des heures de travail s'écrit donc :

$$\Delta h|_{RM} \approx \frac{\partial \widetilde{h}}{\partial f}(-\alpha\ P) + \frac{\partial h}{\partial y} E_u \Delta\ \widetilde{U}\Big|_{RM}$$

L'approximation de la variation des heures non-cliniques s'obtient de la même façon. ■

Sachant que la variation d'utilité est positive, l'hypothèse de normalité des loisirs est alors suffisante pour prédire l'effet de la rémunération mixte sur les heures de travail totales. Si l'on suppose, en outre, que consommation et loisir sont des substituts nets, le modèle à effort exogène permet de prédire sans ambiguïté l'effet de la rémunération mixte sur les choix de pratique, résumé dans la Proposition 2.1.

Proposition 2.1. *Si l'effort (nombre d'actes par heure) est constant, le passage à la rémunération mixte devrait :*

- *Diminuer les heures de travail totales, si les loisirs sont des biens normaux ;*
- *Augmenter les heures de travail non clinique et diminuer les heures de travail clinique, si loisir et consommation sont des substituts nets.*

Preuve Le signe de la variation des heures totales s'obtient directement à partir de l'expression (2.6). Si le loisir est un bien normal, on a : $\frac{\partial h}{\partial y} < 0$. Par définition de la fonction de dépense et sachant que l'utilité s'accroît dans le passage à la rémunération mixte, le second terme du membre de droite est donc négatif. Par définition, la demande Hicksienne est décroissante de son prix. La rémunération des actes étant le prix du loisir, on a donc : $\frac{\partial \widetilde{h}}{\partial f} > 0$. Puisque le prix et le taux de réduction sont positifs, le premier terme du membre de droite est par conséquent négatif et : $\Delta h|_{RM} < 0$.

Les fonctions de demande Hicksiennes sont homogènes de degré 0. L'équation d'Euler de la demande Hicksienne d'heures de travail non clinique est donc :

$$\frac{\partial \widetilde{h}^{nc}}{\partial P_{nc}}.P_{nc} + \frac{\partial \widetilde{h}^{nc}}{\partial P_l}.P_l + \frac{\partial \widetilde{h}^{nc}}{\partial P_X}.P_X = 0$$

L'hypothèse que loisir et consommation sont des substituts nets se traduit alors par : $\frac{\partial \widetilde{h}^{nc}}{\partial f} < 0$. Dans l'expression (2.7), le premier terme est donc positif. Si les heures non cliniques – modélisées comme une forme de loisir – sont un bien normal, le second terme est également positif et $\Delta h^{nc}|_{RM} > 0$.

Les heures totales de travail incluent les heures cliniques et non cliniques. La contrainte d'allocation du temps journalier peut donc s'écrire en termes d'heures totales comme : $T = h + l$, et donc : $\Delta l|_{RM} = -\Delta h|_{RM} > 0$. Ainsi, la décomposition de la contrainte de temps selon les différents types de temps de travail ($T = h^c + h^{nc} + l$) permet de prédire le signe des heures cliniques : $\Delta h^c|_{RM} = -\Delta l|_{RM} - \Delta h^{nc}|_{RM} < 0$. ■

Si l'effort est, comme c'est en général le cas dans la littérature théorique, supposé constant quel que soit le mode de rémunération, l'analyse prédit donc un succès mitigé à la réforme. En particulier, la diminution des heures totales de travail constitue un grave effet pervers au regard des problèmes de liste d'attente rencontrés au Québec. Comme le montre le Tableau 2.3 le nombre d'actes par heures a cependant subi d'importantes variations suite à la réforme. La prochaine section propose donc une extension du modèle, capable d'intégrer les ajustements en termes de marge intensive.

2.2.3 Arbitrage qualité/quantités

Nous adoptons ici une démarche identique à celle utilisée dans le modèle à effort exogène. Afin de mettre en évidence la substitution entre les marges intensive et extensive, nous nous intéressons cependant au programme d'optimisation sous contrainte non-linéaire (2.4), dans lequel l'effort – donc les prix – est une variable endogène.

Pour chaque variable de pratique β, $\beta = \{X, l, h^{nc}, h^c, e\}$, les conditions du premier ordre du programme définissent les demandes optimales comme une fonction implicite des paramètres du mode de rémunération : $\beta = \beta(f, y)$. Le nombre d'actes optimal s'en déduit par la contrainte technologique (ii) : $A(f, y) = h^c(f, y)\ e(f, y)$. Comme précédemment, l'effet du passage à la rémunération mixte sur les choix de pratique peut être décomposé entre effets substitution et revenu :

$$\Delta\beta|_{RM} \approx \frac{\widetilde{\beta}}{\partial f}\Delta f + \frac{\partial\beta}{\partial y}E_U\ \Delta\widetilde{U}\Big|_{RM}$$

Preuve Pour les médecins qui choisissent de l'adopter, l'impact de la rémunération mixte sur les choix optimaux correspond à : $\Delta\beta|_{RM} = \beta(f_{RM}, y_{RM}) - \beta(f_{RA}, y_{RA})$. Pour tout système de rémunération C ($C \in \{RA, RM\}$), la fonction de dépense correspond par définition à : $y_C = E(f_C, \widetilde{U_C})$. L'effet du passage à la rémunération mixte peut donc être approximé par l'expression :

$$\Delta\beta|_{RM} \approx \frac{\partial\beta}{\partial f}\Delta f + \frac{\partial\beta}{\partial y}\left[E_f \Delta f + E_U \left.\Delta\widetilde{U}\right|_{RM}\right] = \left[\frac{\partial\beta}{\partial f} + \frac{\partial\beta}{\partial y}E_f\right]\Delta f + \frac{\partial\beta}{\partial y}E_U \Delta \left.\widetilde{U}\right|_{RM} \tag{2.8}$$

Blomquist (1989) propose une analyse systématique de la théorie du consommateur sous contrainte non-linéaire. En particulier, le lemme de Shephard et la décomposition de Slutsky peuvent être facilement adaptés à ce cas.

En adoptant les notations proposées par l'auteur, soit $g(f, \beta) = X - f.e\,[T - h^o - l]$ la contrainte budgétaire. Le lemme de Shephard dans le cas non-linéaire s'écrit : $\dfrac{\partial E}{\partial f} = \dfrac{\partial g}{\partial f} = g'_f$. Définissant les demandes Hicksiennes, $\widetilde{\beta}(f, U)$,comme les solutions du programme de minimisation de la dépense pour un niveau d'utilité U donné, l'équation de Slutsky devient par ailleurs : $\dfrac{\partial\beta}{\partial f} = \dfrac{\partial\widetilde{\beta}}{\partial f} - \dfrac{\partial\beta}{\partial y}.g'_f$.

Ensemble, ces résultats impliquent donc que : $\dfrac{\partial\beta}{\partial f} + \dfrac{\partial\beta}{\partial y}.E_f = \dfrac{\partial\widetilde{\beta}}{\partial f}$, d'où provient le résultat par substitution dans (2.8). ■

a) Effets revenu

La modélisation que nous avons adopté suggère un certain nombre d'hypothèses quant à la forme de la fonction d'utilité qui permettent de circonscrire les signes des effets revenus.

D'une part, notre analyse considère l'effort comme un critère de qualité des soins dispensés, puisqu'il constitue un indicateur du temps consacré aux soins et à leur explication, de la vigilance du médecin, de la pertinence du diagnostic, etc. Pour toutes ces raisons, le niveau de santé atteint par les patients est supposé décroissant de l'effort (voir Section 2.2.1). Dans la mesure où le niveau de santé affecte négativement le bien-être des médecins, l'effort devrait donc apparaître comme un mal dans leur fonction d'utilité.

D'autre part, un nombre significatif de médecins consacrent une partie importante de leur temps aux activités non-cliniques (Tableau 2.3) et ce y compris lorsque, comme sous la rémunération à l'acte, celles-ci ne donnent lieu à aucune rémunération. Cette caractéristique nous a conduit à considérer les heures de travail non-cliniques comme une forme particulière de loisir, au sens où elles participent à accroître le bien-être des médecins. Prenant acte des résultats obtenus par la plupart des travaux empiriques consacrés à l'offre de travail (Pencavel, 1986), nous supposons que tous les loisirs, le loisir pur comme les heures de travail non-clinique, sont des biens normaux. Sous ces hypothèses, l'effet de l'introduction de la rémunération mixte sur les variables de pratique dépend alors exclusivement de la sensibilité des demandes compensées aux variations de prix.

***Lemme* 2.1.** *Si :*

- *Les deux types de loisir sont des biens normaux ;*
- *L'effort est un mal ;*

alors la sensibilité des choix de pratique aux variations induites par la rémunération mixte ne dépend que de la sensibilité au prix des demandes Hicksiennes.

Preuve Nous utilisons ici les résultats présentés dans la preuve de la Proposition 2.1. On a ainsi : $E_U > 0$ par définition de la fonction de dépense, $\Delta\widetilde{U}\Big|_{RM} > 0$ en raison du passage volontaire à la RM et $\Delta f|_{RM} = -\alpha P \leq 0$. Les hypothèses présentées dans le Lemme 2.1 fournissent en outre la forme des effets revenu.

Si l'effort est un mal, sa demande non-compensée est décroissante du niveau de revenu. On a alors :

$$\Delta e|_{RM} \approx \frac{\partial \widetilde{e}}{\partial f} \underbrace{\Delta f}_{<0} + \underbrace{\frac{\partial e}{\partial y}}_{<0} \underbrace{E_U}_{>0} \underbrace{\Delta\, \widetilde{U}\Big|_{RM}}_{>0}$$

Une condition suffisante pour que l'effort diminue avec le passage à la rémunération mixte est donc que : $\dfrac{\partial \widetilde{e}}{\partial f} > 0$.

Si les deux types de loisir sont des biens normaux, on a :

$$\Delta l|_{RM} \approx \frac{\partial \widetilde{l}}{\partial f}\underbrace{\Delta f}_{<0} + \underbrace{\frac{\partial l}{\partial y}}_{>0}\underbrace{E_U}_{>0}\underbrace{\Delta\, \widetilde{U}\Big|_{RM}}_{>0}, \text{ et } \Delta h^{nc}|_{RM} = \frac{\partial \widetilde{h}^{nc}}{\partial f}\underbrace{\Delta f}_{<0} + \underbrace{\frac{\partial h^{nc}}{\partial y}}_{>0}\underbrace{E_U}_{>0}\underbrace{\Delta\, \widetilde{U}\Big|_{RM}}_{>0}$$

et $\dfrac{\partial \widetilde{l}}{\partial f} < 0, \dfrac{\partial \widetilde{h}^{nc}}{\partial f} < 0$ sont donc des conditions suffisantes pour que les deux types de loisir augmentent simultanément dans le passage à la rémunération mixte.

Par la contrainte d'allocation du temps (ii), la sensibilité des heures de travail totales au revenu hors-tavail est : $\dfrac{\partial h^c}{\partial y} = \dfrac{\partial}{\partial y}\left[T - l - h^{nc}\right] = -\dfrac{\partial l}{\partial y} - \dfrac{\partial h^{nc}}{\partial y} < 0$. L'effet du passage à la rémunération mixte peut donc être résumé par :

$$\Delta h^c|_{RM} \approx \frac{\partial \widetilde{h}^c}{\partial f}\underbrace{\Delta f}_{<0} + \underbrace{\frac{\partial h^c}{\partial y}}_{<0}\underbrace{E_U}_{>0}\underbrace{\Delta\, \widetilde{U}\Big|_{RM}}_{>0}$$

et $\dfrac{\partial \widetilde{h}^c}{\partial f} > 0$ est une condition suffisante pour que les heures cliniques augmentent avec le passage à la rémunération mixte. ■

b) Effets prix

L'analyse des effets prix est considérablement complexifiée par la non-linéarité de la contrainte budgétaire. Sa linéarisation, fondée sur la définition de prix virtuels, permet cependant de retrouver les résultats standards de la théorie du consommateur.[32]

On note $\pi_\alpha, \alpha = \{l, h^{nc}, e\}$ les prix virtuels associés aux variables de pratique, c'est à dire les prix tels que les fonctions de demande $\beta(f, y)$ résultent d'une contrainte linéaire dans les prix virtuels. Ils correspondent par définition à : $\pi_l = fe$, $\pi_{nc} = fe$, $\pi_e = -fh^c$ et le programme d'optimisation (2.4) est alors formellement équivalent au programme

[32] A notre connaissance, l'utilisation de cette technique d'analyse remonte au travail de Becker & Lewis (1973) étudiant l'arbitrage entre quantité et qualité des enfants dans les choix de fertilité. Cette technique de linéarisation utilisant les prix virtuels a ensuite fait l'objet d'analyses systématiques par Edlefsen (1981) puis Blomquist (1989).

linéaire :

$$\begin{aligned} Max \quad & \widetilde{U}\left(X, l, h^{nc}, e\right) \\ s.c. \quad & X + \pi_l l + \pi_{nc} h^{nc} = y + \pi_e e \end{aligned} \tag{2.9}$$

Preuve Soient γ_1 le prix implicite de la consommation et L le lagrangien associé au programme (2.4). Les conditions du premier ordre s'écrivent :

$$\begin{aligned} \frac{\partial \widetilde{L}}{\partial X} &= \frac{\partial \widetilde{U}}{\partial X} - \gamma_1 = 0 && \Rightarrow \quad \gamma_1 = \widetilde{U} m_X \\ \frac{\partial \widetilde{L}}{\partial l} &= \frac{\partial \widetilde{U}}{\partial X} - \gamma_1 fe = 0 && \Rightarrow \quad \frac{\widetilde{U} m_l}{\widetilde{U} m_X} = fe \\ \frac{\partial \widetilde{L}}{\partial h^{nc}} &= \frac{\partial \widetilde{U}}{\partial h^{nc}} - \gamma_1 fe = 0 && \Rightarrow \quad \frac{\widetilde{U} m_{nc}}{\widetilde{U} m_X} = fe \\ \frac{\partial \widetilde{L}}{\partial e} &= \frac{\partial \widetilde{U}}{\partial e} + \gamma_1 f[T - h^{nc} - l] = 0 && \Rightarrow \quad \frac{\widetilde{U} m_e}{\widetilde{U} m_X} = -fh^c \\ \frac{\partial \widetilde{L}}{\partial \gamma_1} &= y + fe[T - h^{nc} - l] - X = 0 \end{aligned} \tag{2.10}$$

Par définition, les prix virtuels sont les prix de la contrainte linéaire telle que les fonctions de demande qui en résultent sont identiques aux fonctions de demande issues de ces CPO. Dans le cas classique d'une optimisation sous contrainte linéaire, la demande optimale du consommateur pour le bien x satisfait $TMS_{x,x_0} = p_x$, où x_0 désigne le bien numéraire. En appliquant ce résultat aux CPO (2.10), les prix virtuels, notés $\pi_\alpha, \alpha = \{l, h^{nc}, e\}$ sont donc définis par les conditions :

$$\pi_l = TMS_{l,X} = fe\,;\ \pi_o = TMS_{nc,X} = fe\,;\ \pi_e = TMS_{e,X} = -fh^c \tag{2.11}$$

Il convient également de noter que, au voisinage de l'équilibre, les demandes Hicksiennes et Marchaliennes concordent. Dans cet intervalle, les prix virtuels peuvent donc s'écrire en fonction des demandes Hicksiennes, $\pi_l = f\widetilde{e}\,;\ \pi_o = f\widetilde{e}\,;\ \pi_e = -f\widetilde{h}^c$.

Les demandes qui résultent du programme linéaire (2.9) sont alors mécaniquement identiques à celles qui résolvent le programme non linéaire (2.4). ■

Par définition, les fonctions de demande qui apparaissent dans la décomposition de l'effet du passage à la rémunération mixte (2.8) sont solutions du programme linéaire. En particulier, les fonctions de demande Hicksiennes peuvent donc être exprimées comme

des fonctions implicites des prix virtuels :

$$\widetilde{l} = \widetilde{l}(\pi_l, \pi_{nc}, \pi_e, \widetilde{U}) \, ; \, \widetilde{h}^{nc} = \widetilde{h}^{nc}(\pi_l, \pi_{nc}, \pi_e, \widetilde{U}) \, ; \, \widetilde{e} = \widetilde{e}(\pi_l, \pi_{nc}, \pi_e, \widetilde{U}) \tag{2.12}$$

Sous l'hypothèse que les heures totales de travail de tout médecin qui choisit la rémunération mixte ouvrent droit à un *per diem* ($h^c + h^{nc} \geq \bar{h}$), les prix des deux types de loisir sont identiques ($\pi_l = \pi_{nc} = f.e$). Le loisir total, $L = T - h^c$, est donc un agrégat Hicksien. En s'appuyant sur cette propriété, le modèle est analysé en termes d'abord d'arbitrage entre les heures totales de loisir et l'effort, puis d'allocation du loisir optimal entre les différentes occupations.

On note $\pi_L = \pi_l = \pi_{nc}$ le prix virtuel du loisir total. Les fonctions de demande Hisksiennes s'écrivent en fonction de ce prix : $\widetilde{e}(\pi_L, \pi_e, \widetilde{U})$ et $\widetilde{L}(\pi_L, \pi_e, \widetilde{U})$. Comme nous l'avons vu plus haut () la littérature d'économie de la santé admet communément qu'une augmentation du taux de rémunération des actes tendra à accroître simultanément les heures de travail clinique et l'effort (en diminuant le temps consacré à chaque patient). Ces résultats proviennent de diverses études théoriques et/ou empiriques dont l'analyse isole l'une des variables de pratique (heures ou effort). En intégrant dans l'analyse l'arbitrage entre les marges intensive et extensive, il apparaît cependant que ce résultat n'est valide que pour certaines valeurs des élasticités-prix croisées entre ces variables.

Proposition 2.2. *Les demandes compensées de l'effort et des heures cliniques sont croissantes du taux de rémunération des actes si :*

- *Condition nécessaire :* $\eta_{h^c,\pi_e} = \eta_{e,\pi_L} < 1$ *;*
- *Condition suffisante :* $(1 - \eta_{h^c,\pi_e})^2 = (1 - \eta_{e,\pi_L})^2 > \eta_{e,\pi_e}\eta_{h^c,\pi_L}$

Preuve Les prix virtuels permettent de contourner la non-linéarité de la contrainte budgétaire due à l'endogénéité des prix. Ils dépendent donc des variables de pratique et leur réaction aux variations du taux de rémunération des actes, au voisinage de l'équilibre, correspond donc à :

$$\frac{\partial \pi_L}{\partial f} = \frac{\partial \pi_l}{\partial f} = \frac{\partial \pi_{nc}}{\partial f} = \widetilde{e} + f\frac{\partial \widetilde{e}}{\partial f} \, ; \, \frac{\partial \pi_e}{\partial f} = -\left[\widetilde{h}^c + f\frac{\partial \widetilde{h}^c}{\partial f}\right]$$

En utilisant ce résultat, l'effet compensé d'une variation du taux de rémunération des actes, f, sur l'effort optimal s'écrit :

$$\frac{\partial \widetilde{e}(\pi_L, \pi_e, \widetilde{U})}{\partial f} = \frac{\partial \widetilde{e}}{\partial \pi_L}\left[\widetilde{e} + f\ \frac{\partial \widetilde{e}}{\partial f}\right] + \frac{\partial \widetilde{e}}{\partial \pi_e}\left[\widetilde{h}^c + f\ \frac{\partial \widetilde{h}^c}{\partial f}\right]$$

Quelques manipulations algébriques permettent de faire apparaître les expressions définissant les prix virtuels, présentés en (2.11). En notant η les élasticités-prix compensées, l'expression précédente s'écrit alors :

$$\eta_{e,f} = \frac{\eta_{e,\pi_L} + \eta_{e,\pi_e} + \eta_{e,\pi_e}\eta_{h^c,f}}{1 - \eta_{e,\pi_L}} \tag{2.13}$$

En procédant de la même façon, on obtient l'expression de l'élasticité prix compensée du loisir total : $\eta_{L,f} = \eta_{L,\pi_L}(1 + \eta_{e,f}) + \eta_{L,\pi_e}(1 + \eta_{h^c,f})$. Par la contrainte d'allocation du temps, cette expression s'écrit de façon équivalente en termes d'heures cliniques. Sachant que $\widetilde{L} = T - \widetilde{h}^c$, on a en effet $\frac{\partial \widetilde{L}}{\partial f} = -\frac{\partial \widetilde{h}^c}{\partial f} \Leftrightarrow \frac{\partial \widetilde{L}}{\partial f}\ \frac{f}{\widetilde{L}} = -\frac{\partial \widetilde{h}^c}{\partial f}\ \frac{f}{\widetilde{L}}$ et donc : $\widetilde{L}\ \eta_{L,f} = -\widetilde{h}^c\ \eta_{h^c,f}$. Après substitutions, l'élasticité-prix compensée des heures cliniques devient :

$$\eta_{h^c,f} = \frac{\eta_{h^c,\pi_L} + \eta_{h^c,\pi_e} + \eta_{h^c,\pi_L}\eta_{e,f}}{1 - \eta_{h^c,\pi_e}} \tag{2.14}$$

Par définition des demandes compensées, les effets prix propres sont négatifs : $\frac{\partial \widetilde{e}}{\partial \pi_e} \leq 0$ et $\frac{\partial \widetilde{h}^c}{\partial \pi_L} = -\frac{\partial \widetilde{L}}{\partial \pi_L} \geq 0$. Lorsque l'analyse inclut deux variables de pratique, et sous hypothèse de substitution nette entre la consommation et les deux types de loisir, les équations d'Euler imposent donc que les effets croisés soient positifs : $\frac{\partial \widetilde{e}}{\partial \pi_L} \geq 0$ et $\frac{\partial \widetilde{h}^c}{\partial \pi_e} = -\frac{\partial \widetilde{L}}{\partial \pi_e} \leq 0$. Comme le prix virtuel de l'effort est négatif ($\pi_e = -f\ \widetilde{h}^c$), toutes les élasticités-prix ($\eta_{\beta,\pi} = \frac{\partial \widetilde{\beta}}{\partial \pi_\beta}\ \frac{\pi_\beta}{\widetilde{\beta}}$, $\beta = \{h^c, e\}$) sont positives.

Les demandes compensées d'effort et d'heures cliniques de travail réagissent positivement aux variations de prix si : $\eta_{h^c,f} > 0$ et $\eta_{e,f} > 0$. Dans ce cas, les numérateurs des expressions (2.13) et (2.14) sont tous deux positifs. Il est donc nécessaire que les dénominateurs le soient également, c'est à dire que : $1 - \eta_{e,\pi_L} > 0$ et $1 - \eta_{h^c,\pi_e} > 0$. La matrice de Slutsky est symétrique, les effets prix croisés sont donc reliés selon l'expression : $\frac{\partial \widetilde{L}}{\partial \pi_e} = \frac{\partial \widetilde{e}}{\partial \pi_L}$. En utilisant les résultats issus de la contrainte d'allocation du temps, $\frac{\partial \widetilde{L}}{\partial \pi_e} = -\frac{\partial \widetilde{h}^c}{\partial \pi_e}$, cette propriété se traduit en termes d'élasticités prix sous la forme : $-\frac{\partial \widetilde{h}^c}{\partial \pi_e}\ f\ \frac{h^c}{h^c} = \frac{\partial \widetilde{e}}{\partial \pi_L}\ f\ \frac{e}{e} \Leftrightarrow \eta_{h^c,\pi_e} = \eta_{e,\pi_L}$. Au total, la condition nécessaire s'écrit donc : $\eta_{h^c,\pi_e} = \eta_{e,\pi_L} < 1$.

Après substitution de (2.14) dans (2.13) et simplifications, les signes des élasticités-prix peuvent être étudiés à partir des relations suivantes :

$$\eta_{e,f} = \frac{\eta_{e,\pi_e}\left(1+\eta_{h^c,\pi_L}\right)+\eta_{e,\pi_L}\left(1-\eta_{e,\pi_L}\right)}{\left(1-\eta_{e,\pi_L}\right)^2-\eta_{e,\pi_e}.\eta_{h^c,\pi_L}}$$
$$\eta_{h^c,f} = \frac{\eta_{h^c,\pi_L}\left(1+\eta_{e,\pi_e}\right)+\eta_{h^c,\pi_e}\left(1-\eta_{h^c,\pi_e}\right)}{\left(1-\eta_{h^c,\pi_e}\right)^2-\eta_{e,\pi_e}.\eta_{h^c,\pi_L}} \tag{2.15}$$

La condition nécessaire garantit que les dénominateurs sont positifs. En s'appuyant sur la symétrie de la matrice de Slutsky, une condition suffisante est alors que les numérateurs soient positifs, *i.e.* : $(1-\eta_{h^c,\pi_e})^2 = (1-\eta_{e,\pi_L})^2 > \eta_{e,\pi_e}\eta_{h^c,\pi_L}$. ■

Sous les conditions décrites dans la Proposition 2.2, le déplacement le long d'une courbe d'utilité induit par un accroissement du taux de rémunération des actes (rotation de la contrainte budgétaire) conduit à une augmentation des demandes optimales d'effort et d'heures de travail clinique ; donc à une réduction du temps consacré au loisir pendant les semaines de travail, L. La modélisation que nous avons adopté introduit cependant une distinction importante entre les allocations possibles de ce temps de loisir. Les heures qui ne sont pas consacrées à la réalisation d'actes médicaux peuvent en effet être utilisées à des activités productives si elles prennent la forme d'heures de travail non-clinique. Accroître le temps consacré à ces activités est, au demeurant, l'un des objectifs qui a présidé à l'introduction de la rémunération mixte (Section 2.1.3). Au-delà de la réaction du temps total de loisir au variables de rémunération, l'allocation du temps entre les loisirs participe donc à la détermination du profil de pratique des médecins.

Proposition 2.3. *Les configurations de signes possibles sont résumées dans le Tableau 2.5. Les effets prix du loisir ne sont simultanément négatifs que sous les conditions (1a), (2b), (3b), (4a), (5) et (6a,b).*

Preuve En utilisant l'expression des demandes Hicksiennes du programme linéarisé (2.12), l'effet compensé d'une variation de prix sur la demande d'heures de travail non-cliniques s'écrit :

Tableau 2.5 – Déterminants théoriques de l'allocation du loisir

Cas	η_{l,π_e}	η_{h^{nc},π_e}	$(\eta_{h^c,\pi_L} - \eta_{e,\pi_e})$	$\eta_{l,f}$	$\eta_{h^{nc},f}$
(1)	+	−	+	$+/-^a$	−
(2)	+	−	−	−	$+/-^b$
(3)	−	+	+	−	$+/-^b$
(4)	−	+	−	$+/-^a$	−
(5)	−	−	+	−	−
(6)	−	−	−	$+/-^a$	$+/-^b$

[a] Négatif si : $\eta_{l,\pi_e}(\eta_{h^c,f} - \eta_{e,f}) < (1 - \eta_{e,f})\eta_{l,p_x}$.
[b] Négatif si : $\eta_{h^{nc},\pi_e}(\eta_{h^c,f} - \eta_{e,f}) < (1 - \eta_{e,f})\eta_{l,p_x}$.

$$\frac{\partial \widetilde{h}^{nc}}{\partial f} = \frac{\partial \widetilde{h}^{nc}}{\partial \pi_l}\left[\widetilde{e} + f\ \frac{\partial \widetilde{e}}{\partial f}\right] + \frac{\partial \widetilde{h}^{nc}}{\partial \pi_{nc}}\left[\widetilde{e} + f\ \frac{\partial \widetilde{e}}{\partial f}\right] - \frac{\partial \widetilde{h}^{nc}}{\partial \pi_e}\left[\widetilde{h}^c + f\ \frac{\partial \widetilde{h}^c}{\partial f}\right] \tag{2.16}$$

Allocation du loisir. Un raisonnement identique s'applique à le demande compensée de loisir pur. D'après la définition des prix virtuels au voisinage de l'équilibre, ces expressions peuvent être converties en termes d'élasticités compensées. L'arbitrage qui guide l'allocation du loisir est alors décrit par le système :

$$\begin{aligned}
\eta_{l,f} &= [\eta_{l,\pi_l} + \eta_{l,\pi_{nc}} + \eta_{l,\pi_e}] + \eta_{e,f}\,[\eta_{l,\pi_l} + \eta_{l,\pi_{nc}}] + \eta_{l,\pi_e}\eta_{h^c,f} \\
\eta_{h^{nc},f} &= [\eta_{h^{nc},\pi_{nc}} + \eta_{h^{nc},\pi_{nc}} + \eta_{h^{nc},\pi_e}] + \eta_{e,f}\,[\eta_{h^{nc},\pi_{nc}} + \eta_{h^{nc},\pi_{nc}}] + \eta_{h^{nc},\pi_e}\eta_{h^c,f}
\end{aligned}$$

soit encore :

$$\begin{aligned}
\eta_{l,f} &= (1 + \eta_{e,f})\,(\eta_{l,\pi_l} + \eta_{l,\pi_{nc}} + \eta_{l,\pi_e}) + \eta_{l,\pi_e}\,(\eta_{h^c,f} - \eta_{e,f}) \\
\eta_{h^{nc},f} &= (1 + \eta_{h^{nc},f})\,(\eta_{h^{nc},\pi_{nc}} + \eta_{h^{nc},\pi_{nc}} + \eta_{h^{nc},\pi_e}) + \eta_{h^{nc},\pi_e}\,(\eta_{h^c,f} - \eta_{e,f})
\end{aligned}$$

En notant p_x le prix de la consommation et sachant que les demandes compensées sont homogènes de degré 0 dans les prix, les équations d'Euler s'écrivent : $\dfrac{\partial \widetilde{\beta}}{\partial \pi_l}\pi_l + \dfrac{\partial \widetilde{\beta}}{\partial \pi_{nc}}\pi_{nc} + \dfrac{\partial \widetilde{\beta}}{\partial \pi_e}\pi_e = \dfrac{\partial \widetilde{\beta}}{\partial p_x}p_x$, $\beta \in \{l, h^{nc}\}$. Après substitution de ce résultat dans le système d'équations précédent, les expressions qui gouvernent l'allocation du loisir sont :

$$\eta_{l,f} = -\eta_{l,p_x}(1 + \eta_{e,f}) + \eta_{l,\pi_e}(\eta_{h^c,f} - \eta_{e,f}) \tag{2.17}$$

$$\eta_{h^{nc},f} = -\eta_{h^{nc},p_x}(1 + \eta_{e,f}) + \eta_{h^{nc},\pi_e}(\eta_{h^c,f} - \eta_{e,f}) \tag{2.18}$$

Les signes des variations compensées du loisir et des heures de travail non-clinique sont déduits de ces expressions.

Etude des signes. Si la consommation et les deux types de loisir sont des substituts Hicksiens, on a : $\eta_{l,p_x} > 0$ et $\eta_{h^{nc},p_x} > 0$. Par ailleurs, la manipulation des expressions (2.15) permet d'obtenir :

$$\eta_{h^c,f} - \eta_{e,f} = \frac{\eta_{e,\pi_e} - \eta_{h^c,\pi_L}}{(1 - \eta_{h^c,\pi_e})^2 - \eta_{e,\pi_e}.\eta_{h^c,\pi_L}}$$

Sous les conditions de la Proposition 2.2, le dénominateur de cette expression est positif. La preuve de cette proposition a en outre établit que $\eta_{L,\pi_e} = \eta_{l,\pi_e} + \eta_{h^{nc},\pi_e} < 0$. Les élasticités des heures de travail non-clinique et du loisir pur ne peuvent donc pas être simultanément positifs.

L'ensemble de ces résultats est utilisé pour construire le Tableau 2.5, dont les quatre premières colonnes décrivent les signes possibles des termes du membre de droite de (2.17) et (2.18), les deux dernières colonnes les signes induits des élasticités compensées. ■

En vertu du Lemme 2.1, l'étude des signes des effets compensés permet de mettre en évidence les conditions suffisantes à ce que la rémunération mixte ait l'effet traditionnellement attendu sur les variables de pratique : une diminution de l'effort et des heures de travail clinique ainsi qu'un accroissement des heures de travail non-clinique. Comme l'indiquent nos résultats, résumés ci-dessous, seules quelques configurations spécifiques des préférences confirment ces attentes.

Résumé

Sous les hypothèses de la Proposition 2.2 et du Lemme 2.1 :

- *Le passage à la rémunération mixte devrait diminuer l'effort ainsi que les heures de travail clinique ;*
- *Dans les cas (1a), (2b), (3b), (4a), (5) et (6a,b) du Tableau 2.5, cette baisse des heures cliniques est partagée entre une augmentation des heures de travail non-clinique et une augmentation du loisir pur ;*
- *En conséquence, le temps total de travail diminue ;*
- *Dans tous les autres cas, les effets substitution et revenu agissent en sens opposé, et l'effet de la rémunération mixte sur l'allocation du loisir est ambigu.*

L'analyse économétrique, consacrée à l'estimation des préférences des médecins, laisse aux comportements observés le soin de trancher ces ambiguïtés.

2.3 Modèle économétrique

Conformément au cadre de notre analyse théorique, le modèle économétrique est spécifié en termes de maximisation d'utilité, selon la forme réduite (2.4). L'estimation s'appuie sur les comportements de pratique annuels des médecins spécialistes du Québec. Les variables de pratique décrites ci-dessus sont donc désormais définies sur une base annuelle plutôt que journalière. Afin de permettre une analyse plus fine de l'allocation du loisir, nous distinguons en particulier le loisir pris pendant les semaines de travail des semaines de loisir dans l'année, notées S ($S = 52 - W$).[33] Les heures de travail (clinique et non-clinique) sont alors mesurées par la moyenne annuelle des heures de travail hebdomadaires réalisées pendant les semaines de travail. Les actes sont, quant à eux, mesurés en termes de quantités annuelles. La distinction introduite par la rémunération mixte entre actes facturables et non-facturables repose en grande partie sur les caractéristiques techniques des actes pratiqués.[34] Nous tenons donc compte de la possibilité d'utilités (ou de désutilités) marginales différentes en incluant séparément ces deux types d'actes dans les préférences des médecins.

En résumé, la démarche économétrique adoptée consiste donc à estimer des préférences de la forme : $U = U(S, l, h^{nc}, AF, ANF, X)$, d'où résultent les choix de pratique optimaux des médecins en termes de :

(i) temps hebdomadaire de travail clinique (actes médicaux), h^c ;

(ii) temps hebdomadaire de travail non-clinique (administration, enseignement), h^{nc} ;

(iii) semaines de travail annuelles, W ;

(iv) quantités d'actes facturables réalisés pendant l'année, AF ;

(v) quantités d'actes non-facturables réalisés pendant l'année, ANF.

[33] Les résultats empiriques obtenus par Hanoch (1980) et Blank (1988) confirment l'existence d'une substitution imparfaite entre ces deux types de loisir.

[34] Une grande partie des actes non-facturables est constituée, par exemple, des visites de contrôle qui suivent la délivrance de soins.

Ce modèle structurel est identifié grâce aux variations de prix induites par l'introduction de la rémunération mixte. Comme nous l'avons vu, la contrainte budgétaire qui en résulte présente différentes non-linéarités, issues à la fois de la combinaison des modes de rémunération mixte et à l'acte le long de la contrainte budgétaire efficiente et de l'endogénéité des prix. Suivant une tradition récente en économétrie de l'offre de travail (van Soest, 1995), cette difficulté est surmontée en discrétisant l'ensemble de choix. L'estimation des paramètres de la fonction d'utilité repose alors sur un modèle de choix discret, dont la spécification est présentée dans la Section 2.3.1. Nous apportons ensuite un certain nombre de modifications à ce cadre de base, commandées par les spécificités de la base de données utilisée (Section 2.3.2).

2.3.1 Modèle de choix discrétisé : éléments de base

Pour chaque variable de pratique, nous considérons un nombre fini de niveaux possibles entre lesquels les médecins choisissent. Le nombre de niveaux retenu pour chaque variable de pratique est donc un élément important de la mise en oeuvre du modèle. Notre objectif en la matière est de recouvrir au mieux le large éventail des choix de pratique observés dans l'échantillon. Nous avons donc fait le choix de conserver un nombre important de niveaux pour chaque variable de pratique. Plus précisément, nous avons retenu $N_c = 5$ niveaux de discrétisation pour les heures de travail clinique, $N_{nc} = 5$ niveaux pour les heures de travail non-clinique, $N_w = 5$ niveaux pour les semaines de travail, $N_{AF} = 6$ niveaux pour les actes facturables et $N_{ANF} = 6$ niveaux pour les actes non-facturables. Comme l'illustre le Tableau 2.6 cette stratégie assure une assez large diffusion de l'échantillon entre les niveaux retenus.

Une alternative consiste alors en une combinaison particulière de variables de pratique, c'est à dire un ensemble de valeurs : $j = \{c_j, nc_j, w_j, ANF_j, AF_j\}$ désignant respectivement le ${c_j}^{eme}$ niveau d'heures de travail clinique, $c_j \in \{1, ..., N_c\}$, le ${nc_j}^{eme}$ niveau d'heures de travail non-clinique, etc .... L'ensemble des niveaux de discrétisation

Tableau 2.6 – Distribution de l'échantillon entre les niveaux de discrétisation

Heures				**Semaines**		**Actes**[a]			
h^c		h^{nc}		W		AF		ANF	
0	3.12%	0	66.89%	0	0.12%	0	36.15%	0	60.16%
20	12.96%	15	27.22%	10	0.53%	140000	55.38%	100000	28.03%
40	55.16%	30	4.34%	20	0.82%	280000	7.96%	200000	10.22%
60	24.92%	45	1.19%	30	1.36%	420000	0.39%	300000	1.50%
80	3.26%	60	0.30%	40	26.43%	560000	0.09%	400000	0.08%
100	0.57%	75	0.04%	50	70.75%	700000	0.02%	500000	0.01%
120	0.01%	90	0.02%	-	-	840000	0.02%	-	-
-	-	105	0.01%	-	-	-	-	-	-

[a] En Dollars constants (base 1996).

Note. Pour chaque variable de pratique (heures de travail clinique, h^c, heures de travail non-clinique, h^{nc}, semaines de travail, W, actes facturables, AF, et actes non-facturables, ANF), pourcentage d'observations pour lesquelles le choix observé discrétisé concorde avec le niveau de discrétisation correspondant.

définit donc un ensemble d'alternatives J, de dimension : $dim(J) = N_c \times N_o \times N_w \times N_{NBA} \times N_{BA} = 4500$.

L'estimation du modèle consiste à retenir les valeurs des paramètres de la fonction d'utilité qui maximisent la vraisemblance de l'alternative effectivement choisie. La mise en oeuvre de l'estimation nécessite donc de spécifier la forme de la fonction d'utilité, qui guide le choix au sein de l'ensemble J. On note V_j l'utilité annuelle que retire un médecin représentatif des choix de pratique dans l'alternative j. Une hypothèse devenue classique dans la littérature (McFadden, 1974) consiste à prendre en compte les erreurs de mesure propres à chaque alternative en décomposant l'utilité, V, entre une part déterministe, u_j, et une erreur de mesure indépendante entre les alternatives, ϵ_j : $V_j = u_j + \epsilon_j$.

La partie déterministe de l'utilité est spécifiée selon une fonction d'utilité translog, qui constitue une approximation du second ordre de toute fonction d'utilité correctement spécifiée et permet de prendre en compte une grande variété de profils de substitu-

tion (Christensen, Jorgenson & Lau, 1975). Formellement, la composante déterministe de la fonction d'utilité peut être définie, sous forme condensée, comme[35] :

$$\begin{aligned} u_j &= \mathbf{G}'\ Z_j + Z_j'\ \mathbf{B}\ Z_j + \gamma_{ANF}\ \ln ANF_j \\ &+ \mathbf{B_{ANF}}'\ Z_j\ \ln ANF_j + \beta_{ANF}\ (\ln ANF_j)^2 \end{aligned} \tag{2.19}$$

Notations Z_j désigne le vecteur colonne des variables de pratique associées à l'alternative j, à l'exception des actes non-facturables :

$$Z_j = \left[\ln(h_j^{nc}), \ln(R - W_j), \ln(T - h_j^{nc} - h_j^c), \ln(AF_j), \ln(X_j)\right]'$$

où T est le nombre d'heures totales disponibles dans une semaine (égal à $7 \times 24 = 168$ dans l'application) et R $(= 52)$ le nombre total de semaines disponibles dans l'année.

Les paramètres à estimer sont contenus dans les matrices :

$$\mathbf{B} = \begin{pmatrix} \beta_{nc} & \beta_{nc}^S & \beta_{nc}^l & \beta_{nc}^{AF} & \beta_{nc}^x \\ \beta_S^{nc} & \beta_S & \beta_S^l & \beta_S^{AF} & \beta_S^x \\ \beta_l^{nc} & \beta_l^S & \beta_l & \beta_l^{AF} & \beta_l^x \\ \beta_{AF}^{nc} & \beta_{AF}^S & \beta_{AF}^l & \beta_{AF} & \beta_{AF}^x \\ \beta_x^{nc} & \beta_x^S & \beta_x^l & \beta_x^{AF} & \beta_x \end{pmatrix}; \mathbf{G} = \begin{pmatrix} \gamma_{nc} \\ \gamma_S \\ \gamma_l \\ \gamma_{AF} \\ \gamma_x \end{pmatrix}; \mathbf{B_{ANF}} = \begin{pmatrix} \beta_{ANF}^{nc} \\ \beta_{ANF}^S \\ \beta_{ANF}^l \\ \beta_{ANF}^{AF} \\ \beta_{ANF}^x \end{pmatrix}$$

auxquels s'ajoutent γ_{ANF} et β_{ANF}. La matrice $\mathbf{B}$ est symétrique par définition, de sorte que : $\beta_k^j = \beta_j^k$ $\forall k \neq j$ tels que $j, k \in \{nc, ANF, S, AF, x\}$. ■

Compte tenu de cette spécification, un médecin choisit l'alternative j si : $V_j \geq V_k, \quad \forall k \neq j$. Sa contribution individuelle à la vraisemblance est donc la probabilité de cet évènement. Si les ϵ_j, $j \in J$ sont supposés i.i.d. selon une Gumbel (distribution à

[35] Dans ce qui suit, l'indice propre aux individus, i, est négligé par souci de clarté aussi souvent que possible.

valeurs extrêmes de type I), cette probabilité s'écrit[36] :

$$\begin{aligned} P(j) &= P[V_j \geq V_k, \quad \forall k \neq j] \\ &= P[\epsilon_j \geq u_k - u_j + \epsilon_k, \quad \forall k \neq j] \\ &= \frac{e^{u_j}}{\sum_{k=1}^{J} e^{u_k}} \end{aligned} \tag{2.20}$$

L'estimation des paramètres de la fonction d'utilité nécessite donc de connaître le niveau d'utilité atteint par l'individu dans chaque alternative. Pour une valeur donnée des paramètres, l'utilité correspond au niveau de bien-être associé au comportement de pratique dans l'alternative j, tel que décrit par la fonction d'utilité (2.19). Comme cette utilité dépend du niveau de consommation offert par le revenu de pratique, l'estimation du modèle requiert en particulier de générer le revenu issu de la pratique dans chaque alternative. Pour ce faire, nous utilisons la modélisation de la contrainte budgétaire présentée dans la Section 2.1, en nous appuyant sur les équations (2.1) à (2.3) pour calculer le niveau de consommation associé à toute combinaison particulière des variables de pratique.

L'ensemble de ces éléments définit un Logit polytomique. Contrairement à d'autres modèles, cette spécification impose en particulier que les termes d'erreur soient indépendants entre les alternatives. Ces derniers ne peuvent donc pas prendre en compte l'hétérogénéité inobservable propre aux individus. Comme l'a souligné l'analyse de la Section 2.2, les choix des médecins en matière de mode de rémunération reposent pourtant de façon cruciale sur la forme de leurs préférences. Il est donc particulièrement important de prendre en compte l'hétérogénéité individuelle dans un modèle destiné à analyser ce choix. On peut par exemple s'attendre à ce que les individus en début de carrière tendent à réaliser un nombre important d'actes et d'heures de travail. Ce profil de pratique étant mieux rémunéré sous la rémunération à l'acte que sous la rémunération

[36] Voir, par exemple, Train (2003, p.78) pour une dérivation complète des probabilités d'un Logit multinomial.

mixte, les individus les plus jeunes montreraient alors une plus forte propension à rester à la rémunération à l'acte. Ce type d'hétérogénéité observable est introduit dans la Section 2.4.1.

Outre ces caractéristiques individuelles observables, nous tenons compte de l'hétérogénéité inobservable en estimant la distribution des préférences des médecins de l'échantillon, plutôt que les préférences elles-mêmes. A cette fin, un certain nombre des coefficients de la fonction d'utilité (2.19) sont supposés être aléatoires. Les statistiques descriptives présentées dans la Section 2.1.3 suggèrent que les médecins appelés à choisir la rémunération mixte diffèrent de ceux qui resteront à la rémunération à l'acte principalement en termes d'heures consacrées au travail non-clinique et de quantité d'actes non-facturables réalisés. C'est donc sur l'utilité marginale de chacune de ces variables que nous permettons aux préférences de se distinguer. Dans la fonction d'utilité (2.19), les termes linéaires associés aux heures de travail non-clinique et aux actes non-facturables sont ainsi supposés suivre des lois normales : $\gamma_k \equiv N(\overline{\gamma}_k, \sigma_k)$, $k \in \{nc, ANF\}$, indépendantes entre elles et indépendantes des ϵ_j, $\forall j$. La moyenne et l'écart-type de ces variables aléatoires sont estimés conjointement avec les paramètres déterministes de la fonction d'utilité. Pour ce faire, les contributions individuelles à la vraisemblance (2.20) doivent être adaptées afin de tenir compte de l'incertitude sur les préférences. Conditionnellement aux valeurs prises par γ_{nc} et γ_{ANF}, la contribution à la vraisemblance de l'individu considéré correspond au logit polytomique décrit ci-dessus. La contribution inconditionnelle de l'individu i, qui a choisit l'alternative j_i, est alors :

$$l_i = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(j_i)\ \phi\left(\frac{\gamma_{nc} - \overline{\gamma}_{nc}}{\sigma_{nc}}\right) \phi\left(\frac{\gamma_{ANF} - \overline{\gamma}_{ANF}}{\sigma_{ANF}}\right) d\gamma_{nc}\ d\gamma_{ANF}$$

où ϕ est la fonction de densité de la loi normale centrée réduite.

L'estimation du modèle requiert donc, désormais, le calcul d'une intégrale bidimen-

sionnelle. Pour faire l'économie de l'importante charge de calcul de l'intégration numérique, nous utilisons une méthode d'intégration par simulation. Les intégrales sont alors approximées par la valeur moyenne de $l_i|\{\gamma_{nc}, \gamma_{ANF}\}$ calculée sur r tirages aléatoires dans les distributions de γ_{nc} et γ_{ANF}. Les tirages sont réalisés en utilisant des séquences d'Halton, qui permettent de minimiser la variance de simulation pour un nombre donné de tirages, r (Train, 1999). Cette méthode d'estimation correspond alors au Maximum de Vraisemblance Simulé, qui est asymptotiquement équivalent à l'estimateur du maximum de vraisemblance exact tant que $\sqrt{r}$ s'accroît plus rapidement que la taille de l'échantillon (Gourieroux & Monfort, 1993).

2.3.2 Aspects spécifiques

Les données utilisées pour réaliser l'estimation du modèle imposent un certain nombre d'adaptations de l'architecture de base présentée ci-dessus. Une première difficulté est imputable au niveau de finesse choisit pour la discrétisation des variables de pratique. La cohérence de la modélisation des choix discrétisés oblige en effet à former l'espace des choix en considérant l'ensemble des combinaisons possibles entre les niveaux discrétisés des variables de pratique. L'ensemble de choix inclut en conséquence un nombre important d'alternatives qui violent les contraintes de faisabilité auxquelles font face les médecins. A titre d'exemple, le modèle devrait théoriquement autoriser le choix d'une alternative dans laquelle le nombre d'actes réalisés est maximum, mais où le nombre d'heures de travail clinique est nul. Ce type de choix n'est à l'évidence jamais observé ; et s'avère, en réalité, indisponible. Afin d'alléger l'estimation, nous nous limitons donc au sous-ensemble des alternatives techniquement réalisables – au sens où elles sont choisies au moins une fois par un médecin au cours de la période d'observation – noté $J^C \subset J$. Cette stratégie d'estimation revient à conserver le même sous-ensemble d'alternatives quelle que soit l'alternative choisie. Comme le montre McFadden (1978) cette propriété de "conditionnement uniforme" assure que l'estimation du modèle Logit reste convergente malgré la réduction de l'espace de choix. La méthode de sélection que

nous utilisons est encore plus restrictive, puisqu'elle repose sur une distribution dégénérée (voir (2.21) ci-dessous) qui se limite à réaliser l'estimation sur un sous-ensemble d'alternatives constant. L'estimation du modèle selon cette stratégie reste donc convergente.

Preuve Nous adaptons ici au cas spécifique qui est le nôtre la preuve fournie par Train (2003, ch. 3) du résultat de McFadden (1978).

Nous nous intéressons à l'estimation des paramètres qui déterminent le choix à l'intérieur de l'ensemble J mais en s'appuyant sur un sous-ensemble $K \in J$. Soit $q(K|\, j)$ la probabilité que le sous-ensemble K soit utilisé pour évaluer la contribution individuelle à la vraisemblance lorsque l'alternative choisie est j. La propriété de *conditionnement uniforme* qualifie les situations dans lesquelles cette probabilité est constante pour tous les choix $j \in K$. Dans notre cas, cette distribution est dégénérée et s'écrit :

$$q(J^C \big|\, j) = \begin{cases} 1 & \text{si } K \equiv J^C \quad \forall j \in J^C \\ 0 & \text{si } K \equiv J^C \quad \forall j \notin J^C \\ 0 & \text{si } K \neq J^C \quad \forall j \in J \end{cases} \tag{2.21}$$

La probabilité inconditionnelle de choisir l'alternative j est notée P_j, définie en (2.20). Soit $P(j|\, J^C)$ la probabilité de choisir l'alternative j dans le sous-ensemble de choix J^C. En utilisant la règle de Bayes, ces probabilités sont liées par la probabilité jointe de sélectionner le sous-ensemble J^C et que l'alternative j soit choisie :

$$P(j, J^C) = q(J^C \big|\, j)\ P_j = P(j|\, J^C)\ Q(J^C) \tag{2.22}$$

où $Q(J^C) = \sum_{j \in J^C} P_j\ q(J^C \big|\, j)$ est la probabilité marginale de sélectionner le sous-ensemble J^C parmi l'ensemble des alternatives contenues dans J. La relation (2.22) se simplifie selon :

$$P(j|\, J^C) = \frac{q(J^C \big|\, j)\ P_j}{\sum_{j \in J^C} P_j\ q(J^C|_j)} = \frac{P_j}{\sum_{j \in J^C} P_j} \tag{2.23}$$

où le second membre de droite s'obtient en utilisant le conditionnement uniforme retenu dans notre application (2.21).

Dans une spécification en termes de Logit mixte, l'utilité est définie conditionnellement à l'hétérogénéité inobservable : $u_j = u_{j|\epsilon}$ puis intégrée dans sa distribution. Les probabilités conditionnelles de l'hétérogénéité inobservable s'écrivent :

$$P_{j|\epsilon} = \frac{e^{u_{j|\epsilon}}}{\sum_{i \in J} e^{u_{i|\epsilon}}}$$

Après substitution dans (2.23), on obtient :

$$P(j|\,J^C,\epsilon) = \frac{p_{j|\epsilon}}{\sum\limits_{j\in J^C} P_{j|\epsilon}} = \frac{\dfrac{e^{u_{j|\epsilon}}}{\sum\limits_{i\in J} e^{u_{i|\epsilon}}}}{\sum\limits_{k\in J^C} \dfrac{e^{u_{k|\epsilon}}}{\sum\limits_{i\in J} e^{u_{i|\epsilon}}}} = \frac{e^{u_{j|\epsilon}}}{\sum\limits_{i\in J^C} e^{u_{i|\epsilon}}}$$

La probabilité conditionnelle correspond donc à la probabilité d'un Logit multinomial associée au choix de l'alternative j dans l'ensemble de choix J^C. La contribution individuelle à la vraisemblance correspond à l'intégrale de cette probabilité conditionnelle dans la distribution de l'hétérogénéité. Sous hypothèse de normalité, on a :

$$P(j|\,J^C,\epsilon) = \int_{-\infty}^{\infty} \frac{e^{u_{j|\epsilon}}}{\sum\limits_{i\in J^C} e^{u_{i|\epsilon}}}\ \phi(\epsilon)d\epsilon$$

La maximisation de la fonction de vraisemblance formée à partir de ces contributions est convergente. La réduction de l'ensemble de choix restreint cependant la quantité d'information utilisée et affecte donc l'efficacité des estimateurs. ■

La seconde difficulté est imputable à la distinction introduite par la rémunération mixte entre actes facturables et non-facturables. Si ces deux types d'actes sont rémunérés et, par conséquent, observés sous la rémunération à l'acte, les actes non-facturables sont par définition inobservables sous la rémunération mixte. Il ne donnent en effet lieu à aucune rémunération spécifique (voir Section 2.1.3 ci-dessus) et ne sont donc pas déclarés par les médecins. Un médecin qui a opté pour la rémunération mixte peut cependant être en partie rémunéré selon le mode de rémunération à l'acte, dès lors que certaines de ses heures de travail ne sont pas couvertes par un *per diem*. Dans ce cas, les actes non-facturables réalisés sont rémunérés à taux plein, donc observés. Soit m le niveau d'actes non-facturables discrétisés réalisé dans une période de travail non couverte par un *per diem*, par un médecin qui a choisit la rémunération mixte (noté $d_i = 1$). Le choix associé, ANF_m, constitue alors un plancher pour les actes non-facturables effectivement réalisés par cette observation. Ainsi, les actes non-facturables observés sont ANF_m alors même que les actes non-facturables effectifs sont $NBA \in \{NBA_m, NBA_{m+1}, \ldots, NBA_{N_{NBA}}\}$. Pour ces observations, la fonction de vraisemblance doit donc incorporer l'incertitude quant au niveau effectivement choisi. Par définition des variables discrétisées, les niveaux

de discrétisation d'une variable de pratique sont mutuellement exclusifs. La contribution à la vraisemblance d'un individu qui exerce sous la rémunération mixte et dont les choix observés sont $\{Z_j, ANF_m\}$ est donc la somme des probabilités de choix parmi les ANF_j :

$$\begin{aligned} P\left(j|_{d_i=1}\right) &= P\left[\{Z_j, ANF_m\}|_{d_i=1}\right] = P\left(Z_j, ANF_m\right) \cup P\left(Z_j, ANF_{m+1}\right) \cup \ldots \cup P\left(Z_j, ANF_{N_{ANF}}\right) \\ &= \sum_{l=m}^{N_{ANF}} \frac{\exp\left(u(Z_j, ANF_l)\right)}{\sum_{k=1}^{J^C} e^{u_k}} \\ &= \sum_{l=m}^{N_{ANF}} \frac{\exp\left(\mathbf{G}'\ Z_j + Z_j'\mathbf{B}Z_j + \gamma_{ANF} \ln ANF_l + \mathbf{B_{ANF}}'\ Z_j\ \ln ANF_l + \beta_{ANF}(\ln ANF_l)^2\right)}{\sum_{k=1}^{J^C} e^{u_k}} \end{aligned}$$

$$P\left(j|_{d_i=1}\right) = \frac{\exp(\mathbf{G}'\ Z_j + Z_j'\mathbf{B}Z_j)}{\sum_{k=1}^{J^C} e^{u_k}} \sum_{l=m}^{N_{ANF}} \exp\left(\gamma_{ANF} \ln ANF_l + \mathbf{B_{ANF}}'\ Z_j \ln ANF_l + \beta_{ANF}(\ln ANF_l)^2\right)$$

Pour les médecins qui ont choisit la rémunération mixte, les probabilités sont donc corrigées de façon à prendre en compte l'incertitude quant à l'alternative selectionnée au sein du sous-ensemble d'alternatives choisi. Pour les médecins qui sont restés à la rémunération à l'acte, en revanche, les actes non-facturables sont observables en toutes circonstances. Leur contribution reste donc conforme à l'expression (2.20). Au total, la probabilité que l'individu i choisisse l'alternative j devient donc :

$$P(j_i) = \left(\frac{e^{u_{j_i}}}{\sum_{k=1}^{J^C} e^{u_k}}\right)^{1-d_i} P\left[\{Z_j, ANF_m\}|_{d_i=1}\right]^{d_i} \tag{2.24}$$

En résumé, nous estimons un logit mixte dont la vraisemblance s'écrit : $\prod_{i=1}^{N} l_i$ où l_i est décrit par (2.21) et $P(j_i)$, la probabilité de choisir l'alternative j_i, par (2.24). La spécification retenue comprend deux paramètres aléatoires et 25 paramètres constants.

L'estimation comporte donc 29 paramètres à estimer – auxquels s'ajoutent les paramètres d'hétérogénéité individuelle, voir Section 2.4.1 – en utilisant 12 842 observations du comportement de pratique des médecins spécialistes du Québec. Le modèle est estimé par la méthode du maximum de vraisemblance simulé. Pour chaque paramètre aléatoire, nous réalisons 20 tirages d'Halton spécifiques à chaque individu. La fonction de vraisemblance est alors évaluée, pour chaque tirage, en calculant le niveau d'utilité atteint par l'individu dans chacune des J^C alternatives. Compte tenu de ces caractéristiques, l'estimation du modèle requiert d'importantes capacités tant de calcul que de mémoire.[37] L'estimation est rendue possible par la parallélisation des calculs (Swann, 2002), qui consiste à répartir l'évaluation de la vraisemblance entre plusieurs processeurs (20 ici). Le programme, présenté dans l'Annexe (Section 2.A), a été développé en langage Ox (Doornik & Ooms, 2001 ; Cribari-Neto & Zarkos, 2003).

2.4 Résultats : les vertus de la flexibilité

Le modèle est identifié empiriquement grâce aux variations de prix induites par la rémunération mixte. A cette fin, des observations sur les comportements de pratique avant et après la réforme ont été recueillies. Ces données doivent subir un certain nombre de transformations pour fournir le pendant empirique de la modélisation que nous avons adopté. Les résultats d'estimation et, en particulier, les simulations réalisées montrent que la liberté d'adoption de la rémunération mixte est un élément clé de son succès.

[37] En utilisant les niveaux de discrétisation présentés dans le Tableau 2.6, chaque itération nécessite ainsi le calcul de plus de 640 niveaux d'utilité par individu.

2.4.1 Présentation des données

Les données que nous utilisons recouvrent les comportements de pratique et un certain nombre de caractéristiques individuelles de l'ensemble des médecins spécialistes exerçant au Québec entre 1996 et 2002. Elles résultent de la combinaison des informations fournies par deux institutions Québécoises : le *Collège des médecins du Québec* (CMQ) et la *Régie d'Assurance Maladie du Québec* (RAMQ).

Le CMQ est l'organisation professionnelle représentative des médecins du Québec. Il réalise chaque année une enquête auprès de ses membres, destinée à recueillir de l'information sur leurs caractéristiques individuelles (telles que la spécialisation, l'âge ou le sexe), les caractéristiques institutionnelles et géographiques de leur établissement de rattachement ainsi que l'allocation de leur temps de travail. Les médecins sont ainsi appelés à évaluer le temps qu'ils consacrent à leur travail en termes d'heures (nombre moyen d'heures hebdomadaires) et de semaines (nombre annuel), puis la répartition de ces heures, en pourcentage, entre le temps consacré aux patients – activités cliniques – et le temps consacré respectivement à l'enseignement, aux activités administratives et à la recherche. Malgré la remarquable stabilité du questionnaire au cours des années, la question portant sur les semaines de travail n'apparaît qu'en 1996, 1997, 1998 et 2002. Ce changement temporaire dans la collecte des données nous oblige donc à abandonner les observations couvrant la période 1999-2001, pour lesquelles les semaines de travail sont manquantes. La rémunération mixte ayant été introduite au quatrième trimestre de l'année 1999, nous sommes donc conduits à abandonner les 3 années qui suivent immédiatement la réforme. Les choix de pratique que nous conservons après la réforme résultent donc d'une longue période d'ajustement au nouveau mode de rémunération.

La RAMQ est l'organisation publique en charge de la rémunération des médecins au Québec. A ce titre, elle reçoit de chaque praticien une déclaration décrivant son activité professionnelle, à partir de laquelle la rémunération est calculée. Les données de prix et de productivité (nombre d'actes réalisés) que nous en obtenons, sur une

base trimestrielle, sont donc très peu sujettes à des problèmes d'erreur de mesure. Ces données administratives offrent en particulier une description parfaite du mode de rémunération sous lequel se déroule la pratique. Conformément au modèle théorique, nous nous concentrons sur le choix entre la rémunération à l'acte et la rémunération mixte. Seuls les médecins (62.68% de l'échantillon, correspondant à 12,819 observations de décisions annuelles) dont l'intégralité du revenu provient de l'un ou l'autre de ces modes de rémunération sont donc conservés dans l'estimation.

Les deux ensembles de données sont combinés grâce à un identifiant codé, propre à chaque médecin. Cette variable nous permet également de retracer les choix multiples d'un même individu entre les périodes. Les choix de pratique des 4544 médecins spécialistes retenus sont donc observés en panel de 1996 à 1998 et en 2002.

2.4.2 Construction des variables

Ces données doivent subir un certain nombre de transformations avant de fournir des mesures en adéquation avec notre modélisation de la marge intensive (heures et semaines de travail), de la marge extensive (quantités d'actes) et des paramètres de la contrainte budgétaire.

a) Heures de travail : mesures de la marge extensive

Pour les années retenues dans l'estimation (1996-1998 et 2002), le nombre de semaines de travail est directement disponible. Les semaines de loisir annuelles – qui sont l'argument de la fonction d'utilité estimée – sont calculées par différence, sur la base de 52 semaines par an.

Les variables d'heures hebdomadaires sont calculées en multipliant les heures to-

tales par le pourcentage de temps consacré à chaque activité. Le pourcentage consacré aux activités cliniques permet ainsi, sans autre transformation, d'obtenir une mesure du nombre d'heures cliniques réalisées. D'après notre définition – qui recouvre celle des activités admissibles au *per diem* sous la rémunération mixte, voir Section 2.1.3 – les heures non-cliniques regroupent l'enseignement et les tâches administratives, mais excluent le temps consacré à la recherche. Notre mesure des heures totales de travail diffère donc de celles qu'ont déclaré les médecins, puisqu'elles correspondent à la somme des heures de travail clinique et non-clinique et ignorent par conséquent le temps consacré à la recherche.

Prendre en compte les changements de rémunération introduits par la rémunération mixte nécessite en outre de ventiler les variables de pratique selon le mode de rémunération sous lequel elles ont exercées (comme l'indique, par exemple, l'expression (2.3)). Un médecin qui a choisi la rémunération mixte est en effet rémunéré selon le mode de rémunération à l'acte pour toutes les heures de pratique qui n'ont pas donné lieu au versement d'un *per diems*. Dans le cadre de notre modélisation en termes de choix discrets, cette disposition impose de prédire le nombre de *per diem* reçus dans chaque alternative. Pour ce faire, nous utilisons une approximation fondée sur la part des heures de travail qui peuvent, compte tenu des restrictions imposées par la rémunération mixte, être admises au *per diem*[38]. Cette proportion, notée θ_i pour chaque individu i, est définie formellement par l'expression : $\theta_i = \dfrac{\overline{d}.N_i}{h^c + h^o}$, où N_i, le nombre de *per diems* hebdomadaire moyen, est définit par (2.2).

[38]la solution exacte, mais extrêmement coûteuse en termes de complexité du modèle, consiste à distinguer les variables de pratique selon le mode de rémunération sous lequel elles ont été exercées. Nous avons fait le choix de la simplicité.

b) Actes : mesure de la marge intensive

Outre les heures de travail clinique, notre analyse intègre l'ajustement des marges intensives par le biais de la quantité d'actes accomplis. Chaque médecin réalise en général une grande variété d'actes, qui différent considérablement tant en termes de temps qu'en termes d'effort (attention, expertise, ...). Les taux de rémunération des actes sont ajustés en conséquence et reflètent, au moins en partie, cette diversité. Afin d'obtenir une mesure du nombre d'actes qui soit à la fois unique et fidèle à l'intensité de la pratique, nous avons donc construit un indice de quantités où, pour chaque type d'acte (correspondant, formellement, à un code d'acte dans la taxinomie de la RAMQ), la quantité délivrée est pondérée par son taux de rémunération.[39]

Pour constituer une mesure fiable du nombre d'actes fourni, les indices de quantités doivent être préservés des variations dues à l'évolution des prix. Les pondérations sont donc maintenues constantes, en utilisant le prix des actes à une année de base (1996). Au cours des six années d'observation, pourtant, de nombreux actes apparaissent ou deviennent, au contraire, obsolètes en raison du progrès des connaissances médicales. Une seconde année de base est donc utilisée (2000) et les indices de quantité prennent alors la forme d'indices de Laspeyres chaînés.[40] Les dispositions de la rémunération mixte nous conduisent, par ailleurs, à considérer deux mesures d'actes, selon qu'ils sont ou non facturables sous ce mode de rémunération. Bien que cette distinction engendre également des différences dans les taux de rémunération des actes (voir Section 2.1.3), les pondérations utilisées sont également maintenues constantes pour les deux indices (égales au prix qui rémunère les actes, à l'année de base, sous le mode rémunération à

[39] A titre d'illustration, un dermatologue qui réalise 4 visites primaires et 6 visites de contrôle totaliserait, en l'absence de pondération, 10 actes. Une visite primaire nécessite pourtant un entretien approfondi avec le patient ainsi qu'un diagnostic complet, et dure en moyenne 45 minutes, tandis qu'une visite de contrôle se limite en général à une vingtaine de minutes. Les rémunérations de ces actes reflètent ces différences, puisqu'elles sont respectivement, en 1996, de 47$ et 16.50$.

[40] Le chaînage permet de convertir les indices calculés selon les prix de la seconde année en indices basés sur la première. Diewert (1993) propose une présentation détaillée de cette technique.

l'acte).

Formellement, le nombre d'actes réalisés par le médecin i à la période t, $A_i^t = \{ANF_i^t, AF_i^t\}$ est alors mesuré par la variable :

$$A_i^t = \begin{cases} \sum\limits_{a \in \mathscr{A}} A_{a,i}^t \; p_{a_s}^{1996} & \text{si } 1996 \leq t < 2000 \\ \sum\limits_{a \in \mathscr{A}} \left(A_{a,i}^t \; p_{a_s}^{2000} \right) \dfrac{\sum\limits_{a \in \mathscr{A}} A_{a,i}^{2000} \; p_{a_s}^{1996}}{\sum\limits_{a \in \mathscr{A}} A_{a,i}^{2000} \; p_{a_s}^{2000}} & \text{si } 2000 \leq t \leq 2002 \end{cases} \tag{2.25}$$

Notations $p_{a_s}^t$ désigne le prix, à la période t et sous la rémunération à l'acte, de l'acte a lorsqu'il est réalisé par un médecin de la spécialité s ; $A_{a,i}^t$ le nombre d'actes de type a réalisés par le médecin i à la période t. Les pondérations restant inchangées que les actes soient facturables ou non, la variable A_i^t désigne indifféremment les premiers, $A_i^t = ANF_i^t$, ou les seconds, $A_i^t = AF_i^t$. Seul le groupe d'actes considéré, $\mathscr{A}$, s'en trouve affecté. Il regroupe l'ensemble des actes pour lesquels le taux de réduction sous la rémunération mixte, α, est strictement positif, dans le calcul de l'indice d'actes facturables : $\mathscr{A}_F = \{a : \alpha_a > 0\}$; et l'ensemble des actes pour lesquels le taux de réduction sous la rémunération mixte est strictement nul, dans le calcul de l'indice d'actes non-facturables : $\mathscr{A}_{NF} = \{a : \alpha_a = 0\}$ ■

c) Prix des actes : simulation du revenu potentiel

Le niveau de consommation (*i.e.* le revenu réel) associé à chaque alternative est calculé en utilisant la contrainte budgétaire développée dans la Section 2.1. Pour un mode de rémunération $d_i \in \{0; 1\}$ donné, le revenu potentiel correspond donc au bénéfice tiré des variables de pratique, décrit par la contrainte (2.3).

La partie du revenu qui provient de la rémunération des actes résulte, en particulier, du produit entre le nombre d'actes réalisés et le prix des actes sous le mode de rémunération choisi. Ce calcul nécessite ainsi de disposer d'une variable reflétant les taux de

rémunération des actes, à chaque période et sous chaque mode de rémunération, qui soit cohérente avec la mesure utilisée pour les quantités (nombre d'actes réalisés, voir ci-dessus). A cette fin, les prix des actes sont agrégés sous forme d'indices de prix.

Dans le calcul de ces indices, le prix est pondéré par le nombre moyen d'actes à l'une des années de base (1996 ou 2000), reflétant ainsi la valorisation monétaire d'un profil de pratique représentatif. Le choix de la pondération répond au souci d'isoler les mesures de prix des variations de pratique dues au passage à la rémunération mixte. Ainsi, seuls les médecins qui sont restés à la rémunération à l'acte sont pris en compte pour calculer les quantités moyennes à l'année de base. De plus, ces même pondérations sont utilisées pour calculer l'indice de prix sous la rémunération à l'acte comme sous la rémunération mixte.

En notant p_s^t l'indice de prix auquel font face les médecins de la spécialité s à la période t (où $p = P$ pour la rémunération à l'acte et $p = (1-\alpha)\ P = PF$ pour la rémunération mixte), le revenu tiré des actes réalisés est alors mesuré par : $A_i^t\ \Delta P_s^t$.

Preuve Les indices de prix de la spécialité s à la période t, p_s^t, sont mesurés par :

$$p_s^t = \begin{cases} \sum\limits_{a_s \in \mathscr{A}} \overline{A}_{a_s}^{\,1996}\ p_{a_s}^t & \text{si } 1996 \leq t < 2000 \\ \sum\limits_{a_s \in \mathscr{A}} (\overline{A}_{a_s}^{\,2000}\ p_{a_s}^t)\ \dfrac{\sum\limits_{a_s \in \mathscr{A}} \overline{A}_{a_s}^{\,1996}\ p_{a_s}^{2000}}{\sum\limits_{a_s \in \mathscr{A}} \overline{A}_{a_s}^{\,2000}\ p_{a_s}^{2000}} & \text{si } 2000 \leq t \leq 2002 \end{cases} \tag{2.26}$$

où $\overline{A}_{a_s}^{\,t}$ désigne le nombre moyen d'actes de type a réalisés, à la période t, par les médecins de la spécialité s qui sont restés à la rémunération à l'acte pendant l'ensemble de la période d'observation (1996-2002). Ces pondérations sont utilisées pour calculer l'indice de prix sous la rémunération à l'acte, $p_s^t = P_s^t$, comme sous la rémunération mixte $p_s^t = (1-\alpha_s^t)\ P_s^t = PF_s^t$. Seul le groupe d'actes considéré, $\mathscr{A}$, est adapté en fonction de l'indice calculé ($\mathscr{A} = \mathscr{A}_F$ dans le calcul de l'indice de prix des actes facturables, $\mathscr{A} = \mathscr{A}_{NF}$ dans le calcul de l'indice de prix des actes non-facturables). La seconde équation dans (2.26) reflète le chaînage entre les deux années de base.

La rapport $\Delta P_s^t = \dfrac{\sum\limits_{a_s} \overline{A}_{a_s}^{\,1996}\ p_{a_s}^t}{\sum\limits_{a_s} \overline{A}_{a_s}^{\,1996}\ p_{a_s}^{1996}}$ mesure donc la revalorisation nominale subie par le panier

d'actes entre la première année de base (1996) et l'année en cours (t). Le revenu tiré des actes correspond au nombre d'actes mesurés aux prix de 1996 et réévalués selon cette mesure : $A_i^t\ \Delta P_s^t = \sum_a A_{a,i}^t\ p_{a_s}^{1996}\ \Delta P_s^t$. ■

Dans le cas de la rémunération à l'acte ($d_i = 0$) cette quantité suffit à décrire le revenu potentiel du médecin i. La quantité totale d'actes réalisés correspond en effet, dans ce cas, à la somme des actes facturables et non-facturables ; et le revenu potentiel dans l'alternative j sous la rémunération à l'acte est la valeur monétaire de l'ensemble des actes réalisés, soit : $X_{j,t}^{RA} = \left(AF_j^t + ANF_j^t\right)\ \Delta P^t$.

Le revenu sous la rémunération mixte, quant à lui, tient compte tant des actes réalisés que des heures de travail. Il concorde cependant avec le revenu de la rémunération à l'acte pour la partie de la pratique qui n'est pas incluse dans un *per diem*. Le calcul du revenu nécessite donc, dans ce cas, de ventiler les variables de pratique en fonction du mode de rémunération sous lequel elles ont exercées. Nous adoptons pour ce faire l'approximation présentée plus haut, fondée sur la proportion des heures de travail qui sont admissibles à un *per diem*, θ. Le revenu potentiel associé à l'alternative j pour un médecin qui a choisi la rémunération mixte est donc : $X_{j,t}^{RM} = S_j N_j D + \theta_{j,t}\ BA_j\ PF^t + (1-\theta_{j,t})(BA_j + NBA_j)P^t$ où le nombre de *per diems* dans l'alternative j, N_j, est définit par (2.2).

Les variables ainsi définies nous permettent donc de calculer le revenu potentiel dans l'alternative j sous chacun des modes de rémunération disponibles. Dans la mesure où le passage à la rémunération mixte est un choix volontaire de la part du médecin (ou, en tout cas, supposé tel, voir Note (20)), nous retenons dans chaque alternative le revenu maximum parmi ceux qui résultent des modes de rémunération disponibles (qui se réduisent à la rémunération à l'acte jusqu'en 1999). Le choix du mode de rémunération est donc une variable implicite de notre modèle, représenté par la valorisation optimale des choix de pratique. Cette stratégie consiste en effet à retenir le choix de rémunération individuellement rationnel, puisque les variables de pratique sont fixes dans

une alternative donnée. Si, comme on doit s'y attendre, l'utilité marginale du revenu est positive, les médecins doivent donc, pour des choix de pratique donnés, adopter le mode de rémunération qui maximise le revenu potentiel. La rémunération potentielle dans l'alternative j est donc : $X_{j,t} = \max\left\{X_{j,t}^{RA}; X_{j,t}^{RM}\right\}$, où $X_{j,t}^{RM} = 0$ si $t \leq 1999$.

d) Plafonds et taux de rémunération : simulation du revenu effectif

Dans l'analyse traditionnelle de la théorie du consommateur, c'est par l'intermédiaire de la consommation que le revenu influence le bien-être des agents. A ce titre, la fonction d'utilité que nous estimons ne dépend pas du revenu potentiel mais du revenu effectif, qui correspond au revenu effectivement versé aux praticiens. Comme nous l'avons vu plus haut (Section 2.1.2), ces quantités diffèrent sensiblement en raison, notamment, des mesures de différenciation et de plafonnement des rémunérations. Le revenu ajusté qui en résulte doit en outre être corrigé de l'inflation afin de fournir une mesure réelle plutôt que nominale de la consommation.

Les dispositions qui gouvernent le calcul des taux de rémunération différenciée font intervenir un large éventail de caractéristiques individuelles (telles que la région et la ville de pratique) dont certaines nous sont indisponibles. Nos données contiennent cependant le niveau de revenu trimestriel de chaque médecin avant et après application du taux de rémunération. Pour chacun d'entre eux, nous utilisons donc une approximation du taux individuel, τ_i, calculée à partir du rapport moyen, sur l'ensemble de la période, entre ces deux niveaux de revenu.

Les seuils au-delà desquels la rémunération est soumise aux mesures de plafonnement sont propres aux spécialités de pratique, et constituent une information publique, fournie par la RAMQ (le détail de ces montants est fourni dans la Section 2.1.2). Les modalités de leur mise en œuvre nous obligent cependant à définir un plafond propre à chaque individu, fondé sur son profil de pratique moyen au cours de la période. Ces dis-

positions d'application dépendent en effet de façon très importante de l'établissement médical où s'est déroulée la pratique. Ainsi, les revenus issus des activités en urgence sont exclus de l'assiette de calcul du plafond de 1996 à 2001. A partir de 2001, cette exclusion s'étend à l'ensemble des revenus liés à la pratique en hôpital. Pour tenir compte des frais professionnels, les revenus provenant de la pratique en cabinet privé sont en outre diminués d'une proportion fixe.

Un traitement exact de ces dispositions aurait, une fois encore, nécessité de distinguer les variables de pratique en fonction de l'établissement ou elles s'exercent, multipliant d'autant le nombre d'alternatives. Par souci de simplicité, nous avons plutôt choisi d'ajuster le niveau des plafonds en utilisant la ventilation moyenne du revenu entre les établissements. Cette méthode permet alors de définir un plafond virtuel, propre à l'individu, correspondant au plafond auquel le revenu de l'individu est effectivement soumis compte tenu de la répartition de ses activités entre les établissements. La mesure de plafond utilisée, $C_{i,t}$, est alors formellement définie par :

$$C_{i,t} = \frac{\widetilde{C}_{s,t}}{s_i^e + s_i^p \ (1 - a_s)}$$

Notations Les dispositions légales qui gouvernent l'application des plafonds sont décrites par le seuil applicable à la spécialité du professionnel, s, à la période t, noté $\widetilde{C}_{s,t}$ et le taux de réduction appliqué aux revenus en cabinet privé, a_s ($a_s = 35\%$ pour toutes les spécialités à l'exception de la radiologie diagnostique qui bénéficie d'une réduction de $a_s = 75\%$).

La répartition des activités du médecin i sur l'ensemble de la période est décrite par la ventilation de son revenu, R_i, entre le revenu issu de la pratique en cabinet privé, R_i^p, et les revenus hors cabinet privé admissibles au plafond (*i.e.* excluant les revenus à l'urgence, ainsi que les revenus hospitaliers à partir de 2001), R_i^e. Pour chaque individu, cette décomposition permet de définir les parts de la pratique réalisées dans chaque type d'établissement comme : $s_i^e = R_i/R_i^e$ et $s_i^p = R_i/R_i^p$. Le revenu du médecin i à la période t est alors soumis au plafond si : $s_i^e.R_{i,t} + s_i^p.R_{i,t}(1-a) \geq \widetilde{C}_{s,t}$. Le plafond virtuel qui s'applique au revenu global est donc : $R_{i,t} \geq \dfrac{\widetilde{C}_{s,t}}{s_i^e + s_i^p.(1-a)} \equiv C_{i,t}$. ■

Tableau 2.7 – Prédiction de la consommation effective

Variable	**Coefficient**	**(Ecart-type)**
Revenu prédit	0.97***	(0.005)
Constante	43081.77***	(4336.695)
R^2	0.83	

Niveaux de signification : *** 10%, ** 5%, *** 1%.

Note. Régression linéaire. La variable endogène est la consommation effective observée du médecin, la variable *Revenu prédit* correspond à la consommation effective simulée pour les choix observés.

L'ensemble de ces variables fournit, pour un revenu potentiel donné, le revenu effectif de chaque observation dans chaque alternative selon l'expression (2.1). Ce revenu potentiel nominal est enfin converti en revenu réel, en utilisant les données d'inflation fournies par *Statistique Canada*.[41] Le taux d'inflation annuel moyen pour l'ensemble de la période est de 1.92%.

L'ensemble de ces variables permet de construire la contre-partie empirique de la contrainte budgétaire des médecins. Suivant en cela une longue tradition en économie appliquée de l'offre de travail (Blundell & Macurdy, 1999), ces variables sont en effet utilisées pour simuler le niveau de consommation dans toute alternative. Pour l'alternative choisie, la comparaison avec le revenu effectivement obtenu par le praticien fournit une évaluation de la qualité de cette prévision. A cette fin, le Tableau 2.7 présente les résultats de la régression de la consommation effective observée sur sur son niveau prédit par le modèle dans l'alternative choisie. La qualité de la modélisation peut être évaluée selon deux dimensions. D'un part, la consommation prédite recouvre une large proportion des variations de la consommation effective (83%). D'autre part, le pouvoir explicatif de la consommation prédite, mesuré par son coefficient dans la regression (0.97), est très proche d'une prévision parfaite (pour laquelle le coefficient serait égal à 1, indiquant que toute variation de la consommation prédite recouvre une variation

[41] Les données sont librement disponibles sur le site web de l'institution. Nous utilisons l'indice des prix à la consommation annuel pour le Québec.

identique de la consommation effective observée).

2.4.3 Résultats d'estimation

Cette contrainte budgétaire "empirique" permet d'évaluer la fonction d'utilité des médecins dans chaque alternative. L'estimation du modèle présenté dans la Section 2.3 consiste à retenir la combinaison de paramètres qui rendent optimale l'alternative choisie. A titre préliminaire, le modèle est estimé sur le sous-échantillon des chirurgiens. Ces préférences participent à lever les indéterminations du modèle théorique de la Section 2.2, en fournissant une appréciation de l'élasticité empirique des choix de pratique aux variations des incitations. Elles permettent également d'anticiper l'effet sur l'offre de soins produit par tout changement des paramètres de la contrainte budgétaire. Les modalités d'instauration de la rémunération mixte choisies par les autorités du Québec peuvent ainsi être comparées à des dispositions alternatives, telles que sa surpression ou sa généralisation.

a) Préférences estimées

Les chirurgiens représentent 9.65% (1237 observations) des observations de l'échantillon, regroupant 495 individus. Avec un taux d'adhésion à la rémunération mixte de 60% en 2002 (voir Tableau 2.1), ce sous-échantillon présente notamment l'avantage de regrouper des individus aux choix de rémunération très variables. Le Tableau 2.8 présente les profils de pratique moyens au sein de cette spécialité selon le mode de rémunération choisi (ou imposé avant 1999). Cette variabilité des choix de rémunération semble s'appuyer sur une importante diversité des comportements de pratique, qui recouvre les différences commentées plus haut (Section 2.3). Une exception notable est l'apparition d'un effet de sélection en termes d'heures de travail non-clinique. Cet aspect est intégré dans le modèle économétrique où l'hétérogénéité inobservable vis-à-vis de

Tableau 2.8 – Profil de pratique des chirurgiens

		h	h^c	h^{nc}	W	AF^a	ANF^a
Médecins	Avant 1999	59.19	49.12	10.07	45.52	172.46	16.96
RM	2002	52.44	46.65	5.78	44.03	117.68	9.57
Médecins	Avant 1999	54.02	46.18	7.84	44.86	142.63	33.95
RA	2002	53.03	48.87	4.15	45.26	159.90	35.21
Total		54.43	47.05	7.38	44.94	147.19	29.61

[a] En milliers de Dollars constants (base 1996).

Note. *Moitié supérieure* : Profil de pratique moyens des chirurgiens qui ont obtenu une partie de leur rémunération sous la rémunération mixte au cours de la période, avant (première ligne) et après (deuxième ligne) l'avoir adoptée. *Moitié inférieure* : profil de pratique des chirurgiens dont 100% du revenu provient de la rémunération à l'acte, avant (troisième ligne) et après (dernière ligne) l'introduction de la réforme.

ces heures de travail est prise en compte par un coefficient aléatoire.

Les résultats d'estimation des préférences des chirurgiens, décrites par la fonction d'utilité (2.19), sont présentées dans le Tableau 2.9. L'hétérogénéité inobservable est introduite progressivement dans les spécifications 2 (heures de travail non-cliniques) et 3 (actes non-facturables). Comme on pouvait s'y attendre, la prise en compte de l'hétérogénéité inobservable permet au modèle de décrire de mieux en mieux les préférences des individus de l'échantillon. Nous retenons en conséquence la spécification 3, qui incorpore l'hétérogénéité par rapport aux heures de travail non-clinique et aux actes non-facturables.

La qualité de l'estimation peut être appréciée par la capacité des préférences estimées à recouvrir la distribution des pratiques réelles. A cette fin, le Tableau 2.10 compare les prédictions du modèle estimé pour le comportement en 2002 (colonne centrale) aux comportements observés à la fois sur l'ensemble de la période (première colonne) et en 2002, qui est l'unique année de réforme incluse dans l'échantillon. Dans l'ensemble, le modèle recouvre avec une précision très satisfaisante les variations du comportement de pratique. La diminution de l'effort et l'accroissement important de la consommation en 2002, par rapport à leur niveau sur l'ensemble de la période, sont en particulier

Tableau 2.9 – Paramètres estimés de la fonction d'utilité Translog

	Specification 1		**Specification 2**		**Specification 3**	
	Coef. Estimé	t de Student	Coef. Estimé	t de Student	Coef. Estimé	t de Student
$\gamma^{nc}, \overline{\gamma}^{o}$	9.100	9.48***	9.543	9.82***	9.547	9.78***
σ_{nc}	.	.	1.040	8.14***	0.879	9.18***
γ^{L}	3.526	2.58***	3.540	2.59***	3.568	2.61***
γ^{l}	207.237	17.19***	206.923	16.43***	206.409	15.78***
$\gamma^{ANF}, \overline{\gamma}^{ANF}$	5.716	6.50***	5.720	6.47***	7.703	6.23***
σ_{ANF}	.	.	.	.	0.979	2.65***
γ^{AF}	4.011	6.79***	3.989	6.83***	3.965	6.38***
γ^{X}	-1.195	1.23	-1.251	1.29*	-1.324	1.29*
β_{L}^{nc}	-0.060	2.03**	-0.061	2.05**	-0.061	2.06**
β_{l}^{nc}	-1.545	8.03***	-1.552	7.95***	-1.558	7.91***
β_{ANF}^{nc}	-0.036	5.19***	-0.038	5.30***	-0.037	5.19***
β_{AF}^{nc}	-0.001	0.13	-0.005	0.51	-0.002	0.16
β_{X}^{nc}	-0.017	1.39*	-0.011	0.88	-0.014	0.95
β_{l}^{L}	-0.147	0.54	-0.151	0.55	-0.158	0.58
β_{ANF}^{L}	-0.014	1.31*	-0.014	1.32*	-0.016	1.36*
β_{AF}^{L}	-0.012	0.88	-0.012	0.90	-0.013	0.92
β_{X}^{L}	0.013	0.77	0.014	0.81	0.014	0.85
β_{ANF}^{l}	-0.071	1.30*	-0.072	1.32*	-0.076	1.39*
β_{AF}^{l}	-0.538	5.64***	-0.542	5.56***	-0.546	5.63***
β_{X}^{l}	0.084	0.69	0.093	0.77	0.095	0.78
β_{AF}^{ANF}	-0.008	2.29**	-0.009	2.59***	-0.007	1.63*
β_{x}^{ANF}	0.050	1.16	0.060	1.43*	0.075	1.74**
β_{X}^{AF}	-0.029	0.50	-0.019	0.33	-0.029	0.45
β^{nc}	-0.644	11.98***	-0.876	16.42***	-0.834	15.60***
β^{L}	-1.009	10.38***	-1.007	10.36***	-1.006	10.34***
β^{l}	-21.356	17.28***	-21.323	16.51***	-21.262	15.85***
β^{ANF}	-0.373	11.49***	-0.381	11.71***	-0.395	11.31***
β^{AF}	-0.227	13.26***	-0.235	13.65***	-0.333	6.67***
β^{X}	0.069	1.37*	0.070	1.42*	0.081	1.46*

Niveaux de signification : *** 10%, ** 5%, *** 1%.

Note. Logit mixte, estimé par le maximum de vraisemblance simulé. La forme fonctionnelle estimée est décrite par (2.19). L'hétérogénéité inobservable est prise en compte par le coefficient de la partie linéaire de la fonction d'utilité, en supposant une distribution normale : $\gamma_k \equiv N(\overline{\gamma}_k, \sigma_k)$. *Spécification 1* : Logit multinomial ; *Spécification 2* : heures de travail non-clinique ; *Spécification 3* : heures de travail non-clinique et actes non-facturables.

Tableau 2.10 – Qualité du modèle estimé

	Observé Ensemble	Prédit 2002	Observé 2002
Heures hebdomadaires totales	54.62	55.92	53.04
_____ cliniques (h^c)	47.21	48.77	48.70
_____ non cliniques (h^{nc})	7.42	7.16	4.33
Semaines (W)	45.96	46.28	45.71
Actes[a] totaux	165.59	167.55	163.05
_____ facturables (AF)	144.52	145.03	144.55
____ non facturables (ANF)	210.66	225.20	185.07
Effort $\left(e = \frac{ANF + AF}{h^c * W}\right)$	76.33	74.23	73.24
Revenu annuel[a] (X)	169.29	228.93	222.92

[a]En milliers de Dollars. Les actes sont mesurés en Dollars constants (base 1996).

Note. Comportements de pratique moyens observés sur l'ensemble de la période (première colonne) ou en 2002 (dernière colonne) et comportements de pratique moyens prédits par le modèle estimé pour l'année 2002 (colonne centrale).

correctement prédites à partir des préférences estimées. Les prédictions du modèle restent cependant fortement influencées par les comportements réels qui ont permis son estimation. Ainsi, le modèle prévoit difficilement la chute importante des heures non cliniques observée en 2002 (4.3 heures en moyenne pour cette année, contre plus de 7 pour l'ensemble de la période) et tend à prédire un comportement proche de celui qui a valu pendant l'ensemble de la période. Cette imprécision se reporte sur les heures de travail totales, légèrement surestimées elles-aussi.

Les propriétés dérivées de ces préférences estimées tendent également à en confirmer la validité. En raison de la forme analytique de la fonction d'utilité Translog, il faut cependant noter que les effets marginaux dépendent non seulement des paramètres de la fonction d'utilité, mais également du niveau des variables de pratique. Nous fournissons donc une évaluation des propriétés locales moyennes des préférences estimées, évaluées en utilisant pour chaque individu le niveau observé des variables de pratique. Nous

utilisons pour ce faire l'année 1998, exempte de la rémunération mixte. Le Tableau 2.11 présente en particulier la distribution de l'échantillon en termes d'utilité marginale en fonction du sexe des individus. L'estimation par discrétisation, adoptée ici, exclu *a priori* les points intérieurs de l'ensemble budgétaire. La cohérence de la méthode nécessite donc que l'utilité marginale du revenu soit positive (van Sœst, 1995, p.68). Quel que soit le sexe des individus, cette hypothèse est respectée par une écrasante majorité des observations (première colonne, 99.7% au total). L'utilité marginale du loisir (deuxième colonne) reflète une forte préférence des chirurgiens en faveur du travail. Ce résultat est assez intuitif au regard de la part qu'occupe le travail dans une semaine-type d'activité (60 heures de travail hebdomadaires en moyenne, Tableau 2.8). Les préférences à l'égard des heures de travail non-clinique (dernière colonne), enfin, confirment assez largement le cadre adopté dans l'analyse théorique, considérant ces heures de travail comme une forme particulière de loisir. Un écart important apparaît en fonction du sexe, les femmes manifestant une préférence beaucoup plus forte pour ces activités d'administration et d'enseignement.

L'effet propre de la rémunération mixte sur l'offre de soins dépend de la réponse des variables de pratique à la variation simultanée du taux de rémunération des actes et du *per diem*. La prochaine section en propose une évaluation, à partir de simulations de dispositifs alternatifs de rémunération. Elles permettent en particulier de lever les ambiguïtés mises en évidence par l'analyse théorique.

Tableau 2.11 – Utilités marginales

	X	l	h^{nc}
Femme	99.86	27.67	45.91
Homme	99.74	29.20	39.91
Total	99.77	28.85	41.27

Note. Proportion des individus de l'échantillon pour lesquels l'utilité marginale du revenu (1° colonne), du loisir (2°) et des heures non-cliniques (3°) est positive, en fonction du sexe. En %.

TABLEAU 2.12 – VARIATIONS INDUITES PAR L'INTRODUCTION DE LA RÉMUNÉRATION MIXTE

	RA	RM volontaire	Variation
Heures hebdomadaires totales	54.29	55.92	3 %
——— cliniques (h^c)	47.21	48.77	3.3 %
——— non cliniques (h^{nc})	7.08	7.16	1.1 %
Semaines (W)	45.99	46.28	.6 %
Actes[a] totaux	176.70	167.55	-5.2 %
——— facturables (AF)	152.33	145.03	-4.8 %
—— non facturables (ANF)	24.36	22.52	-7.6 %
Effort $\left(e = \frac{ANF + AF}{h^c * W}\right)$	81.38	74.23	-8.8 %
Revenu annuel[a] (X)	161.92	228.93	41.4 %

[a]En milliers de Dollars. Les actes sont mesurés en Dollars constants (base 1996).

Note. Comportements de pratique moyens prédits par le modèle pour l'année 2002 selon que le schéma de rémunération correspond à la rémunération à l'acte (colonne de gauche) ou au dispositif de rémunération mixte existant (colonne centrale). La variation (dernière colonne) correspond au taux de variation relatif entre la première et la deuxième colonne.

b) Simulations

L'estimation d'un modèle structurel permet en effet de générer les réponses optimales des médecins à tout changement hypothétique de la contrainte budgétaire. les paramètres estimés permettent d'identifier l'alternative qui rend maximale l'utilité sous la contrainte budgétaire supposée ; les comportements de pratique simulés correspondent alors aux choix de pratique inclus dans cette alternative. Les simulations sont réalisées pour l'année 2002, qui est la seule année d'existence de la rémunération mixte incluse dans l'échantillon.

Le Tableau 2.12 compare les comportements de pratique simulés de l'ensemble des médecins selon que la rémunération mixte existe dans ses dispositions actuelles (colonne centrale) ou que le dispositif pré-réforme a été maintenu, contraignant l'ensemble des

médecins de l'échantillon à conserver la rémunération à l'acte (première colonne). Par définition, seuls le médecins qui choisissent librement d'adopter la rémunération mixte lorsqu'elle est disponible sont affectés par ce changement. La variation induite par l'introduction de la rémunération mixte (dernière colonne) correspond donc à l'effet propre de la réforme sur les comportements de pratique des chirurgiens du Québec. La rémunération mixte engendre d'abord un important accroissement de revenu (plus de 41%) pour les médecins qui la choisissent. Comme nous l'avons vu (Tableau 2.3), ces médecins se caractérisent par un niveau de revenu plus faible à heures de travail comparables. En ce sens, il semble donc que la rémunération mixte parvienne à rétablir l'équité des rémunérations offertes aux médecins indépendemment de la diversité de leur pratique.

En termes d'offre de soins, la réforme affecte principalement le niveau d'effort consentit par les médecins. Le léger accroissement des heures de travail clinique (+1.1%) s'accompagne en effet d'une baisse beaucoup plus importante du nombre d'actes fournis (−5.2%), facturables comme non facturables. Les médecins consacrent en conséquence un temps plus important à la réalisation de chacun des actes (+8.8%). Les heures de travail non-cliniques sont, quant à elles, peu affectées par le passage à la rémunération mixte. L'augmentation simultanée des heures de travail clinique et non-clinique engendre cependant un accroissement non négligeable (+3%) des heures passées au travail. Enfin, les semaines de travail apparaissent assez insensibles au mode de rémunération conformément au résultat obtenu par les études classiques consacrées au sujet (Sloan, 1975). Dans l'ensemble, ces résultats ne satisfont que partiellement les objectifs poursuivis lors de l'introduction de la rémunération mixte. Si le temps consacré aux actes en constitue une mesure adéquate, l'objectif de promotion de la qualité des soins semble être atteint. La rémunération mixte ne provoque, en revanche, qu'une diversification très modérée des activités de pratique. Au regard de l'important effet de sélection remarqué plus haut sur cette variable (Tableau 2.8), la rémunération mixte semble donc avoir pour seul effet de rémunérer les médecins qui manifestent une préférence marquée pour les activités non-cliniques.

Tableau 2.13 – Variations induites par une rémunération mixte obligatoire

	RA	RM obligatoire	Variation
Heures hebdomadaires totales	54.29	48.50	-11.9 %
_____ cliniques (h^c)	47.21	41.71	-13.2 %
_____ non cliniques (h^{nc})	7.08	6.78	-4.3 %
Semaines (W)	45.99	46.51	1.1 %
Actes[a] totaux	176.70	160.03	-10.4 %
_____ facturables (AF)	152.33	143.66	-6.0 %
____ non facturables (ANF)	24.36	16.38	-48.8 %
Effort $\left(e = \frac{ANF + AF}{h^c * W}\right)$	81.38	82.48	1.3 %
Revenu annuel[a] (X)	161.92	190.72	15.1 %

[a]En milliers de Dollars. Les actes sont mesurés en Dollars constants (base 1996).

Note. Comportements de pratique moyens prédits par le modèle pour l'année 2002 selon que le schéma de rémunération correspond à la rémunération à l'acte (colonne de gauche) ou à un dispositif qui contraint l'ensemble des médecins à adopter la rémunération mixte (colonne centrale). La variation (dernière colonne) correspond au taux de variation relatif entre la première et la deuxième colonne.

D'une façon générale, ces variations – quantitativement faibles – se traduisent par une importante augmentation du coût du système de santé en termes de rémunération des médecins. L'introduction de la rémunération mixte s'avère donc un instrument puissant de rééquilibrage des rémunérations entre médecins, accompagnée d'une légère amélioration de la qualité. Elle apparaît ainsi comme un instrument plus efficace sur le plan des objectifs politiques que de l'efficience économique. Cette conclusion quant à l'efficacité de la combinaison d'instruments choisie doit cependant être nuancée par l'effet des modalités de mise en œuvre adoptées.

Afin de l'évaluer, le Tableau 2.13 présente les changements qui auraient prévalu si la rémunération mixte avait été rendue obligatoire pour l'ensemble des chirurgiens. La plupart des résultats précédents sont alors renversés, selon une ampleur démultipliée. A l'exception du revenu, qui continue à augmenter quoique beaucoup plus légèrement,

l'offre de soins est en effet réduite dans toutes ses dimensions par cette version de la réforme. Une très importante diminution des heures de travail clinique, associée à une baisse substantielle mais moins marquée du nombre d'actes délivrés, conduit ainsi à une diminution du temps consacré aux patients (accroissement de l'effort). La baisse des heures de travail non clinique vient par ailleurs renforcer celle du temps de travail clinique, qui résultent en une baisse importante du temps de travail des médecins.

Bien que les instruments utilisés semblent quantitativement inappropriés à promouvoir l'efficience de l'offre de soins, la rémunération mixte tire donc un parti important de la liberté d'adoption laissée aux médecins.

2.5 Conclusion

L'étude des effets attendus et observés de l'introduction de la rémunération mixte sur les choix de pratique des médecins a permis, dans ce chapitre, d'approfondir l'analyse quant à l'influence des rémunérations sur l'arbitrage qualité/quantité de l'offre de soins. Combinant une rémunération fixe et un taux de rémunération des actes réduit, la rémunération mixte est explicitement destinée à encourager la diversification des activités médicales des médecins et l'amélioration de la qualité des soins fournis aux patients. Ce changement dans les incitations s'accompagne potentiellement d'un effet de sélection, puisque les médecins du Québec peuvent, après l'introduction de la réforme en 1999, conserver l'ancien système de rémunération.

L'analyse théorique de l'effet attendu de son adoption a mis en évidence les propriétés des préférences des médecins suffisantes à ce que la réforme parvienne à atteindre ses objectifs. Sous les conditions traditionnellement retenues quant au goût pour le loisir (bien normal) comme pour la consommation (utilité marginale positive), la réaction des médecins à la réforme dépend uniquement de la sensibilité des choix de pratique aux

variations de prix. En raison de l'arbitrage entre marge intensive (temps consacré aux patients) et marges extensives (temps de travail et nombre d'actes) cet effet est plus ambigu que ne le laissent présager les travaux existants. Certaines configurations des préférences (valeurs relatives des élasticités croisées) assurent cependant que le passage à la rémunération mixte aboutisse à un double accroissement des heures de travail consacrées à l'enseignement ou à l'administration des établissements, et du temps consacré à chaque patient. Ces effets positifs en termes de diversification et d'amélioration de la qualité peuvent s'accompagner d'une réduction du temps total de travail des médecins.

L'analyse économétrique du comportement des médecins, observé de 1996 à 1998 et en 2002, permet de lever ces indéterminations pour les médecins appartenant à l'échantillon retenu. Les préférences des médecins sont estimées grâce à une base de données originale, combinant des données d'enquête sur le temps consacré au travail par les médecins et des données administratives sur les quantités d'actes délivrés, et leur valorisation monétaire. Les résultats permettent de simuler le comportement optimal des médecins sous divers modes de rémunération. En comparaison des choix qu'aurait engendré le maintien de la seule rémunération à l'acte, la rémunération mixte a eu pour effet d'augmenter légèrement l'ensemble des heures de travail et, surtout, d'accroître de façon substantielle le temps consacré à chaque acte médical. L'effet le plus important reste cependant un accroissement considérable du revenu versé aux médecins, suite à la prise en compte d'activités qui étaient jusqu'alors exclues des rémunérations. En ce sens, la rémunération mixte s'avère être une réforme coûteuse dont les effets sur la santé sont modestes, mais permettant de promouvoir l'équité des rémunérations entre les médecins.

Les instruments choisis par les autorités en charge de la politique de santé au Québec (montant de la rémunération fixe et taux de réduction de la rémunération des actes) conduisent donc à des résultats très mitigés au regard de l'efficience économique. La rémunération mixte tire cependant un bénéfice considérable du caractère volontaire de son adoption. Si l'instauration obligatoire du nouveau mode de rémunération avait été

préférée au dispositif en vigueur, en effet, la réforme aurait eu pour effet d'abaisser l'offre de soins dans toutes ses dimensions, du nombre d'heures de travail au temps consacré à chaque acte, et aurait simultanément accru le revenu versé aux praticiens. S'ils conduisent à douter de l'efficacité des niveaux de rémunération choisis, nos résultats abondent par conséquent dans le sens des revendications de plus en plus fréquentes en faveur d'une plus grande liberté dans les choix de rémunération.[42]

Ces conclusions proviennent d'une première estimation, utilisant le sous-ensemble des médecins chirurgiens. Bien que cette population constitue à de nombreux égards (diversité de la pratique, variété des choix de rémunération) un échantillon représentatif de la population des médecins, elles ne sauraient donc être définitives sans que soit évalué l'effet de la rémunération mixte sur l'ensemble des spécialités de pratique. Il faut noter, en particulier, que l'offre de pratique des chirurgiens est en partie rationnée par la disponibilité des équipements dans les établissements hospitaliers. Les médecins de cette spécialité peuvent n'être en conséquence que partiellement maîtres du nombre d'actes réalisés. Pour dépasser cette limite potentielle, il s'agirait alors d'estimer les préférences de tous les médecins en activité afin de simuler la réponse optimale des choix de pratique à la réforme dans l'ensemble du Québec.

Si elles sont confirmées, ces conclusions disqualifient moins la combinaison d'instruments incluse dans le dispositif de rémunération mixte que les niveaux de rémunération choisis pour son application. Comme l'a montré l'analyse théorique, en effet, l'efficacité de la réforme dépend assez largement des propriétés locales des préférences des médecins, et une plus grande efficacité du dispositif n'est pas à exclure *a priori*. L'estimation

[42] «*Il est urgent de compléter le paiement à l'acte par d'autres éléments de rémunération, selon le type d'exercice ou les efforts effectués en termes de qualité des soins. De plus en plus de médecins, notamment parmi les plus jeunes, sont prêts à une telle évolution. Pourquoi ne pas leur offrir ce choix, tout en permettant à ceux qui le souhaitent de continuer à être payés uniquement à l'acte ?*»"Sécu : la solitude de l'assuré" P-Y. Geoffard, *Libération* (3 octobre 2005). Les articles, cités plus haut, de Encinosa, Gaynor & Rebitzer (1997) et Barro & Beaulieu, (2003) mettent en évidence l'efficacité de ce type d'auto-sélection.

d'un modèle structurel permet de simuler les choix de pratique qui résultent de tout système de rémunération. Un prolongement naturel de notre analyse consisterait par conséquent à chercher la combinaison optimale entre rémunération fixe et taux de rémunération des actes, c'est à dire les niveaux de rémunération (fixe et variable) tels que l'amélioration des soins soit maximale pour un coût minimum. Une seconde lacune, difficile à combler, tient à ce que notre analyse néglige la décision de groupe qui préside à l'adoption de la rémunération mixte. Si cet aspect pourrait être intégré à l'analyse théorique, les clauses de confidentialité nous interdisent en revanche d'accéder aux informations quant au groupe d'appartenance des médecins. Une investigation empirique s'avère de ce fait irréalisable.

Enfin, notre analyse laisse de côté deux aspects importants de l'effet des incitations sur les comportements de pratique, que les données dont nous disposons pourraient permettre d'appréhender. D'une part, il semble établi que les variations dans les taux de rémunération des actes sont susceptibles de donner lieu à un phénomène de demande induite. Cet effet a déja, au Québec, été observé dans le passé (Rochaix, 1993) et pourrait modifier de façon importante l'analyse coûts–bénéfices de l'effet de la réforme. Outre les variables de revenu, les données administratives fournies par la *Régie de l'Assurance Maladie du Québec* contiennent également des informations détaillées sur le nombre de patients traités et le nombre de visites réalisées par chaque médecin. L'effet de l'introduction de la rémunération mixte sur ces mesures de la demande de soins pourrait donc fournir une évaluation de l'ampleur de la demande induite engendrée par la réforme. Outre l'introduction de la rémunération mixte, la période couverte par nos données contient également, d'autre part, un certain nombre d'ajustements (montants et modalités d'application) dans les mesures de plafonnement qui s'appliquent au revenu des médecins. Compte tenu de l'originalité de cette mesure, connaître son effet sur les choix de pratique participerait à approfondir la compréhension du rôle des incitations dans la politique de santé. La mise en œuvre du dispositif s'appuie sur une distinction des activités médicales en fonction, notamment, de l'établissement de pratique. Un traitement adéquat de cet aspect nécessiterait une discrétisation spécifique des variables de pratique, et doit donc faire l'objet de nouvelles investigations.

Annexes

2.A Programme Ox

Le programme, écrit en langage Ox, a été adapté à partir du programme GAUSS développé par Train.[43] La structure du programme et les noms de variables ont été autant que possible conservées. Nous décrivons dans un premier temps les éléments spécifiques à notre version. Les matrices requises ont été crées à l'aide du logiciel STATA. Les commandes qui contrôlent la parallélisation requièrent le package OxMPI (Doornik, Shephard & Hendry, 2004).

2.A.1 Lexique des variables et matrices utilisées

```
 //* Code Files *//


par.ox
--------
Contains all usefull controls for the current estimation.
debug.ox
--------
Contains parameters controling extensive printing in debug mode
logit.ox
--------
Main code for one processor running
logitMPI.ox
--------
Main code for parallelized running


 //* Input files *//


xb.mat
--------
File containing NOBS rows padded with individual caracteristics values (0 or 1 for sex)
```

[43] Le programme est librement disponible sur la page personnelle de l'auteur, sous le titre "Mixed Logit Estimation for Panel Data using Maximum Simulated Likelihood".

```
for each variable
you want to include in the estimation.
cons.mat
--------
File containing the constant variables values in each alternative.
Must contain NALT*sumc(CONS) elements.
yvec.mat
--------
Contains the id of chosen alternative for each observation.
RM.mat
--------
Contains choice uncertainty for each observation. 0 if choice is certain. Otherwise, contains
the number of
alternatives subsequent to the one designated in yvec.mat between which the observation
may have chosen.
rescale.mat
--------
Contains 2 columns : the first indicates the id of the variable (1 for the first, 2 for
the second, ...) to
be rescaled the second the rescale factor. rows(rescale)n can be lower than NVAR.
xmat.mat
--------
Contains the values of non-constant variables in each alternative for each observation.
times.mat
--------
Identifies the number of times for which each of the NP agents chose an alternatives-eg.
the number of choices
each consumer made. For example, if the first agent faced 3 choice situations (e.g., made
a choice in each of three time periods)
and the second agent faced 7 choice situations, then TIMES is 3, 7, .... The sum of TIMES
over its NP elements
must equal the number of observations NOBS.


 //* par.ox *//


PREDICT
--------
Controls prevision behavior in the last likelihood calculation

 0 = Never compute predicted probability for the sample
 1 = Compute predicted probability for each individual in the sample at estimated parameters
 2 = Predict only mode : only compute predicted probabilities at starting values
NPRED
--------
The number of variables involved in prediction for PREDICT!=0. The matrix of predicted
values for each observation
witll contain NPRED columns, each (c) resulting from alternative j estimated probability
product with variable c.
Variables must be sorted according to NPRED order in Xmat.

RESCALE_START
```

```
--------
Controls starting values rescaling : starting values will be rescaled according to variables
rescale factor if 1.
Usefull while restarting an interrupted estimation.
CENSOR
--------
XXX
CONS
--------
1*NVAR vector identifying parameters associated with constant variable. A variable is
constant if its value
is the same for each observation. Constant variable matrix thus contains only one row,
used for every observation.
idXB
--------
Vector containing one row per individual caracteristics variable. If the variable is of
dummy type, put the
number of single values for which you a want a parameter estimate (e.g. 1 for sexe).
sumc(iDXB) must be the number of columns in xb.mat
iDB
--------
NVAR*1 vector associating parameters with individual caracteristics variables.
No interaction between parameter and individual variables when 0. If >0, put the id of
variable in iDXB.
As an example, if you want the second and fourth parameters among seven to be estimated
in interaction with sexe :
iDXB=<1>, iDB<0 ; 1 ; 0 ; 1 ; 0 ; 0 ; 0> ;
START
--------
Controls starting values : if string, starting values are loaded from "start.mat". The
file must contains
NFC+NN*2 values. Otherwise, put every numerical value, which will be used as default starting
for all parameters.
TITLE
--------
String containing the title for current estimation
PRINT
--------
Controls printing during iteration. Intermediate results will appear every PRINT iteration.
HALT
--------
Controls halton sequence. Loaded from file HMNAME if 1, created if 0.
DRAWS
--------
Controls simulation draws
 1 = SIMPLE
 2 = HALTON

QUICK
--------
Controls the gradient computation method used for optimisation.
 0 = Numerical Derivatives (Robust but time consumming)
 1 = Analytical Gradient (Quicker)
GRAD
```

```
--------
XXX
GRADOBS
--------
XXX


 //* MPI management *//


tag
--------
tag is the variable used by the master for sending tasks to nodes
 10 = Compute likelihood
 20 = Likelihood computation for probabilities prediction
 30 = Numerical derivatives for hessian calculation
seek
--------
One row per sequentially loaded file. Each contains the adress in the corresponding file
where loading has
previously stopped.


 //* debug.ox *//


DEBUG
--------
Controls extensive printing. Main calculation steps printed if 1.

debug_ll
--------
Commands a session without optimisation. Likelihood computation will be performed only
once.
debug_err
--------
If DEBUG, the matrix of random draws will be printed once.
debug_v
--------
If DEBUG, informations about the matrix containing utility based on fixed coefficients
will be printed once.
debug_ev
--------
If DEBUG, informations about the matrix containing utility will be printed once.
debug_expev
--------
If DEBUG, informations about the matrix containing exp(utility) will be printed once.
debug_denom
--------
If DEBUG, the denominator in individual contribution to likelihood will be printed once.
debug_p00
--------
If DEBUG, the probability of chosen alternative will be printed once.
```

```
debug_y
--------
If DEBUG, the vector of NALT dummies for choice will be printed once.
debug_p1
--------
If DEBUG, the sum over alterantive probabilities will be printed once (should be 1).
debug_mpi
--------
Nodes and master prints communiactions steps.
debug_halt
--------
Number of drawn Halton sequences to print if DEBUG.


 //* logit.ox : Program specific variables *//
IDXB
--------
Describe xb.mat : number of variable for each dummy in first column, position of
the first in the second.
IDB
--------
Describe parameters : individual caracteristic to be intercted with in first column,
number of estimated coeficients in the second.
```

2.A.2 Programme

```
#include <oxstd.h>
#import <maximize>
#include <par.ox>
#include <debug.ox>
static decl   XMAT, XCONS, YVEC, TIMES, YPERM, XB, RM, RSCLMAT, IDXB, IDB, MATCENSOR;
const decl   IDA = IDNC, HMNAME = "hm15.asc";
decl      id, numproc, procname, SIM=0, HM, tag;
decl      IDCONS, IDSPEC, SPEC, NEVAR, N, MyN, MyObs;
       start();
       data();
       rescale(rsclmat);
       rescB(b, const rsclmat);
       logitll( b, ll, score, hess);
       halton();
       haltonserial(n, s);
       cdfinvn(p);
       compll(b, ll, score, hess);
       score(b);
       ll(b, ll, score, hess);
       ScoreContributions(const func, vP, const avScore);
#include <single.ox>
//#include <mpi.ox>
main()
  {if (SINGLE)
     single();
```

```
26    else
27      mpi();
28    decl BETA;
29
30    OxMPIBarrier();        // Nodes synchronization for sequential printing
31    if (id==0)             // Master Prints session title
32      {println("");
33      println(date(), " ", time());
34      println(TITLE);
35      println(NP, " Individuals, ", NOBS, " observations, ", NALT, " Alternatives.");
36      println("");}
37
38    OxMPIBarrier();        // Nodes synchronization before names printing
39
40                    // Will contain people allocated to nodes
41    decl nodesNP = zeros(numproc,1);
42
43    if (id==0)              // Allocates observations between nodes
44                    // (fixing for non cylindered panel data)
45      {decl   allocate = cumulate(reshape(loadmat("times.mat",1),NP,1)),
46           i = 1, node = 1, aimed = floor(NOBS/(numproc-1)), done = 0;
47      while (i <= NP-1)
48           {if (aimed < allocate[i][0])
49                    // Number of raws read allocated to node
50           {nodesNP[node][0] = i - sumc(nodesNP[0 : node-1][0]);
51           done = allocate[i-1][0];
52           allocate = allocate - done;
53           node = node + 1;}
54         i = i + 1;}
55
56                    // Master in charge with the residual
57    nodesNP[0][0] = NP - sumc(nodesNP);}
58
59    OxMPIBcast(&nodesNP,0);    // Sends people vector to nodes
60    MyN = nodesNP[id][0];
61    println("Process ", id+1, " of ", numproc," on ", procname, ", MyN = ", MyN);
62    OxMPIBarrier();
63    data();                // Data loading
64
65    tag = 10 + QUICK;
66    if(id == 0)
67      {BETA = start();      // Master loads and rescales starting values
68      if (RESCALESTART*RESCALE)
69        {rescB(&BETA, RSCLMAT);
70        print("Rescaled ");}
71      print("Starting values : ", BETA);
72      decl lik;
73      if (PREDICT!=2)        // No computation in predict only mode
74
75          //* One likelihood calculation for running time indication *//
76
77        {println("");
78        println("Starting calculations...");
79                    // Calculations starting time
```

```
80          decl timeconv0 = OxMPIWtime(), time0 = time();
81          compll(BETA, &lik, 0, 0);
82          decl timeconv1 = OxMPIWtime();
83          println("");      // end time
84          println("Calculation started at ", time0, " finished at ", time());
85          print("First calculated likelihood : ", lik, " in ", timeconv1-timeconv0, "s");
86
87            //* Optimization *//
88
89          if (debugll==0)
90            {decl std;
91            MaxControl(NITER,1);
92            println("");
93            print("Starting optimization");
94            if (QUICK)
95              print(" using analytical gradient");
96            time0 = time();
97            timeconv0 = OxMPIWtime();
98            decl optim = MaxBFGS(compll, &BETA, &lik, 0, 1-QUICK);
99            if (optim == 4 )
100             println("Estimation failed, you should try a DEBUG session");
101           timeconv1 = OxMPIWtime();
102           println("Estimation started at ", time0, " finished at ", time());
103           println(MaxConvergenceMsg(optim), " obtained in ", timeconv1-timeconv0 , "s"
104 );
105           //* Estimation results *//
106
107           println("");
108           println("Preparing results...");
109
110           tag = 30;      // First derivatives calculation
111           decl H = score(BETA);
112           if (ROBUST)      // Numerical 2nd derivatives for robust variances
113             Num2Derivative(compll, BETA, &std);
114           else
115             std = H;
116
117           decl sigma = sqrt((diagonal(invertgen(std, 3)*H*invertgen(std,3)))');
118           decl t = BETA./sigma;
119           decl PrintB = BETA;
120
121           if (RESCALE)   // Descale before printing and saving
122             {decl descl = ones(rows(RSCLMAT), 2);
123                    // Create descaling matrix
124             descl[][0] = RSCLMAT[][0];
125             descl[][1] = 1 ./ RSCLMAT[][1];
126                    // Descale parameter and standard errors values
127             rescB(&PrintB, descl);
128             rescB(&sigma, descl);}
129
130           savemat("sigma.mat", sigma, 1);
131           savemat("beta.mat", PrintB, 1);
132
133                    // Results printing
```

```
134           decl printer = PrintB   sigma   t;
135
136           println(" Parameter Estimated Standard");
137           println(" value Error t");
138           println(" -------------------------");
139           println("Fixed Coefficients :");
140
141           decl k = 0, estimated = < >, pr = 0, pr1 = -1;
142           while (k <= NFC-1)
143             {pr1 = pr + IDB[k][columns(IDB)-1]-1;
144             print(constant(IDFC[k],IDB[k][columns(IDB)-1],1)   printer[pr : pr1][]);
145             estimated = estimated | constant(IDFC[k],IDB[k][columns(IDB)-1],1);
146             pr = pr + IDB[k][columns(IDB)-1];
147             k = k+ 1;}
148           if (NNC > 0)
149             {println("Normally distributed coefficients : ");
150             println(shape(IDNC | IDNC, 2*NNC,1)   printer[pr1 + 1 : ][]);
151             estimated = estimated | shape(IDNC | IDNC, 2*NNC,1);}
152           if (ROBUST)
153             println("Uses robust standard errors.");
154           else
155             println("Uses non robust standard errors. WARNING not reliable for random
156 coefficients.");
157           print("Final loglikelihood : ", lik);
158           savemat("results.mat", estimated   PrintB   sigma   t,1);}
159         }
160       if (PREDICT!=0)        // Predicted probabilities at BETA while PREDICT=1 or 2
161         {if (GRAD)
162           {tag = 50 + GROBS;
163           decl G = score(BETA);
164           println("");
165           savemat("gradient.mat", G, 1);}
166         SIM = 1;
167         println("");
168         println("Preparing forecasts...");
169         tag = 20;         // Send job to nodes
170         OxMPIBcast(&tag,0);
171                     // Mean over observations
172         if (RESCALE)      // Descale before forecasts
173           {decl DSCLMAT = RSCLMAT[][];
174           DSCLMAT[][1] = 1 ./ DSCLMAT[][1];
175           rescale(DSCLMAT);
176           rescB(&BETA, DSCLMAT);}
177         decl FC;
178         compll(BETA, &FC, 0, 1-QUICK);
179         FC = FC ./NOBS;
180                     // Print prediction results
181         decl printer = cumulate(ones(NPRED,1))   FC   (FC[][0]-FC[][1])./FC[][1];
182         println("");
183         println(" Variable Predicted Actual Rel. Error");
184         println("----------------------------");
185         println(printer);
186         savemat("forecast.mat",printer,1);}
187
```

```
188       tag = 99;           // Announce the end to nodes
189       OxMPIBcast(&tag,0);
190       println("End");}
191     else
192       {while(tag!=99)        // Nodes loops on calculations
193         {if (debugmpi)
194           println("Process ", id+1, " waiting for tag...");
195         OxMPIBcast(&tag,0);// Stop looping when tag==99
196         if (debugmpi)
197           println("Process ", id+1, " received tag : ", tag);
198         if (tag!=99)
199           {decl LL;      // Parameters in current iteration
200           OxMPIBcast(&BETA,0);
201           if (debugmpi)
202             println("Process ", id+1, " last received beta : ", BETA[NVAR-1]);
203           if (tag==20)   // Likelihood calculated for prediction
204             SIM=1;
205
206           if (tag!=30)    // Likelihood for My Obs at received BETA
207             ll(BETA, &LL, 0, QUICK);
208           if (tag==30)   // Numerical 1st derivative
209             score(BETA);
210           }
211         else           // Computations stops when tag=99
212           {if (debugmpi)
213               println("Process ", id+1, " about to exit...");}
214         }
215       }
216     if (debugmpi)
217       println("Process ", id+1, " has finished jobs");
218
219     OxMPIFinalize();        // Break communications
220   }
221           //* Data loading *//
222   data()
223     {NEVAR = NFC + 2*NNC;      // Number of estimated parameters
224   //    NVAR = columns(IDFC)+columns(IDNC);
225
226     SPEC = (CONS .== 0);      // Dummies identifying specific and constant variables
227     decl VARS = SPEC*ones(NVAR,1), VARC = CONS*ones(NVAR,1);
228     if (id == 0 && DRAWS == 2)    // Master generates halton sequences
229       HM = halton();
230   //* Nodes specific sequential loading *//
231     decl    seek = zeros(7,1),    // Opens files to be read
232         ftimes = fopen("times.mat", "r"), fxmat = fopen("xmat.mat", "r"),
233         fy = fopen("yvec.mat", "r"), frm = fopen("RM.mat", "r");
234
235     if (DRAWS == 2)
236       decl fhalt = fopen("temphalt.mat", "r");
237     if (sumc(iDB)!= 0)
238       decl fxb = fopen("xb.mat", "r");
239     if (CENSOR == 1)
240       decl fcensor = fopen("censor.mat", "r");
241     decl master = 1;        // For coherency with people allocation, loading
```

```
242   while (master >= 0)        // must start with nodes (master=1), finish with
243                     // master (master=0)
244     {for(decl i = 0; i <= (numproc-2)*master; i=i+1)
245       {if (id == i + master )
246         {if (MyN>0)      // Debug the case where master's MyN=0
247           {println("");
248           println("Loading process ", id+1, " datas...");
249
250                 // Load each individual's number of choices in process i
251           fseek(ftimes,"c", seek[0]);
252           fscan(ftimes, "
253           seek[0] = fseek(ftimes);
254           MyObs = sumc(TIMES[]);
255           print(" MyObs : ", MyObs);
256 //           println("TIMES ok");
257
258                 // Load specific variables values in process i
259           fseek(fxmat,"c", seek[1]);
260           fscan(fxmat, "
261           seek[1] = fseek(fxmat);
262 //           println("XMAT ok");
263
264                 // Load chosen alternative id in process i
265           fseek(fy,"c", seek[2]);
266           fscan(fy, "
267           seek[2] = fseek(fy);
268 //           println("YVEC ok");
269
270                 // Load RM in process i
271           fseek(frm,"c", seek[3]);
272           fscan(frm, "
273           seek[3] = fseek(frm);
274 //           println("RM ok");
275
276                 // Distribute generated halton sequences to nodes
277           if (DRAWS == 2)
278             {fseek(fhalt,"c", seek[4]);
279             fscan(fhalt,"
280             seek[4] = fseek(fhalt);
281 //              println("HM ok");
282             }
283                 // Load individual specific variables
284           if (sumc(iDB)!= 0)
285             {fseek(fxb,"c", seek[5]);
286             fscan(fxb,"
287             seek[5] = fseek(fxb);
288 //              println("XB ok");
289             }
290                 // Load censoring variable
291           if (CENSOR == 1)
292             {fseek(fcensor,"c", seek[6]);
293             fscan(fcensor,"
294             seek[6] = fseek(fcensor);
295 //              println("MATCENSOR is ", rows(MATCENSOR), " * ", columns(MATCENSOR));
```

```
296                 }
297               }
298             }
299                       // Seek of the loaded files sent to all procs
300         OxMPIBcast(&seek,i+master);}
301       master = master - 1;}    // Switch to master's loading
302     fclose(frm);
303     fclose(fy);
304     fclose(fxmat);
305     fclose(ftimes);
306     if (DRAWS==2)
307       fclose(fhalt);
308     if (sumc(iDB)!= 0)
309       fclose(fxb);
310   //* Common simultaneous loading *//
311
312     if (RESCALE)           // Load rescaling matrix
313       {RSCLMAT = loadmat("rescale.mat",1);
314       RSCLMAT = shape(RSCLMAT,2,NVAR)';}
315                       // Create dependent variable permutation matrix : 1 if alternative
316   is chosen; 0 otherwise
317     YPERM = zeros(MyObs,NALT);
318     decl i = 0;
319     while (i <= MyObs-1)
320       {YPERM[i][(YVEC[i]-1) : (YVEC[i]-1+RM[i])] = 1;
321       i = i + 1;}
322   //    println("iDB ", iDB);
323     IDB = iDB   ones(NVAR,1);             // Individual Caracteristics pointer
324     if (sumc(IDB[][0])!= 0)
325       {IDXB = idXB   zeros(rows(idXB),1);
326                       // Augments BETA dimensions with individual intercepts
327       decl i = 0;
328       while (i <= NVAR-1)
329         {if (IDB[i][0]!=0)
330           {decl col = 0;
331           while (col <= columns(IDB)-2)
332             {IDB[i][columns(IDB)-1] = IDB[i][columns(IDB)-1] + IDXB[IDB[i][col]-1][0];
333             // Number of b[i] estimated
334             NEVAR = NEVAR + IDXB[IDB[i][col]-1][0];
335             col = col + 1;}}
336         i = i + 1;}
337
338       i = 1;              // Variables adress in XB
339       while (i <= rows(IDXB)-1)
340         {IDXB[i][1] = IDXB[i-1][0] + IDXB[i-1][1];
341         i = i + 1;}}
342   //    println("IDXB :",IDXB);
343   //    println("IDB ",IDB);
344
345     if (MyN>0)              // Check RM treatment
346       {decl temp = sumc((sumc(YPERM'))')-sumc(RM)-MyObs;
347       if (VERBOSE)
348         {println("");
349         if (temp==0)
```

```
350           println("RM correctly treated in process ", id+1 ) ;
351         else
352           println("YPERM contains", temp , "more choices than what RM indicates in process
353 ", id+1 ) ;
354         }
355
356    IDCONS = IDSPEC = zeros(NVAR,1) ;
357    decl k = 0, idc = 1, ids = 1 ;
358    while (k <= NVAR-1)        // Identify the position of constant and specific variables
359      {if (CONS[k])
360        {IDCONS[k] = idc ;
361        idc = idc + 1 ;}
362      else if(SPEC[k])
363        {IDSPEC[k] = ids ;
364        ids = ids + 1 ;}
365      else
366        {if (VERBOSE)
367          println("Variable ", k+1, " is identified neither as constant nor as specific") ;}
368      k = k + 1 ;}
369    if (DEBUG)
370      {if (id==0)
371        println("IDCONS ", IDCONS, "IDSPEC ", IDSPEC) ;}
372                    // Load constant variables values
373    decl X1 = loadmat("cons1.mat", 1), X2 = loadmat("cons2.mat", 1) ;
374    XCONS = X1 | X2 ;
375    if (DEBUG && id==0)
376      {println("X1 contains : ", rows(X1)) ;
377      println("X2 contains : ", rows(X2)) ;
378      println("VARC : ", VARC, " NVAR : ", NVAR) ;}
379
380    if (VERBOSE && id==0)      // Master checks files
381                    // TIMES file
382        {decl temp=loadmat("times.mat",1) ;
383        temp = reshape(temp,NP,1) ;
384        println("") ;
385        if (sumc(temp)==NOBS)
386          println("TIMES file seems to be correct.") ;
387        else
388          println("Check TIMES file : does not fit NOBS.") ;
389                    // XCONS files
390        println("") ;
391        if (rows(XCONS) == VARC*NALT)
392          println("Constant variables correctly loaded.") ;
393        else
394          println("Constant variables file contains ", rows(XCONS), " elements, it should
395 have ", VARC*NALT) ;
396        println("") ;
397        if (sumc(iDB)!= 0)
398          println(columns(XB), " Individual caracteristics variables loaded.") ;
399        }
400      }
401                    // Reshape constant variables vector to its true dimensions   VARC
402 rows, NALT columns
403    XCONS = shape(XCONS, NALT, VARC)' ;
```

```
404   rescale(RSCLMAT);        // Rescale variables values using RSCLMAT
405 }
406         //* Starting values (master only) *//
407 start()
408   {decl b;
409   if (isstring(START))      // Starting values loaded if START="Y"
410     {b = loadmat(START,1);
411     b = reshape(b,NEVAR,1);
412     if (VERBOSE)
413       {println("");
414       println("Last estimation values loaded as starting");}
415     }
416   else               // START taken as default for all parameters otherwise
417     {b = ones(NEVAR,1)*START;
418     if (VERBOSE)
419       {println("");
420       println("Default value used as starting");}
421     }
422   return b;
423 }
424         //* Rescale datas using rsclmat *//
425 rescale(rsclmat)
426   {if (RESCALE)
427     {decl km;
428     if (id==0)          // Master prints the current operation
429       {if (VERBOSE)
430         {println("");
431         println("Rescaling data...");}
432       if (DEBUG)
433         {if (rsclmat[][0] > NVAR)
434           println("RSCLMAT identifies a variable that is not in the data set.");
435         println(" ");
436         println(" Variable Mult. Factor");
437         print(rsclmat[][0]   rsclmat[][1]);}
438       }
439
440     decl j = rows(rsclmat)-1;
441     decl i = 0;
442     while (i <= j)        // Rescale the variables
443       {if (CONS[rsclmat[i][0]-1])
444         {XCONS[IDCONS[rsclmat[i][0]-1]-1][] =
445           XCONS[IDCONS[rsclmat[i][0]-1]-1][] * rsclmat[i][1];}
446       else if (SPEC[rsclmat[i][0]-1])
447           {km = (IDSPEC[rsclmat[i][0]-1]-1)*NALT;
448           if (DEBUG)
449             println("km = ", km);
450           XMAT[][(km) :(km+NALT)-1] =
451             XMAT[][(km) :(km+NALT)-1] * rsclmat[i][1];}
452       i = i + 1;}
453     if (id==0)
454       println(" ...done");
455     }
456 }
457         //* Rescale parameters (master only) *//
```

```
458 rescB(b, const rescl)
459   {decl j = rows(rescl)-1, idb = 0, i = 0, colB = 0, maxB = columns(IDB)-1;
460   while (i <= NFC-1)        // Rescale fixed starting values
461     {decl ki = 0, l = 0;
462     while (l <= j)
463       {if (rescl[l][0] .== IDFC[i])
464         {decl nbB = 0;
465         while (nbB <= maxB-1)
466           {colB = IDB[IDFC[i]-1][nbB]-1;
467           if (0 <= colB)
468             ki = ki + IDXB[colB][0];    //
469           nbB = nbB + 1;}
470         (b[0])[idb : idb + ki ] =(b[0])[idb : idb + ki ] / rescl[l][1];
471         idb = idb + ki + 1;}
472       l = l + 1;}
473     i = i + 1;}
474
475   i = 0;                // Rescale normal starting values
476   while (i <= NNC-1)
477     {decl l = 0;
478     while (l <= j)
479       {if (rescl[l][0] .== IDNC[i])
480         {(b[0])[idb] =(b[0])[idb] / rescl[l][1];
481         (b[0])[idb + 1] =(b[0])[idb + 1] / (rescl[l][1]);
482         idb = idb + 2;}
483       l = l + 1;}
484     i = i + 1;}
485 }
486         //* Nodes coordination for likelihood comptutation (master only) *//
487
488 compll(b, LIK, score, hess)
489   {if (debugmpi)
490     decl timell = time();
491
492   if (debugmpi)
493     println("Process ", id+1, " sending tag : ", tag );
494
495   OxMPIBcast(&tag,0);      // Send tag to nodes
496   OxMPIBcast(&b,0);        // Send iteration parameters to nodes
497
498   if (debugmpi)
499     println("Process ", id+1, " last received b in compll : ", b[NVAR-1] );
500   decl LL;             // Iteration likelihood is saved in LL
501   LIK[0] = ll(b, &LL, score, hess);
502
503   if (debugmpi)
504     println("Calculated likelihood : ", LIK[0], " in : ", timespan(timell));
505
506   return 1;
507 }
508         //* LogLikelihood computation *//
509
510 ll(b, ll, score, hess)
511   {decl Mylik, Myscore = zeros(NEVAR,1);
```

```
512   if (debugmpi)
513     println("Process ", id+1, " last received b in ll : ", b[NVAR-1]);
514
515   if (MyN>0)             // Each process computes its observations loglikelihood if positive
516     {logitll(b, &Mylik, score, hess);
517     if (tag == 10 + QUICK || tag == 30)
518       Mylik = sumc(Mylik);
519     if (score)
520       Myscore = Myscore + sumc(score[0])';}
521
522   if (debugmpi)
523     println("Process ", id+1, " Mylik : ", Mylik );
524
525   if (score)
526     (score[0]) = OxMPIReduce(Myscore, MPISUM, 0);
527                    // LogLikelihoods summed over nodes and sent to master
528   return (ll[0]) = OxMPIReduce(Mylik, MPISUM, 0);
529 }
530           //* Score or Hessian computation *//
531 score(b)
532   {decl score;
533   OxMPIBcast(&tag,0);
534   OxMPIBcast(b,0);
535   if (MyN>0)             // Each node computes MyObs's hessian,
536     {if (ROBUST)         // stored in score
537       {ScoreContributions(logitll, b, &score);
538       if (tag!= 50+GROBS)
539         score = score'*score;}
540     else
541       {decl lik;
542       if (tag == 50+GROBS)
543         logitll(b, &lik, &score, 0);
544       else
545         logitll(b, &lik, 0, &score);}}
546   if (debugmpi)
547     println("Process ", id+1, " H first row : ", score[0][]);
548
549                    // Return sum over nodes hessians
550   return OxMPIReduce(score, MPISUM, 0);
551 }
552           //* Individual contributions to likelihood computation *//
553
554 logitll(b, LL, score, hess)
555   {decl X, err;
556   if (SIM)
557     decl PP=zeros(NPRED,2);      // For each predicted variable, contains the sum of
558 predicted levels
559   if (debugmpi)
560     println("Process ", id+1, " last received b in logitll : ", b[NVAR-1]);
561
562   decl time0 = time();
563   decl v = zeros(MyObs,NALT);
564   decl p0 = zeros(MyN,1);      // Simulated probability
565   decl der = zeros(MyN,NEVAR);// Jacobian matrix
```

```
566   decl derobs = zeros(MyObs,NEVAR);
567   decl p0obs = zeros(MyObs,1);
568   decl maxB = columns(IDB)-1, k = 0, idb = 0;
569 //* Adds variables with fixed coefficients *//
570
571   while (k <= NFC-1)
572   //* Constructs the relevant alternatives variables *//
573     {if (CONS[IDFC[k]-1])
574       X = XCONS[IDCONS[IDFC[k]-1]-1][].* ones(MyObs,NALT);
575     else if (SPEC[IDFC[k]-1])
576          {decl km = (IDSPEC[IDFC[k]-1]-1)*NALT;
577          X = XMAT[][(km) :(km+NALT-1)];}
578   //* Alternative constant *//
579     v = v + b[idb].*X[][ : ];    // Matrix (NOBS * NALT)
580   //* Interaction terms *//
581     idb = idb + 1;
582     decl nbB = 0;
583     while (nbB <= maxB-1)
584       {decl colB = IDB[IDFC[k]-1][nbB]-1;
585       if (0 <= colB)
586         {decl ki = 0;
587         while (ki <= IDXB[colB][0]-1)
588           {v = v + b[idb + ki].*XB[][IDXB[colB][1] + ki].*X[][ : ];
589           ki = ki + 1;}
590         idb = idb + ki;}
591       nbB = nbB + 1;}
592     k = k + 1;}
593
594   if (DEBUG*debugv)
595     {println("v has ", rows(v), " rows and ", columns(v), " columns" );
596     println("v first individual : ", v[0 :TIMES[0]-1][]);
597     println("ev first element : ", exp(v[0][0]));
598     debugv = 0;
599     }
600
601 //* Loop on individuals   : random coefficients *//
602   decl rd = 0, n = 0;
603   while (n <= MyN-1)
604
605   //* Loads random draws for simulation *//
606
607     {if (DRAWS == 1)        // Random draw
608       {decl NECOL = max(NVAR-NFC,1);
609       err = rann(NREP,NECOL);}
610
611     else if (DRAWS == 2)    // Halton sequence for individual n
612       {err = HM[(NREP*n) :(NREP*n+NREP-1)][];
613       if (DEBUG*debugerr)
614         println("HM = ", HM);}
615
616     if (DEBUG*debugerr*DRAWS)
617       {println("n = ", n, " err :", err);
618       debugerr = 0;}      // err has NREP rows and NECOL columns for each person
619
```

```
620      decl p00 = ones(NREP,1);// Simulated probabilities for individual n
621                      // Score vector for individual n, one row per simulation
622      decl g = zeros(NREP,NEVAR);
623  //      idb = idb + 1;
624  //      print("idb : ", idb);
625    //* loop over individual i observations *//
626      decl t=1, ev;
627      while (t<=TIMES[n])
628        {decl kmm = rd + t -1;
629      ev = v[kmm][];      // Contains utillity derived from fixed coefficients for each
630  alt (columns)
631      decl k = 0;
632      while (k <= NNC-1) // Adds variables with normal coefficients
633           {if (CONS[IDNC[k]-1])
634             X = XCONS[IDCONS[IDNC[k]-1]-1][];
635           else if (SPEC[IDNC[k]-1])
636             {decl km = (IDSPEC[IDNC[k]-1]-1)*NALT;
637             X = XMAT[kmm][(km) :(km+NALT-1)];}
638                      // Matrix (NREP * NALT)
639        ev = ev + (b[idb+(2*k)] + (b[idb+(2*k)+1] .* err[][k])) .*X[][ : ];
640          if (DEBUG*debugev)
641            {println("ev is ", rows(ev), " * ", columns(ev));
642            println("ev first raw : " , ev[0][]);
643            debugev = 0;
644            }
645          k = k+1;}
646                      // Individual i dummies for choice at period t
647        decl y = YPERM[kmm][];
648        if (DEBUG*debugy)
649          {println("y", y);
650          debugy = 0;}
651
652        ev = exp(ev);      // Probabilities based on exp(U)
653        if (DEBUG*debugexpev)
654          {println("exp(ev) has " , rows(ev), " r ", columns(ev), " c" );
655          debugexpev = 0;}
656                    // Probability denominator
657        decl denom;
658        if (CENSOR==1)
659          denom = (ev * MATCENSOR[kmm][])';    // Matrix (1 * NREP)
660        else
661          denom = (sumr(ev))'; // Matrix (1 * NREP)
662        if (DEBUG*debugdenom)
663          {println("denom = " , denom);
664          debugdenom = 0;}
665
666                    // Chosen alternative probability
667        if (CENSOR == 1)
668          {p0obs[kmm] = meanc(((sumr(ev.*y.* MATCENSOR[kmm][]))'./denom)');
669          p00 = p00.*((sumr(ev.*y.* MATCENSOR[kmm][]))'./denom)';}
670        else
671          {p0obs[kmm] = meanc(((sumr(ev.*y))'./denom)');
672          p00 = p00.*((sumr(ev.*y))'./denom)';}
673        if (DEBUG*debugp00)
```

```
674           {println("p00", p00);
675           debugp00 = 0;}
676                     // Row vector of NALT choice probabilities
677         decl p1;
678         if (CENSOR == 1)
679           p1 = ((ev .* MATCENSOR[kmm][])./denom'); // Matrix (NREP * NALT), row sum should
680 be unitary
681         else
682           p1 = (ev./denom'); // Matrix (NREP * NALT), row sum should be unitary
683         if (DEBUG*debugp1)
684           {println("p1 :", p1);
685           println("sum : ", sumr(p1));
686           debugp1 = 0;}
687
688
689         if (SIM)        // In prediction mode : increments predicted levels with individual
690 n contribution
691           {decl k = 0, F = zeros(NPRED,NALT), p = meanc(p1);
692           while (k < NPRED)
693             {if (CONS[k])
694               F[k][] = XCONS[(IDCONS[k]-1)][];
695             else if (SPEC[k])
696               F[k][] = XMAT[kmm][((IDSPEC[k]-1)*NALT) : ((IDSPEC[k]-1)*NALT+NALT-1)];
697             k = k + 1;}
698           F = exp(F);
699           decl CORRECT = (F .!= 1);
700           F = F .* CORRECT;
701                     // Column vector containing the sum of predicted choices
702 //           if (CENSOR == 1)
703 //             PP[][0] = PP[][0] + F*p' .* MATCENSOR[kmm][];
704 //           else
705             PP[][0] = PP[][0] + F*p';}
706
707 //* Gradient specific calculations *//
708
709         if (hess || score || DEBUG)
710   //* Increment gradient with calculated probabilities *//
711           {decl idb = 0, k = 0;
712         //* Constructs the relevant alternatives variables *//
713           while (k <= NFC-1)
714             {if (CONS[IDFC[k]-1])
715               X = XCONS[IDCONS[IDFC[k]-1]-1][];
716             else if (SPEC[IDFC[k]-1])
717               {decl km = (IDSPEC[IDFC[k]-1]-1)*NALT;
718               X = XMAT[kmm][(km) :(km+NALT-1)];}
719
720                     // Alternative constant
721             g[][idb] = g[][idb] - sumc((p1.*X[][])')
722             + sumc(((ev.*X[][]).*y)')./(sumc((ev.*y)'));
723                     // Interaction terms
724             idb = idb + 1;
725             decl nbB = 0;
726             while (nbB <= maxB-1)
727               {decl colB = IDB[IDFC[k]-1][nbB]-1;
```

```
728                if (0 <= colB)
729                  {decl ki = 0;
730                  while (ki <= IDXB[colB][0]-1)
731                    {g[][idb] = g[][idb] - sumc((p1.*X[][]*XB[kmm][IDXB[colB][1] + ki])');
732                    g[][idb] = g[][idb]+ sumc((ev.*X[][]*XB[kmm][IDXB[colB][1] + ki].*y)')
733 ./(sumc((ev.*y)'));
734                    ki = ki + 1;}
735                  idb = idb + ki;}
736                nbB = nbB + 1;}
737              k = k+1;}
738
739           k = 0;        // Normally distributed coefficients
740           while (k <= NNC-1)
741             {if (CONS[IDNC[k]-1])
742               X = XCONS[IDCONS[IDNC[k]-1]-1][];
743             else if (SPEC[IDNC[k]-1])
744               {decl km = (IDSPEC[IDNC[k]-1]-1)*NALT;
745               X = XMAT[kmm][(km) :(km+NALT-1)];}
746             g[][NFC+(2*k)] = g[][NFC+(2*k)]    + sumc((X[][ : ].*y)');
747             g[][NFC+(2*k)+1] = g[][NFC+(2*k)+1] + sumc((err[][k].*X[][ : ].*y)');
748             k = k+1;}
749           derobs[kmm][] = meanc(p0obs[kmm].*g);
750           }
751         t = t+1;}
752
753       p0[n] = meanc(p00);      // Individual i contribution to likelihood
754                     // Increment hessian factor
755       der[n][] = der[n][] + meanc(p00.*g);
756       rd = rd + TIMES[n];      // Next individual address
757       n = n + 1;}
758
759
760     LL[0] = log(p0);        // Returned likelihood vector
761
762     if (SIM)
763       {decl i = 0;
764       YPERM[][]=0;
765       while (i<=MyObs-1)
766         {YPERM[i][(YVEC[i]-1) : (YVEC[i]-1)] = 1;
767         i = i + 1;}
768       for (decl k = 0; k < NPRED; k = k + 1)
769         {decl F;
770         if (CONS[k])      // Construct the vector of actual values
771           {F = XCONS[(IDCONS[k]-1)][];
772           F = exp(F);
773           decl CORRECT = (F .!= 1);
774           F = F .* CORRECT;
775           PP[k][1] = sumc((F*YPERM')');}
776         else if (SPEC[k])
777           {F = XMAT[][((IDSPEC[k]-1)*NALT) : ((IDSPEC[k]-1)*NALT+NALT-1)];
778           F = exp(F);
779           decl CORRECT = (F .!= 1);
780           F = F .* CORRECT;
781           PP[k][1] = sumc((diagonal(F*YPERM')'));}}
```

```
782                    // Sum of predicted and actual levels returned to ll
783      LL[0] = PP;
784      return 1;}
785
786  //* Hessian & Gradient : Final results *//
787    if (hess || score || DEBUG)
788      {decl det, H;
789      der = der ./p0;        // First derivative matrix
790
791      if (hess || DEBUG)      // Return hessian matrix (NVAR*NVAR) if required
792        {H = der'*der;
793        if (hess)
794          (hess[0]) = H;
795        det = determinant(H);
796        if (det == 0)      // Check Hessian determinant
797          println(id+1, " : Singular Hessian!");}
798      if (score)           // Return score matrix (NVAR*1) if required
799        {if (tag == 51)
800          (score[0]) = derobs./p0obs; //
801        else
802          (score[0]) = der;}
803
804      if (DEBUG)           // Print results for debugging
805        {println("Process : ", id+1);
806        decl sg = (sumc(der))';
807        print("Iteration LogLikelihood ", sumc(log(LL[0])));
808        println("First Derivatives matrix has ", rows(der), " rows and ", columns(der),
809  " columns");
810        println("First Derivatives matrix, diagonal : ", (diagonal(sg*sg'))');
811        decl invhess = invertgen(H, 3);
812        println("Hessian determinant : ", det);
813        println("Inverse Hessian first row : ", invhess[0][]);
814        println(" Iteration Standard ");
815        println(" Parameters Errors Gradient");
816        println("--------------------");
817        println(b   sqrt(diagonal(invhess))'   sg);
818        println(" ");
819        println("Likelihood calculated in ", timespan(time0), "ms");}
820      }
821    return 1;}
822        //* Halton sequences generation (master only) *//
823
824  halton()
825    {decl hm = < >;
826    if (HALT==1)           // Halton sequence loaded from HMNAME file
827    {hm = loadmat(HMNAME, 1);
828    println("Halton sequences loaded in process ", id+1, ".");}
829
830    else
831      {if (HALT==0)        // Halton sequence created
832      println("Creating Halton sequences in process ", id+1, " ....");
833      else
834        println("HALT must be 0 or 1. Default : Halton sequence created.");
835
```

```
836                     // Number of random estimated parameters (set to 1 if 0)
837       decl NECOL = max(NVAR-NFC,1);
838                     // Prime numbers vector
839       decl prim = < 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61,
840  67, 71, 73, 79, 83, 89, 97, 101, 103, 107, 109, 113 >;
841       if (VERBOSE)
842         println("Halton sequences are based in primes : ", prim[0 :NECOL-1]);
843       print;
844
845       decl h = 1, hm1;      // Sequences generation
846       while (h <= NECOL)
847         {hm1 = haltonserial(10+NREP*NP, prim[h]);
848       if ((h <= NNC) || (h > (NNC+NUC+NTC)))
849           {hm1 = cdfinvn(hm1);
850                     // Truncation for inverse-normal extreme values
851       hm1 = hm1.*(hm1 .<= 10) + 10 .* (hm1 .> 10);
852     hm1 = hm1.*(hm1 .>= -10) -10 .* (hm1 .< -10);}
853
854         hm = hm   hm1[10 :rows(hm1)-1][];
855       h = h + 1;}
856       println("Finished Halton sequences.");
857
858       if (DEBUG*debughalt)
859         println("Halton sequences : ", hm[0 :debughalt][]);
860
861       if (SAVH)           // Sequences saved in specified file if required
862         {savemat(HMNAME, hm, 1);
863         println("Halton sequences saved.");}
864
865                     // Sequences saved in temporary file for nodes distribution
866       savemat("temphalt.mat", hm, 1);
867       }
868     return hm;
869  }
870         //* Halton serial numbers (master only) *//
871  haltonserial(n,s)          // Use the pattern described in Train, "Halton Sequences for
872  Mixed Logit." http ://elsa.berkeley.edu/wp/train0899.pdf
873     {decl j, y, x;
874                     // Create n+1 Halton numbers including the initial zero
875     decl k = floor(log(n + 1) ./ log(s));
876     decl phi = < 0 >, i = 1;
877     while (i <= k)
878       {x = phi;
879       j = 1;
880       while (j < s)
881         {y = phi + (j / s^i);
882       x = x | y;
883       j = j + 1;}
884       phi = x;
885       i = i + 1;}
886     x = phi;
887     j = 1;
888     while ((j < s) && (rows(x) < (n+1)))
889       {y = phi + (j / s^i);
```

```
890       x = x | y;
891       j = j + 1;}
892
893     phi = x[1 :n];          // Starting at the second element gets rid of the initial zero
894     return phi;
895   }
896         //* Inverse Normal function (master only) *//
897   cdfinvn(p)
898     {decl   p0 = -0.322232431088,      q0 = 0.0993484626060,
899         p1 = -1.0,             q1 = 0.588581570495,
900         p2 = -0.342242088547,      q2 = 0.531103462366,
901         p3 = -0.0204231210245,      q3 = 0.103537752850,
902         p4 = -0.453642210148*1e-4,   q4 = 0.38560700634*1e-2;
903     if ((p > 1.0) || (p < 0.0))
904       {println("Error : Probability is out of range.");
905     break;}
906                       // Create masks for p = 0 or p = 1
907     decl mask0 = (p .== 0), mask1 = (p .== 1),inf0 = mask0 .* (-1e+300),inf1 = mask1 .*
908   (1e+300);
909
910                       // Create masks for handling p > 0.5 and p >= 0.5
911     decl maskgt = (p .> 0.5), maskeq = (p!= 0.5);
912     decl sgn = (maskgt .== 0) * (-1) + maskgt;
913                       // Convert p > 0.5 to 1-p
914     decl pn = ( maskgt - p) .* sgn + mask1 + mask0;
915                       // Computation of function for p < 0.5
916     decl y=sqrt(sqrt((-2*log(pn)).^2));
917     decl norms = y + ((((y*p4+ p3).*y + p2).*y + p1).*y + p0)./
918   ((((y*q4 + q3).*y + q2) .*y + q1).*y + q0);
919
920                       // Convert results for p > 0.5 and p = 0.5
921     norms=((norms.*sgn).*maskeq).*(1-mask0)
922         .*(1-mask1)+mask0.*inf0+mask1.*inf1;
923
924     return norms;
925   }
926         //* Michael Creel's routine for numerical derivatives of a vector
927   (http ://ideas.repec.org/c/boc/bocode/x981001.html)*//
928   const decl SQRTEPS =1E-8;    // appr. square root of machine precision
929   const decl DIFFEPS1=5E-6;      // Rice's formula : log(DIFFEPS)=log(MACHEPS)/3
930   static dFiniteDiff1(const x)
931     {return max( (fabs(x) + SQRTEPS) * SQRTEPS, DIFFEPS1);
932   }
933   ScoreContributions(const func, vP, const avScore)
934     {decl i, cp = rows(vP), left, right, fknowf = FALSE, p, h, f, fm, fp, v;
935                       // get 1st derivative by central difference
936     for (i = 0; i < cp; i++)
937       {p = double(vP[i][0]);
938       h = dFiniteDiff1(p);
939       vP[i][0] = p + h;
940       right = func(vP, &fp, 0, 0);
941       if(i==0)
942         v = new matrix[rows(fp)][cp];
943       vP[i][0] = p - h;
```

```
944       left = func(vP, &fm, 0, 0);
945       vP[i][0] = p;    // restore original parameter
946       if (left && right)
947                       // take central difference
948         v[][i] = (fp - fm) / (2 * h);
949       else
950         return FALSE;}
951
952     avScore[0] = v;
953     return TRUE;
954   }
```