Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de Docteur

Délivré par l'**Université Montpellier 1**

Préparée au sein de l'école doctorale **Économie Gestion**
Et de l'unité de recherche **LAMETA**

Spécialité: **Économie**

Présentée par **Fattouma SOUISSI BENREJAB**

Les régressions Gini-PLS:

Une application aux inégalités des revenus agricoles européens

Soutenue le 07 juillet 2016 à 10h devant le jury composé de

Arthur Charpentier	Professeur	UQAM	(Rapporteur)
Maria Noel Pi Alperin	Directrice de recherche	LISER, Luxembourg	(Examinatrice)
Jules Sadefo Kamdem	Professeur	Université de Guyane	(Examinateur)
Dorothée Buccanfuso	Professeure	Université Sherbrooke	(Rapporteur)
Stéphane Mussard	Professeur	Université de Nîmes	(Directeur)
Françoise Seyte	Maître de conférences HDR	Université de Montpellier	(Codirectrice)

UNIVERSITÉ DE MONTPELLIER

« L'Université n'entend donner aucune approbation ni improbation aux opinions émises dans cette thèse ; ces opinions doivent être considérées comme propres à leur auteur ».

Remerciements

Il faut être patient pour devenir " chercheur" .

Après 55 mois de travail épuisant, nous cueillons le fruit de cette thèse. Je tiens à remercier avec émotion toutes les personnes qui ont contribué à l'élaboration de ce travail.
En premier lieu, je voudrais exprimer ma reconnaissance, ma vive gratitude ainsi que le témoignage de mon profond respect à mes directeurs Monsieur Stéphane Mussard et Madame Françoise Seyte pour avoir accepté l'encadrement de cette thèse. Leurs conseils judicieux, leurs accueils chaleureux, leurs gentillesses et leurs bonnes qualités humaines m'ont été très précieux pour la réalisation de ce travail. J'espère que ce travail est digne de la confiance que Monsieur Stéphane Mussard m'a accordé à la fin de mon master M2 CPPER. Après mon travail de re-programmation de la macro Dagum, concernant la décomposition des revenus en sous groupes et mes résultats de mémoire sur la construction des régressions Gini-PLS (2011), Monsieur Mussard a estimé que je pouvais poursuivre mes études doctorales sur le thème des nouvelles méthodes de régression Gini-PLS. Madame Seyte, en tant que spécialiste des inégalités et des tests statistiques, a accepté la co-direction de cette thèse. Cette thèse n'aurait pas pu aboutir sans les conseils judicieux de ma co-directrice Madame Seyte, en particulier les tests de significativité des modèles. Son travail de correction, dans un lapse de temps très court, pour perfectionner la thèse est très marqué. Malgré ses nombreuses charges, Monsieur Mussard m'a élégamment guidé dans mes efforts. Son travail est très original, bien évidemment au niveau des programmations, simulations et de la partie purement théorique. Je lui remercie infiniment pour ses corrections de rédaction, des modèles et des équations. Il m'a fait apprendre de la patience pour la programmation. Nos modèles n'auraient pas pu être vérifiés sans l'aide de Monsieur Ricco Rakotomalala, à qui j'adresse mes remerciements les plus distingués.

Je suis très reconnaissante aux personnes qui ont aimé nos nouveaux algorithmes. Je remercie vivement Monsieur Michel Terraza pour ses conseils judicieux, en particulier pour le plan de la thèse. J'adresse mes remerciement les plus sincères à Monsieur Benoît Mulkay de m'avoir accordé de son temps pour m'accueillir dans son bureau. Ses conseils pertinents étaient déterminants dans la finalisation de la thèse. Je tiens à remercier Monsieur Michel Tenenhaus d'avoir apprécié la méthode, de nous avoir encouragé et aidé. Je remercie également Monsieur François Benhmad pour ses conseils judicieux et ses encouragements. Je remercie tous les membres de mon comité de thèse Messieurs François Benhmad, Michel Terraza, Jules Sadefo Kamdem, Thierry Blayac, Benoît Mulkay. Je remercie aussi Messieurs Michel Simioni et Philippe

Van Kerm pour leurs conseils.
En plus de la collaboration de plusieurs intellectuels, la disponibilité d'équipement informatique puissant et le libre accès aux plateformes scientifiques ont été déterminantes dans l'élaboration de ce travail. Je remercie tous les membres du LAMETA. Je remercie vivement Monsieur Emmanuel Sol de m'avoir aidé à installer des logiciels, récupérer mes mails, et déverrouiller l'ordinateur, je remercie également Monsieur Olivier Gadea. Je tiens à remercier les chargés de documentation Madame Patricia Modat et Monsieur Laurent Garnier pour leurs aides. Je n'oublie pas de remercier notre directeur adjoint Monsieur Thierry Blayac. Je remercie particulièrement Madame Thésy Pothet de m'avoir élégamment guidée dans mes démarches administratives. Je remercie également tout le corps administratif du LAMETA Madame Isabelle Romestan, Madame Irène Blondeau. Je n'oublie pas tous les membres du LAMETA qui m'ont aidée.
Mes plus vifs remerciements s'adressent également à l'EDEG. Je remercie les directeurs Messieurs Daniel Serra et Patrick Sentis d'avoir validé mon projet de recherche et mes ré-inscriptions annuelles. Je remercie également le secrétariat de l'EDEG et particulièrement Madame Sabine De Béchevel, qui malgré ses différentes taches, m'a aimablement guidé dans mes procédures administratives. Je n'oublie pas de remercier le corps administratif de la DRED. Mes remerciements s'adressent également à la présidence de l'Université Montpellier et en particulier à Monsieur Philippe Augé.
Je remercie vivement mes rapporteurs Monsieur Arthur Charpentier et Madame Dorothée Boccanfuso pour leurs conseils judicieux concernant les questions d'ordre technique, notamment les liens entre les approches de régressions Gini et PLS avec l'ACP, sur le fond et la forme de la thèse. Je remercie également mes examinateurs (Madame Maria Noel Pi Alperin et Monsieur Jules Sadefo Kamdem) de l'intérêt qu'il ont accordé à cette thèse.
Pour la même occasion, je voudrais adresser mes sincères remerciements et ma respectueuse gratitude aux membres de jury qui ont accepté de juger le présent travail et qui m'ont honoré de leur présence.
Je n'oublie pas de remercier tous mes enseignants, particulièrement, les professeurs intervenants à l'EDEG, les professeurs et les enseignants-chercheurs de master M2 CPPER à l'Université Montpellier 1.
J'adresse mes remerciements à tous ceux qui ont contribué de près ou de loin à l'élaboration de ce travail et qui n'ont pas été cités.

Sommaire

Introduction générale 1

1 Les modèles de régression PLS et Gini 11

2 Construction des régressions Gini-PLS 41

3 Étude des inégalités des revenus agricoles européens 79

Conclusion générale 117

Annexes 123

Introduction générale

" *L'économétrie est un vaste domaine en pleine expansion.*"
[*Greene, 2005, p. 5*].

L'explication rigoureuse des phénomènes économiques observés nécessite la mobilisation de plusieurs disciplines, en particulier, les mathématiques, les statistiques et l'économie. Les approches utilisées sont assez diverses et relèvent d'une méthodologie qui se résume en quatre points : (i) repérer les problèmes ; (ii) mettre en place des hypothèses susceptibles d'aboutir aux phénomènes observés ; (iii) tester ces hypothèses ou confronter les données collectées à un modèle ou une forme fonctionnelle pré-définie ; (iv) interpréter les résultats obtenus et conclure.

L'économie est un vaste domaine. Les approches varient en passant par des calculs simples (des ratios ou des pourcentages, par exemple) aux relations très complexes entre les variables économiques (des liens de causalités, par exemple). Trois principales branches quantitatives sont utilisées : la statistique, l'analyse des données et l'économétrie.

Les calculs statistiques servent essentiellement à décrire les données, à tester des hypothèses ou encore la robustesse[1] d'un modèle. Quant à l'analyse des données, elle retrace des statistiques multivariées, son intérêt est de traiter les observations afin de déduire des liens entre les variables ou les groupes

1. Un modèle est dit robuste lorsqu'il est valable quelles que soient les circonstances.

de variables. L'analyse des données a pris de l'ampleur avec les techniques d'analyse en composantes principales et d'analyse des correspondances multiples. La méthode d'analyse en composantes principales est intéressante. Les interprétations qui en découlent sont effectuées sur la base de distinction des groupes d'individus ayant les mêmes caractéristiques, en les projetant par des axes principaux. La description seule ne résout pas l'ensemble des problèmes rencontrés en économie. Pour cette raison, la recherche des liens de causalité entre les variables explique mieux les interactions entre les variables de la sphère économique.

Les objectifs des modèles économétriques se résument essentiellement en trois points : (i) trouver des estimateurs très proches des paramètres de la population (ii) tester des théories (iii) effectuer des prévisions.
Le modèle économétrique de base est celui des moindres carrés ordinaires (MCO). Il s'agit d'un modèle linéaire qui consiste à établir un lien entre une variable dépendante (y) et une variable explicative (x). Le modèle de régression par moindres carrés généralisées (MCG) est une forme de généralisation des MCO pour estimer les paramètres d'une relation linéaire entre une variable à expliquer (y) et plusieurs régresseurs (x). Ces modèles de régression linéaires nécessitent des hypothèses fondamentales, sans lesquelles, les régressions MCO ou MCG pourraient conduire à des résultats erronés. Les estimateurs respectant les hypothèses de base sont considérés comme les meilleurs estimateurs linéaires non biaisés ("Best Linear Unbiased Estimator : BLUE "). Aux estimateurs "BLUE" se rajoute le critère qualité d'ajustement du modèle, qui informe de la fiabilité des paramètres estimés. Il est à noter que les régressions MCO et MCG sont simplificatrices de la réalité observée. L'estimation des paramètres d'un modèle nécessite une adaptation des modèles théoriques aux données, lorsqu'au moins une hypothèse n'est pas vérifiée. Ainsi, le manque de précision dans l'explication des phénomènes économiques observés a poussé les chercheurs à utiliser ou à développer de nouvelles approches de régression de plus en plus sophistiquées. Bien que chaque méthode résout un problème particulier, l'objectif est de garantir l'exactitude des résultats et une meilleure évaluation des phénomènes observés. Il est à signaler que la multitude des mo-

dèles de régression est imputable aux problèmes des données et à l'inexistence d'approche générale permettant de résoudre tous les problèmes.

En 1983, Wold et alii. ont développé la régression par moindres carrés partiels (" Partial Least Squares regression" : PLS),[Wold et alii. (1983a, 1983b)].[2] La régression PLS est un modèle économétrique fondé sur l'analyse en composantes principales. Cette méthode permet de résoudre quelques insuffisances des modèles économétriques standards. La forte corrélation entre les régresseurs est purgée à l'aide des composantes principales orthogonales. Ces composantes de la régression PLS, construites comme combinaisons linéaires des variables explicatives, permettent de pallier au problème de faible taille d'échantillon[3]. Un autre avantage de cette régression est sa capacité à estimer un modèle en cas de données manquantes dans l'échantillon[4]. L'estimation des coefficients de régression avec la méthode PLS est efficace pour obtenir des estimateurs cohérents, de point de vue amplitude et signes des paramètres estimés. Cependant, elle ne peut pas s'affronter à d'autres difficultés comme l'endogénéité des variables explicatives. L'endogénéité peut provenir des valeurs aberrantes, des erreurs de mesure, d'une double causalité entre les régresseurs et la variable dépendante, ou de l'omission de certaines variables. Pour pallier à ces problèmes, certaines approches économétriques sont proposées, en particulier la régression sur variables instrumentales et les régressions Gini. Nous nous intéressons dans ce qui suit aux régressions Gini.

La régression Gini initiée par Olkin et Yitzhaki (1992) a pour objectif d'ajuster les paramètres estimés. Cette méthode de régression repose sur l'indice de Gini, un indice alternatif à la variance. Il existe deux régressions sur indice de Gini (nommées régressions Gini) : l'approche par minimisation ou paramétrique et l'approche semi-paramétrique. Dans la régression Gini semi-paramétrique les coefficients sont obtenus à partir d'une moyenne pondérée des tangentes issues de toutes les paires d'observations. Cette régression permet

2. Voir aussi, Tenenhaus M. (1998) pour une revue de littérature.

3. On parle de faible taille de l'échantillon lorsque le nombre de variables explicatives dépasse le nombre d'observations. Dans ce cas, les MCO ne peuvent pas s'appliquer.

4. Lorsque la matrice des données est incomplète, les calculs matriciels ne peuvent pas s'effectuer.

de réduire l'influence des valeurs aberrantes, [Olkin et Yitzhaki, (1992)].

La régression Gini par minimisation ou paramétrique introduite par Olkin et Yitzhaki en 1992 concerne la minimisation de l'indice de Gini des résidus au lieu de la variance traditionnelle. Les estimateurs sont ainsi déterminés à partir de la norme $l1$ (appelée encore moindres déviations) [5] alors que les moindres carrés ordinaires (MCO) sont issus de la norme $l2$ [6], ainsi une meilleure robustesse est obtenue lorsque les erreurs dévient de la distribution normale. Les approches de régression Gini fournissent un meilleur ajustement en présence de valeurs aberrantes. Ils peuvent affecter l'amplitude et les signes des coefficients estimés de façon excessive [Choi, (2009)]. La contamination [7]9 des données peut aussi exclure la possibilité d'obtenir une inférence valide lorsque les variances des coefficients estimés sont très grandes (voire infinies), on parle d'un problème d'instabilité, voir Dixon (1950), et John (1995). La force de la régression Gini est de pouvoir estimer les paramètres d'un modèle sans recourir aux moments d'ordre 2, (Voir Yitzhaki et Schechtman (2013) pour un aperçu de la méthodologie du Gini.)

Les deux objectifs de notre thèse sont de proposer : (i) de nouveaux modèles de régression " Gini-PLS" [8] univariés et multivariés [9] et (ii) une nouvelle approche de régression Gini-PLS basée sur la décomposition en sources de revenus (RISD-Gini-PLS) [10].

5. Moindres déviations (MD) (norme $l1$) : les estimateurs sont obtenus par la somme des valeurs absolues des écarts.

6. Norme $l2$ ou moindres carrés (MC) : les estimateurs sont obtenus par minimisation des carrés des résidus, lorsque les erreurs sont normalement distribuées.

7. Données contaminées : ou déformées en présence de valeurs aberrantes, de multicollinéarité ou d'endogénéité des variables explicatives.

8. Les modèles Gini-PLS sont issus de la combinaison des régressions Gini et PLS.

9. Dans les modèles univariés, on a une seule variable dépendante, alors que dans les modèles multivariés le nombre de variables cibles est au moins égal à deux.

10. Notre approche RISD-Gini-PLS : Regression based Income Source Decomposition-Gini-PLS nous permet d'estimer les contributions des facteurs des exploitations à l'inégalité totale des revenus agricoles européens

Il est à signaler que les conséquences des deux modèles Gini et PLS sont assez similaires. En présence de valeurs aberrantes et de corrélations importantes entre les variables explicatives, les estimations sont biaisées, pire, des inversions de signes sont fréquemment rencontrées ce qui rend problématique l'interprétation des résultats d'une équation de régression. Le problème des valeurs aberrantes peut être équivalent, sous certaines conditions, aux problèmes d'erreurs de mesure que les modèles standards PLS ou Gini sont incapables de traiter. Les modèles de régression Gini-PLS, que nous proposons, sont des substituts aux MCO, MCG, ils permettent de résoudre simultanément les problèmes d'endogénéité, d'erreurs de mesure, de multi-colinéarité, de valeurs aberrantes, de faible taille de l'échantillon et de données manquantes. Nous construisons les algorithmes Gini-PLS univariés et multivariés, respectivement. Un autre apport de la thèse consiste à insérer les paramètres estimés des régressions (Gini, PLS et Gini-PLS) dans une mesure d'inégalité, à savoir l'indice du Gini absolu. Les paramètres estimés sont donc convertis en contributions des sources à l'inégalité totale des revenus agricoles européens. Malgré le budget important consacré aux aides des exploitations agricoles européennes et les différentes réformes de la Politique Agricole Commune (PAC), les inégalités des rémunérations persistent. Il est indéniable que le concept d'inégalité est considéré comme un phénomène abstrait. Cependant, il existe différentes études qui le simplifient. Les premiers travaux se résument en quelques représentations graphiques, en particulier la courbe de concentration de Lorenz. Cette courbe permet de décrire la répartition d'une variable (revenu, santé, bien-être) dans la population. À partir de cette courbe, les chercheurs ont calculé des ratios des indicateurs comme l'indice du Gini. Ces ratios évaluent plus précisément les inégalités notamment car ils sont décomposables et permettent d'établir des liens de causalités avec d'autres variables économiques.

Notre application de ces régressions Gini-PLS concerne l'estimation des contributions des sources de revenus aux inégalités des rémunérations agricoles des pays européens. Les résultats obtenus permettent de cerner l'impact des réformes de la PAC sur les inégalités. Dans les modèles univariés, la variable dépendante correspond aux revenus des produits bruts des exploitations agri-

coles européennes, alors que les revenus des productions agricoles et d'élevage représentent les deux variables dépendantes utilisées dans les modèles multivariés. Ces estimations permettent de distinguer simultanément l'impact des réformes de la PAC et des orientations technico-économiques des exploitations agricoles sur les inégalités.

En réalité, les problèmes de mesure des inégalités sont dûs non seulement à l'interaction avec les composantes de l'environnement économique (liens entre inégalité et croissance économique, par exemple), mais aussi aux interférences de certaines variables socio-économiques, comme les fortes corrélations entre les sources de revenus. Les approches économétriques sont donc plus adaptées pour répondre à ces problèmes, notamment la régression PLS qui reste valide en cas de corrélation excessive entre les régresseurs.

Les modèles de Morduch et Sicular (2002) et ceux de Cowell et Fiorio (2011), pour l'estimation des contributions des sources à l'inégalité totale, sont fondés sur les recherches de Rao (1969) et Shorrocks (1982) concernant la décomposition des inégalités en sources. Les contributions des sources à l'inégalité des revenus sont déduites à partir des coefficients estimés régressant les revenus des individus sur les sources de revenus. Ces travaux utilisent les estimations par les méthodes économétriques standards telles que les MCO. Il est à signaler qu'en présence de valeurs aberrantes et de fortes corrélations entre les régresseurs, les coefficients estimés ainsi que les contributions des sources de revenus aux inégalités totales sont aussi biaisées.

Notons par ailleurs que les approches économétriques standards ne permettent pas d'obtenir des estimations fiables des inégalités. Dans la plupart des situations, les bases de données concernant les rémunérations des individus mettent en évidence des problèmes de multicolinéarité, d'erreurs de mesure et de valeurs aberrantes. Cependant, avec la régression PLS ou Gini, ces difficultés sont résorbées.

Notre travail s'oriente dans le même sens que les travaux de Morduch et

Sicular (2002) et de Cowell et Fiorio (2011). L'idée est simple, une fois les rémunérations estimées à l'aide d'autres variables, les paramètres du modèle sont insérés dans une mesure d'inégalité tel que l'indice du Gini absolu. Les contributions des différentes variables à l'inégalité totale de production sont donc déduites.

Le premier chapitre intitulé **Les modèles de régression PLS et Gini** s'intérese à deux axes primordiaux de l'économétrie : (i) les problèmes de valeurs aberrantes, de multi-colinéarité et des erreurs de mesure (ii) les approches de régressions PLS et Gini qui remédient à quelques insuffisances des modèles standards des régressions.

Dans la première section, nous décrivons les trois principales difficultés économétriques (i) qui, en plus de la faible qualité d'ajustement, aboutissent à des incohérences de signes des paramètres estimés par MCO. Nous présentons dans la seconde section la régression PLS1 [11] avec ses différentes étapes, qui sert à modéliser le lien entre une variable dépendante et un ou plusieurs régresseur(s). Trois situations font appel à l'utilisation de la régression PLS1 : la présence de fortes corrélations entre les variables explicatives ; l'absence d'observations dans l'échantillon (c'est à dire, lorsque la matrice des données est incomplète) ; et la faible taille de l'échantillon. La troisième section s'intéresse à la version classique de la régression PLS2 multivariée, elle peut ainsi s'affronter aux mêmes difficultés que le modèle PLS1 en estimant les liens entre deux blocs de variables. La quatrième section détaille les régressions Gini (paramétrique et semi-paramétrique), capables de pallier aux problèmes de valeurs aberrantes et d'erreurs de mesures.

Dans le second chapitre : Construction des régressions Gini-PLS, nous proposons de nouvelles méthodes de régression Gini-PLS qui sont des combinaisons des régressions Gini et PLS. La première section s'intéresse aux modèles univariés (Gini-PLS1) où la variable dépendante est un vecteur. La seconde section s'intéresse aux régressions Gini-PLS multivariées (Gini-PLS2), qui généralisent les régressions univariées. Nous proposons deux variantes pour

11. Régression PLS univariée

chacune des deux régressions : les modèles Gini1-PLS utilisant l'opérateur Co-Gini, les modèles Gini2-PLS basés sur les coefficients de la régression Gini semi-paramétrique. Ils sont capables d'estimer les paramètres d'une régression linéaire / non linéaire. Ces modèles ne nécessitent pas de modification préalable de la base de données et aboutissent à des résultats probants en présence de valeurs aberrantes, de multi-colinéarité et d'endogénéité. Dans les deux premières sections de ce chapitre, les algorithmes sont présentés, puis les propriétés relatives à la régression PLS sont exposées.

La troisième section s'intéresse aux des simulations de Monte Carlo, pour chacun des modèles. Pour chacun des modèles il est montré qu'en présence d'erreurs de mesures (ou valeurs aberrantes) le biais d'atténuation tend vers zéro de sorte que les estimateurs convergent vers leurs vraies valeurs en limite de probabilité. [12].

Le troisième chapitre intitulé **Étude des inégalités des revenus agricoles européens** consiste à tester les modèles précédents avec des données réelles : les données agrégées sur les productions agricoles européennes. Nous nous intéressons aux inégalités des revenus entre les agriculteurs européens et notamment à l'impact des mesures de la Politique Agricole Commune sur les inégalités.

La première section décrit les modèles économétriques pour l'estimation des contributions des sources à l'inégalité totale des revenus. La première partie s'intéresse au modèle de Morduch et Sicular (2002)(RISD-MCO) [13]. Dans la sexconde section, nous proposons une approche de régression basée sur la décomposition des revenus en source dérivée des régressions Gini-PLS qui permet également de RISD-Gini-PLS, dans laquelle nous insérons les paramètres estimés des régressions Gini-PLS dans un indice d'inégalités afin de capter les contributions des sources de rémunérations à l'inégalité totale. Le challenge étant de trouver les variances estimées de ces contributions à l'inégalité totale. En utilisant l'indice de Gini absolu, nous montrons que ces variances peuvent être estimées.

La seconde section de ce chapitre s'intéresse aux comparaisons des résul-

12. Notons que nous avons programmé toutes les régressions et simulations présentées dans cette thèse à l'aide du logiciel GAUSS

13. Regression based-Income Source Decomposition utilisant la régression MCO

tats des différentes régressions (MCO, Gini, PLS et Gini-PLS). La comparaison des résultats des différentes régressions est faite à l'aide des tests de la qualité d'ajustement, d'auto-corrélation, d'hétéroscédasticité, etc. La troisième section comporte les tableaux illustratifs des différentes statistiques déduites.

Revenons tout d'abord aux modèles de régressions PLS et Gini.

Chapitre 1

Les modèles de régression PLS et Gini

Sommaire

1.1 Problématiques : valeurs aberrantes, erreurs de mesure et multi-colinéarité 14
1.2 Régression PLS1 (univariée) 19
1.3 Régression PLS2 (multivariée) 24
1.4 Covariance Gini et régression Gini 29

Introduction

Dans la plupart des domaines (économie, finance, biologie, médecine, etc.), les modèles économétriques sont très utilisés. La recherche d'estimateurs robustes s'avère donc cruciale pour aboutir à des résultats crédibles aussi bien en terme d'interprétations que de mise en place de politiques efficaces. Malgré les transformations remarquables et les tentatives d'améliorations des modèles économétriques, les méthodes fournies à travers la littérature ne répondent pas à toutes les difficultés. Depuis le début des années 1980 de nouvelles approches de régressions ont été développées pour résoudre séparément un ensemble de problèmes.

Les deux principaux objets de ce premier chapitre sont : de mettre en lumière les problèmes d'erreurs de mesure, de multicollinéarité et de valeurs aberrantes et d'énumérer quelques méthodes de régression pouvant pallier à ces trois problèmes.

Certaines techniques de régression ont été proposées dans la littérature, à savoir : la régression sur Gini pour résoudre les problèmes de valeurs aberrantes, la méthode des variables instrumentales pour pallier à l'endogénéité et la régression PLS pour résorber les problèmes de multi-colinéarité, de faible taille de l'échantillon et de données manquantes (c'est à dire lorsque la matrice des données est incomplète.

L'utilisation des modèles de régression PLS permet d'estimer des coefficients en présence de corrélations entre les variables explicatives. Les fortes corrélations sont purgées à l'aide de composantes orthogonales. Lorsque le nombre de régresseurs dépasse la taille de l'échantillon, la régression sur ces composantes orthogonales résout le problème des MCO, puisque les composantes sont des combinaisons linéaires des différentes variables explicatives. Un autre avantage de la régression PLS est sa capacité d'estimer des paramètres en absence d'observations.

D'un autre côté, les régressions Gini (paramétrique et semi-paramétrique)

permettent de pallier aux valeurs aberrantes, et aux erreurs de mesure via l'emploi de la matrice de rang. Les régressions PLS et Gini permettent aussi d'éviter les problèmes des signes inappropriés des coefficients estimés.

Commençons par les trois problèmes essentiels de l'économétrie.

1.1 Problématiques : valeurs aberrantes, erreurs de mesure et multi-colinéarité

La présence des problèmes [1] dans les bases de données, affecte considérablement les résultats des estimateurs standards par moindres carrés. Le diagnostic préalable des données s'avère très important. Ainsi, avant de construire le modèle économétrique, il faudrait systématiquement se demander si les hypothèses du modèle que l'on veut utiliser seront respectées ; ou encore si le traitement des données contaminées pourrait résoudre le problème. Les difficultés liées à la présence de valeurs aberrantes, de multi-colinéarité et d'erreurs de mesure sont très connus en économétrie. Pour cette raison, nous en discutons très brièvement avant d'introduire les régressions PLS et Gini.

1.1.1 Valeurs aberrantes : "Outliers "

Le problème des valeurs aberrantes " est très ancien" [Vasyechko et alii. (2005)]. Il existe différentes méthodes de détection des points aberrants. Chaque méthode possède sa propre description. La définition commune présente le point aberrant comme une déviation remarquable de l'ensemble des observations. Selon Ramsawmy et alii., un point aberrant peut "paraître inconsistant par rapport aux autres données",[Ramsawmy et alii. (2000)]. Dans cette perspective, les observations dans les queues de distribution peuvent être considérées comme valeurs aberrantes.

La présence de valeurs aberrantes dans l'échantillon est délicate aussi bien pour la modélisation économétrique que pour les analyses statistiques. La plupart des chercheurs (tels que Choi (2009), Furusjö et alii. (2006), Knorr et Ng (1998), Planchon (2005), Ramsawmy et alii. (2000), Vasyechko et alii.(2005)) admettent que les observations extrêmes ont un impact très net sur l'amplitude et les signes des coefficients estimés. Dans les modèles MCO, par exemple, les valeurs aberrantes amplifient les carrés des résidus. Ainsi, les interprétations

1. Par exemple : les valeurs aberrantes, la forte corrélation entre les régresseurs, l'endogénéité, la faible taille de l'échantillon, les données manquantes.

qui en découlent seront erronées. Les valeurs aberrantes étalent aussi le kurtosis de la distribution et augmentent la variance entre les individus de l'échantillon. Dans cette perspective, Ramsawmy et alii. (2000) qualifient les point atypiques de "nuisance qui affecte le processus d'inférence". Choi (2009) rajoute que cette "nuisance" peut produire des artéfacts statistiques.

Vu l'impact néfaste de la présence de valeurs aberrantes dans l'échantillon, les point de vue des chercheurs sur la manipulations des observations extrêmes ont divergé. Certains chercheurs comme Ramsawmy et alii. (2000) obligent à l'identification rapide et à l'élimination des valeurs aberrantes. D'autres, comme Vasyechko et alii. (2005) font appel au "traitement préalable des données" [2]. Nous pouvons signaler à ce niveau que le retrait d'observations aberrantes pourrait masquer des informations importantes.

La détection des valeurs aberrantes s'avère donc cruciale. "Il existe certaines méthodes algébriques, graphiques et probabilistes pour détecter les valeurs aberrantes " [Vasyechko et alii. (2005)]. Les méthodes de repérage des valeurs aberrantes ne convergent pas sur la même idée. Certaines d'entre elles sont fondées sur des représentations graphiques : citons par exemple la méthode box-plot ou diagramme en boîtes de Tukey capable de détecter les observations aberrantes, [Tukey, (1977)]. D'autres reposent sur des calculs de distances entre les observations ou par rapport aux queues de distribution. L'approche de Knorr et Ng (1998), par exemple est basée sur le concept de distance entre les observations voisines. Cette méthode ne requiert aucune connaissance sur la forme de distribution des données. Cependant, la spécification de la distance nécessite plusieurs itérations et elle ne donne aucune idée sur le rang des valeurs aberrantes. En particulier, un point ayant peu d'observations voisines peut être considéré comme fort outlier. Le problème de spécification des valeurs aberrantes est d'autant plus prononcé dès que le nombre d'observations requises croît exponentiellement avec la dimension de l'échantillon, [Knorr et Ng (1998)]. Il est important de rajouter que la méthode de régression sur Gini proposée par Olkin et Yitzhaki en 1992 revêt d'une importance majeure dans

2. remplacer les points aberrants par des caractéristiques de tendance centrale comme la moyenne ou la médiane.

la résolution des problèmes d'observations aberrantes.

Parmi les tests des valeurs aberrantes, nous pouvons citer le test de Dixon et le T^2 de Hotelling[3]. Lorsque le nombre de mesures est faible, il est difficile de rejeter un point aberrant en utilisant le test de Dixon car ce test repose sur le calcul des écarts moyens par rapport à l'étendue de l'échantillon, [Tsay (1988)]. La statistique T^2 de Hotelling pourrait être employé même pour les échantillons de faible taille. Furusjö et alii. indiquent que la régression PLS est intéressante pour la détection des valeurs aberrantes à l'aide de la statistique T^2 de Hotelling [Furusjö et alii. (2006)].

Dans ce qui suit, nous allons nous intéresser uniquement au T^2 de Hotelling pour détecter les valeurs aberrantes. Cette statistique est très utilisée avec la régression PLS, donc elle est compatible avec une taille faible de l'échantillon. Cette statistique est en sorte la probabilité qu'une observation appartienne à l'intervalle de confiance estimé. Le T^2 de Hotelling est détaillé dans la section de la régression PLS.

Lors des modélisations économétriques, on peut rencontrer des problèmes au niveau des observations, comme les valeurs aberrantes, ou bien des problèmes au niveau de la matrice des variables explicatives comme l'endogénéité.

1.1.2 Problème d'erreurs de mesure (endogénéité)

L'endogénéité indique la présence de fortes corrélations entre une ou plusieurs variables explicatives et le terme d'erreur. Le problème d'endogénéité semble être très fréquent dans les bases de données des revenus. Le test de Hausmann est une méthode de détection d'endogénéité. Nous constatons que ce test s'applique aux échantillons de grande taille. Pour les échantillons de faible taille, les erreurs de mesure[4] peuvent être repérées via l'existence des corrélations entre les variables explicatives et les résidus.

Pour détecter l'endogénéité des variables explicatives lors d'une régression,

3. *Cf.* aussi Rousseeuw et Leroy (2003), Hawkins et alii (1984) pour d'autres tests.
4. On désigne par erreurs de mesure, l'endogénéité pour les échantillons de faible taille.

il suffit de valider l'hypothèse d'exogénéité des variables x_k suivante :

$$cov(x_k, \hat{\varepsilon}) \neq 0.$$

Il s'agit donc de récupérer les résidus du modèle $\hat{\varepsilon}$ et d'effectuer la régression suivante :

$$x_k = \theta\hat{\varepsilon} + u,$$

Où u est le terme d'erreur de cette régression. On sait que :

$$\theta = \frac{cov(x_k, \hat{\varepsilon})}{\sigma^2(\hat{\varepsilon})},$$

avec $\sigma^2(\hat{\varepsilon})$ est la variance de $\hat{\varepsilon}$. Si $\theta \neq 0$, alors $cov(x_k, \hat{\varepsilon}) \neq 0$, donc la variable x_k est endogène. Elle peut être complétée par un test de significativité (T de Student par exemple).

Dans la suite, nous nous intéressons uniquement au $R(x)$ comme dans les travaux de Durbin (1954) et de Yitzhaki et Schechtman (2004).

1.1.3 Problème de multi-colinéarité

La multi-colinéarité est un problème très récurrent en économétrie. Ainsi, la présence de fortes corrélations entre les variables explicatives biaise les paramètres estimés par MCO. Par ailleurs, ces paramètres ne reflètent pas les vraies contributions des variables en question.

L'estimateur des moindres carrés est considéré comme l'estimateur ayant la plus faible variance. Néanmoins, la présence de fortes corrélations entre les variables explicatives est à l'origine de grandes variances des paramètres estimés, [Greene, (2005)].

La présence de fortes corrélations entre les variables explicatives peut engendrer des inversions de signes des paramètres estimés ou encore "des erreurs standards dont les niveaux seront importants pour les variables concernées avec des statistiques de Student très faibles pour ces variables", un coefficient d'ajustement du modèle (R^2) élevé alors que les coefficients estimés ne sont pas significatifs, [De Bourmont, (2012)].

Parmi les méthodes de détection de multi-colinéarité, nous pouvons citer

la statistique VIF (Variance Inflation Factor) définie comme suit :

$$VIF_k = \frac{1}{1 - R_k^2}$$

$\forall k = 1, \cdots, K$ variables explicatives, R_k^2 est le coefficient de détermination multiple entre les variables explicatives.

$$x_1 = \beta_2 x_2 + \cdots + \beta_K x_K : R_1^2$$

$$\vdots$$

$$x_K = \beta_1 x_1 + \cdots + \beta_K x_{K-1} : R_K^2$$

Si $VIF > 10$, la multi-colinéarité est forte.

Nous pouvons reprocher à cette statistique sa dépendance du coefficient de détermination du modèle (R^2). Ce coefficient (R^2) dépend du pouvoir explicatif du modèle de régression en question. La matrice des corrélations est la solution dans ce cas pour repérer les fortes corrélations [5].

La régression PLS permettent de résoudre les problèmes de multi-colinéarité à l'aide des composantes orthogonales.

Nous constatons que ces trois problèmes (valeurs aberrantes, multi-colinéarité et endogénéité) sont nuisibles à toute estimation robuste. Ces problèmes préconisent l'utilisation de méthodes de littérature importantes basées sur les régressions PLS, Gini ou variables instrumentales. Nous nous intéressons dans ce qui suit uniquement aux régressions PLS et Gini [6].

Nous commençons par la régression PLS. Ses différentes étapes et ses propriétés prouvent son efficacité face aux problèmes de faible taille de l'échantillon, de fortes corrélations et de données manquantes.

5. Il y a d'autres possibilités pour détecter la multi-colinéarité entre toutes les paires de variables explicatives, De Bourmont, (2012).

6. Dans le second chapitre de la thèse, le rang de x est un instrument de x pour les régressions Gini-PLS

1.2 Régression PLS1 (univariée)

La régression PLS résout quelques insuffisances des modèles économétriques standards. La régression PLS se distingue par sa capacité de purifier les données de la forte corrélation (multi-colinéarité) à l'aide de ses composantes orthogonales. La faible taille de l'échantillon [7] ne constitue pas un handicap à la régression PLS grâce aux composantes orthogonales présentés sous forme de combinaisons linéaires des différentes variables explicatives. L'absence d'observations dans un échantillon pourrait être traitée à l'aide de la régression PLS.

La régression PLS1 consiste à modéliser une variable dépendante (prédite), le vecteur colonne y de taille n, sur des composantes orthogonales $t_1, \ldots, t_h$ (vecteur colonne de taille n), où $t_h^\intercal$ est la transposée de t_h. Soit X la matrice des variables explicatives (régresseurs) x_k $(k = 1, \ldots, K)$ de taille $n \times K$. Dans ce qui suit, $x_{i,k}$ est définie comme la $i^{\text{ème}}$ observation $(i = 1, \ldots, n)$ du $k^{\text{ème}}$ régresseur.

Les variables x_k sont supposées centrées tout au long de l'analyse pour faciliter l'exposition des principaux résultats.

1.2.1 Les étapes de la régression PLS1

La régression PLS1 consiste à expliquer la variance de y à l'aide des composantes orthogonales des variables latentes $t_1, \ldots, t_h$ issues des régresseurs x_k. Le modèle estimé est donc purifié de la multicolinéarité. Les composantes sont issues d'un programme de maximisation (étape 1 de l'algorithme) qui améliore les corrélations entre la variable dépendente y et les régresseurs x_k.

• **Étape 1 :** La contribution de chaque régresseur x_k à la variable dépendante y est donné par le vecteur colonne des coefficients w_1 (de taille K) qui maximise la covariance entre X et y :

$$\max \ \text{cov}(Xw_1, y) \text{ s.t. } \|w_1\| = 1 \ .$$

Le Lagrangien est donné par :

7. Lorsque le nombre d'observations dépasse le nombre de variables explicatives

$$\begin{aligned} L &= \text{cov}(Xw_1, y) - \lambda(w_{(1)1}^2 + w_{(1)2}^2 + \cdots + w_{(1)K}^2 - 1) \\ &= \text{cov}(Xw_1, y) - \lambda(w_1^\intercal w_1 - 1) \ . \end{aligned} \tag{1.1}$$

La solution est :

$$w_{1k} = \frac{\text{cov}(x_k, y)}{\sqrt{\sum_{k=1}^{K} \text{cov}^2(x_k, y)}} \ , \ \forall k = 1, \ldots, K \ .$$

•**Remarque : lien entre Analyse en Composantes Principales (ACP) et régression PLS**

La première composante t_1 est exprimée de la même façon en PLS qu'en ACP :

$$t_1 = Xw_1$$

Les coefficients w_1 diffèrent selon la méthode utilisée.
La méthode d'ACP est non supervisée, elle détermine les coefficient w_1 comme suit :

$$w_{1k} = \arg \max_{\|w_1\|=1} \|Xw\|^2 = \arg \max_{\|w_1\|=1} w^\intercal X^\intercal X w,$$

alors que dans la régression PLS, les coefficients w_1 sont les solutions du problème d'optimisation suivant :

$$w_{1k} = \arg \max_{\|w_1\|=1} \{< y, Xw >\} = \arg \max_{\|w_1\|=1} w^\intercal X^\intercal y y^\intercal X w$$

•

La première composante de la régression PLS1 est déterminée ainsi :

$$t_1 = w_{11}x_1 + \cdots + w_{1k}x_k + \cdots + w_{1K}x_K \ .$$

Même si les régresseurs sont parfaitement corrélés, la variable dépendante y peut être régressée sur t_1 par MCO :

$$y = c_1 t_1 + \varepsilon_1 \ .$$

Le modèle entier, avec tous les régresseurs x_k, est déduit de la décomposition de la composante t_1 :

$$y = c_1 \left(w_{11} x_1 + w_{12} x_2 + \cdots + w_{1K} x_K \right) + \varepsilon_1 \ .$$

Dès que c_1 est estimé par MCO, quelques problèmes très connus surviennent lorsque les données sont contaminés de valeurs aberrantes (outliers) : la variance des coefficients estimés augmente proportionnellement à l'intensité des outliers et des signes contradictoires des coefficient estimés peuvent être enregistrés. Si les outliers sont retirés de l'échantillon, ou en général lorsqu'il y a des données manquantes dans l'échantillon, la première composante est donnée par :

$$t_{1i} = \frac{\sum_{k:\exists x_{ik}} w''_{1k} x_{ik}}{\sum_{k:\exists x_{ik}} (w''_{1k})^2} \ , \ \forall i = 1, \ldots, n \tag{1.2}$$

où

$$w''_{1k} = \frac{w'_{1k}}{\sqrt{\sum_{k=1}^{K} (w'_{1k})^2}} \ \text{ et } \ w'_{1k} = \frac{\sum_{i:\exists x_{ik}, y_i} x_{ik} y_i}{\sum_{i:\exists x_{ik}, y_i} (y_i)^2} \ .$$

Le nombre optimal de composantes est déterminé à l'aide de la validation croisée. Nous en discuterons dans la section (**1.2.2**).

• **Étape 2 :** La seconde composante t_2 est telle que $t_1 \perp t_2$. L'orthogonalité est captée à l'aide des régressions partielles de chaque x_k sur t_1 dans le but d'extraire l'information qui est indépendante de t_1, *i.e.* les résidus obtenus par MCO $\hat{u}_{(1)k}$ pour tout $k = 1, \ldots, K$ tels que :

$$\begin{cases} \text{régresser } x_1 \text{ sur } t_1 : x_1 = \beta_{11} t_1 + u_{(1)1} \\ \quad \vdots \\ \text{régresser } x_K \text{ sur } t_1 : x_K = \beta_{1K} t_1 + u_{(1)K} \ . \end{cases}$$

Le vecteur de poids w_2 lié à la seconde composante t_2 est issu de la maximisation du lien entre $\hat{U}_{(1)}$ et $\hat{\varepsilon}_1$, où $\hat{U}_{(1)}$ est la matrice $n \times K$ comportant les

colonnes des vecteurs $\hat{u}_{(1)k}$:

$$\max \operatorname{cov}(\hat{U}_{(1)}w_2, \hat{\varepsilon_1}) \text{ s.t. } \|w_2\| = 1 \implies w_{2k} = \frac{\operatorname{cov}(\hat{u}_{(1)k}, \hat{\varepsilon}_1)}{\sqrt{\sum_{k=1}^{K} \operatorname{cov}^2(\hat{u}_{(1)k}, \hat{\varepsilon}_1)}} , \ \forall k = 1, \ldots, K .$$

La seconde composante t_2 est obtenue comme suit :

$$\begin{aligned} t_2 = \sum_{k=1}^{K} w_{2k}\hat{u}_{1k} &= w_{21}(x_1 - \hat{x}_1) + \cdots + w_{2K}(x_K - \hat{x}_K) \\ &= w_{21}(x_1 - \hat{\beta}_{11}t_1) + \cdots + w_{2K}(x_K - \hat{\beta}_{1K}t_1) \ . \end{aligned}$$

Les valeurs estimées des x_k notées $\hat{x_k}$ sont obtenues à l'aide du modèle avec la composante t_1 :

$$y = c_1 t_1 = c_1(w_{11}x_1 + \cdots + w_{1K}x_K)$$

Le modèle complet avec t_1 orthogonale à t_2 est donné par :

$$y = c_1 t_1 + c_2 t_2 + \varepsilon_2 \ .$$

• **Étape h :** L'aléa estimé (résidu) [8]

$\hat{\varepsilon}_{h-1}$ des étapes $h-1$ sont liés à l'aide des résidus partiels $\hat{U}_{(h-1)}$ obtenus en régressant chaque x_k sur les composantes $t_1, \ldots, t_{h-1}$, pour tout $k = 1 \ldots, K$. En maximisant la covariance entre $\hat{\varepsilon}_{h-1}$ et $\hat{U}_{(h-1)}$, on obtient la composante t_h :

$$t_h = \sum_{k=1}^{K} w_{hk}\hat{u}_{(h-1)k} = \sum_{k=1}^{K} \frac{\operatorname{cov}(\hat{u}_{(h-1)k}, \hat{\varepsilon}_{h-1})}{\sqrt{\sum_{k=1}^{K} \operatorname{cov}^2(\hat{u}_{(h-1)k}, \hat{\varepsilon}_{h-1})}} \cdot \hat{u}_{(h-1)k} \ .$$

La régression MCO de y sur $t_1, t_2, \ldots, t_h$ donne :

$$y = c_1 t_1 + \cdots + c_h t_h + \varepsilon_h \ . \tag{1.3}$$

Même si les régresseurs sont fortement corrélés, l'emploi de la régression MCO

8. Dans les régressions PLS : l'aléa estimé correspond au résidu $\hat{\varepsilon} = y - \hat{y}$.

est possible lorsque l'hypothèse du plein rang est satisfaite (si $n \geqslant h$). Le nombre de composantes pertinentes t_h est déduit du pouvoir prédictif associé à chaque composante, appelé validation croisée.

1.2.2 Validation croisée

La validation croisée, ressemble beaucoup au "Jackknife", elle permet de retenir un nombre optimal de composantes. Soit $\hat{y}_{(h)_i}$ la valeur prédite de y_i mesurée à l'aide du modèle (1.3), qui est estimé à l'aide de toutes les observations $i = 1, \ldots, n$ et avec h composantes. D'un autre côté, soit $\hat{y}_{(h)_{-i}}$ la valeur prédite de y_i calculée à l'aide du modèle (1.3), qui est estimé avec h composantes mais sans la i^{me} observation. Le processus est alors de faire des boucles pour tout i allant de 1 à n.[9] La qualité de prédiction du modèle est mesurée à l'aide des carrés des différences entre la variable dépendante et ses prédictions, ce que l'on appelle la somme des carrés des erreurs prédites (PRESS) :

$$PRESS_h = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(h)_{-i}}\right)^2 \ .$$

La somme des carrés des résidus (RSS) du modèle avec $h-1$ composantes est :

$$RSS_{h-1} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(h-1)_i}\right)^2 \ .$$

Le ratio $\frac{PRESS_h}{RSS_{h-1}}$ indique si les prédictions du modèle avec t_h composantes sont meilleures ou non. Dans le cas où $PRESS_h$ est proche de RSS_{h-1}, il n'y a pas d'amélioration de la prédiction suite à l'utilisation de la composante t_h. La statistique suivante est alors calculée :

$$Q_h^2 = 1 - \frac{PRESS_h}{RSS_{h-1}} \ .$$

Si le modèle avec h composantes donne des prédictions meilleures, alors $\sqrt{PRESS_h}$ est suffisamment faible.Plus précisément, la composante t_h est retenue si $\sqrt{PRESS_h} \leqslant$

9. Les observations peuvent être éliminées bloc par bloc au lieu d'une par une, voir Tenenhaus (1998), p. 77.

$0.95\sqrt{RSS_h}$, c'est à dire, lorsque $Q_h^2 \geqslant 0.0975 = (1 - 0.95^2)$. Pour tester la composante t_1, la somme des carrés est calculée :

$$RSS_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \ .$$

Nous pouvons conclure que la régression PLS1 est intéressante pour répondre à quelques limites des régressions standards par moindres carrés, tels que la faible taille de l'échantillon, la multicollinéarité et l'absence d'observations. La régression PLS est très répandue dans plusieurs domaines. Ces mêmes difficultés sont résolus à l'aide de la régression PLS multivariée (PLS2).

Les résultats des régressions Gini-PLS multivariés sont différents des résultats des régressions univariés effectuée variable par variable. Les régressions multivariées tiennent compte des corrélations pouvant exister entre les variables dépendantes. Les propriétés de ces modèles, leurs différentes étapes de calcul et les aides à interprétations sont détaillées dans ce chapitre.

1.3 Régression PLS2 (multivariée)

La régression PLS2 est une généralisation de la régression PLS1. Elle permet de retrouver le lien de causalité entre deux matrices. Il existe différentes versions de la régression PLS2. Nous présentons dans ce qui suit la version "classique qui tient compte des observations manquantes". [Voir aussi Tenenhaus, (2009)].

La régression PLS2 est un algorithme qui permet de régresser les variables dépendantes Y_l $(\forall l = 1, \ldots, q)$ sur des composantes orthogonales $t_1, \ldots, t_h$. Les variables Y_l sont mises en colonne dans la matrice des variables dépen-

dantes $Y \equiv Y_{(1)}$, de taille (n, q). De la même manière que la méthode PLS1, l'algorithme PLS2 purge le modèle de la multi-colinéarité. La contribution de chaque variable X_k $(\forall k = 1, \ldots, K)$ à la variance de chaque Y_l est ainsi captée à travers les composantes t_h. Nous noterons y_{li} la variable Y_l observée sur l'individu $i = 1, \ldots, n$. Les variables Y_l et X_k sont centrées afin de faciliter l'analyse.

1.3.1 Étapes de la régression PLS2

Les étapes de la régression PLS2 sont similaires à celles que nous avions précédemment pour l'algorithme PLS1. Il s'agit ici de tenir compte du rôle tenu par plusieurs variables dépendantes y_l, l'algorithme PLS devenant un cas particulier de PLS2.

• **Étape h_0 :** La régression PLS2 consiste à modéliser le lien entre une matrice de variables dépendantes et une autre matrice de variables explicatives. L'objet de l'étape d'initialisation (étape h_0) est de trouver les composantes qui maximisent le lien entre les variables dépendantes d'une part et les variables explicatives d'autre part. L'étape h_0 que nous décrivons est nouvelle par rapport à l'algorithme PLS1. Il s'agit de trouver les pondérations W_h permettant de déterminer les composantes orthogonales t_h. L'étape h_0 se répètera à l'intérieur de chaque étape h.

↪ Initialisation d'une boucle :

↪ **[R]**épéter jusqu'à convergence de W_h. [10]

↪ Définir le vecteur u_h comme la première colonne de $Y_{(h-1)}$, avec pour tout $h \geqslant 1$, $Y_{(h)} = Y_{(h-1)} - t_h c_h^\intercal$.

↪ Régresser chaque colonne de la matrice $\hat{U}_{(h)}$ sur le vecteur u_h, on obtient le vecteur de taille K, $W_h = \hat{U}_{(h-1)}^\intercal u_h / u_h^\intercal u_h$, avec pour $h = 1$, $W_1 = X^\intercal u_1 / u_1^\intercal u_1$.

↪ Le vecteur W_h est normé : $\|W_h\| = 1$.

↪ Régresser chaque ligne de la matrice $\hat{U}_{(h)}$ sur le vecteur W_h, on obtient le vecteur de taille n, $t_h = \hat{U}_{(h-1)} W_h / W_h^\intercal W_h$, avec pour $h = 1$, $t_1 = XW_1 / W_1^\intercal W_1 = XW_1$.

10. Ici, le but derrière les boucles est de trouver la plus grande valeur propre.

$\hookrightarrow$ Régresser chaque colonne de la matrice $\hat{Y}_{(h)}$ sur le vecteur t_h, on obtient le vecteur de taille q, $c_h = Y_{(h-1)}^{\intercal} t_h / t_h^{\intercal} t_h$, avec pour $h = 1$, $c_1 = Y^{\intercal} t_1 / t_1^{\intercal} t_1$. Pour $h \geqslant 2$, $\hat{\varepsilon}_{(h-1)} \equiv Y_{(h)}$, avec $\hat{\varepsilon}_{(h)}$ la matrice contenant en colonnes les résidus des régressions ci-dessus.

$\hookrightarrow$ Régresser chaque ligne de la matrice $\hat{Y}_{(h)}$ sur le vecteur c_h, on obtient le nouveau vecteur de taille n, $u_h = Y_{(h-1)} c_h / c_h^{\intercal} c_h$, avec pour $h = 1$, $u_1 = Y c_1 / c_1^{\intercal} c_1$.

$\hookrightarrow$ Revenir à **[R]**.

- **Étape 1 :**

La première composante t_1 se détermine par :

$$t_1 = W_{(1)1} X_1 + \cdots + W_{(1)k} X_k + \cdots + W_{(1)K} X_K \ .$$

On effectue la régression par MCO de chaque Y_l sur t_1, on obtient en notant $\varepsilon_{(1)} \equiv Y_{(2)}$ la matrice $n \times q$ contenant en colonnes les vecteurs des termes d'erreur observés $\varepsilon_{(1)l}$:

$$Y_l = c_{1l} t_1 + \varepsilon_{(1)l}, \ \forall l = 1, \ldots, q.$$

Le vecteur des coefficients estimés $c_1 = (c_{11}, \ldots, c_{1l}, \ldots, c_{1q})^{\intercal}$ permet de retrouver le modèle complet avec tous les régresseurs X_k :

$$Y_l = c_{1l} \left(W_{(1)1} X_1 + W_{(1)2} X_2 + \cdots + W_{(1)K} X_K \right) + \varepsilon_{(1)l}, \ \forall l = 1, \ldots, q.$$

Dans le cas où il y a de valeurs manquantes [11] :

$$t_{1_i} = \frac{\sum_{k:\exists X_{ik}} W''_{(1)k} X_{ik}}{\sum_{k:\exists X_{ik}} (W''_{(1)k})^2}$$

où

$$W''_{(1)k} = \frac{W'_{(1)k}}{\sqrt{\sum_{k=1}^{K} (W'_{(1)k})^2}} \ .$$

11. Lorsqu'il manque des observations y_i ou x_{ik}, les calculs matriciels se font à l'aide des valeurs existantes.

et

$$W'_{(1)k} = \frac{\sum_{i:\exists X_{ki},y_i} X_{ki}y_i}{\sum_{i:\exists X_{ki},y_i}(y_i)^2} \ .$$

• **Etape 2 :** La deuxième composante t_2 doit être construite de sorte que $t_1 \perp t_2$. L'orthogonalité permet de respecter l'hypothèse de plein rang des MCO et d'extraire la multi-colinéarité du modèle. On effectue K régressions partielles des K régresseurs sur t_1 afin d'isoler tout ce que t_1 ne peut pas expliquer, autrement dit, les résidus $\hat{U}_{(1)k}$ pour tout $k = 1, \ldots, K$:

$$\begin{cases} \text{on régresse } X_1 \text{ sur } t_1 : X_1 = \beta_{(1)1}t_1 + U_{(1)1} \\ \quad \vdots \\ \text{on régresse } X_K \text{ sur } t_1 : X_1 = \beta_{(1)p}t_1 + U_{(1)K} \ . \end{cases}$$

Le vecteur poids W_2 associé à la seconde composante t_2 est déterminé conformément à la procédure expliquée à l'étape h_0. La seconde composante t_2 se détermine alors de la manière suivante :

$$\begin{aligned} t_2 &= \sum_{k=1}^{K} W_{(2)k}\hat{U}_{(1)k} = W_{(2)1}(X_1 - \hat{X}_1) + \cdots + W_{(2)K}(X_K - \hat{X}_K) \\ &= W_{(2)1}(X_1 - \hat{\beta}_{(1)1}t_1) + \cdots + W_{(2)K}(X_K - \hat{\beta}_{(1)K}t_1) \ . \end{aligned}$$

Les MCO sont appliqués en considérant les vecteurs t_1 et t_2 comme régresseurs. Pour chaque Y_l, le modèle s'écrit :

$$Y_l = c_{1l}t_1 + c_{2l}t_2 + \varepsilon_{(2)l}, \ \forall l = 1, \ldots, q.$$

On trouve donc $\hat{c}_{2l}$ (le paramètre $\hat{c}_{1l}$ reste constant du fait de l'orthogonalité des régresseurs t_h). L'hypothèse de plein rang est respectée si $n \geqslant 2$.

• **Etape h :** La matrice $\hat{\varepsilon}_{(h-1)}$ de taille $n \times q$ contenant en colonnes les vecteurs des résidus $\hat{\varepsilon}_{(2)k}$ de l'étape $h-1$ est liée à la matrice des résidus partiels $\hat{U}_{(h-1)}$ afin d'obtenir les nouveaux poids W_h (voir étape h_0). La régression de chaque Y_l sur $t_1, t_2, \ldots, t_h$ s'écrit :

$$Y_l = c_{1l}t_1 + \cdots + c_{hl}t_h + \varepsilon_{(h)l}, \ \forall l = 1, \ldots, q.$$

Le nombre de composantes du modèle est ainsi élevé jusqu'à t_h lorsque la différence de qualité d'ajustement avec le modèle à $h+1$ composantes n'est pas significative.

1.3.2 Validation croisée

Comme pour l'algorithme PLS1, la validation croisée permet de trouver le nombre optimal de composantes à retenir. Pour tester une composante t_h, on calcule la prédiction du modèle avec h composantes comprenant l'observation i, $\hat{Y}_{lh_i}$, puis sans l'observation i, $\hat{Y}_{lh_{(-i)}}$, pour chaque colonne Y_l de Y, $l = 1, \ldots, q$. L'opération est répétée pour tout i variant de 1 à n : on enlève à chaque fois l'observation i et on ré-estime les modèles pour chaque $l = 1, \ldots, q$. Pour mesurer la qualité prédictive du modèle l, on mesure l'écart entre la variable prédite et la variable observée :

$$PRESS_{h,l} = \sum_i \left(Y_{il} - \hat{Y}_{lh_{(-i)}} \right)^2, \ \forall l = 1, \ldots, q.$$

La somme des carrés résiduels obtenue avec le modèle à $(h-1)$ composantes est :

$$RSS_{h-1,l} = \sum \left(Y_{il} - \hat{Y}_{l(h-1)_i} \right)^2, \ \forall l = 1, \ldots, q.$$

Le critère somme des carrés des résidus $RSS_{h,l}$ (Residual Sum of Squares) du modèle à h composante et $PRESS_{h,l}$ (PRedicted Error Sum of Squares) sont comparés. Si le modèle avec la composante t_h améliore la prédictabilité du modèle, leur rapport augmente. La statistique suivante est alors calculée :

$$Q_h^2 = 1 - \frac{\sum_{l=1}^{q} PRESS_{h,l}}{\sum_{l=1}^{q} RSS_{h-1,l}} \ .$$

On retrouve alors la même règle de décision de l'algorithme PLS1. La composante t_h est retenue si : $Q_h^2 \geqslant 0,0975$. Elle améliore dans ce cas prévision de chaque variable Y_l. Pour la significativité de la première composante t_1, on utilise :

$$RSS_{0,l} = \sum_{i=1}^{n} \left(Y_{i,l} - \bar{Y}_l \right)^2 \ .$$

Cette procédure est un test de significativité globale des q modèles $\hat{Y}_l$. Un test moins restrictif consiste à valider une composante t_h lorsqu'au moins un $Q^2_{hl} \geqslant 0,0975$, où :

$$Q^2_{hl} = 1 - \frac{PRESS_{h,l}}{RSS_{h-1,l}} \ .$$

Les deux règles sont :

[R1] : t_h est significative lorsque $Q^2_h \geqslant 0,0975$ [12] ;

[R2] : t_h est significative si : $\exists l \in \{1, \ldots, q\} : Q^2_{hl} \geqslant 0,0975$.

Les régressions PLS (univarié et multivarié) fondées sur des composantes orthogonales servent à trouver des liens de causalités entre deux groupes de variables. Les coefficients obtenus sont robustes en absence d'observations, en présence de fortes corrélations entre les variables explicatives et lorsque le nombre de variables explicatives dépasse la taille de l'échantillon.

La régression PLS2 comme la régression PLS1 ne peuvent pas résoudre les difficultés liées à l'endogénéité comme les valeurs aberrantes et les erreurs de mesure. Une régression par moindres déviations pourrait résoudre le problème. Le lien avec la régression sur indice de Gini peut maintenant être introduit.

1.4 Covariance Gini et régression Gini

Le concept Gini ou la différence moyenne du Gini, initiée par Gini en 1912, est une caractéristique de dispersion très répandue dans le domaine de distribution des revenus. La spécificité de cet indicateur réside dans ses calculs simples. L'indice de Gini utilise la distance euclidienne entre toutes les paires de l'échantillon [Gini (1912,1914)]. Les interprétations qui en découlent sont faciles comparativement à la variance qui élève au carré les écarts entre les individus. Il est à noter qu'il existe plusieurs approches qui dérivent du Gini, [13]allant des statistiques de dispersion aux approches de régression, plus précisément la théorie des valeurs aberrantes. [Lerman et Yitzhaki (1989b), Yitzhaki et Schechtman (2013)].

Les travaux de Shechtman et Yitzhaki (1987) et d'Olkin et Yitzhaki (1992) se sont basés sur la méthode de covariance Gini ou du Co-Gini.

12. Cette valeur est la même pour la validation croisée dans PLS1.
13. "More than a dozen alternative ways spelling Gini" [Yitzhaki (1998), Yitzhaki (2003)]

La covariance Gini (Co-Gini) a été introduite par Yitzhaki et Schechtman (2004) :

$$\mathrm{cog}(x, y) := \mathrm{cov}(x, R(y)), \tag{1.4}$$

où $R(y)$ est le vecteur rang de y.[14] Habituellement, deux principales approches ont été employées pour analyser la relation entre deux variables aléatoires x et y : soit la covariance usuelle $\mathrm{cov}(x, y)$ et le coefficient de corrélation de Pearson, qui dépend de l'analyse de la variance ou la covariance entre les vecteurs rangs de x et y (ou leurs fonctions de distributions cumulées), autrement dit, $\mathrm{cov}(R(x), R(y))$. Le Co-Gini est un mélange des deux approches. Il permet d'employer une nouvelle statistique de corrélation, très proche du coefficient de corrélation de Pearson, l'indice de corrélation *Gini* $\Gamma_{xy} := \mathrm{cov}(x, R(y))/\mathrm{cov}(x, R(x))$.[15] Il prend des valeurs dans l'intervalle $[-1, 1]$, il est insensible aux transformations monotones des x et aux transformation linéaires de y, et il est nul si et seulement si les variables x et y sont indépendantes, [voir Yitzhaki, (2003)]. Comme démontré par Yitzhaki et Schechtman (2013), bien que le coefficient de corrélation de Pearson est utile, quelques difficultés peuvent apparaître. Par exemple, les coefficients de corrélation de Pearson comparés aux valeurs ± 1 x et y (0 s'il n'y a aucun lien entre eux). Les deux interprétations peuvent induire en erreur car les distributions multivariées peuvent présenter naturellement une gamme de $[\pm\frac{1}{3}]$ des coefficients de Pearson possibles. D'autre part, les variables aléatoires x et y peuvent être reliées par une transformation monotone même si leur coefficient de corrélation de Pearson est proche de 0. Comme le Co-Gini est un compromis entre la covariance et l'approche rang, il est, tel que, un meilleur candidat pour supprimer les valeurs aberrantes de l'échantillon. En revanche, les régressions MCO peuvent donner des coefficients instables en présence de valeurs aberrantes (la variance des coefficients estimés ainsi que la variance des résidus peut tendre

14. Le vecteur rang est obtenu en remplaçant les valeurs de y par leurs rangs (la plus petite valeur de y est de rang égal à 1 et le rang de la plus grande valeur de y est égal à n).

15. Il existe deux coefficients de corrélation *Gini* (qui ne sont pas symétriques) $\Gamma_{xy} := \mathrm{cov}(x, R(y))/\mathrm{cov}(x, R(x))$ et $\Gamma_{yx} := \mathrm{cov}(y, R(x))/\mathrm{cov}(y, R(y))$ de la même manière que les Co-Gini $\mathrm{cov}(x, R(y))$ et $\mathrm{cov}(y, R(x))$. Les deux corrélations *Gini* sont égales si les distributions sont invariantes suite à une transformation linéaire. Notons que dans la suite, un seul Co-Gini est utilisé : $\mathrm{cog}(x, y) = \mathrm{cov}(x, R(y))$.

vers l'infini). Par conséquent, le test t de Student pour évaluer la force de la corrélation entre les variables, peut indiquer si les paramètres sont significatifs.

Il est à signaler que le Co-Gini (que nous notons "cog") est calculé entre chaque paire de variables aléatoires. Le Co-Gini est basé sur le calcul des covariances. Il est défini comme une somme pondérée des pentes de la courbe de régression : Le Co-Gini est la covariance entre une variable et la distribution cumulative d'une autre variable. Le Co-Gini est un concept très proche des mesures de corrélations les plus répandues, comme les coefficients de corrélation de Pearson et de Spearman. Le coefficient de corrélation de Spearman est défini comme la covariance entre les fonctions de répartitions de x et de y : $cov(F(x), F(y))$: il s'agit d'une corrélation entre les rangs. Vu sa robustesse, le concept rang est utilisé dans l'ACP. Le Co-Gini se place entre l'ACP et le coefficient de corrélation de Spearman :
$cog(x, y) = cov(x, F(y))$ et $cog(y, x) = cov(y, F(x))$. où F(x) et F(y) sont respectivement les fonctions de répartitions de x et de y [16]

Il existe deux méthodes de régression qui peuvent s'interpréter comme la différence moyenne de Gini : une régression semi-paramétrique basée sur le Co-Gini et une régression basée sur la minimisation des résidus de l'indice du Gini. L'indice Co-Gini et la régression Gini semi-paramétrique sont liés à travers l'estimateur empirique de F proportionnel au rang R.

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1(x \leq x_i)$$

avec :

$$\hat{F}(x_i) = \frac{R(x_i)}{n}$$

Dans ce qui suit, nous allons nous intéresser à ces deux modèles de régression Gini qui permettent de pallier aux problèmes des valeurs aberrantes et des erreurs de mesure.

The sum of the squared residuals is sensitive to extreme values ...

16. Il est à noter que ces deux Co-Gini sont différents et peuvent avoir des signes différents.

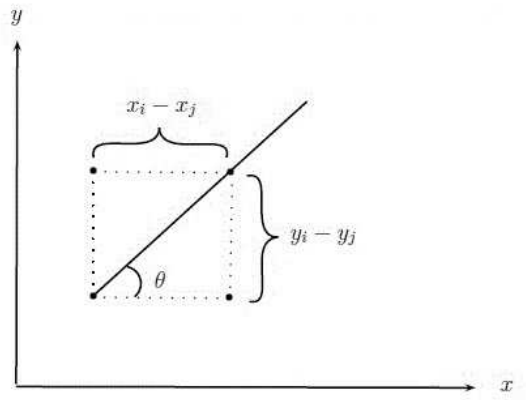
Olkin and Yitzhaki (1992, p. 185)

1.4.1 Régression Gini semi-paramétrique

La régression Gini semi-paramétrique consiste à trouver la pente de la courbe de régression en ayant recours à une analyse géométrique. Dans le modèle avec une seule variable explicative, $y_i = \alpha + \beta_G x_i + \varepsilon_i$, Olkin et Yitzhaki (1992) montrent que la pente est une moyenne pondérée des tangentes ($\tan\theta$) issues de toutes les paires d'observations possibles (i, j). Soit les valeurs des x classées dans un ordre croissant : $x_1 \leqslant \cdots \leqslant x_n$. Ainsi :

$$\tan\theta \equiv \tau_{ij} := \frac{y_i - y_j}{x_i - x_j},\ \forall j < i\ ,\ j = 1, \ldots, n\ .$$

Figure 1 : Pente de la régression Gini semi-paramétrique



La droite de régression Gini est donnée par :

$$\hat{\beta}_G = \sum_{j<i} v_{ij}\tau_{ij},\ \text{avec}\ v_{ij} = \frac{x_i - x_j}{\sum_{j<i}(x_i - x_j)},\ \text{et}\ \sum_{j<i} v_{ij} = 1. \qquad (1.5)$$

Olkin et Yitzhaki (1992) montrent que l'équation (1.5) permet de capter les pentes de plusieurs méthodes. Par exemple, le poids suivant,

$$w_{ij} = \frac{(x_i - x_j)^2}{\sum_{j<i}(x_i - x_j)^2},$$

entraine $\hat{\beta}_{MCO} = \sum_{j<i} w_{ij}\tau_{ij}$.
Dans la régression MCO, les poids w_{ij} sont quadratiques. En conséquence, les valeurs aberrantes amplifient les valeurs de la pente. Au contraire, les $\hat{\beta}_G$ sont basés sur des poids v_{ij} qui sont moins sensibles aux valeurs aberrantes.

La régression Gini semi-paramétrique peut s'interpréter comme une somme pondérée des pentes de la courbe de régression. Cette régression est similaire à la régression MCO. Il s'agit de remplacer la covariance et la variance de la MCO respectivement par le Co-Gini et le Gini. Le coefficient de régression Gini semi-paramétrique est le ratio des deux covariances :

$$\beta^G_{y,x} := \text{cov}(y, F(x))/\text{cov}(x, F(x)) \ .$$

La covariance $\text{cov}(x, F(x))$ est toujours positive, donc le signe de $\beta^G_{y,x}$ est déterminé par $\text{cov}(y, F(x))$. Une transformation monotonique de la variable explicative x n'affecte pas sa fonction de répartition $(F(x))$, mais peut faire varier l'amplitude du coefficient estimé $(\beta^G_{y,x})$.

Les paramètres estimés à l'aide de la régression Gini semi-paramétrique sont différents des coefficients estimés à l'aide de la régression MCO. iétés d'optimalité. L'inconvénient des paramètres estimés à l'aide de la régression Gini semi-paramétrique réside dans le fait que les coefficients de régression ne peuvent pas être déduites explicitement, mais peuvent être calculés numériquement. Dans le cadre de la régression Gini multiple, on peut combiner la régression semi-paramétrique à la méthode de régression par MCO où quelques variables explicatives peuvent être traitées en utilisant la régression MCO et d'autres variables traitées en utilisant la régression Gini semi-paramétrique. Cette flexibilité permet d'évaluer le choix de la méthode de régression, Yitzhaki et Schechtman (2013).

La régression Gini semi-paramétrique multiple dépend de la matrice rang de x $(R(x))$ et de sa transposée $R^\intercal(X)$. La matrice rang contient dans ses colonnes les vecteurs rangs $R(x_k)$ des régresseurs x_k, $k = 1, \ldots, K$. L'estimateur du Gini

semi-paramétrique est un vecteur de taille $K \times 1$ donné par :

$$\hat{\boldsymbol{\beta}}_G = (R^\intercal(X)X)^{-1}R^\intercal(X)y \ .$$

L'approche Gini semi-paramétrique est intéressante vu qu'elle requiert moins d'hypothèses que la MCO, par exemple l'hypothèse de linéarité est relâchée. Elle ne nécessite pas de spécification de la forme fonctionnelle pour le modèle de régression. Cependant, en présence de multi-colinéarité, de la même manière que la MCO, la régression Gini semi-paramétrique ne peut pas s'appliquer. En effet, la matrice X doit être de plein rang colonne , autrement $R^\intercal(X)X$ n'est pas inversible.

Durbin (1954) a montré qu'il y a convergence des $R(x)$ [17] avec les variables instrumentales. Les travaux de Yitzhaki et Schechtman (2004) sont aussi basés sur l'utilisation du rang de x ($R(x)$) comme instrument. [18] Si $R(x)$ est un instrument fort, il sera adapté pour corriger les erreurs de mesure. Dans les travaux de Yitzhaki, Pudalov et Schechtman (2011), le rang des variables instrumentales ($(R(z)$) est utilisé pour traiter l'endogénéité. "La technique des variables instrumentales permet d'obtenir une estimation convergente des paramètres du modèle," [Robin, (2000)]. Une variable instrumentale est très corrélée avec le régresseur, elle n'est corrélée ni avec le terme d'erreur, ni avec la variable à expliquer. Le choix des variables instrumentales est crucial pour obtenir des estimateurs robustes. Il est possible de tester la validité des instrument (c'est-à-dire de voir si l'instrument est fort ou faible), ainsi que la sur-identification des instruments [Voir aussi Greene (2005)]. "L'objectif est de construire des prédicteurs linéaires de y basés sur x. La prédiction théorique est notée comme suit :"[Voir aussi Yitzhaki et Schechtman (2013)]

$$\hat{y} = \alpha + \beta x \ .$$

Le résidu du modèle est défini ainsi :

17. $R(x)$: est la matrice rang de x où dans chaque colonne de x la plus petite observation prend la valeur 1 et la plus grande prends la valeur n (avec n est la taille de l'échantillon).
18. $R(x)$ est aussi employé pour résoudre les problèmes de valeurs aberrantes.

$$\varepsilon = y - \hat{y} = y - \alpha + \beta x \ .$$

En appliquant la covariance, on obtient :

$$\text{cov}(y, x) = \text{cov}(\alpha + \beta x + \varepsilon, x) \tag{1.6}$$

$$\text{cov}(y, x) = \text{cov}(\alpha, x) + \text{cov}(\beta x, x) + \text{cov}(\varepsilon, x) \tag{1.7}$$

$$\text{cov}(y, x) = \beta \text{cov}(x, x) + \text{cov}(\varepsilon, x) \tag{1.8}$$

En imposant l'hypothèse de normalité des résidus ($\varepsilon \perp x$), la covariance entre les résidus du modèle et les régresseurs est nulle ($\text{cov}(\varepsilon, x) = 0$). On obtient :

$$\begin{aligned} \text{cov}(y, x) &= \beta \text{cov}(x, x) \\ \beta &= \frac{\text{cov}(y, x)}{\text{cov}(x, x)} \end{aligned}$$

La structure du β_G est équivalente à celle des β_{MCO} ; La contrainte de normalité des résidus provient de la minimisation de la variance des erreurs, Yitzhaki et Schechtman (2013).

La régression Gini semi-paramétrique peut se présenter ainsi :

$$\beta_G = \frac{\text{cov}(y, F(x))}{\text{cov}(x, F(x))} \tag{1.9}$$

En utilisant la propriété de covariance, on obtient :

$$\text{cov}(\varepsilon_G, F(x)) = 0 \tag{1.10}$$

où ε_G est le résidu de la régression Gini semi-paramétrique. Ainsi, la droite des prédicteurs linéaires ($\hat{y} = \alpha + \beta x$) peut être déduite. α est déterminé en imposant l'hypothèse supplémentaire suivante : la droite de régression passe par le point moyen de l'échantillon. Il est possible d'utiliser le critère de minimisation de la somme des déviations absolues des résidus par rapport à la constante α. Il en résulte que la droite de régression passe par la médiane ou n'importe quel quantile de la distribution des résidus. Le point important ici est de distinguer le critère utilisé pour déterminer la pente de celui utilisé pour déterminer le terme constant (α). "Le coefficient de régression Gini est une somme pondérée de la courbe de régression. " [19][Yitzhaki et Schechtman (2013)]

Les deux régressions MCO et Gini semi-paramétrique peuvent s'exprimer comme une moyenne pondérée des mêmes pentes, car les pentes entre observations adjacentes sont déterminées par les données. Les interprétations de la régression Gini sont donc plus faciles que les interprétations des régressions MCO.

Le choix de la méthode de régression est actuellement un choix du schéma de pondération. Les deux schémas de pondération sont donnés par les propriétés des distributions des variables explicatives. Ces schémas de pondération dépendent de deux facteurs : le premier est le rang de l'observation. Le poids maximal est lié à l'observation médiane de la variable explicative, le poids décroit systématiquement lorsque l'observation décroît de la médiane *Cf Yitzhaki et Schechtman (2013)*. Cette propriété est la même dans les deux régressions MCO et Gini. Le second facteur qui affecte les poids est la distance entre les observations adjascentes (δx). La différence entre les méthodes MCO et Gini, est le poids attaché à la distance δx. Dans la régression Gini, le poids est basé sur δx, alors que dans la régression MCO, le poids est basé sur δx^2. Ceci explique le fait que les coefficients de régression MCO sont plus sensible aux

19. Il est à noter que la régression Gini semi-paramétrique est similaire à la régression MCO lorsque les variables explicatives suivent une loi Uniforme.

valeurs aberrantes que les coefficients de régression Gini.

La régression Gini semi-paramétrique, est une moyenne pondérée des pentes. Cette régression Gini n'est pas une approche d'optimisation [20] et ne nécessite pas d'hypothèse de linéarité de la courbe de régression. Il existe une autre méthode de régression Gini, très proche de la régression MCO : l'approche paramétrique (nommée aussi approche Gini par minimisation).

1.4.2 Approche par minimisation (ou régression Gini paramétrique)

La régression Gini paramétrique est basée sur la minimisation des résidus de la différence moyenne du Gini. Comme le montrent Olkin et Yitzhaki (1992), la condition de normalité est retrouvée à partir des β_G ($\text{cov}(x, F(\varepsilon)) = 0$).

Les estimateurs obtenus par l'approche de minimisation Gini n'imposent pas de forme fonctionnelle bien déterminée. Il est à noter que les estimateurs Gini sont consistants, et que tous les concepts de la régression MCO peuvent se traduire dans la régression Gini. Ainsi les concepts de variance et de covariance de la régression MCO peuvent se traduire en différence moyenne de Gini et Co-Gini dans la régression Gini. Les estimateurs des régressions Gini sont consistants surtout en présence de valeurs aberrantes. Dans le cas des erreurs de mesure, la variable rang constitue un instrument robuste, [Yitzhaki et Schechtman, (2013)].

La régression Gini est employée pour éviter les biais relatifs à la présence de valeurs aberrantes. Dans le cas de la régression paramétrique, Olkin et Yitzhaki (1992) montrent que le même $\hat{\beta}_G$ minimise l'indice de Gini des résidus $G(e)$ si le modèle est linéaire. Soit le vecteur des résidus estimé par la reletion linéaire : $e := y - \hat{\beta}_G x$. Ainsi,

$$\hat{\beta}_G = \arg\min_{\beta_G} G(e) = \arg\min_{\beta_G} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |e_i - e_j| \ . \tag{1.11}$$

20. C'est aussi le point fort des méthodes de calcul d'expectiles [Gneiting, (2012)].

Ainsi, une expression équivalent $\hat{\beta}_G$ est dérivée : [21]

$$\hat{\beta}_G = \frac{\text{cov}(y, R(x))}{\text{cov}(x, R(x))} \ .$$

21. Si l'hypothèse de linéarité est relâchée, les approches Gini paramétrique et semi-paramétrique ne sont pas nécessairement équivalentes.

Conclusion

L'importance des modèles économétriques réside dans leurs applications. Dans la plupart des domaines (finance, économie, écologie, biologie, médecine, agronomie, etc.), la robustesse des estimateurs aboutit à des résultats probants en terme d'interprétations. Pour cette raison, deux éléments s'avèrent très importants : le diagnostic préalable des données et le choix du modèle en question.

Les difficultés qui découlent de la multi-colinéarité, des valeurs aberrantes et de l'endogénéité ont conduit à une vaste littérature basée sur les régressions PLS ou Gini. Chaque régression a été conçue pour obtenir des estimations précises des valeurs prédites de la variable dépendante et, dans une moindre mesure, à interpréter les signes et les grandeurs des coefficients estimés. La régression PLS à son tour résout les problèmes liés à la multi-colinéarité, à la faible taille de l'échantillon et aux données manquantes. De même, la régression sur Gini résorbe l'endogénéité via les variables rangs. La régression Gini résout aussi le problème d'observations extrêmes.

Les problèmes de valeurs aberrantes, de multi-colinéarité et d'erreurs de mesure sont traitées séparément à l'aide des régressions PLS ou Gini. À notre connaissance, il n'existe pas de méthode de régression qui résout simultanément ces difficultés. Ce constat nous permettra de proposer dans le second chapitre de nouveaux modèles de régression Gini-PLS.

Chapitre 2

Construction des régressions Gini-PLS

Sommaire

2.1 Les régressions univariées Gini-PLS1 44
2.2 Les régressions multivariées Gini-PLS2 60
2.3 Simulations 70

Introduction

Valeurs aberrantes, erreurs de mesure et multi-colinéarité biaisent les résultats des régressions MCO. Ainsi, leurs détections sont nécessaires pour aboutir à des résultats fiables. La coexistence de ces trois difficultés pose problème. À notre connaissance, aucune approche n'a été conçue pour résoudre simultanément ces trois difficultés. Comme nous l'avons vu dans le chapitre précédent, il existe certaines régressions qui résolvent séparément les problèmes déjà présentés, en particulier, les régressions PLS et Gini. Notre intérêt derrière les combinaisons des régressions Gini et PLS, est d'obtenir des régressions capables de résoudre simultanément les fortes corrélations, les erreurs de mesure, la faible taille de l'échantillon, les valeurs aberrantes et les données manquantes.

Le modèle de régression Gini1-PLS1 a pour objet de régresser une variable dépendante sur des composantes orthogonales à partir de l'opérateur covariance Gini (co-Gini). Des poids sont issus de la maximisation du co-Gini entre les régresseurs et la variable dépendante. Les caractéristiques principales de la régression Gini1-PLS1 sont l'emploi des vecteurs rang des régresseurs [1] comme instruments comme l'a suggéré Durbin (1954), dans le but de traiter les problèmes des corrélations des valeurs aberrantes. L'approche de construction du modèle de régression Gini2-PLS1 concerne la conception des poids à partir des coefficients de pente de la régression Gini semi-paramétrique adaptés pour traiter les valeurs aberrantes. Les deux régressions Gini-PLS1 [2] ont pour point commun l'opérateur co-Gini.

Nous proposons aussi une généralisation de ces régressions dans le cas multivarié pour estimer une matrice de variables dépendantes en fonction d'une matrice de variables explicatives, dans le même but que les régressions univariées(pour la résolution des problèmes déjà cités). Les régressions Gini1-PLS2 et Gini2-PLS2 permettent de résoudre les mêmes problèmes.

Il est indéniable que notre travail s'aligne avec la littérature qui met l'accent

1. $R(x)$: la matrice rang de x où les éléments de x sont ordonnés, la plus petite valeur de x prends la valeur 1 et la plus grande valeur de x prend n (où n représente la taille de l'échantillon est de taille n)
2. La notation Gini-PLS1 englobe les deux régressions Gini1-PLS1 et Gini2-PLS1.

sur l'amélioration des régressions PLS, *par exemple*, Bastien, Esposito Vinzi et Tenenhaus (2005) sur la régression PLS généralisée dans le cas des variables ordinales, Chung et Keles (2010) sur la régression PLS pour atteindre simultanément la sélection des variables et la réduction de dimension, Russolillo (2012) pour les données définies à différentes échelles et en cas de non-linéarité entre les variables, Bry *et alii.* (2013) pour la régression PLS avec des modèles linéaires généralisés basés sur l'algorithme du score de Fisher pour fournir des estimations fiables dans le cas de multiples variables dépendantes. Nous abordons dans un premier temps les régressions univariées (Gini-PLS1) et dans un second temps les modèles multivariés Gini-PLS (Gini-PLS2).

2.1 Les régressions univariées Gini-PLS1

Les régressions Gini-PLS résolvent simultanément les problèmes suivants : outliers, faible taille de l'échantillon ($n \leqslant p$), corrélations excessives entre les régresseurs, et des données manquantes. Nous proposons deux régressions : Gini1-PLS1 et Gini2-PLS1.

2.1.1 La régression Gini1-PLS1

Les étapes de régression Gini1-PLS1

La conception de la régression Gini1-PLS1 repose sur les poids du Co-Gini maximisant le lien entre y et les régresseurs x_k dans le but de limiter l'influence des valeurs aberrantes. Dans ce qui suit, les régresseurs x_k et la variable dépendante y sont supposés être centrés. La matrice rang des régresseurs standardisés est notée par $R(X)$ (avec $R(x_k)$ en colonnes).

• **Étape 1 :** Comme défini dans le chapitre précédent, les coefficients estimés de la régression Gini sont moins sensibles aux valeurs aberrantes, grâce à l'utilisation de l'opérateur Co-Gini (cog). En conséquence, le nouveau vecteur de poids w_1 qui maximise le lien du Co-Gini entre la variable à prédire y et les régresseurs x_k est dérivée comme suit.

Proposition 1. *La solution du programme suivant :*

$$\max \operatorname{cov}(y, R(X)w_1) \ , \ \ s.t. \ \ \|w_1\| = 1 \tag{2.1}$$

est :

$$w_{1k} = \frac{\operatorname{cog}(y, x_k)}{\sqrt{\sum_{k=1}^{K} \operatorname{cog}^2(y, x_k)}} \ , \ \forall k = 1, \ldots, K \ . \tag{2.2}$$

Démonstration. à partir de la définition du Co-Gini (2.1) et de la solution connue du programme de maximisation (2.1) de la régression PLS1, les résultats sont :

$$w_{1k} = \frac{\operatorname{cov}(y, R(x_k))}{\sqrt{\sum_{k=1}^{K} \operatorname{cov}^2(y, R(x_k))}} \ , \ \forall k = 1, \ldots, K \ .$$

On remarque que, si la variable prédite y a été standardisée, ainsi $\text{cov}(y, R(x_k))$ peut représenter le coefficient de régression MCO des $R(x_k)$ sur y. ∎

Nous déterminons maintenant la composante t_1 (la procédure reste valide avec les données manquantes, voir Eq. (1.2) Section 1.2). La variable y est régressée par MCO sur la composante t_1, qui est construite de la manière habituelle comme une combinaison linéaire entre les variables :

$$t_1 = \sum_{k=1}^{K} w_{1k}\, x_k \implies y = c_1 t_1 + \varepsilon_1 \ .$$

• **Étape 2 :** Comme dans la régression classique PLS1, la régression MCO est utilisée pour implémenter les régressions partielles, hormis le fait que les régresseurs sont remplacés par leurs rangs. Nous verrons dans ce qui suit que ces vecteurs rang sont actuellement des instruments. Soit $\beta_1 := (\beta_{11}, \ldots, \beta_{1k}, \ldots, \beta_{1K})$ les vecteurs de taille $K \times 1$ où les éléments β_{1k} sont des coefficients résultants de la régression MCO de chaque $R(x_k)$ sur t_1 :

$$R(x_k) = \beta_{1k} t_1 + u_{(1)k} \ , \ \forall k = 1, \ldots, K \ . \tag{2.3}$$

Le second vecteur de poids w_2 est donc déduit à partir du Co-Gini entre l'aléa estimé de la régression totale de y sur t_1 ($\hat{\varepsilon}_1$) et les résidus des régressions partielles ($\hat{u}_{(1)j}$). Soit la matrice $\hat{U}_{(1)}$ dont les vecteurs sont $\hat{u}_{(1)k}$:

$$\max \text{cov}(\hat{\varepsilon}_1, R(\hat{U}_{(1)})w_2) \ , \text{ s.c. } \|w_2\| = 1 \implies w_{2k} = \frac{\text{cog}(\hat{\varepsilon}_1, \hat{u}_{(1)k})}{\sqrt{\sum_{k=1}^{K} \text{cog}^2(\hat{\varepsilon}_1, \hat{u}_{(1)k})}} \ , \ \forall k = 1, \ldots, K \ .$$

La seconde composante t_2 permet d'estimer le modèle complet :

$$t_2 = \sum_{k=1}^{K} w_{2k} \hat{u}_{(1)k} \implies y = c_1 t_1 + c_2 t_2 + \varepsilon_2 \ .$$

• **Étape h :** Les régressions partielles sont déduites en ajoutant l'influence

de la composante t_{h-1} [3] :

$$R(x_k) = \beta_{1k}t_1 + \cdots + \beta_{h-1k}t_{h-1} + u_{(h-1)k} \ , \ \forall k = 1, \ldots, K \ .$$

Ainsi, après maximisation, nous obtenons :

$$w_{(h)k} = \frac{\operatorname{cog}(\hat{\varepsilon}_{h-1}, \hat{u}_{(h-1)k})}{\sqrt{\sum_{k=1}^{K} \operatorname{cog}^2(\hat{\varepsilon}_{h-1}, \hat{u}_{(h-1)k})}} \ , \ \forall k = 1, \ldots, K \ ,$$

$$t_h = \sum_{k=1}^{K} w_{hk} \ \hat{u}_{(h-1)k} \implies y = c_1t_1 + \cdots + c_ht_h + \varepsilon_h \ .$$

L'algorithme s'arrête lorsque la validation croisée rejette la composante t_{h+1}.

2.1.2 La régression Gini2-PLS1

Les étapes de la régression Gini2-PLS1

Dans la régression Gini1-PLS1, l'orthogonalité est présente entre les composantes t_h issues de la maximisation du Co-Gini et de l'utilisation des vecteurs rangs dans les régressions partielles. Une autre possibilité est de bénéficier des coefficients de la régression Gini semi-paramétrique. Ceci représente une autre méthode pour minimiser l'influence des outliers en employant l'opérateur Co-Gini operator. En particulier, nous commençons avec les régressions Gini suivantes :

$$y = \delta_{G1k}x_k + e_k, \ \forall k = 1, \ldots, K \ .$$

Où e_k est l'aléa de la régression Gini.

- **Étape 1 :** L'élément w_{1k} du vecteur poids w_1 représente le lien entre x_k

3. Nous proposons aussi une version modifiée de la régression Gini1-PLS1 où les régressions partielles s'expriment en fonction de x et non $R(x)$

et y (qui est $\hat{\delta}_{G1k}$) comme fraction de tous les liens possibles : [4]

$$\hat{\delta}_{G1k} = \frac{\text{cog}(y, x_k)}{\text{cog}(x_k, x_k)} \implies w_{1k} = \frac{\hat{\delta}_{G1k}}{\sqrt{\sum_{k=1}^{K}(\hat{\delta}_{G1k})^2}} \ , \ \forall k = 1, \ldots, K \ .$$

La première composante t_1 est :

$$t_1 = \sum_{k=1}^{K} w_{1k} x_k \ .$$

Comme dans la régression Gini1-PLS1, l'opérateur Co-Gini permet de purifier le modèle des outliers, étant donné que chaque $\hat{\delta}_{G1k}$ est déterminé à partir d'une moyenne pondérée des tangentes $\hat{\delta}_{G1k}$, pour lesquels les poids $v_{ij} = \frac{x_{ji} - x_{jr}}{\sum_{r<i}(x_{ji} - x_{jr})}$ ne sont pas quadratiques (Eq. 1.5). Le modèle entier estimé par MCO est :

$$y = c_1 t_1 + \varepsilon_1 \ .$$

La suite de l'algorithme est équivalente à PLS1 sauf que les poids sont déduits à partir des coefficients de régression Gini semi-paramétrique.

- **Étape $h \geqslant 2$:**

Les régression partielles sont implémentées par MCO sans pour autant utiliser ls vecteurs rang comme dans Gini1-PLS1 :

$$x_k = \beta_{1k} t_1 + \cdots + \beta_{h-1k} t_{h-1} + u_{(h-1)k} \ , \ \forall k = 1, \ldots, K \ . \tag{2.4}$$

Les poids sont construits à partir des régressions Gini semi-paramétrique des $\hat{\varepsilon}_{h-1}$ sur chaque $\hat{u}_{(h-1)k}$:

$$\hat{\varepsilon}_{h-1} = \delta_{Ghk} \hat{u}_{(h-1)k} + \nu_{(h-1)k} \implies \hat{\delta}_{Ghk} = \frac{\text{cog}(\hat{\varepsilon}_{h-1}, \hat{u}_{(h-1)k})}{\text{cog}(\hat{u}_{(h-1)k}, \hat{u}_{(h-1)k})} \ .$$

4. Notons que les vecteurs poids w_k peuvent être aussi déterminés à partir de la minimisation de l'indice de Gini des résidus, *c'est-à-dire.* par la régression Gini paramétrique même si le lien entre y et x_k n'est pas linéaire.

Les poids sont donnés par :

$$w_{hk} = \frac{\hat{\delta}_{Ghk}}{\sqrt{\sum_{k=1}^{K}(\hat{\delta}_{Ghk})^2}} \ , \ \forall k = 1, \ldots, K. \tag{2.5}$$

Le modèle entier est estimé par MCO :

$$t_h = \sum_{k=1}^{K} w_{hk} \ \hat{u}_{(h-1)k} \implies y = c_1 t_1 + c_2 t_2 + \cdots + c_h t_h + \varepsilon_h \ .$$

L'expression (2.4) est utilisée pour maintenir l'orthogonalité $t_1 \perp \cdots \perp t_h$, comme dans la régression PLS1 standard.

Propriétés

Les régressions PLS sont fondées sur des propriétés mathématiques tels que l'orthogonalité et entre autres, les conditions de normalisation. Ces propriétés montrent que les régressions Gini1-PLS1 et Gini2-PLS1 sont très proches de PLS1, tel que récapitulé dans la Tableau 1 ci-dessous.

Proposition 2. *Les propriétés des régressions PLS1, Gini1-PLS1, et Gini2-PLS1 sont les suivantes.*

Tableau 1 : Propriétés

Propriétés	PLS1	Gini1-PLS1	Gini2-PLS1
(o) $t_1 \perp t_2 \perp \cdots \perp t_h$	✓	✓	✓
(i) $w_\ell^\intercal \hat{\beta}_\ell = 1, \ \forall \ell \in \{1, \ldots, h\}$	✓	✓$(\ell > 1)$	✓
(ii) $w_h^\intercal \hat{U}_{(\ell)}^\intercal = \mathbf{0}, \ \forall \ell \geqslant h \geqslant 1$ *	✓	✓$(h > 1)$	✓
(iii) $w_h^\intercal \hat{\beta}_\ell = 0, \ \forall \ell > h \geqslant 1$	✓	✓$(h > 1)$	✓
(iv) $w_h^\intercal w_\ell = 0, \ \forall \ell > h \geqslant 1$	✓	✓$(h > 1)$	×
(v) $t_h^\intercal \hat{U}_{(\ell)} = \mathbf{0}, \ \forall \ell \geqslant h \geqslant 1$	✓	✓	✓
(vi) $\hat{U}_{(h)} = \hat{U}_{(0)} \prod_{\ell=1}^{h} (\mathbb{I} - w_\ell \beta_\ell^\intercal), \ \forall h \geqslant 1$	✓	×	✓
(vii) Valeurs manquantes	✓	✓	✓
(viii) Valeurs aberrantes	×	✓	✓
(ix) Petits échantillons $(n < k)$	✓	✓	✓

* $\mathbf{0}$ est le vecteur de zéros de taille $1 \times K$.

Démonstration. Les propriétés mathématiques (o)–(vi) sont développées dans ce qui suit. La robustesse des valeurs manquantes (vii) est donnée par Eq. (1.2) et est valide pour toutes les régressions. L'estimation par MCO de $y = c_1 t_1 + \cdots + c_h t_h$ est possible pour tout $n \geqslant h$. Par conséquent, la bonne performance des petits échantillons (ix) est atteinte lorsque le nombre de composantes h est le plus faible que possible. Finalement, la robustesse des valeurs aberrantes (viii) sont étudiés dans la partie des simulations ci-après (Section 2.3). ∎

Preuve de la Proposition 2

1. Propriétés (o)–(vi) de la régression PLS1 : Voir Tenenhaus (1998).

2. Propriétés (o)–(vi) de la régression Gini1-PLS1.

(o) $t_1 \perp \cdots \perp t_h$:

La preuve est faite par induction mathématique. Nous suivons Tenenhaus (1998, p. 101) pour PLS1, sauf que dans notre cas, $\hat{U}_{(0)} := R(X)$ (les résidus sont issus des vecteurs rang Eq.(2.3)). D'une part, nous montrons que $t_1 \perp t_2$:

$$t_1 \perp t_2 \iff t_1^\intercal t_2 = t_1^\intercal \underbrace{\hat{U}_{(1)} w_2}_{t_2} = 0,$$

lorsque $t_1^\intercal \hat{U}_{(1)} = \mathbf{0}$, où $\mathbf{0}$ est le vecteur (ligne) nul de taille K. Supposons que la supposition suivante soit vraie :

$$\textbf{[h]} : t_1 \perp t_2 \perp \cdots \perp t_h \ .$$

Nous avons montré que **[h+1]** est vraie, *c'est-à-dire*, t_{h+1} est orthogonale à toutes les composantes $t_1, \ldots, t_h$. La relation **[h]** implique $t_h^\intercal \hat{U}_{(h)} = \mathbf{0}$, par conséquent :

$$t_h^\intercal t_{h+1} = t_h^\intercal \hat{U}_{(h)} w_{(h+1)} = 0 \ .$$

Selon les étapes 2–h, les régressions partielles impliquent, pour tout $k = 1, \ldots, K$,

$$R(x_k) = \hat{\beta}_{1k} t_1 + \hat{u}_{(1)k} = \hat{\beta}_{1k} t_1 + \hat{\beta}_{2k} t_2 + \hat{u}_{(2)k} = \cdots = \sum_{r=1}^{h} \hat{\beta}_{rk} t_r + \hat{u}_{(h-1)k} \ . \quad (2.6)$$

La relation (2.6) implique $\hat{U}_{(h)} = \hat{U}_{(h-1)} - t_h\hat{\beta}_{(h)}^\intercal$, où $t_h\hat{\beta}_{(h)}^\intercal$ est la matrice de taille $n \times K$ contenant $\hat{\beta}_{hk}t_h$ en colonnes, pour tout $k = 1, \ldots, K$. Puisque **[h]** implique $t_{h-1}^\intercal \hat{U}_{(h-1)} = \mathbf{0}$ et $t_{h-1}^\intercal t_h = 0$, nous obtenons

$$\begin{aligned} t_{h-1}^\intercal t_{h+1} &= t_{h-1}^\intercal \hat{U}_{(h)} w_{(h+1)} \\ &= t_{h-1}^\intercal \left(\hat{U}_{(h-1)} - t_h\hat{\beta}_{(h)}^\intercal\right) w_{(h+1)} \\ &= \left(t_{h-1}^\intercal \hat{U}_{(h-1)} - t_{h-1}^\intercal t_h\hat{\beta}_{(h)}^\intercal\right) w_{(h+1)} = 0 \ . \end{aligned}$$

En utilisant **[h]**, nous obtenons

$$\begin{aligned} t_{h-2}^\intercal t_{h+1} &= t_{h-2}^\intercal \left(\hat{U}_{(h-1)} - t_h\hat{\beta}_{(h)}^\intercal\right) w_{(h+1)} \\ &= t_{h-2}^\intercal \left(\hat{U}_{(h-2)} - t_{h-1}\hat{\beta}_{(h-1)}^\intercal - t_h\hat{\beta}_{(h)}^\intercal\right) w_{(h+1)} \\ &= \left(t_{h-2}^\intercal \hat{U}_{(h-2)} - t_{h-2}^\intercal t_{h-1}\hat{\beta}_{(h-1)}^\intercal - t_{h-2}^\intercal t_h\hat{\beta}_{(h)}^\intercal\right) w_{(h+1)} = 0 \ . \end{aligned}$$

Enfin, **[h]** donne

$$\begin{aligned} t_1^\intercal t_{h+1} &= t_1^\intercal \left(\hat{U}_{(h-1)} - t_h\hat{\beta}_{(h)}^\intercal\right) w_{(h+1)} \\ &= \left(t_1^\intercal \hat{U}_1 - t_1^\intercal \sum_{r=2}^{h} t_r\hat{\beta}_{(r)}^\intercal\right) w_{(h+1)} = 0 \ . \end{aligned}$$

(i) $w_\ell^\intercal \hat{\beta}_\ell = 1, \ \forall \ell \in \{2, \ldots, h\}$:

Soit $\hat{\beta}_h$ le vecteur colonne dont les éléments sont $\hat{\beta}_{hk}$ pour tout $k = 1, \ldots, K$. Les composantes t_h sont données par $w_h^\intercal \hat{U}_{(h-1)}^\intercal = t_h^\intercal$, et ainsi

$$w_{(h)}^\intercal \hat{\beta}_h = w_h^\intercal \frac{\hat{U}_{(h-1)}^\intercal t_h}{t_h^\intercal t_h}. \tag{2.7}$$

Pour $h = 1$, nous avons $w_1^\intercal X^\intercal = t_1^\intercal$, et ainsi $w_1^\intercal \hat{\beta}_1 = w_1^\intercal \frac{R(X)^\intercal t_1}{t_1^\intercal t_1} \neq 1$ si $R(X) \neq X$. Pour $h > 1$, nous avons $w_h^\intercal \hat{U}_{(h)}^\intercal = t_h^\intercal$. À partir de l'expression (2.6), nous avons

$$\hat{\beta}_h = \frac{\hat{U}_{(h)}^\intercal t_h}{t_h^\intercal t_h}, \tag{2.8}$$

et ainsi $w_h^\intercal \hat{\beta}_h = w_h^\intercal \frac{\hat{U}_{(h)}^\intercal t_h}{t_h^\intercal t_h} = \frac{t_h^\intercal t_h}{t_h^\intercal t_h} = 1$.

(ii) $w_h^\intercal \hat{U}_{(\ell)}^\intercal = \mathbf{0},\ \forall \ell \geqslant h > 1$:

La relation (2.6), $R(x_k) = \hat{\beta}_{1k} t_1 + \hat{\beta}_{2k} t_2 + \cdots + \hat{u}_{(h-1)k}$, donne

$$\hat{U}_{(h-1)} - t_h \hat{\beta}_h^\intercal = \hat{U}_{(h)}. \tag{2.9}$$

Pour $h = \ell = 1$, nous avons $R(X) \equiv \hat{U}_{(0)} = t_1 \hat{\beta}_1^\intercal + \hat{U}_{(1)}$, et ainsi

$$w_1^\intercal \hat{U}_{(1)}^\intercal = w_1^\intercal \left(R(X)^\intercal - \hat{\beta}_1 t_1^\intercal \right) = w_1^\intercal \left(\hat{\beta}_1 t_1^\intercal + \hat{U}_{(1)}^\intercal - \hat{\beta}_1 t_1^\intercal \right) = w_1^\intercal \hat{U}_{(1)}^\intercal \neq \mathbf{0}.$$

Pour $h = \ell > 1$, utilisant (i), nous déduisons à partir de (2.9) que

$$\begin{aligned} w_h^\intercal \hat{U}_h^\intercal &= w_h^\intercal \left(\hat{U}_{(h-1)}^\intercal - \hat{\beta}_h t_h^\intercal \right) \\ &= w_h^\intercal \hat{U}_{(h-1)}^\intercal - w_h^\intercal \hat{\beta}_h t_h^\intercal \\ &= t_h^\intercal - t_h^\intercal = \mathbf{0}. \end{aligned}$$

Pour tout $\ell > h > 1$, les expressions (i) et (2.8) donnent

$$\begin{aligned} w_h^\intercal \hat{U}_{(\ell+1)}^\intercal &= w_h^\intercal \left(\hat{U}_{(\ell)} - t_{\ell+1} \hat{\beta}_{\ell+1}^\intercal \right)^\intercal \\ &= w_h^\intercal \hat{U}_{(\ell)}^\intercal - w_h^\intercal \frac{\hat{U}_{(\ell)}^\intercal t_{\ell+1}}{t_{\ell+1}^\intercal t_{\ell+1}} t_{\ell+1}^\intercal \\ &= w_h^\intercal \hat{U}_{(\ell)}^\intercal - w_h^\intercal \hat{U}_{(\ell)}^\intercal (t_{\ell+1} t_{\ell+1}^\intercal)^{-1} t_{\ell+1} t_{\ell+1}^\intercal \\ &= \mathbf{0}. \end{aligned}$$

(iii) $w_h^\intercal \hat{\beta}_\ell = 0,\ \forall \ell > h > 1$:

Grâce aux relations (i) et (2.8), nous obtenons

$$w_h^\intercal \hat{\beta}_\ell = w_h^\intercal \frac{\hat{U}_{(\ell)}^\intercal t_\ell}{t_\ell^\intercal t_\ell}.$$

Pour $\ell > h > 1$, la relation (ii) donne ces résultats.

(iv) $w_h^\intercal w_\ell = 0,\ \forall \ell > h > 1$:

La relation (ii) donne $w_h^\intercal \hat{U}_{(\ell)}^\intercal = \mathbf{0}$, pour tout $\ell > h > 1$. Ainsi,

$$w_h^\intercal w_\ell = w_h^\intercal \frac{\hat{U}_{(\ell-1)}\hat{\varepsilon}_{\ell-1}}{\left(\sum_{k=1}^K \operatorname{cog}^2\left(\hat{u}_{(\ell-1)k}^\intercal, \hat{\varepsilon}_{\ell-1}\right)\right)^{\frac{1}{2}}} = w_h^\intercal \frac{\hat{U}_{(\ell-1)}\hat{\varepsilon}_{\ell-1}}{\left\|\hat{U}_{(\ell-1)}\hat{\varepsilon}_{\ell-1}\right\|} = 0.$$

(v) $t_h^\intercal \hat{U}_\ell = \mathbf{0},\ \forall \ell \geqslant h \geqslant 1$:

Eqs. (2.9) et (2.6) impliquent

$$\begin{aligned} t_h^\intercal \hat{U}_{(\ell)} &= t_h^\intercal \left(\hat{U}_{(\ell-1)} - t_\ell \hat{\beta}_\ell^\intercal\right) \\ &= t_h^\intercal \left(R(X) - \sum_{\ell=1}^{h} t_\ell \hat{\beta}_\ell^\intercal\right) \\ &= t_h^\intercal \hat{U}_{(h)} = \mathbf{0}. \end{aligned}$$

(vi) $\hat{U}_{(h)} \neq \hat{U}_{(0)} \prod_{\ell=1}^{h} \left(\mathbb{I} - w_\ell \beta_\ell^\intercal\right),\ \forall h \geqslant 1$:

Soit $h = 1$, puisque $\hat{U}_{(0)} \equiv R(X)$, à partir de l'équation (2.9) nous avons

$$\begin{aligned} \hat{U}_{(1)} &= R(X) - t_1 \hat{\beta}_1^\intercal \\ &= R(X) - X w_1 \beta_1^\intercal \\ &\neq R(X)\left(\mathbb{I} - w_1 \beta_1^\intercal\right), \text{ si } X \neq R(X) \\ &\neq R(X) \prod_{\ell=1}^{h+1} \left(\mathbb{I} - w_\ell \beta_\ell^\intercal\right). \end{aligned}$$

3. Propriétés (o)–(vi) de la régression Gini2-PLS1.

Toutes les propriétés (o)–(v) sont obtenues de la même manière que ceux de la

régression Gini1-PLS1. Quant à la propriété (vi), soit $h = 1$, lorsque $\hat{U}_{(0)} \equiv X$:

$$\begin{aligned}
\hat{U}_{(1)} &= R(X) - t_1 \hat{\beta}_1^{\intercal} \\
&= X - X w_1 \beta_1^{\intercal} \\
&= X \left(\mathbb{I} - w_1 \beta_1^{\intercal} \right) \\
&= X \prod_{\ell=1}^{h+1} \left(\mathbb{I} - w_\ell \beta_\ell^{\intercal} \right) .
\end{aligned}$$

□

2.1.3 Propriétés et aides à interprétations

Les interprétations de la régression PLS sont possibles tant qu'elles sont de bonne qualité. Pour évaluer la qualité d'ajustement de la régression PLS, les statistiques suivantes sont calculées : les redondances sur les variables explicatives, les redondances sur la variable dépendante, l'importance de la variable dans la projection (VIP), et le contrôle des valeurs aberrantes, [Tenenhaus, (1998)].

• *La redondance sur les intrants* X expliquée par l'ensemble des composantes $t_1, \ldots, t_h$ est donné par :

$$Rd(X; t_1, \ldots, t_h) := \frac{1}{p} \sum_{\ell=1}^{h} \sum_{k=1}^{K} \mathrm{cor}^2(x_k, t_\ell) =: \sum_{\ell=1}^{h} Rd(X; t_\ell), \tag{2.10}$$

où $\mathrm{cor}(\cdot, \cdot)$ dénote le coefficient de corrélation de Pearson, [Tenenhaus, (1998)]. La redondance sur X d'une composante t_h, *i.e.* $Rd(X; t_h)$, donne la part de variance de X expliquée par t_h, tandis que $Rd(X; t_1, \ldots, t_h)$ donne la somme des parts de variance issues de t_1 à t_h. Le calcul des redondances est une pratique courante de la régression PLS puisque toutes les étapes sont fondées sur des régressions partielles de chaque x_k sur $t_1, \ldots, t_h$. Ces statistiques permettent de comparer les régressions PLS1 et Gini2-PLS1, puisque ces deux régressions dépendent des mêmes régressions partielles par MCO. Les régressions partielles dans le modèle Gini1-PLS1 sont basées sur le vecteur rang, de sorte que la redondance sur les vecteurs rang est calculée comme suit :

$$Rd(R(X); t_1, \ldots, t_h) := \frac{1}{K} \sum_{\ell=1}^{h} \sum_{k=1}^{K} \mathrm{cor}^2(R(x_K), t_\ell) =: \sum_{\ell=1}^{h} Rd(R(X), t_\ell). \tag{2.11}$$

• *La redondance sur la variable dépendante* y expliquée par l'ensemble des composantes $t_1, \ldots, t_h$ est donnée par :

$$Rd(y; t_1, \ldots, t_h) := \frac{1}{K} \sum_{\ell=1}^{h} \mathrm{cor}^2(y, t_\ell) =: \sum_{\ell=1}^{h} Rd(y; t_\ell). \tag{2.12}$$

Ceci donne la part de la variance de y expliquée par $t_1, \ldots, t_h$. À partir des régressions MCO de y sur toutes les composantes t_h est la même pour toutes les régressions (PLS et Gini-PLS), la redondance sur y permet de comparer la qualité des régressions Gini-PLS1.

• *L'importance des variables dans la projection (VIP)* donne les prédicteurs pertinents x_k de y qui peuvent être sélectionnés. Le pouvoir prédictif de x_k dans un modèle avec h composantes est donné par

$$VIP_{hk} := \sqrt{\frac{K\sum_{\ell=1}^{h} Rd(y;t_\ell)w_{\ell k}^2}{Rd(y;t_1,\ldots,t_h)}}\ ,\ \text{tel que}\ \sum_{k=1}^{K} VIP_{hk}^2 = K. \tag{2.13}$$

Les prédicteurs x_k les plus importants sont les régresseurs pour lesquels $VIP_{hk} > 1$. Comme la statistique VIP est fondée sur la redondance de y, elle est comparable d'un modèle à l'autre. Cependant, la statistique VIP de la régression PLS1 n'est pas fiable en présence d'outliers. Prenons *les erreurs de mesure* pour illuster ce fait. Soit $\tilde{x}_k$ la variable explicative avec erreurs de mesure tel que $\tilde{x}_k = x_k + u$, avec u est le terme d'erreur. Comme nous le verrons ci-dessus, les poids des régression Gini1-PLS1 et Gini2-PLS1 (sans erreurs de mesure) sont, respectivement,

$$w_{1k}^{Gini1} = \frac{\operatorname{cog}(y,x_k)}{\sqrt{\sum_{k=1}^{K}\operatorname{cog}^2(y,x_k)}}\ ;\ w_{1k}^{Gini2} = \frac{\frac{\operatorname{cog}(y,x_k)}{\operatorname{cog}^2(x_k,x_k)}}{\sqrt{\sum_{k=1}^{K}\left(\frac{\operatorname{cog}(y,x_k)}{\operatorname{cog}(x_k,x_k)}\right)^2}}. \tag{2.14}$$

Notons $\tilde{w}_{1k}^{Gini1}$ et $\tilde{w}_{1k}^{Gini2}$ les poids contaminés par les erreurs de mesure. Le vecteur rang $R(x_k)$ n'est pas sensible à l'augmentation des transformations monotones. Il reste aussi invariant si l'erreur de mesure u présente de faibles écart-types (ce n'est pas nécessairement le cas d'une valeur aberrante qui affecte considérablement le rang d'une observation donnée). Par conséquent, les poids restent constants $w_{1k}^{Gini1} = \tilde{w}_{1k}^{Gini1}$.[5] Comme les poids w_{1k}^{PLS1} dépendent des x_k, ils peuvent être sensibles aux erreurs de mesure dans x_k (ils sont in-

5. Le vecteur rang est homogène de degré zéro en x_k, $R(x_k) = R(\lambda x_k)$ pour $\lambda > 0$, aussi bien qu'invariant par translation, $R(x_k) = R(x_k + a_k)$ avec $a_k = (a, a, \ldots, a) \in \mathbb{R}^n$. Ainsi, la standardisation des variables x_k permet de purger les erreurs des données de la forme suivante $\tilde{x}_k = \lambda x_k + a$, comme dans PLS1 et Gini2-PLS1.

sensibles si les erreurs ne sont pas corrélées avec x_k). Soit $V\tilde{I}P$ les statistiques incorporées dans les erreurs de mesure. En prenant la dérivée des w_{1k}^{PLS1}, puis des manipulations algébriques simples donnent [6]

$$\text{si } w_{1k}^{PLS1} > 0 \text{ et si } \text{cov}(x_k, u) \lesseqgtr 0 \implies V\tilde{I}P_{1k} \lesseqgtr VIP_{1k}.$$

Si l'erreur de mesure est positivement (négativement) corrélée avec x_k, la statistique VIP inhérente à t_1 est plus grande (petite) ainsi que sa vraie valeur. Dans la régression Gini2-PLS1, l'inverse est obtenu [7]

$$\text{si } w_{1k}^{Gini2} > 0 \text{ et si } \text{cov}(x_k, u) \lesseqgtr 0 \implies V\tilde{I}P_{1k} \gtreqless VIP_{1k}.$$

Pour conclure, les erreurs de mesure dans x_k impliquent que tous les poids $\tilde{w}_{\ell k}$ des régressions PLS1 et Gini2-PLS1 sont biaisés, de sorte que les statistiques VIP peuvent produire des interprétations absurdes. Une variable peut être considérée comme importante par le fait que $V\tilde{I}P_{hk} > 1$, alors qu'en réalité $VIP_{hk} < 1$ (ou inversement). Comme nous l'avons trouvé à l'aide des simulations [voir Mussard et Souissi-Benrejab (2015)] en présence d'outliers dans les données, bien que la régression Gini1-PLS1 semble être appropriée pour neutraliser les contaminations des données, la régression Gini2-PLS1 est utile pour traiter les corrélations des valeurs aberrantes avec les x_k.

• *Vecteur rang et erreurs de mesure.* Dans le but de faire une comparaison complète des différents modèles de régression lorsque les valeurs aberrantes perturbent les données, nous examinons à nouveau le cas particulier des erreurs de mesure. Prenons deux composantes t_1, t_2 et admettons $w_h^\star := \prod_{\ell=1}^{h-1}(\mathbb{I} - w_h\hat{\beta}_\ell^\intercal)w_h$, où $\mathbb{I}$ est la matrice identité de taille $K \times K$. Ainsi, la régression PLS1 peut s'exprimer comme suit :

$$y = c_1t_1 + c_2t_2 = (c_1w_{11}^\star + c_2w_{21}^\star)x_1 + \cdots + (c_1w_{1K}^\star + c_2w_{2K}^\star)x_K \ . \qquad \text{(PLS1)}$$

6. Pour simplifier, nous mettons $\text{cov}(x_k + u, y) - \text{cov}(x_k, y) \approx \frac{\partial \text{cov}(x_k, y)}{\partial x_k}$. Sur cette base, nous calculons les dérivées des poids w_{1k} et nous déduisons la variation des VIP_{1k}. Notons que si les poids sont négatifs, l'inverse est obtenu : si $w_{1k}^{PLS1} < 0$ et si $\text{cov}(x_k, u) \lesseqgtr 0 \implies V\tilde{I}P_{1k} \gtreqless VIP_{1k}$.

7. Notons que si le poids est négatif : si $w_{1k}^{Gini2} < 0$ et si $\text{cov}(x_k, u) \lesseqgtr 0 \implies V\tilde{I}P_{1k} \lesseqgtr VIP_{1k}$.

La régression PLS1 est définie avec toutes les variables explicatives $x_1, \ldots, x_K$, sans autres variables ou instruments. En revanche, Yitzhaki et Schechtman (2004, 2013) montrent que les régressions Gini peuvent s'interpréter comme les régressions MCO avec variables instrumentales sauf qu'aucune hypothèse n'est postulée (*par exemple* les hypothèses de linéarité et de normalité). L'emploi des vecteurs rang comme instrument a été suggéré par Durbin en 1954. Afin de définir le rôle des vecteurs rang, la régression Gini1-PLS1 peut être ré-écrite comme suit :

$$\begin{aligned} y &\approx c_1 t_1 + c_2 t_2 \\ &= c_1 \left(\sum_{k=1}^{K} w_{1k} x_k \right) + c_2 \left(\sum_{k=1}^{K} w_{2k} (R(x_k) - \hat{\beta}_{1k} t_1) \right) \\ &= \sum_{k=1}^{K} \left([c_1 w_{1k} - c_2 \sum_{k=1}^{K} (w_{2k} \hat{\beta}_{1k} w_{1k})] x_k + c_2 w_{2k} R(x_k) \right) . \end{aligned} \qquad \text{(Gini1-PLS1)}$$

Dans la régression Gini2-PLS1, les coefficients $\hat{\delta}_{Gh}$ dépendent de l'opérateur Co-Gini et par conséquent des $R(x_k)$. Puisque le Co-Gini est invariant à l'accroissement des transformations monotones, Gini2-PLS1 est également doté de bonnes propriétés pour éliminer les valeurs aberrantes :

$$\begin{aligned} y &\approx c_1 t_1 + c_2 t_2 \\ &= c_1 \left(\sum_{k=1}^{K} w_{1k} x_k \right) + c_2 \left(\sum_{k=1}^{K} w_{2k} (x_k - \hat{\beta}_{1k} t_1) \right) \qquad \text{(Gini2-PLS1)} \\ &= \sum_{k=1}^{K} \left[c_1 \frac{\hat{\delta}_{G1k}}{\sqrt{\sum_{k=1}^{K} (\hat{\delta}_{G1k})^2}} + c_2 \frac{\hat{\delta}_{G2k}}{\sqrt{\sum_{k=1}^{K} (\hat{\delta}_{G2k})^2}} - c_2 \sum_{k=1}^{K} \frac{\hat{\delta}_{G2k}}{\sqrt{\sum_{k=1}^{K} (\hat{\delta}_{G2k})^2}} \, \beta_{1k} w_{1k} \right] x_k . \end{aligned}$$

Bien que l'équation de régression PLS1 soit déterminée uniquement à l'aide des variables explicatives, les régressions Gini1-PLS1 et Gini2-PLS1 présentent en plus le vecteur rang de chaque régresseur, afin de limiter l'influence des outliers.

Définissons les *erreurs de mesure*. Pour simplifier, prenons une seule composante t_1 et deux régresseurs, tel que $\tilde{x}_1 = x_1 + u_1$ et $\tilde{x}_2 = x_2 + u_2$, où u_1, u_2 représentent les termes d'erreurs de mesure. Comme on le voit dans la conta-

mination des poids w_k, nous avons aussi $\tilde{w}_1 = w_1 + \omega_1$, $\tilde{w}_2 = w_2 + \omega_2$, et $\tilde{y} = y + v$, avec ω_1, ω_2 et v les termes d'erreurs de mesure. Suite à la littérature sur ce sujet, la régression MCO implique que $\hat{c}_1$ doit être biaisé. Les erreurs de mesure τ de la composante t_1 sont déduites comme suit :

$$\begin{aligned}\tilde{t}_1 &= \tilde{w}_1\tilde{x}_1 + \tilde{w}_2\tilde{x}_2 \\ &= t_1 + \underbrace{w_1u_1 + \omega_1x_1 + \omega_1u_1 + w_2u_2 + \omega_2x_2 + \omega_2u_2}_{\tau} = t_1 + \tau.\end{aligned}$$

Alors,

$$\hat{c}_1 = \frac{\text{cov}(\tilde{y}, \tilde{t}_1)}{\text{cov}(\tilde{t}_1, \tilde{t}_1)} = \frac{\text{cov}(y + v, t_1 + \tau)}{\text{cov}(t_1 + \tau, t_1 + \tau)}.$$

Soit $\sigma_{\tau y} := \text{plim } (\frac{1}{n}\tau^\intercal y)$, $\sigma_{t_1 v} := \text{plim } (\frac{1}{n}t_1^\intercal v)$, $\sigma_{\tau v} := \text{plim } (\frac{1}{n}\tau^\intercal v)$, $\sigma_{\tau t_1} := \text{plim } (\frac{1}{n}\tau^\intercal t_1)$, $\sigma^2_\tau := \text{plim } (\frac{1}{n}\tau^\intercal\tau)$, $\sigma^2_{t_1} := \text{plim } (\frac{1}{n}t_1^\intercal t_1)$, et ainsi

$$\text{plim } \hat{c}_1 = \frac{c_1\sigma^2_{t_1} + \sigma_{t_1 v} + \sigma_{\tau y} + \sigma_{\tau v}}{\sigma^2_{t_1} + \sigma_{\tau^2} + 2\sigma_{\tau t_1}}. \tag{2.15}$$

Si toutes les erreurs ne sont pas corrélées ($\sigma_{t_1 v} = \sigma_{\tau y} = \sigma_{\tau v} = \sigma_{\tau t_1} = 0$), alors nous retrouvons le cas des *erreurs de mesures classiques*, qui est,

$$\text{plim } \hat{c}_1 = c_1\frac{\sigma^2_{t_1}}{\sigma^2_{t_1} + \sigma^2_\tau} =: c_1\lambda,$$

où c_1 est biaisé à partir du biais d'atténuation $\lambda = \frac{\sigma^2_{t_1}}{\sigma^2_{t_1}+\sigma^2_\tau} \in [0, 1]$. Supposons que les erreurs de mesure produisent uniquement dans les régresseurs et qu'elles ne sont pas corrélées avec y *i.e.* $\sigma_{\tau y} = \sigma_{t_1 v} = \sigma_{\tau v} = 0$, ainsi :

$$\text{plim } \hat{c}_1 = \frac{c_1\sigma^2_{t_1}}{\sigma^2_{t_1} + \sigma^2_\tau + 2\sigma_{\tau t_1}}.$$

Comme démontré dans la section précédente concernant les VIP, si les vecteurs rang restent les mêmes, les poids w_k sont invariants dans la régression Gini1-PLS1 et non dans la régression Gini2-PLS1. Par conséquent, $\sigma_{\tau t_1}$ reste faible dans la régression Gini1-PLS1 puisque le terme d'erreur τ est de moindre importance, en effet $\omega_1 = \omega_2 = 0$. Ce n'est pas nécessairement le cas de la

régression Gini2-PLS1. D'autre part, si $\sigma_{\tau t_1} < 0$, la régression Gini2-PLS1 peut donner un biais d'atténuation de moindre importance pour que la qualité d'ajustement de la régression Gini2-PLS1 soit meilleure que Gini1-PLS1.
Grâce à l'opérateur Co-Gini, les termes d'erreurω_1, ω_2 sont atténués afin que les deux régressions Gini-PLS1 soient proches des résultats du cas d'erreurs de mesures classiques, et, elles donnent des résultats meilleurs que la régression PLS1. Nous montrons dans le papier [Mussard et Souissi-Benrejab (2015)] que des résultats similaires peuvent être obtenus pour la présence de valeurs aberrantes uniquement dans une observation.

• *La détection des outliers*peut être étudiée à l'aide du T^2 de Hotelling modifié, qui est une statistique de Fisher :

$$T^2 = \frac{n^2(n-h)}{h(n^2-1)(n-1)} \sum_{\ell=1}^{h} \frac{t_{\ell i}^2}{\text{var}(t_\ell)} \sim \mathcal{F}(h, n-h),$$

Ce qui revient à tester les hypothèses suivantes :

$$H0 : T_{calcul}^2 < \mathcal{F}(h, n-h),$$

$$H1 : T_{calcul}^2 \geq \mathcal{F}(h, n-h),$$

où $\text{var}(t_\ell)$ est la variance simple de t_ℓ. Le T^2 de Hotelling modifié indique si les outliers ont été enlevé, par conséquent, il justifie le recours aux régressions Gini-PLS.

Les régressions y par y donne des résultats différents de la régression multivariée de tous les y ensemble vue l'existence des corrélations possibles entre les y.

Voyons maintenant les régressions multivariées lorsque la variable dépendante correspond à une matrice.

2.2 Les régressions multivariées Gini-PLS2

L'approche Gini-PLS permet de résoudre simultanément les problèmes suivants : les valeurs aberrantes, la faible taille des échantillons ($n \leqslant K$), la corrélation excessive entre les régresseurs, et les valeurs manquantes. [8]

2.2.1 La régression Gini1-PLS2

Le premier algorithme Gini1-PLS2 est la généralisation à q variables Y_l de l'algorithme Gini1-PLS1 introduit par Mussard et Souissi-Benrejab (2015). Son objet est de permettre la construction de poids W_h dont le rôle est de maximiser le lien entre chaque Y_l et les régresseurs X_k, tout en limitant l'influence exercée par les outliers.

Les poids W_h sont déterminés par convergence, comme nous l'avons expliqué à la Section précédente, pour la méthode traditionnelle PLS2. Nous reprenons cette étape en introduisant la régression Gini semi-paramétrique.

• **Étape h_0 :** Les pondérations W_h permettant de déterminer les composantes orthogonales t_h à l'aide de la régression Gini. L'étape h_0 se répétera à l'intérieur de chaque étape h.

↪ Initialisation d'une boucle :

↪ **[R]**épéter jusqu'à convergence de W_h.

↪ Définir le vecteur u_h comme la première colonne de $Y_{(h-1)}$, avec pour tout $h \geqslant 1$, $Y_{(h)} = Y_{(h-1)} - t_h c_h^{\intercal}$.

↪ On maximise le Co-Gini de chaque colonne de la matrice $\hat{U}_{(h)}$ sur le vecteur u_h, on obtient pour $h \geqslant 2$:

$$W_{(h)k} = \frac{\text{cog}(u_h, \hat{U}_{(h)k})}{\sqrt{\sum_{k=1}^{K} \text{cog}^2(u_h, \hat{U}_{(h)k})}} \ , \ \forall k = 1 \ldots, K \ .$$

Pour $h = 1$:

$$W_{(1)k} = \frac{\text{cog}(Y_1, X_k)}{\sqrt{\sum_{k=1}^{K} \text{cog}^2(Y_1, X_k)}} \ , \ \forall k = 1 \ldots, K \ .$$

8. Les codes GAUSS des algorithmes sont disponibles sur demande.

On obtient le vecteur de taille K, $W_h = (W_{(h)1}, \ldots, W_{(h)K})^\intercal$.

$\hookrightarrow$ Le vecteur W_h est normé : $\|W_h\| = 1$.

$\hookrightarrow$ Régresser par MCO chaque ligne de la matrice $\hat{U}_{(h)}$ sur le vecteur W_h, on obtient le vecteur de taille n, $t_h = \hat{U}_{(h-1)}W_h/W_h^\intercal W_h$, avec pour $h = 1$, $t_1 = XW_1/W_1^\intercal W_1 = XW_1$.

$\hookrightarrow$ Régresser par MCO chaque colonne de la matrice $Y_{(h)}$ sur le vecteur t_h, on obtient le vecteur de taille q, $c_h = Y_{(h-1)}^\intercal t_h/t_h^\intercal t_h$, avec pour $h = 1$, $c_1 = Y^\intercal t_1/t \intercal_1 t_1$. Pour $h \geqslant 2$, $\hat{\varepsilon}_{(h-1)} \equiv Y_{(h)}$, avec $\hat{\varepsilon}_{(h)}$ la matrice contenant en colonnes les résidus des régressions ci-dessus.

$\hookrightarrow$ Régresser par MCO chaque ligne de la matrice $Y_{(h)}$ sur le vecteur c_h, on obtient le nouveau vecteur de taille n, $u_h = Y_{(h-1)}c_h/c_h^\intercal c_h$, avec pour $h = 1$, $u_1 = Yc_1/c_1^\intercal c_1$.

$\hookrightarrow$ Revenir à **[R]** où la maximisation du Co-Gini sera à nouveau effectuée afin de limiter l'influence des valeurs aberrantes.

• **Etape 1 :** On régresse par MCO chaque Y_l $(l = 1, \ldots, q)$ sur la composante t_1, dont on rappelle sa construction :

$$t_1 = \sum_{k=1}^{K} W_{(1)k}X_k \implies Y_l = c_{1l}t_1 + \varepsilon_{(1)l},\ \forall l = 1, \ldots, q.$$

La validation croisée, qui reste la même que celle de l'algorithme PLS2, permet de savoir si t_1 est significative. Si t_1 est significative, on passe à l'étape suivante.

• **Etape 2 :** On régresse le vecteur rang de chaque régresseur $R(X_k)$ [9] sur la composante t_1 par MCO afin de récupérer les vecteurs des résidus $\hat{U}_{(1)j}$:

$$R(X_k) = \beta_{(1)k}t_1 + U_{(1)k}\ ,\ \forall k = 1, \ldots, K\ .$$

On utilise à ce niveau l'étape h_0 afin de trouver les nouveaux poids W_2. La procédure est initialisée en posant le vecteur $u_2 = \hat{\varepsilon}_{(1)1}$, d'où la maximisation

9. Nous proposons dans le troisième chapitre une modification de cette régression, il s'agit de remplacer les variables $R(x_k)$ par les variables (x_k) nous nommons l'algorithme de régression Gini1'-PLS2.

du Co-Gini suivante :

$$\max \operatorname{cog}(\hat{\varepsilon}_{(1)1}, \hat{U}_{(1)} W_2) \text{ , s.c. } \|W_2\| = 1 \implies W_{(2)k} = \frac{\operatorname{cog}(\hat{\varepsilon}_{(1)1}, \hat{U}_{(1)k})}{\sqrt{\sum_{k=1}^{K} \operatorname{cog}^2(\hat{\varepsilon}_{(1)1}, \hat{U}_{(1)k})}}.$$

Une fois la convergence de W_2 obtenue (étape h_0), on utilise à présent les composantes t_1 et t_2 pour établir un lien par MCO entre chaque Y_l et les régresseurs X_k :

$$t_2 = \sum_{k=1}^{K} W_{(2)k} \hat{U}_{(1)k} \implies Y_l = c_{1l} t_1 + c_{2l} t_2 + \varepsilon_{(2)l} \ .$$

La validation croisée permet de savoir si t_2 est significative.

• **Etape h :** Les régressions partielles par MCO sont réitérées en rajoutant l'influence de t_{h-1} :

$$R(X_k) = \beta_{(1)k} t_1 + \cdots + \beta_{(h-1)k} t_{h-1} + U_{(h-1)k} \ , \ \forall k = 1, \ldots, K \ ,$$

Après maximisation (étape h_0), les poids sont initialisés par :

$$W_{(h)k} = \frac{\operatorname{cog}(\hat{\varepsilon}_{(h-1)1}, \hat{U}_{(h-1)k})}{\sqrt{\sum_{k=1}^{K} \operatorname{cog}^2(\hat{\varepsilon}_{(h-1)1}, \hat{U}_{(h-1)k})}} \ , \ \forall k = 1, \ldots, K \ ,$$

puis après convergence,

$$t_h = \sum_{k=1}^{K} W_{(h)k} \cdot \hat{U}_{(h-1)k} \implies Y_l = c_{1l} t_1 + \cdots + c_{hl} t_h + \varepsilon_{(h)l} \ .$$

La procédure s'arrête lorsque la validation croisée indique que la composante t_{h+1} n'est pas significative. L'algorithme Gini-PLS1 est valable si toutes les composantes t_h sont orthogonales deux à deux.

2.2.2 L'algorithme Gini2-PLS2

Nous avons montré dans la section précédente que les poids W_h n'ont pas d'importance dans la détermination de l'orthogonalité des composantes t_h. On

peut alors, à la première étape, obtenir des poids qui reposent non plus sur la maximisation du Co-Gini, mais par des pondérations qui sont des coefficients de pente estimés par indice de Gini ($\hat{\delta}_{G,k}$), minimisant par définition l'influence des outliers. Ensuite, contrairement à l'algorithme Gini1-PLS2, nous construisons la seconde composante t_2 en utilisant aussi la régression Gini et sa propriété d'orthogonalité $\text{cog}(e, x) = 0$.

• **Étape h_0 :** Les pondérations W_h permettent de déterminer les composantes orthogonales t_h à l'aide de la régression Gini. L'étape h_0 se répétera à l'intérieur de chaque étape h.

↪ Initialisation d'une boucle :

↪ **[R]**épéter jusqu'à convergence de W_h.

↪ Définir le vecteur u_h comme la première colonne de $Y_{(h-1)}$, avec pour tout $h \geqslant 1$, $Y_{(h)} = Y_{(h-1)} - t_h c_h^\intercal$.

↪ Pour déterminer le poids associé à variable $\hat{U}_{(h)k}$, on utilise la contribution du bêta Gini de la régression de chaque $\hat{U}_{(h)k}$ sur u_h, pour $k = 1, \ldots, K$:

$$\hat{U}_{(h)k} = \hat{\delta}_{G,k}\hat{u}_h \implies W_{(h)k} = \frac{\hat{\delta}_{G,k}}{\sqrt{\sum_{k=1}^{K}(\hat{\delta}_{G,k})^2}} \ , \ \forall k = 1, \ldots, K \ .$$

Avec pour $h = 1$:

$$X_k = \hat{\delta}_{G,k}Y_1 \implies W_{(1)k} = \frac{\hat{\delta}_{G,k}}{\sqrt{\sum_{k=1}^{K}(\hat{\delta}_{G,k})^2}} \ , \ \forall k = 1, \ldots, K \ .$$

On obtient le vecteur de taille K, $W_h = (W_{(h)1}, \ldots, W_{(h)K})^\intercal$, tel que W_h est normé : $\|W_h\| = 1$.

↪ Régresser par MCO chaque ligne de la matrice $\hat{U}_{(h)}$ sur le vecteur W_h, on obtient le vecteur de taille n, $t_h = \hat{U}_{(h-1)}W_h/W_h^\intercal W_h$, avec pour $h = 1$, $t_1 = XW_1/W_1^\intercal W_1 = XW_1$.

↪ Régresser par MCO chaque colonne de la matrice $\hat{Y}_{(h)}$ sur le vecteur t_h, on obtient le vecteur de taille q, $c_h = Y_{(h-1)}^\intercal t_h/t_h^\intercal t_h$, avec pour $h = 1$, $c_1 = Y^\intercal t_1/t_1^\intercal t_1$. Pour $h \geqslant 2$, $\hat{\varepsilon}_{(h-1)} \equiv Y_{(h)}$, avec $\hat{\varepsilon}_{(h)}$ la matrice contenant en colonnes les résidus des régressions ci-dessus.

↪ Régresser par MCO chaque ligne de la matrice $\hat{Y}_{(h)}$ sur le vecteur c_h, on

obtient le nouveau vecteur de taille n, $u_h = Y_{(h-1)}c_h/c_h^\intercal c_h$, avec pour $h = 1$, $u_1 = Yc_1/c_1^\intercal c_1$.

$\hookrightarrow$ Revenir à **[R]** où la minimisation du Gini des erreurs sera à nouveau effectuée afin de limiter l'influence des valeurs aberrantes.

- **Etape 1 :** La composante t_1 est déduite de l'étape précédente h_0 :

$$t_1 = \sum_{k=1}^{K} W_{(1)k} X_k \ .$$

La composante t_1 construite à partir des poids $\hat{\delta}_G$ provenant de la régression Gini. La première régression par MCO est :

$$Y_l = c_{1l} t_1 + \varepsilon_{1l}, \ \forall l = 1, \ldots, q.$$

La validation croisée doit indiquer si t_1 est conservée ou non.

- **Etape $h \geqslant 2$:**

On effectue pour toutes les étapes qui suivent des régressions partielles par MCO :

$$X_k = \beta_{(1)k} t_1 + \cdots + \beta_{(h-1)k} t_{h-1} + U_{(h-1)k}, \ \forall k = 1, \ldots, K \ . \tag{2.16}$$

La matrice des erreurs estimées $\hat{U}_{(h-1)}$ permet par convergence de déterminer les nouveaux poids associés à t_h obtenus par régression Gini (voir étape h_0). D'où :

$$t_h = \sum_{k=1}^{K} W_{(h)k} \cdot \hat{U}_{(h-1)k} \implies Y_l = c_{1l} t_1 + c_{2l} t_2 + \cdots + c_{hl} t_h + \varepsilon_{hl}, \ \forall l = 1, \ldots, q.$$

L'algorithme continue avec l'étape h_0 et les régressions partielles mentionnées à l'équation (2.16). La procédure s'arrête lorsque la validation croisée indique que la composante t_h ne permet pas d'améliorer la prévision du modèle.

2.2.3 Propriétés et aides à interprétations

Commentons les deux algorithmes Gini-PLS2 avant de comparer leurs propriétés à celles de l'algorithme PLS2. Les algorithmes Gini-PLS2 reposent sur l'opérateur Co-Gini (noté *cog*). Ce dernier fournit une nouvelle statistique de corrélation, nommée indice de G-correlation $\Gamma = \text{cog}(y,x)/\text{cog}(y,y)$ [10], proche de celui de Pearson, dont les propriétés mathématiques sont les suivantes (voir Yitzhaki, 2003) :

- il est inclus dans l'intervalle $[-1;1]$;
- il est insensible aux transformations monotones de x et aux transformations linéaires de y ;
- il est nul si et seulement si x et y sont des variables indépendantes.

L'algorithme Gini1-PLS2. La mesure du Co-Gini permet par définition de limiter l'influence des valeurs aberrantes par l'intermédiaire du vecteur rang. La technique Gini1-PLS2 permet donc une double prise en compte des valeurs aberrantes. Dans l'étape d'élaboration des pondérations (h_0) les valeurs aberrantes sont éliminées du vecteur Y_1 (ou $\hat{U}_{(h)}$ pour $h \geqslant 2$) et de la matrice X. L'étape 2 repose sur des régressions partielles par MCO en utilisant le vecteur rang des régresseurs X_k (pour tout k) plutôt que les régresseurs eux-mêmes. Ces régressions sont donc proches de celle de Gini dans la mesure ou la variable dépendante est un vecteur rang (la variable indépendante dans la régression Gini semi-paramétrique). Les outliers présents dans chaque régresseur X_k sont ainsi traités.

L'algorithme Gini2-PLS2. L'étape de construction des pondérations (h_0) repose aussi, comme précédemment, sur la mesure du Co-Gini, précisément sur les coefficients de pente de la régression Gini semi-paramétrique (qui sont des Co-Gini normalisés). Ainsi, dès le premier pas de l'algorithme, les valeurs aberrantes présentes à la fois dans Y_1 (ou $\hat{U}_{(h)}$ et dans la matrice des régresseurs X sont traités. Par la suite, la première composante t_1 de la régression PLS est un vecteur rang, ce qui renforce davantage la prise en compte des valeurs aberrantes.

L'avantage de l'algorithme Gini2-PLS2 est la mise en place d'un vecteur

10. $cog(y,y) = cov(y, R(y))$

rang d'une variable latente t_1 dès le début. Ainsi, avant même que la seconde étape de l'algorithme s'amorce, les effets des valeurs aberrantes présents dans X sont purgés. L'algorithme se prête donc bien à des modèles où le nombre de composantes t_h reste faible pour un traitement efficace des valeurs aberrantes. Au contraire, l'algorithme Gini1-PLS2 traite les outliers à chaque étape par l'intermédiaire de la matrice rang de X. Ainsi, pour un modèle dont les composantes t_h sont nombreuses, le modélisateur sait que les valeurs aberrantes présentes dans la matrice X sont traités graduellement à chaque étape.

Le traitement des valeurs aberrantes semblent concerner exclusivement les régresseurs. [Mussard et Souissi-Benrejab, (2015a, 2015b)] ont montré par simulations de Monte Carlo que les versions univariées des algorithmes Gini-PLS1 sont robustes aux outliers présents dans X et/ou y. Nous pouvons noter que les versions Gini1-PLS2 et Gini2-PLS2 peuvent être toutes deux étendues à une régression Gini semi-paramétrique des vecteurs Y_l sur les composantes respectivement, $t_1, \ldots, t_h$ et $T_1, \ldots, t_h$? En notant c^G les coefficients estimés par régression Gini semi-paramétrique, nous aurions les algorithmes modifiés ci-dessous :

$$Y_l = c_{1l}^G t_1 + \cdots + c_{hl}^G t_h + \varepsilon_{hl} \ , \ \forall l = 1, \ldots, q \qquad \text{(Gini1-PLS2')}$$

$$Y_l = c_{1l}^G T_1 + \cdots + c_{hl}^G t_h + \varepsilon_{hl} \ , \ \forall l = 1, \ldots, q. \qquad \text{(Gini2-PLS2')}$$

Ces techniques modifiées renforceraient la diminution de l'influence exercée par les valeurs aberrantes présents dans la matrice Y. Cependant, l'indépendance entre les composantes t_h et les résidus estimés $\hat{\varepsilon}_{hl}$ seraient perdue : $\text{cov}(\hat{\varepsilon}_{hl}, t_h) \neq 0$. L'indépendance serait traduite par $\text{cog}(\hat{\varepsilon}_{hl}, t_h) = 0$, pour tout $h = 1, \ldots, H$.

Les propriétés mathématiques. Les deux algorithmes proposés (Gini1-PLS2 et Gini2-PLS2) sont assez proches de l'algorithme PLS2. Le Tableau 2 suivant permet de résumer leurs points communs.

Tableau 2 : Propriétés

Propriétés	PLS2	Gini1-PLS2	Gini2-PLS2
(o) $t_1 \perp t_2 \perp \cdots \perp t_h$	✓	✓	✓
(i) Valeurs manquantes	✓	✓	✓
(ii) Valeurs Aberrantes	×	✓	✓
(iii) Petits échantillons ($n < k$)	✓	✓	✓
(iv) Erreurs de mesure (endogénéité)	×	✓	✓

Les aides à interprétation. Elles concernent la possibilité au modélisateur de bien expliquer la portée des résultas concernant : les corrélations, les parts de variances expliquées (R^2), la détection des points aberrants. Ainsi, les statistiques habituelles associées à la régression PLS2 standard peuvent être utilisées.

• La redondance (Tenenhaus, 1998). Il s'agit de la part de variance de la matrice X qui est expliquée par l'ensemble des composantes $t_1, \ldots, t_H$. Elle est appelée redondance, notée $Rd(\cdot)$:

$$Rd(X; t_1, \ldots, t_H) := \frac{1}{K} \sum_{h=1}^{H} \sum_{k=1}^{K} \text{cor}^2(X_k, t_h) = \sum_{h=1}^{H} Rd(X, t_h),$$

où cor dénote l'opérateur de corrélation linéaire de Pearson. La redondance est bien adaptée à la régression PLS2 standard où les régressions partielles permettent de lier les régresseurs X_k aux composantes t_h. Pour l'algorithme Gini2-PLS2, cette statistique est cohérente à partir de $h \geqslant 2$ uniquement où les MCO sont utilisés. Pour la seconde étape la régression sur Gini, l'indice de G-corrélation est utilisé. Ainsi la redondance doit être mesurée par :

$$Rd^{G2}(X; t_1, \ldots, t_H) := \frac{1}{K} \sum_{k=1}^{K} \Gamma^2(X_k, t_1) + \frac{1}{K} \sum_{h=2}^{H} \sum_{k=1}^{K} \text{cor}^2(X_k, t_h).$$

Pour l'algorithme Gini1-PLS2, les régressions partielles concernent les vecteurs rangs des régresseurs X_k. Ainsi, la redondance s'écrit :

$$Rd^{G1}(X; t_1, \ldots, t_H) := \frac{1}{K} \sum_{h=1}^{H} \sum_{k=1}^{K} \text{cor}^2(R(X_k), t_h) = \sum_{h=1}^{H} Rd^{G1}(R(X), t_h),$$

La part de variance de la matrice Y expliquée par $t_1, \ldots, t_H$ est donnée par :

$$Rd(Y; t_1, \ldots, t_H) := \frac{1}{K} \sum_{h=1}^{H} \sum_{l=1}^{q} \text{cor}^2(Y_l, t_h) = \sum_{h=1}^{H} Rd(Y, t_h).$$

• La statistique VIP (Variable Importance in the Projection) permet de mesurer la contribution de la variable X_k à l'ensemble des variables dépendantes Y_l, autrement dit, à l'ensemble de la matrice Y.

Le pouvoir explicatif de la variable X_k sur la matrice Y, à travers un modèle à h composantes, est donné par (Tenenhaus, 1998) :

$$VIP_{hk} := \sqrt{\frac{K \sum_{l=1}^{h} Rd(Y; t_l) W^2_{(l)k}}{Rd(Y, t_1, \ldots, t_h)}},$$

tel que,

$$\sum_{k=1}^{K} VIP^2_{hk} = K.$$

" La contribution d'une variable x_k à la construction de la composante t_l est mesurée par le poids w^2_{lk}. Pour chaque l la somme de ces poids sur l'ensemble des K variables x_k vaut 1. Pour mesurer la contribution de la variable x_k à la construction de Y à travers les composantes t_h, il faut prendre en compte le pouvoir explicatif de la composante t_l, mesuré par redondance $Rd(Y; t_l)$. Un même poids w^2_{lk} indique un pouvoir explicatif de la variable x_k sur l'ensemble des variables Y d'autant plus important que la redondance $Rd(Y; t_l)$ est élevée. D'où la formule du VIP utilisée. Le VIP permet de classer les variables x_k en fonction de leurs pouvoir explicatif de Y. Les variables ayant un fort VIP (> 1) sont les plus importantes dans la construction de Y." *Cf. Tenenhaus, (1998).*

• Le coefficient de détermination de chaque Y_l [11] sur les composantes $t_1, \ldots, t_h$ noté $R^2(Y_l; t_1, \ldots, t_h)$ correspond à "la part de variance expliquée par les composantes $t_1, \ldots, t_h$ ou redondance de Y_l par rapport aux composantes $t_1, \ldots, t_h$

11. $l = 1, \cdots, q$ variables dépendantes.

" [Tenenhaus, (1998)].

$$R^2(Y_l; t_1, \ldots, t_h) = \frac{1}{q} \sum_{l=1}^{q} R^2(y_l; t_1, \cdots, t_h)$$

$$R^2(Y_l; t_1, \ldots, t_h) = \frac{1}{q} \sum_{l=1}^{q} \sum_{l=1}^{q} (y_l; t_1, \cdots, t_h)$$

• Le T^2 de Hotelling de l'observation i, calculé en utilisant H composantes est défini par :

$$T_i^2 = \frac{n}{n-1} \sum \frac{t_{hi}^2}{s_h^2}$$

où s_h^2 est la variance de la composante t_h (avec division par $n-1$).
Tracy, Young et Mason (1992) ont montré que cette statistique, calculée pour une nouvelle observation, suit une loi de Fisher Snedecor à H et $n-1$ degrés de liberté. Une observation i est considéré comme atypique si :

$$T_i^2 \geq \frac{H(n^2-1)}{n(n-H)} F_{0.95}(H, n-H).$$

Il s'agit donc de faire le test d'hypothèse suivant :

$$H0 : T_{i^2 calcul} < F_{0.95}(H, n-H).$$

$$H1 : T_{i^2 calcul} \geq F_{0.95}(H, n-H).$$

L'idée est que si après l'utilisation du Gini-PLS, l'échantillon est homogène, alors on a bien purgé l'influence des valeurs aberrantes. (1) Utiliser PLS classique et voir si on a des points aberrants avec le T^2 de Hotelling. Dans l'affirmative, nous utilisons la méthode Gini-PLS. (2) Vérifier que la méthode donne alors un échantillon homogène dans le nouvel espace.

Les propriétés des régressions Gini-PLS ont montré la résolution simultanée des problèmes de multicollinéarité, de valeurs aberrantes, d'endogénéité, d'observations manquantes et de faible taille de l'échantillon. Le paragraphe suivant confirme les propriétés de ces modèles de régression. Nous effectuons dans le paragraphe suivant des simulations de Monte Carlo afin de prouver les

propriétés des régressions Gini-PLS.

2.3 Simulations

Dans ce paragraphe, nous allons simuler des lois Normales, de Weibull, de Poisson, etc., afin de prouver la robustesse des algorithmes Gini-PLS par rapport aux régressions PLS et Gini utilisées séparément

Ces simulations visent à comparer les régressions pour deux situations : (1) lorsque les valeurs aberrantes se produisent à la fois dans la variable dépendante y et dans les régresseurs, et (2) lorsque les valeurs aberrantes se posent dans la matrice X seulement. Les principales conclusions sont les suivantes : il y a une supériorité de Gini1-PLS1 dans le premier cas et de Gini2-PLS1 dans le second.- Afin de vérifier la robustesse des régressions de Gini-PLS1 pour de petits échantillons, deux simulations sont mises en œuvre pour chaque situation : une avec quelques observations $n = 10$ (contamination de 10% de l'échantillon) et l'autre avec un échantillon plus grand de taille $n = 500$ (contamination de 0.2%). Dans tous les cas, la simulation de référence est le suivant.

Simulation Benchmark

• Boucle pour $\theta = 100, \ldots, 10000$ (incrément de 100 : θ est une valeur aberrante) ;

Boucle pour $b = 1, \ldots, B = 1000$ (incrément de 1 : B est le nombre des simulations) ; Générer K variables de distribution normale $X \sim \mathcal{N}(\mu, \Sigma)$ tels que $\mu = \mathbf{0}$ avec $\Sigma_{jj} = 1$, $\Sigma_{jk} = \text{cor}(x_j, x_k) = 0.9\ \forall j \neq k$;
Soit $\varepsilon \sim \mathcal{N}(0, 1)$, fixer un vecteur β et calculer $y = \sum_k \beta_k x_k + \varepsilon$;[12]

[Étape contamination] : un outlier dépend de θ est rajouté à x_k et/ou y ; La statistique Q^2 du modèle avec outliers est calculée ;
Fin b ;
• Fin θ ;

12. Le vecteur β est choisi aléatoirement pour générer les vraies valeurs de y. Les mêmes valeurs de β sont conservés pour tous les b et pour toutes les θ. Les simulations peuvent être performés avec toute β. Comme démontré dans le cas d'erreurs de mesure ci-dessus, les valeurs de β, leurs signes, et leurs amplitudes ont des implications sur les résultats.

Valeurs aberrantes dans x_k et y

Les simulations ont été réalisées avec $K = 5$ régresseurs (le nombre de régresseurs sera discuté dans la section suivante) afin de tester la robustesse des régressions PLS1 (ligne rouge) , Gini1-PLS1 (ligne verte) et Gini2-PLS1 (Ligne bleue).

Étape Contamination : Soit $u \sim \mathcal{N}(0,1)$. Une valeur aberrante u_i est ajoutée à une seule variable explicative x_k et une observation i, ainsi, la valeur contaminée est $\tilde{x}_{ik} = x_{ik} + \theta u_i$ (u_i étant la valeur maximale de u). La valeur aberrante est supposée être indépendante de x_k : $\text{cov}(x_k, u) = 0$. La variable dépendante y est affectée en conséquence : $\tilde{y} = \sum_{k \neq j} \beta_k x_k + \beta_j \tilde{x}_j + \varepsilon$. Plus précisément, l'observation contaminée $\tilde{y}_i = y_i + v_i$ avec $v_i = \beta_j \theta u_i$ telle que v est corrélée avec u, ($\text{cov}(u, v) \neq 0$). Pour chaque valeur de θ, nous effectuons $B = 1000$ simulations. De plus, nous calculons la moyenne de B du pouvoir prédictif de chaque régression. Les chiffres ci-dessous montrent les pouvoirs prédictifs de $\bar{Q}^2$ (moyenne de Q^2 de chaque modèle avec une composante t_1, pour chaque valeur de θ en abscisses ($\theta = 100, \ldots, 10000$). [13]

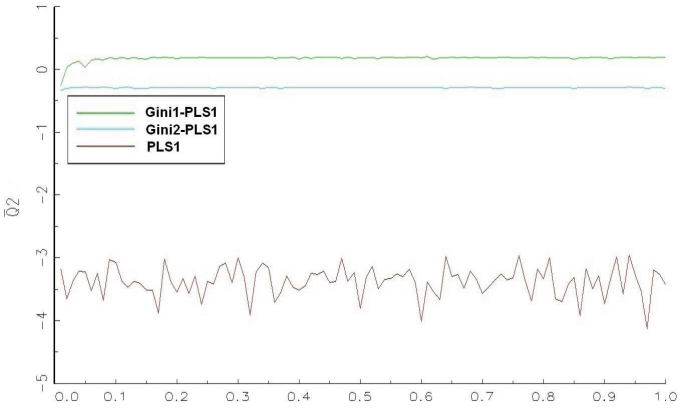


Figure 1a : Contamination de 10% ($\bar{Q}^2 = f(\Theta)$)

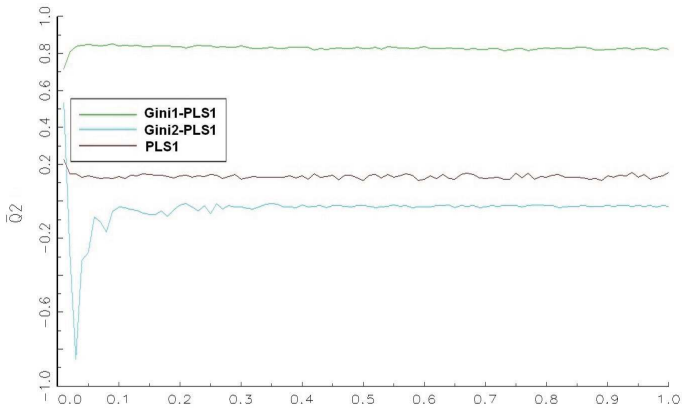


Figure 1b : Contamination de 0.2% ($\bar{Q}^2 = f(\Theta)$)

La figure 1a montre que si l'outlier u_i prend des valeurs de $\theta = 100$ à $\theta = 10000$, alors la capacité de prédiction du modèle Gini1-PLS1 avec une composante t_1 est significative seulement à $\bar{Q}^2 \approx 0,1 \geqslant 0,095$. Un mauvais ajustement est

13. Pour des raisons de simplicité, nous ne présentons que les résultats pour la première composante t_1. Les résultats sont similaires pour t_2 . Notons que dans toutes les figures, la valeur maximale dans l'axe des abscisses est égale à 1, soit, 1×10^4.

enregistré pour PLS1 lorsque $\bar{Q}^2 \approx -3,5$. Gini1-PLS1 est alors robuste lorsque l'amplitude des contaminations est importante (10%). Cependant, lorsque la contamination est faible (0,2%) dans un échantillon plus grand ($n = 500$, Figure 1b), Gini1-PLS1 ajuste mieux les données ($\bar{Q}^2 \approx 0,8$) et PLS1 devient uniquement significatif $\bar{Q}^2 \approx 0,1 \geqslant 0,095$. Pour Gini2-PLS1, $\bar{Q}^2 \approx 0$, il est donc non significatif. Il est à noter que si l'erreur u_i était la valeur minimale de u (désormais négative), les résultats seraient équivalents.

D'autre part, nous avons également simulé le cas où l'erreur u est fortement corrélée avec un régresseur x_k. Par exemple, si la corrélation de Pearson $\text{cor}(u, x_k) = 0,7$, alors nous obtenons les résultats suivants.

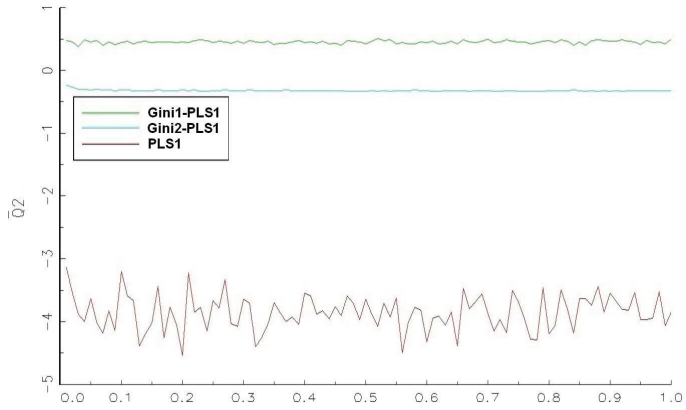


Figure 2a : Contamination de 10% ($\bar{Q}^2 = f(\Theta)$)

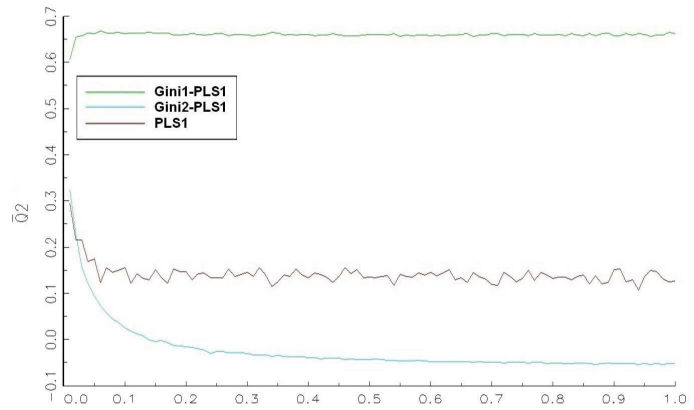


Figure 2b : Contamination de 0.2% ($\bar{Q}^2 = f(\Theta)$)

Dans le cas de 10% de contamination, Gini1-PLS1 est le meilleur modèle. Il montre une statistique $\bar{Q}^2$ autour de $0,5$, alors que Gini2-PLS1 et PLS1 sont proches de $-0,2$ et $-3,5$, respectivement. Dans le cas de $0,2\%$ de contaminations, Gini1-PLS1 est mieux classé ($\bar{Q}^2 \approx 0,65$), suivi de PLS1 ($\bar{Q}^2 \approx 0,15$), et finalement Gini2-PLS1 ($\bar{Q}^2 \approx -0,5$).
Les résultats obtenus à l'aide des simulations coïncident avec les postulats théoriques.

Valeurs aberrantes dans x

La régression Gini2-PLS1 est préférée à Gini1-PLS1 quand une observation donnée x_{ki} est affectée par des erreurs de mesure. Nous vérifions ce résultat pour les simulations portant sur des valeurs aberrantes ou non corrélées avec

les régresseurs. Contrairement à la section précédente, le nombre de régresseurs est significativement élevé : $K = 50$. La simulation de référence est la même que précédemment, sauf que $K = 50$.

1. Étape contamination : la valeur aberrante n'est pas corrélée avec la variable explicative x_k. Une seule observation x_{ik} est contaminée pour un seul régresseur x_k. Soit l'outlier $u \sim \mathcal{N}(0, 1)$. Uniquement la valeur u_i est rajoutée à x_{ik}, avec i est un emplacement aléatoire. L'observation contaminée est $\tilde{x}_{ik} = x_{ik} + \theta u_i$ où $\text{cov}(u, x_k) = 0$ et $\theta = 100, \ldots 10000$ (incrément de 100). Le modèle contaminé est $y = \sum_{k \neq j}^{50} \beta_k x_k + \beta_j \tilde{x}_j + \varepsilon$. Les figures suivantes montrent les moyennes des pouvoirs prédictifs $\bar{Q}^2$ de chaque régression estimée avec une composante t_1 (pour toutes les valeurs de θ en abscisses).

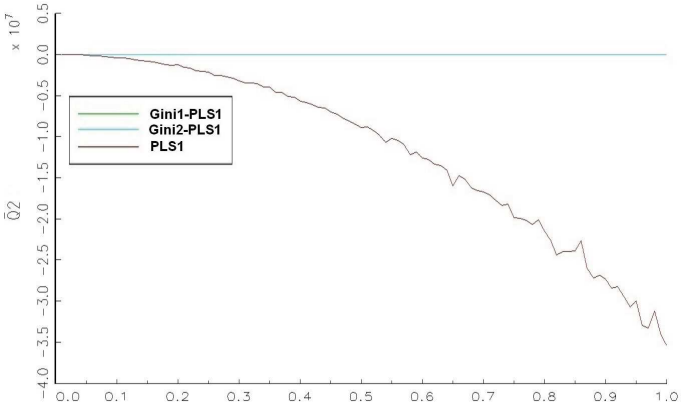


Figure 3a : Contamination de 10% ($\bar{Q}^2 = f(\Theta)$)

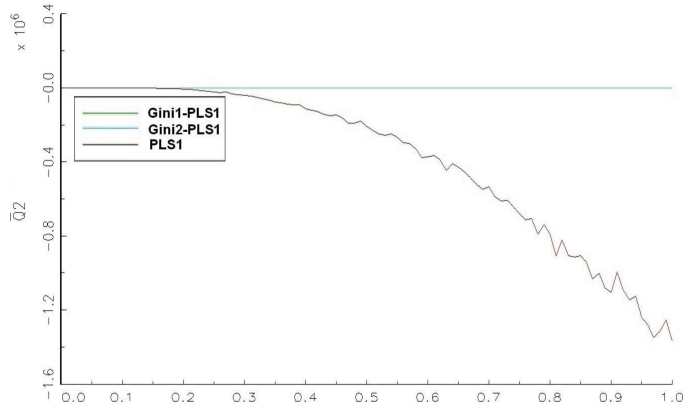


Figure 3b : Contamination de 0.25% ($\bar{Q}^2 = f(\Theta)$)

La moyenne des pouvoirs prédictifs ($\bar{Q}^2$) de la régression PLS1 est très dépendante de l'outlier [dans le cas de petits échantillons (Figure 3a) et de grands échantillons (Figure 3b)]. Les statistiques $\bar{Q}^2$ données par les régressions Gini1-PLS1 et Gini2-PLS1 sont nettement très proches. Maintenant, si nous regardons uniquement les moyennes des Q^2 pour les régressions Gini1-PLS1 et Gini2-PLS1 (Figure 4a) ou uniquement Gini2-PLS1 (Figure 4b), nous trouvons que le modèle Gini2-PLS1 est préférable au Gini1-PLS1 :

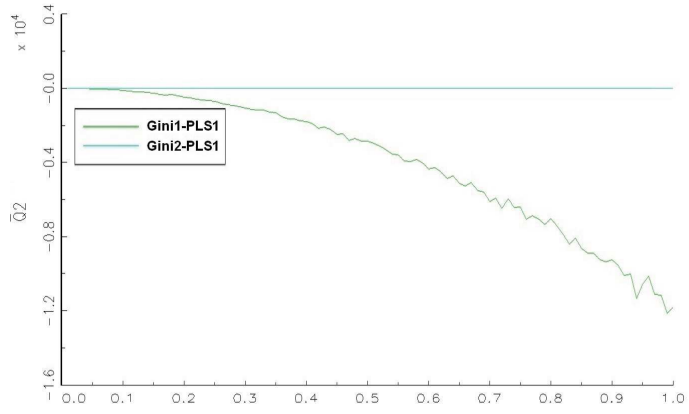


Figure 4a : Contamination de 10% ($\bar{Q}^2 = f(\Theta)$) : Gini1-PLS1 versus Gini2-PLS1

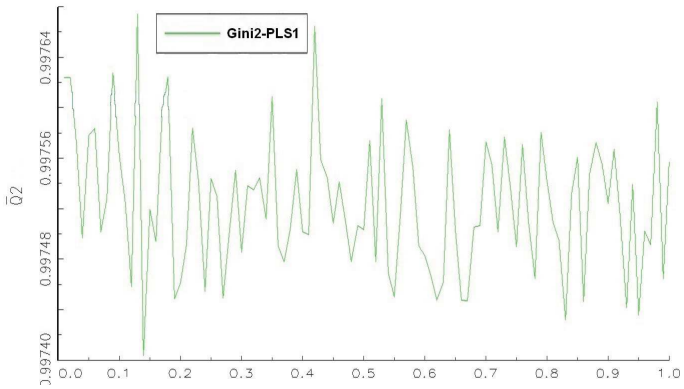


Figure 4b : Contamination de 10% ($\bar{Q}^2 = f(\Theta)$) : Gini2-PLS1

La régression Gini1-PLS1 donne des statistiques $\bar{Q}^2$ négatives, ainsi le pouvoir prédictif issu de la composante t_1 n'est pas significatif (Figure 4a). Néanmoins, la régression Gini2-PLS1 donne des pouvoirs prédictifs assez élevés avec $\bar{Q}^2 \approx 0,9$ (Figure 4b).

2. Étape contamination : L'outlier est corrélé avec toutes les variables explicatives. Un vecteur d'observations aberrantes est défini comme un régresseur quelconque $u := x_k$. L'outlier u_i est rajouté à tous les régresseurs x_k pour la seule observation i, alors que l'observation contaminée est $\tilde{x}_{ki} = x_{ki} + \theta u_i$ pour tout $k = 1, \ldots, K$ (i étant un emplacement aléatoire) avec $\theta = 100, \ldots 10,000$ (incrément de 100). Le modèle $y = \sum_{j=1}^{50} \beta_k \tilde{x}_k + \varepsilon$ est estimé. Pour chaque valeur de θ en abscisses, nous calculons la moyenne (sur B) du pouvoir prédictif de t_1 de chaque modèle ($\bar{Q}^2$).

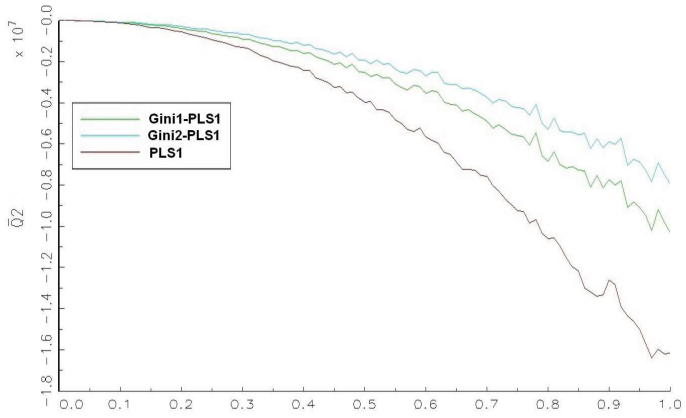


Figure 5a : Contamination de 10% ($\bar{Q}^2 = f(\Theta)$)

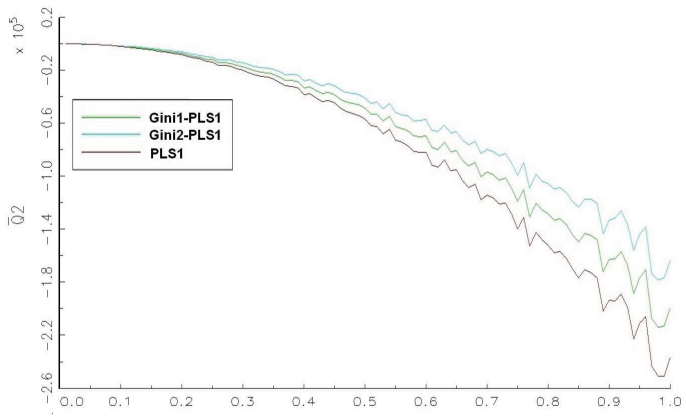


Figure 5b : Contamination de 0,2% ($\bar{Q}^2 = f(\Theta)$)

Comme nous pouvons le voir dans les Figures 5a et 5b, les modèles ne sont pas significatifs car $\bar{Q}^2 \leqslant 0$. Le modèle de régression Gini2-PLS1 reste le meilleur, même si la taille de l'échantillon augmente.

Enfin, pour conclure avec les simulations, il convient de mentionner que les simulations de Monte Carlo ne dépendent pas du processus de génération de données. En effet, d'autres lois ont été étudiées (Poisson, Uniforme, Normale, etc.), mais les résultats sont semblables. En outre, lorsque les valeurs aberrantes perturbent les valeurs de la variable dépendante y seulement, les régressions Gini-PLS sont meilleurs que les régressions PLS classiques.

Remarque : Les simulations des fortes corrélations entre les variables explicatives ont montré que les régressions PLS et Gini-PLS sont intéressantes. En présence simultanée de fortes corrélations et de multicollinéarité, les régressions Gini-PLS sont beaucoup plus intéressantes que les régressions PLS.

Conclusion

Dans les régressions Gini1-PLS, l'utilisation de l'opérateur Co-Gini, qui a deux propriétés fondamentales : il capte notamment une corrélation linéaire entre les rangs des régresseurs et la variable dépendante, et il limite l'influence des valeurs aberrantes, puisque le Co-Gini est fondé sur les vecteurs rangs. Ainsi, les algorithmes Gini1-PLS sont d'excellents candidats lorsque les échantillons sont contaminés. L'une des propriétés importantes des régressions Gini1-PLS réside dans leurs flexibilités. Lors des études de problèmes concrêts comme les contributions des variables socio-économiques aux inégalités de revenus, les régressions finales pourraient s'exprimer en fonction de x au lieu de $R(x)$. Elle sera détaillée dans le chapitre suivant (section 3.1.3).

Les régressions Gini2-PLS garantissent également une bonne correction des valeurs aberrantes quand ils se trouvent seulement au niveau des variables explicatives et lorsque le nombre de variables explicatives est important, ce résultat est confirmé par les simulations [Mussard et Souissi-Benrejab (2015a, 2015b)]. Les poids w_k ne sont pas déduites d'un processus de maximisation comme dans les régressions PLS et Gini1-PLS. Ils sont déterminés à partir de la pente de la droite de régression Gini semi-paramétrique, qui possède de bonnes propriétés concernant l'élimination de l'impact des valeurs aberrantes.

Lorsque les valeurs aberrantes sont de faible taille (en nombre et en intensité), il n'y a pas de différence significative entre les régressions PLS, et Gini-PLS. Nous notons simplement que lorsque l'amplitude des valeurs aberrantes est faible, les régressions PLS sont significatifs. Les régressions Gini1-PLS fournissent de bons résultats en présence simultanée d'erreurs de mesures, d'outliers et de multi-colinéarité.

Les modèles Gini2-PLS restent intéressants pour purifier le modèle des valeurs aberrantes qui existent uniquement au niveau des variables explicatives. Il est important de souligner que les modèles Gini-PLS2 multivariés sont capables de résoudre les mêmes problèmes que les modèles univariés Gini-PLS1. Une autre propriété se rajoute aux modèles multivariés est la distinction de l'endroit des outliers : lorsque les valeurs aberrantes affectent à la fois les va-

riables explicatives et les variables dépendantes, les régression Gini1-PLS2 est robuste ; si les valeurs aberrantes sont localisées dans la matrice des régresseurs, la régression Gini2-PLS2 s'impose, et ce quelle que soit la taille de l'échantillon.

Chapitre 3

Étude des inégalités des revenus agricoles européens

Sommaire

3.1 Modèles économétriques pour l'étude des inégalités 82
3.2 Analyse des disparités des revenus agricoles européens 94

Introduction

Dans ce chapitre, nous proposons une nouvelle approche de régression Gini-PLS basée sur la décomposition des revenus en sources, nommée RISD-Gini-PLS [1], pour estimer les contributions des variables sources à l'inégalité totale des revenus. Il s'agit d'une extension de la méthode de Morduch et Sicular (2002). La différence entre ces deux approches réside dans la méthode d'estimation. Nous reprochons aux auteurs l'usage de la régression par MCO, en particulier lorsque les données sont contaminées (valeurs aberrantes, multi-colinéarité, erreurs de mesure, etc). Nous rappelons à ce niveau que la présence d'une ou de plusieurs contamination(s) des données exige l'utilisation de l'une des régressions présentées dans les deux premiers chapitres.

Pour illustrer notre méthode RISD-Gini-PLS, nous proposons une études des inégalités des revenus agricoles européens dans le cadre de la PAC. Malgré le budget important de la commission européenne consacré aux aides agricoles et les différentes réformes de la PAC, "l'inégalité des revenus agricoles persiste. Elle est de plus très prononcée. Elle reflète les disparités des structures des productions entre les pays européens," [Butault et Lerouvillois, (1999)]. Le recours aux modèles de régressions Gini-PLS est justifié par la faible taille de l'échantillon, l'absence d'observations, la présence de multi-colinéarité et de valeurs aberrantes ou erreurs de mesure. Nous précisons à ce niveau que les contributions des variables sont calculées en deux étapes : la première consiste à estimer les paramètres du modèle, ces paramètres sont introduits par la suite dans l'indice de Gini absolu [2] pour déduire les contributions. Nous proposons aussi les intervalles de confiance pour chacune des contributions. Nous pouvons distinguer deux groupes de modèles de régressions : d'abord les modèles de régressions univariés qui servent à estimer les contributions de variables socio-économiques liées aux exploitations agricoles européennes (main d'œuvre, superficies des cultures , rendements des activités, taille du cheptel, revenus d'activités et subventions) à l'inégalité totale de production, ensuite les modèles multivariés qui estiment les contributions de chaque variable socio-

1. RISD : Regression-based Income Source Decomposition.
2. Le Gini absolu est invariant au centrage des variables, il évite d'interpréter la constante comme source de revenu.

économique aux inégalités des activité agricoles et d'élevage.

Ce chapitre est structuré comme suit : la première section traite régressions basées sur la décomposition en sources de revenus : nous présentons l'approche de Morduch et Sicular (2002) (RISD-MCO) et notre nouvelle approche (RISD-Gini-PLS). La seconde section illustre ces approches à l'aide des données du RICA [3]. Nous présentons les enjeux de la PAC, la base de données RICA et les principaux résultats des différents modèles économétriques. Les résultats des estimations sont convertis en contributions des caractéristiques technico-économiques des exploitations agricoles européennes aux inégalités totales des revenus.

3. RICA : Réseau d'Information Comptable Agricole.

3.1 Modèles économétriques pour l'étude des inégalités

"*Les inégalités économiques et leurs mesures intéressent les économistes et les sociologues* "

[*Cf.* Boisso *et alii.* (1994)]. [4]

Les mesures d'inégalités sont assez variées. Des travaux axiomatiques ont fourni différentes mesures. Par exemple, les mesures de Gini (1914), de Theil (1967), d'Atkinson (1970), de Kakwani (1980a, 1980b), et de Basmann et Slottje (1987, 1988). La plupart des axiomes construits ne décrivent pas fidèlement les inégalités [*Cf.* Shorrocks (1980, 1982, 1983) ; Basmann *et alii.* (1990)]. Pour cette raison, plusieurs travaux sur les inégalités ont été développés ces dernières décennies. En particulier, la littérature insiste sur l'étude de la décomposition des inégalités des revenus. Deux types de décomposition ont été développés, la décomposition en sous-groupes des mesures d'inégalités de revenus et la décomposition en sources, voir par exemple les approches heuristiques de Shorrocks (1980) et Shorrocks (1982), respectivement. La première approche est consacrée à l'inégalité inter et intra-groupe, alors que la seconde essaie d'identifier les sources de revenus et leurs contributions à l'inégalité totale.

Nous nous intéressons dans ce dernier chapitre à la décomposition des inégalités des revenus en sources, aussi nommée décomposition en facteurs. Deux pistes de recherche ont été suggérées dans la littérature : analytique et paramétrique. D'un côté, Rao (1969) a proposé la première approche de la décomposition de l'indice de concentration en sources de revenus. Ce travail est analytique alors que le problème consiste à décomposer le ratio de concentration et de l'exprimer comme une combinaison linéaire des souces de revenus sans recourir à la modélisation économétrique. Ce travail analytique a été développé par la suite par Fei, Ranis et Kuo (1978), Shorrocks (1982), Lerman et Yitzhaki (1985), ou plus récemment par Shorrocks (1999) et Chantreuil et

4. " Whereas economic inequality and its measurement have long been aereas of interests to both economists and sociologists" [*Cf.*Boisso *et alii.* (1994)].

Trannoy (1999) via l'utilisation de la valeur de Shapley (1953) [Voir Shorrocks (1999)], qui a aboutit à un intérêt non négligeable pour la généralisation des indices d'inégalités qu'elle offre. D'un autre côté, Morduch et Sicular (2002) ont proposé une technique paramétrique de décomposition, encore nommée régression basée sur la décomposition en facteurs. Leur approche ne s'éloigne pas beaucoup de l'approche analytique. Cependant, dans la première étape le revenu est estimé à l'aide d'un modèle économétrique, et dans la seconde étape, les sources de revenus estimées sont insérées dans un indice d'inégalité (approche analytique). L'approche économétrique est intéressante, car elle permet aux variables autres que les sources de revenus (tels que l'âge, l'éducation, la main d'œuvre, etc.) d'expliquer l'inégalité totale. En conséquence, ces variables non monétaires peuvent être utilisées pour expliquer l'inégalité totale, comme dans l'approche traditionnelle de Oaxaca (1973). Il est à noter que l'approche économétrique pour l'estimation des contributions des facteurs à l'inégalité totale est fondée sur les travaux de Shorrocks (1982).

3.1.1 Shorrocks (1982)

L'approche de décomposition des inégalités par sources de revenus, introduite par Shorrocks en 1982 est basée sur les travaux de Rao (1969). L'inégalité de revenu $I(y)$ est la somme des contributions (S^k) des K facteurs (sources) à l'inégalité totale ($I(y)$) :

$$I(y) = \sum_{k=1}^{K} S^k$$

La contribution du facteur k à l'inégalité totale ($I(y)$) dépend des sources de revenus (x_k) telles que :

$$\sum_{k=1}^{K} x_k = y.$$

La somme des sources de revenus correspond exactement au revenu y. La contribution relative de chaque facteur x_k à l'inégalité totale est :

$$s^k = \frac{S^k}{I(y)} \forall k = 1, \cdots, K.$$

L'approche de Shorrocks (1982) est déterminante pour tous les travaux

postérieurs. Elle a déclenché toute une piste de recherche en économétrie. Les premiers modèles économétriques basés sur la décomposition des inégalités en sources de revenus sont ceux de Morduch et Sicular (2002).

3.1.2 Morduch et Sicular (2002) (Approche RISD-MCO)

Les modèles de régression de Morduch et Sicular (2002) reposent sur la décomposition des inégalités en sources de revenus proposée par Shorrocks (1982). La régression est beaucoup plus flexible pour décomposer le revenu en attributs (tels que l'âge, l'éducation, la région, etc) plutôt qu'en sources de revenu.
L'équation du revenu proposée par Morduch et Sicular est $\forall x_k = x_1, \cdots, x_K$ sources (facteurs) :

$$y = \beta_0 + \sum_{k=1}^{K} \beta_k x_k + \varepsilon$$

Si on prend l'indice de Gini absolu [5], l'équation d'inégalité totale est :

$$I(y) = 4cov(y, F(y))$$

$$I(y) = 4.\frac{1}{n}.y^{\intercal}.F(y) - \bar{y}.\overline{F(y)}$$

Dans notre cas, l'estimation est effectuée sur les variables centrées d'où :

$$I(y) = 4.\frac{1}{n}.y^{\intercal}.F(y)$$

La part d'inégalité de chaque facteur x_k, calculée à l'aide du Gini absolu, est :

$$a_k(y) = 4.cov(x_k, F(y)); \forall k = 1, \cdots, K.$$

Avec, la contribution du résidu (ε) :

$$a_\varepsilon(y) = 4.cov(\varepsilon, F(y)).$$

5. Le Gini absolu appartient à l'intervalle $[0, +\infty[$.

Puisque les variables x et y sont centrées, alors

$$a_k(y) = 4.\frac{1}{n}.x_k^{\intercal}.F(y); \forall k = 1, \cdots, K.$$

L'inégalité totale se réécrit :

$$I(y) = \sum_{k=1}^{K} a_k(y) + 4.cov(\varepsilon, F(y)).$$

La valeur estimée de la contribution relative de chaque facteur x_k à l'inégalité totale de y est :

$$\widehat{s^k} = \widehat{\beta_k}\left(\frac{a_k(y)}{I(y)}\right); \forall k = 1, \cdots, K.$$

où $a_k(y)$ est la part d'inégalité. Morduch et Sicular (2002) ont critiqué l'approche de Rao car elle ne permet pas de capter l'effet d'une source constante sur l'inégalité totale [Cf. aussi Araar(2006) et Mussard (2007)]. En effet, l'approche économétrique permet de tenir compte du risque d'échantillonnage. En effet, l'écart-type des contributions de chaque attribut est :

$$\sigma(\widehat{s^k}) = \sigma(\widehat{\beta_k})\left(\frac{a_k(y)}{I(y)}\right); \forall k = 1, \cdots, K.$$

La valeur estimée de la contribution relative du résidu à l'inégalité totale est :

$$\widehat{s^{\varepsilon}} = \frac{\sum_{i=1}^{n} a_{ik}(y)\hat{\varepsilon}}{I(y)},$$

avec $a_{ik}(y)$ est un élément du vecteur $a_k(y)$, $\forall i = 1, n$ observations. Son écart-type est :

$$\sigma(\widehat{s_i^{\varepsilon}}) = \sigma(\widehat{\varepsilon_i})\sqrt{\frac{\sum_{i=1}^{n}(a_{ik}(y))^2}{(I(y))^2}}, \forall i = 1, n.$$

L'approche de Mussard *et alii.* (2005) ne diffère pas beaucoup de celle de Morduch et Sicular (2002). Elle s'intéresse aux parts des contributions des composantes ("niveau d'éducation, âge, etc.) à l'inégalité du revenu à l'échelle des ménages". Après avoir décomposé l'indice de Gini en plusieurs composantes,

Mussard *et alii.* (2005) ont estimé l'équation suivante d'inégalité de revenu :

$$y_i = \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon,$$

avec $i = i^{me}$ observation (ménage). Ces décompositions paramétriques de Mussard *et alii.* (2005) et de Chameni (2009) permettent de déterminer les contributions des attributs x_k en mesurant des inégalités intra-groupes et inter-groupes en leur affectant des intervalles de confiance.

La décomposition de Morduch et Sicular (2002) est une extension de l'approche semi-logarithmique de Fields et Yoo (2000). C'est une approche semi-logarithmique du revenu disponible par rapport aux différentes co-variables (variables muettes, variables non-linéaires et variables corrélées) intégrées dans une approche analytique. Le modèle économétrique est intéressant car il permet d'analyser des régresseurs différents des sources de revenu usuelles. En conséquence, d'autres caractéristiques non monétaires (comme l'âge, l'éducation, la main d'œuvre, etc) peuvent être utilisées, comme dans l'approche traditionnelle de Oaxaca (1973). En outre, Wan (2004) met l'accent sur l'estimation des régressions et leurs interprétations.

La contribution estimée de chaque $\hat{s}^k$ permet une meilleure compréhension du rôle des différentes sources de revenu dans le montant global de l'inégalité. De plus, la valeur estimée de la contribution du résidu, notée $\hat{s}^\varepsilon$ permet un contrôle de la qualité de la régression. En outre, cette approche fondée sur la régression généralise les décompositions (naturelles) analysées par Rao (1969), Fei, Ranis et Kuo (1978), Shorrocks (1982, 1983, 1999), Lerman et Yitzhaki (1985, 1989a), et Silber (1993), parmi d'autres.

Dans ce qui suit, nous nous intéressons à l'approche d'estimation basée sur la décomposition en sources de revenu (RISD [6]). Nous proposons une extension des régressions Gini-PLS introduites par Mussard et Souissi-Benrejab (2015a, 2015b). La partie Gini de cette régression est particulièrement intéressante pour gérer les valeurs aberrantes, voir les travaux fondamentaux de

6. RISD : Regression based Income Source Decomposition.

Olkin et Yitzhaki (1992) et de Yitzhaki et Schechtman (2013) pour une vue d'ensemble de la méthodologie Gini. La partie de la régression PLS est utilisée pour faire face aux fortes corrélations entre les sources de revenus, aux observations manquantes et à la faible taille de l'échantillon. Nous montrons aussi que les régressions Gini-PLS basées sur la décomposition des inégalités en sources représentent une extension de la méthode de Morduch et Sicular (2002). [7] Nous traitons l'indice du Gini absolu pour calculer l'inégalité afin d'éviter l'interprétation de la constante de régression comme source de revenu. De plus, nous étudions l'inférence des contributions des sources à l'inégalité totale, et nous fournissons des extensions possibles qui traitent l'endogénéité qui pourrait se produire dans les régressions du revenu.

3.1.3 Une nouvelle approche RISD-Gini-PLS

Ce paragraphe a pour objet d'exposer les régressions Gini-PLS basées sur la décompositions des revenus en sources. C'est une approche "*regression-based income source decomposition*" (RISD) semblable à celle initiée par Morduch et Sicular (2002). L'approche RISD-Gini-PLS est justifiée par la présence de quelques problèmes liés aux régressions. Nous proposons aussi les modèles RISD-PLS et RISD-Gini afin d'effectuer des comparaisons dans la section 3.2.

Avant de passer à l'application, nous rappelons rapidement (pour l'étape 1) les principales différences entre les régressions PLS, Gini, et Gini-PLS. [8]

• **Régression Gini semi-paramétrique :**

Le β_G remplace le β des MCO de l'approche de Morduch et Sicular.

Le coefficient de pente de la régression Gini semi-paramétrique simple ($y = \alpha + \beta_G x$) est :

$$\hat{\beta}_G = \sum_{j<i} v_{ij}\tau_{ij}, \text{ avec } v_{ij} = \frac{x_i - x_j}{\sum_{j<i}(x_i - x_j)}, \text{ et } \sum_{j<i} v_{ij} = 1. \qquad (3.1)$$

7. Nous prouvons aussi que les régressions Gini-PLS sont flexibles avec des transformations de Box-Cox introduites par Wan (2004), ces transformations Box-Cox ne feront pas l'objet d'un travail ultérieur.

8. Toutes les régressions (MCO, Gini, PLS et Gini-PLS) avec les estimations des contributions des régresseurs à l'inégalité totale des revenus, les tests de significativité, les tests d'autocorrélations et d'hétéroscédasticité sont programmés sous Gauss.

- **Régression PLS :**

$$w_{1k} = \frac{\text{cov}(x_k, y)}{\sqrt{\sum_{k=1}^{K} \text{cov}^2(x_k, y)}} \ , \ \forall k = 1, \ldots, K \ .$$

Les régresseurs x_k sont donc affectés d'un poids w_{1k} qui maximise le lien entre y et chaque x_k. La première composante t_1 est donc exprimée comme suit :

$$t_1 = w_{11}x_1 + \cdots + w_{1k}x_k + \cdots + w_{1K}x_K \ .$$

Même si les régresseurs sont parfaitement corrélés, la variable dépendante y peut être régressée sur t_1 par MCO :

$$y = c_1 t_1 + \varepsilon_1 \ .$$

- **Régressions Gini1-PLS1 et Gini1'-PLS1 :**

Nous attribuons aux régresseurs x_k le vecteur de poids w_{1k} suivant :

$$w_{1k} = \frac{cog(y, x_k)}{\sqrt{\sum_{k=1}^{K} cog^2(y, x_k)}} \ , \ \forall k = 1, \ldots, K \ .$$

avec *cog* est l'opérateur Co-Gini (ou covariance Gini). La première composante t_1 est déterminée ainsi :

$$t_1 = \sum_{k=1}^{K} w_{1k} R(x_k).$$

La variable y est régressée par MCO sur la composante t_1 :

$$y = c_1 t_1 + \varepsilon_1.$$

La régression finale s'exprime donc en fonction des variables instrumentales $(R(x_k))$. La régression Gini1-PLS1 est efficace pour contourner les problèmes liés aux valeurs aberrantes, à l'endogénéité, à la multicollinéarité et à la faible taille de l'échantillon. Nous pouvons modifier cette régression en remplaçant les variables $(R(x_k))$ par les variables (x_k) pour exprimer les équations de régression finale en fonction des variables explicatives (x_k). , nous nommons

cette nouvelle régression : Gini1'-PLS1.

$$t_1 = \sum_{k=1}^{K} w_{1k} x_k.$$

La variable y est régressée par MCO sur la composante t_1 s'exprime alors en fonction de x_k.

- **Régression Gini2-PLS1 :**

L'élément w_{1j} du vecteur poids w_1 est déterminé à partir de la régression Gini semi-paramétrique :

$$\hat{\delta}_{G1j} = \frac{\text{cog}(y, x_j)}{\text{cog}(x_j, x_j)} \implies w_{1j} = \frac{\hat{\delta}_{G1j}}{\sqrt{\sum_{j=1}^{p} (\hat{\delta}_{G1j})^2}} \ , \ \forall j = 1, \ldots, p \ .$$

La première composante t_1 est donnée par :

$$t_1 = \sum_{j=1}^{p} w_{1j} x_j \ .$$

L'équation de régression finale s'exprime ainsi :

$$y = c_1 t_1 + \varepsilon_1 \ .$$

- **Régressions multivariées :**

Pour les régressions multivariées (PLS2, Gini1-PLS2 et Gini2-PLS2 : *Cf.* les deux chapitres précédents). La régression Gini1'-PLS2 est obtenue en remplaçant le vecteur des régresseurs $R(x_k)$ par x_k dans les équations de régressions partielles.

Une fois que les paramètres sont estimés à l'aide des régressions PLS, Gini et Gini-PLS, nous pouvons déduire les contributions des différentes variables explicatives et des résidus à l'inégalité totale des revenus, comme suit :

$$\widehat{s^k} = \widehat{\beta_k} \left(\frac{a_k(y)}{I(y)} \right) ; \forall k = 1, \cdots, K.$$

Les écart-types sont :

$$\sigma(\widehat{s^k}) = \sigma(\widehat{\beta_k}) \left(\frac{a_k(y)}{I(y)} \right); \forall k = 1, \cdots, K.$$

La valeur estimée de la contribution relative du résidu à l'inégalité totale est :

$$\widehat{s^\varepsilon} = \frac{\sum_{i=1}^{n} a_{ik}(y)\hat{\varepsilon}_i}{I(y)}, \forall i = 1, n.$$

Son écart-type est :

$$\sigma(\widehat{s^\varepsilon}) = \sigma(\widehat{\varepsilon}) \sqrt{\frac{\sum_{i=1}^{n} (a_{ik}(y))^2}{(I(y))^2}}, \forall i = 1, n.$$

Ces approches peuvent être considérés comme des extensions des travaux précédents de Morduch et Sicular (2002) et de Fields et Yoo (2000).

Relations avec les approches précédentes

L'utilisation des régressions RISD-Gini-PLS est compatible avec les approches précédentes développées par Fields et Yoo (2000) et par Wan (2004). En effet, comme le montre Wan (2004), les décompositions basées sur la régression de Morduch et Sicular (2002) et Fields et Yoo (2000) peuvent être généralisées grâce à la transformation de Box-Cox [9] dans laquelle y est remplacée par $y^{(\lambda)}$ où :

$$y^{(\lambda)} := \frac{y^\lambda - 1}{\lambda}, \ \lambda \geqslant 0,$$

et les régresseurs x_k sont remplacés par $x_k^{(\theta)}$ où :

$$x_k^{(\theta)} := \frac{x_k^\theta - 1}{\theta}, \ \theta \geqslant 0.$$

Indices absolus et estimateurs de variance

Wan (2004) explique comment faire face à des indices relatifs afin de gérer

9. Dans ce cas, la régression Gini (3.1) peut être estimée avec diverses combinaisons de λ et θ pour retrouver le modèle de Wan (2004), qui comprend le modèle de Morduch and Sicular (2002) lorsque $\lambda = \theta = 1$. Ces cas ne seront pas envisagés.

la constante de la régression. Nous montrons que les indices absolus sont également intéressants puisque l'ordonnée à l'origine est nulle, et par conséquent, la décomposition par régression fournit des contributions uniquement pour les variables et les résidus. En effet, soit Y le revenu non centré, ainsi $y = Y - \bar{Y}$. Le modèle centré Gini-PLS (Gini-PLS) peut être réécrit, par exemple avec deux composantes t_1 et t_2 comme :

$$\begin{aligned}\hat{y} &= \hat{c}_1 t_1 + \hat{c}_2 t_2 \\ &= \hat{c}_1 \left(\sum_{k=1}^{K} w_{1k} x_k \right) + \hat{c}_2 \left(\sum_{k=1}^{K} w_{2k} (x_k - \hat{\beta}_{1k} t_1) \right) \\ &= \sum_{k=1}^{K} \underbrace{\left[\hat{c}_1 w_{1k} + \hat{c}_2 w_{2k} - \hat{c}_2 w_{1k} \sum_{j=1}^{K} \hat{\beta}_{1j} w_{2j} \right]}_{\hat{\xi}_k} x_k \equiv \sum_{k=1}^{K} \hat{\xi}_k x_k. \qquad \text{(Gini-PLS)}\end{aligned}$$

Le coefficient $\hat{\xi_k}$ peut s'exprimer comme coefficient de régression de y sur X. On peut distinguer les contributions s^k du facteur k de celle du résidu :

$$J(Y) = J(y) = J(y) \left[\sum_{k}^{K} \hat{s}^k + \hat{s}^{\varepsilon} \right] = \left[\sum_{k=1}^{K} \frac{\sum_{i=1}^{n} \hat{\xi}_k x_{ik}}{J(y)} + \frac{\sum_{i=1}^{n} e_i}{J(y)} \right].$$

La variance estimée nous permet de déduire les intervalles de confiance des contributions s^k :

$$IC_{\hat{s^k}} = [\hat{S^k} \pm 1,96 \sqrt{\widehat{\text{Var}} \left(\hat{\xi_k} \right)}]$$

À cet égard, en partant de la condition d'orthogonalité, la variance des coefficients estimés $\hat{\xi}_k$ est, pour deux composantes :

$$\begin{aligned}\mathrm{Var}\left(\hat{\xi_k}\right) &= \mathrm{Var}\left(\hat{c}_1 w_{1k}\right) + \mathrm{Var}\left(\hat{c}_2 w_{2k} - \hat{c}_2 w_{1k} \sum_{j=1}^{K} w_{2j}\hat{\beta}_{1j}\right) \\ &= w_{1k}^2 \mathrm{Var}\left(\hat{c}_1\right) + \mathrm{Var}\left(\hat{c}_2\left(w_{2k} - w_{1k}\sum_{j=1}^{K} w_{2j}\hat{\beta}_{1j}\right)\right).\end{aligned}$$

Puisque $\hat{\beta}_{1j}$ provient de la régression sur t_1 et comme $\hat{c}_2$ provient de la régression sur t_2, et puisque $t_1 \perp t_2$, on obtient pour tout $j = 1, \ldots, K$:

$$\begin{aligned}\mathrm{cov}\left(\hat{c}_2, \hat{\beta}_{1j}\right) &= \mathbb{E}\left[\left(\hat{c}_2 - c_2\right)\left(\hat{\beta}_{1j} - \beta_{1j}\right)\right] \\ &= \mathbb{E}\left[\left(t_2^\intercal t_2\right)^{-1} t_2^\intercal\ \varepsilon_2\ u_{(1)j}^\intercal\ t_1\left(t_1^\intercal t_1\right)^{-1}\right] \\ &= \mathrm{cov}\left(\varepsilon_2, u_{(1)j}\right)\mathbb{E}\left[\left(t_2^\intercal t_2\right)^{-1} t_2^\intercal t_1\left(t_1^\intercal t_1\right)^{-1}\right] \\ &= 0.\end{aligned}$$

En conséquence, puisque $\hat{c}_2$ et $\hat{\beta}_{1j}$ sont indépendants,

$$\begin{aligned}\mathrm{Var}\left(\hat{\xi_k}\right) = w_{1k}^2\mathrm{Var}\left(\hat{c}_1\right) &+ \mathrm{Var}\left(\hat{c}_2\right)\mathrm{Var}\left(w_{2k} - w_{1k}\sum_{j=1}^{K} w_{2j}\hat{\beta}_{1j}\right) \\ &+ \mathrm{Var}\left(\hat{c}_2\right)\mathbb{E}^2\left[w_{2k} - w_{1k}\sum_{j=1}^{K} w_{2j}\hat{\beta}_{1j}\right] \\ &+ \mathrm{Var}\left(w_{2k} - w_{1k}\sum_{j=1}^{K} w_{2j}\hat{\beta}_{1j}\right)\mathbb{E}^2\left[\hat{c}_2\right].\end{aligned}$$

Il vient :

$$\mathrm{Var}\left(w_{2k} - w_{1k}\sum_{j=1}^{K} w_{2j}\hat{\beta}_{1j}\right) = w_{1k}^2\sum_{j=1}^{K} w_{2j}^2\mathrm{Var}(\hat{\beta}_{1j}) + 2w_{1k}^2\sum_{j<h} w_{2j}w_{2h}\mathrm{cov}\left(\hat{\beta}_{1j}, \hat{\beta}_{1h}\right).$$

Pour $\text{cov}(\hat{\beta}_{1j}, \hat{\beta}_{1h})$, nous avons :

$$\begin{aligned}\text{cov}\left(\hat{\beta}_{1j}, \hat{\beta}_{1h}\right) &= \mathbb{E}\left[\left(\hat{\beta}_{1j} - \beta_{1j}\right)\left(\hat{\beta}_{1h} - \beta_{1h}\right)\right] \\ &= \mathbb{E}\left[\left(t_1^\intercal t_1\right)^{-1} t_1^\intercal \; u_{(1)j} \; u_{(1)h}^\intercal \; t_1 \left(t_1^\intercal t_1\right)^{-1}\right] \\ &= \left(t_1^\intercal t_1\right)^{-1} \text{cov}(u_{(1)j}, u_{(1)h}).\end{aligned}$$

Finalement, nous obtenons :

$$\begin{aligned}\text{Var}\left(\hat{\xi}_k\right) = w_{1k}^2 \text{Var}\left(\hat{c}_1\right) &+ \text{Var}\left(\hat{c}_2\right)\left[w_{1k}^2 \sum_{j=1}^{K} w_{2j}^2 \text{Var}(\hat{\beta}_{1j}) + 2w_{1k}^2 \sum_{j<h} w_{2j} w_{2h} \frac{\text{cov}(u_{(1)j}, u_{(1)h})}{t_1^\intercal t_1}\right] \\ &+ \text{Var}\left(\hat{c}_2\right)\left[w_{2k} - w_{1k} \sum_{j=1}^{K} w_{2j} \beta_{1j}\right]^2 \\ &+ \left[w_{1k}^2 \sum_{j=1}^{K} w_{2j}^2 \text{Var}(\hat{\beta}_{1j}) + 2w_{1k}^2 \sum_{j<h} w_{2j} w_{2h} \frac{\text{cov}(u_{(1)j}, u_{(1)h})}{t_1^\intercal t_1}\right] \hat{c}_2^2.\end{aligned}$$

Donc, un estimateur naturel de cette variance est :

$$\begin{aligned}\widehat{\text{Var}}\left(\hat{\xi}_k\right) = w_{1k}^2 \widehat{\text{Var}}\left(\hat{c}_1\right) &+ \widehat{\text{Var}}\left(\hat{c}_2\right)\left[w_{1k}^2 \sum_{j=1}^{K} w_{2j}^2 \widehat{\text{Var}}(\hat{\beta}_{1j}) + 2w_{1k}^2 \sum_{j<h} w_{2j} w_{2h} \frac{\widehat{\text{cov}}(\hat{u}_{(1)j}, \hat{u}_{(1)h})}{t_1^\intercal t_1}\right] \\ &+ \widehat{\text{Var}}\left(\hat{c}_2\right)\left[w_{2k} - w_{1k} \sum_{j=1}^{K} w_{2j} \hat{\beta}_{1j}\right]^2 \\ &+ \left[w_{1k}^2 \sum_{j=1}^{K} w_{2j}^2 \widehat{\text{Var}}(\hat{\beta}_{1j}) + 2w_{1k}^2 \sum_{j<h} w_{2j} w_{2h} \frac{\widehat{\text{cov}}(\hat{u}_{(1)j}, \hat{u}_{(1)h})}{t_1^\intercal t_1}\right] c_2^2.\end{aligned}$$

Afin de montrer la supériorité des modèles RISD-Gini-PLS, nous les comparons avec les approches (RISD-MCO, RISD-PLS et RISD-Gini), dans le cadre des inégalités des rémunérations agricoles des pays européens. Notons que pour le cas multivarié, il s'agit de récupérer les paramètres estimés et les implémenter pour chaque variable dépendante dans l'indice du Gini absolu. Nous proposons aussi des estimations des intervalles de confiance pour les contributions, qui permettent de se faire une idée de la fiabilité des contributions.

3.2 Analyse des disparités des revenus agricoles européens

Pour illustrer les différents modèles présentés dans les sections précédentes, nous étudions les contributions des sources de revenus et des variables technico-économiques des exploitations agricoles européennes aux inégalités des revenus. Depuis la fondation de la PAC, et malgré ses réformes successives, les inégalités des revenus agricoles[10] entre les pays européens ont persisté, voire se sont légèrement accentués (3.2.3). Pour étudier les causes de ces inégalités, plus précisément les contributions des différentes variables socio-économiques à l'inégalité totale, nous adoptons les approches RISD. Les résultats des régressions permettent de comptabiliser les contributions des différentes sources à l'inégalité totale.

Dans cette section nous présentons les objectifs de la PAC, la base de données et les principaux résultats d'estimations.

3.2.1 Les enjeux de la Politique Agricole Commune (PAC)

Dans ce paragraphe, nous présentons un bref aperçu de l'histoire de la PAC et les traits saillants des principales réformes. Les objectifs de cette première politique commune en Europe, instaurée depuis 1957 (Traité de Rome) et mise en œuvre en 1962 sont : de garantir l'indépendance alimentaire, d'augmenter la productivité agricole et de stabiliser les marchés européens ("approvisionner les consommateurs à des prix raisonnables") et d'améliorer les revenus des agriculteurs, Bureau (2007).

En 1957, le paysage européen était bien différent d'aujourd'hui : l'Europe venait de sortir de la seconde guerre mondiale avec des déficits alimentaires. Ainsi, les premiers objectifs de la PAC étaient d'encourager les agriculteurs européens à produire davantage. Un système d'aides est mis en place. Tous les pays de l'union européenne disposent d'un budget commun dont la redistribution est la même pour tous les pays. Afin d'éviter la concurrence les produits

10. Revenus agricoles moyens pondérés par pays.

agricoles circulent librement à l'intérieur de l'union européenne. La frontière des pays membres de l'UE est protégée via un système de droits de douane, mis en place en 1968 (" des tarifs douaniers communs et un marché unique pour le sucre, la viande bovine et le lait"). Ainsi, les produits agricoles étrangers arrivent plus chers sur les marchés européens, [Bureau, (2007)].

Le système d'aides aux prix garantis adopté depuis la mise en place de la PAC a aboutit à un accroissement remarquable des productions. Suite à la crise de surproduction de 1984, les Ministres européens de l'agriculture se sont mobilisés pour limiter les productions excédentaires. Ils ont instauré un système de quotas pour les productions laitières et de gel obligatoire [11] pour les grandes cultures (céréales, oléagineux et protéagineux (COP)). Ces systèmes d'aides (gel de terre et quotas laitiers) ont été mis en place lors de la réforme de Mac Sharry en 1992. Cette première réforme de la PAC consiste à remplacer les aides aux prix garantis par les aides directes versées aux agriculteurs. Les dépenses communautaires sont plafonnées par type de production, donc les montants des aides se sont réduits. Lors des Accords de Berlin en 1999, la seconde réforme dite l'Agenda 2000 a apporté des nouveautés dans les systèmes de subventions. La PAC est fondée sur deux piliers : le premier concerne les aides directes aux productions et à l'organisation des marchés et le second s'intéresse aux aides pour le développement rural et à de nombreux domaines du secteur agricole (environnement, zones défavorisées etc.).

La réforme de 2003 ou accord de Luxembourg garde les mêmes principes de la PAC et apporte quelques nouveautés. Elle s'intéresse d'une part au découplage des aides, c'est à dire à la dissociation des aides des quantités produites par les agriculteurs, d'autre part aux nouvelles dimensions tels que la certification environnementale, la qualité et sécurité alimentaire, etc. En 2007, la commission européenne a migré vers une PAC rationnelle et moderne, dans le cadre du Bilan de santé. L'intérêt de la PAC est donc d'améliorer son fonctionnement et de l'adapter aux nouveaux défis qui se présentent dans l'union européenne à 27 pays membres [12], et de s'adapter aussi à un contexte interna-

11. Les agriculteur européens sont obligés de geler certaines parcelles, c'est à dire doivent cesser de produire sur une partie de leurs exploitations, et en contre partie, ils reçoivent des aides de la Commission européenne.

12. Les réformes de la PAC se sont accompagnées avec l'élargissement de l'union euro-

tional en pleine mutation, *Cf.* Capeye (2014).

En 2012, les états membres de la commission européenne engagent d'importantes négociations afin de réformer la PAC. La commission a présenté un ensemble de propositions. L'une d'elle pose le principe de verdissement de la PAC : il s'agit de verser 30% des aides directes aux agriculteurs qui remplissent trois conditions : maintenir les pâturages permanents, pratiquer au moins trois cultures distinctes, garantir la présence d'infrastructures agro-écologiques (haies, arbres, mares, murets, etc) et de surfaces d'intérêt environnemental sur au moins 7% des terres, [Ministère de l'agriculture, de l'agroalimentaire et de la forêt, (2015)].

3.2.2 Présentation de la base de données

La base de données est fournie par le réseau d'information comptable agricole (RICA) *Cf.* FADN, (2014). Les données que nous utilisons dans la thèse sont des données agrégées, établies à partir des fiches d'exploitations individuelles. La typologie est faite selon l'orientation technico-économique des exploitations agricoles européennes et selon la dimension économique. "La dimension économique d'une exploitation s'exprime en unité de dimension économique ou en équivalent hectare de blé. Elle est calculée à partir des coefficients de marges brutes standard des différentes productions qui permettent de pondérer, et donc d'additionner des surfaces cultivées et des effectifs animaux" [Desriers et alii.(2000)]. La base de donnée est détaillée par pays et par année. Ainsi, les données sont rapportées au nombre d'exploitations dans chaque pays. Les problématiques de fortes corrélations entre les régresseurs sont omniprésentes dans les bases de données des revenus agricoles européens. Ces bases de données sont très riches en valeurs aberrantes.

Nous nous intéressons aux variables suivantes :

— Variables dépendantes :

(y) : Production totale : recettes des exploitations (€), (y_1) : Production

péenne, en début des années 1990, on compte 12 pays européens, et à partir de 2007, l'europe s'est élargie au 27 pays.

végétale : revenus des activités agricoles des exploitations (€), (y_2) : Production animale : revenus issus de l'élevage (€),

— Variables explicatives :

x_1 : Heures de main d'œuvre (heures travaillée par an), x_2 : Superficie agricole totale utilisée (hectares), x_3 : Superficie d'arbres fruitiers (hectares), x_4 : Superficie agricole totale hors production (jachère et terres gelées [13]) (hectares), x_5 : Total unités de bétail (Unités de Bétail (UB)), x_6 : Superficie fourragère (hectares), x_7 : Produits laitiers (Kilogrammes), x_8 : rendements des céréales (€), x_9 : rendements des cultures protéagineuses (€), x_{10} : rendements des cultures oléagineuses (€), x_{11} : Total subventions sauf sur investissements (€).

3.2.3 Résultats

Il est à noter que depuis la fondation de la Politique Agricole Commune, et malgré ses réformes successives, les inégalités des revenus agricoles entre les pays européens persistent. Les fortes inégalités au niveau des rémunérations sont liées à la production totale des exploitations agricoles, et aux rémunérations végétales et d'élevage. Ces inégalités sont identifiées à l'aide de l'indice du Gini absolu, qui indique un niveau d'inégalité assez élevé, voire un accroissement d'une période à l'autre. D'où proviennent ces inégalités ?
Une fois les paramètres sont estimé à l'aide des régressions (RISD-MCO, RISD-Gini, RISD-PLS, RISD-Gini1'-PLS1 et RISD-Gini2PLS1), nous estimons les contributions des sources de revenus à l'inégalité totale, comme précisé dans les sections précédentes. Les résultats des estimations sont présentés en fonction des réformes les plus importantes de la Politique Agricole Commune (PAC), en particulier la réforme de Mac Sharry (1992) et la réforme de 2003 (Accord de Luxembourg). Les estimations sont effectuées année par année et sur une étendue de vingt ans. Vu la similarité des résultats pour chaque période, nous présentons uniquement les trois années de référence suivantes (1990, 1997 et 2008) : l'année 1990 pour représenter la période précédent la réforme de 1992, l'année 1997 pour voir l'impact de la réforme de Mac Sharry et ses répercus-

13. La jachère est un choix de l'exploitant dans le cadre de l'assolement et rotation des parcelles pour reposer les parcelles et réduire le risque de maladies, alors que les terres gelées est imposé par la Commission européenne pour limiter les productions et en contre partie, les exploitants reçoivent une prime de gel.

sions sur les inégalités, et l'année 2008 pour présenter l'impact de la réforme de 2003 sur les inégalités. Quelles sont les variables qui contribuent beaucoup à l'inégalité totale ? Les réformes-ont-elles changé les contributions ?

Avant la première réforme de Mac Sharry

Le soutien par les prix est indépendant de la taille des exploitations. Cette politique a favorisé l'intensification des cultures. Dans cette période, la PAC s'est focalisée sur les prix des marchés des produits agricoles. Les objectifs sont essentiellement l'amélioration des revenus des agriculteurs et l'encouragement à produire davantage. La Commission européenne fixe des prix garantis permettant aux agriculteurs de vendre leurs produits. Si les quantités produites dépassent la demande sur le marché, la Commission européenne achète les produits excédentaires aux prix garantis. Ces quantités excédentaires sont stockées ou bien subventionnées pour l'exportation. À la fin des années 1980, cette politique s'est traduite par une surproduction, notamment pour les produits laitiers et grandes cultures (céréales par exemple) et des inégalités de revenus importantes. Pour trouver les causes de ces inégalités, nous adoptons la démarche économétrique RISD que nous avons présentée dans les sections précédentes. Les résultats des régressions permettent de déterminer les contributions des différentes sources à l'inégalité totale.

Tableau 1 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (%) (1990)

$\hat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	-15,5	-15,5	0,09*	1,1*	14,5*
x_2	-463,2	-463,1	-0,000006*	0,02*	-21,3*
x_3	-34,6	-0,000000005	0,00000007	0,0001	5,6
x_4	-15,8	-34,6	0,0001***	0,0006*	9,1*
x_5	-82,2	-15,8	0,0001*	0,06*	40,9*
x_6	241,1	-82,3	-0,00003*	0,01*	-2,2*
x_7	262,8	241,1	82,1*	81,2*	35,4*
x_8	197,6	262,8	6,9*	6,2*	2,6*
x_9	0,02	197,6	0,05*	0,4*	8,1*
x_{10}	-16,8	-16,8	0,1*	0,6*	2,7*
x_{11}	26,4	26,4	0,2*	0,1*	-1,2*
Résidus	0	0	10,4	10,2	5,7
Durbin-Watson	1,487	1,457	2,826	2,815	2,403
Test de White	0,157	1,753	2,621	2,614	5,391

Coefficients significatifs à : ***1%, *10%.

Durbin-Watson : Autocorrélation négative : [0 ;0,098] ; Absence d'autocorrélation : [0,497 ;3,503] ; Autocorrélation positive : [3,902 ;4].
Test de White : La statistique F est égale à $4,96$.

Les modèles MCO et Gini sont biaisés. Ils surestiment les contributions des variables liées aux productions de lait et de céréales (ces contributions dépassent les 100%), ceci est dû à la multi-colinéarité des variables explicatives (3.2.3). Voyons ce qui se passe au niveau des régressions PLS1, Gini1'-PLS1 et Gini2PLS1.

Les statistiques VIP (3.2.3) montrent que la variable produits laitiers (x_7) est significative pour la régression PLS1, les variables main d'œuvre (x_1), taille du cheptel(x_5), produits laitiers (x_7)et protéagineux (x_9) sont importantes pour la régression Gini1'-PLS1 et la variable arboriculture est intéressante dans l'algorithme Gini2-PLS1.

Les variables (taille du cheptel (x_5), productions laitières (x_7) et rende-

ments céréaliers(x_8)) accroissent les inégalités d'environ 85% dans le cas de la régression Gini2-PLS1. Ces mêmes variables accroissent les inégalités de revenus à hauteur de 82% dans le cas de la régression PLS1, néanmoins dans ce cas, la variable x_5 qui représente la taille du cheptel n'intervient qu'avec 0.001%. Il est à noter que les régression MCO, Gini, PLS1 et Gini1'-PLS1 montrent des inversions de signe, par exemple les subventions sur activité agricole (variable x_{11}) accroit les inégalités de revenus, sauf pour Gini2-PLS1. Le T^2 de Hotelling (3.2.3) montre qu'il n'existe pas d'outliers dans les trois modèles PLS1, Gini1'-PLS1 et Gini2-PLS1. Cependant, les covariances montrent qu'il existe des erreurs de mesure (endogénéité), ainsi, les corrélations entre les régresseurs x_k et le terme d'erreur ε ne sont pas nulles (3.2.3).
La statistique Q montre qu'il y a une seule composante (t_1) significative pour les trois modèles de régression(elle est de $0,73$, $0,77$ et de $0,46$ pour les trois régressions PLS1, Gini1'-PLS1 et Gini2-PLS1, respectivement). La régression appropriée dans ce cas est repérée à l'aide des statistiques de Durbin-Watson et le test de White montrent l'absence d'autocorrélation des erreurs et l'absence d'hétéroscédasticité dans le cas du modèle Gini2-PLS1. Pour confirmer notre choix du Gini2-PLS1, nous regardons les redondances qui représentent les parts de variances de la composante t_1 expliquées par les différentes variables y et x_k.

Tableau 2 –Redondances (Rd) sur la première composante t_1 (Rd(y,t_1) et Rd(x_k,t_1)) (1990)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,84	0,86	0,65
x_1	0,46	0,48	0,44
x_2	0,35	0,42	0,78
x_3	0,32	0,32	0,34
x_4	0,26	0,27	0,29
x_5	0,86	0,89	0,83
x_6	0,37	0,42	0,69
x_7	0,99	0,98	0,66
x_8	0,17	0,25	0,67
x_9	0,08	0,13	0,45
x_{10}	0,06	0,11	0,42
x_{11}	0,01	0,03	0,34

Le tableau 2 indique que les redondances [14] R^2 (Rd(y,t_1)) des modèles PLS1 et Gini1'-PLS1 sont plus importants que celui du Gini2-PLS1. Cependant, les parts de variances de t_1 expliquées par les régresseurs sont très significatives comparativement aux deux autres modèles (PLS1 et Gini1'-PLS1). De plus, la contribution du résidu pour le modèle Gini2-PLS1 est de l'ordre de 5%. Nous pouvons donc retenir le modèle Gini2-PLS2.

Nous pouvons conclure pour cette période que l'inégalité totale de production est sensible aux productions laitières et aux cultures intensives (Céréales, Oléagineux et Protéagineux), ces activités ont contribué à la hausse de l'inégalité totale à hauteur de 50%. Le mécanisme de soutien par les prix garantis est indépendant de la taille des exploitations. En effet, tous les exploitants reçoivent des aides en fonction de leurs quantités produites. Les rendements à l'hectare des exploitations européennes sont très proches. Donc, les grandes exploitations vont bénéficier de montants d'aides plus élevés que les petites exploitations. Ceci justifie encore la présence des inégalités entre les pays européens.

14. variances expliquées par les variables x_k ou y

À présent supposons que les revenus d'activité agricole soient se scinder en deux composantes : les revenus d'élevage, y_2 et les revenus d'activité végétale y_1. Ainsi, nous estimons les contributions simultanées des mêmes régresseurs à l'inégalité totale de chaque composante de revenu. Pour ce faire, nous utilisons les approches multivariées (RISD-PLS2, RISD-Gini1'-PLS2 et RISD-Gini2-PLS2).

Dans cette même période précédent la réforme de Mac Sharry, les inégalités de revenus d'élevage y_2 et végétale y_1 sont assez élevées.

Tableau 3– Valeurs estimées des contributions des régresseurs à l'inégalité totale (%) (année 1990)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
$\hat{s_k}$	y_1	y_2	y_1	y_2	y_1	y_2
x_1	31,3	18,3	24,8	15,1	28,3	16,2
x_2	15,1	-3,9	-19,6	-5,2	-13,5*	-3,2
x_3	9,1	-2,1	-8,4	-7	10*	-4,2*
x_4	9,6	8,5	1	10,6	11,9*	10,5*
x_5	27,6	35,6	33,2	42,1	28,1**	36,3*
x_6	-16,1	-7,8	-8,4	-4	-14,3*	-6,9*
x_7	30,5	44,6	34,1	43,5	31	45,5
x_8	-3,2	0,4	-11,6	-1,4	5	0,1
x_9	28,5	3,6	44,7	5,2	30,7	3,7
x_{10}	6,6	0,8	-4	-0,5	3,8	0,5
x_{11}	6,3	-0,9	5,8	-0,8	5,8	-0,9
Résidus	3	2,7	2,7	2,4	3,7	2,4
Durbin-Watson [15]	0,009	0,0002	2,911	2,183	2,335	2,878
Test de White [16]	0,9	0,08	10,710	14	11,6	1,3

Coefficients significatifs à : **5%, *10%.

Dans ces trois modèles, les valeurs de la statistique Q sont autour de $0,4$ ($0,35$; $0,36$ et $0,36$ pour PLS1, Gini1'-PLS1, Gini2-PLS1 respectivement). Nous retenons une seule composante optimale. La variance expliquée par la première composante t_1 (Redondance Rd : voir tableau suivant) est autour de 50% pour la variable dépendante y_1 (revenu d'activité végétale). Cette même

composante explique autour de 65% de la variance totale pour la variable y_2(revenu d'activité animale).

Les statistiques VIP des trois modèles PLS2, Gini1'-PLS2 et Gini2-PLS2 3.2.3 montrent que les variables x_5 et x_7 sont importantes. Dans le cas de la régression Gini2-PLS2, les variables x_5 et x_7 accroissent les inégalités des revenus issus de la production animale à hauteur de 80%. Ces mêmes variables accroissent les inégalités des revenus d'activité végétale d'environ 60%.

Dans la période précédent la réforme de Mac Sharry, les variables x_2 (SAU) et x_6 (superficie des fourrages) réduisent l'inégalité totale des rémunérations des activités végétales d'environ 30%, et réduisent l'inégalité totale des revenus d'élevage de 10% environ.

Nous constatons une inversion de signe pour les modèles PLS2 et Gini1'-PLS2 pour la variable céréales. Les rendements des céréales accroissent les inégalités des revenus d'activités végétales d'environ 60% dans le cas de la régression Gini2-PLS2. Le T^2 de Hotelling (3.2.3) montre qu'il existe des valeurs aberrantes pour le modèles PLS2. La régression des x_k sur les résidus montre que les $cov(x_k, \varepsilon)$ ne sont pas nuls, en d'autres termes, la présence d'erreurs de mesure (endogénéité) (voir Tableau exogénéité en Annexes)(voir tableau en Annexes).

Nous retenons ici le modèle de régression Gini2-PLS2 vu l'absence d'autocorrélations et d'hétéroscédasticité, en plus de la cohérence des signes des coefficients estimés.
Dans la période qui précède la première réforme de la PAC, nous pouvons déduire que les produits laitiers et céréalier ont de fortes contributions à l'inégalité totale des revenus. En partant des mêmes variables explicatives (main d'œuvre, subventions, etc), nous pouvons déduire que les modèles multivariés ont des résultats très similaires aux modèles univariés, en terme de contributions des céréales et des produits laitiers à l'inégalité des revenus d'activité végétale et à l'inégalité des revenus d'élevage.

Tableau 4 –Rd(y,t_1) Rd(x_k,t_1) (1990)

Variables	PLS2	Gini1'-PLS2	Gini2-PLS2
$y1$	0,504	0,512	0,490
$y2$	0,658	0,647	0,632
x_1	0,491	0,477	0,457
x_2	0,771	0,783	0,782
x_3	0,356	0,336	0,355
x_4	0,306	0,278	0,321
x_5	0,867	0,850	0,848
x_6	0,700	0,691	0,702
x_7	0,693	0,685	0,664
x_8	0,624	0,657	0,648
x_9	0,392	0,431	0,417
x_{10}	0,368	0,405	0,393
x_{11}	0,346	0,339	0,369

Le mécanisme d'incitation aux productions via le système des prix garantis n'a pas résolu le problème des inégalités. Comment vont évoluer les contributions des sources à l'inégalité totale des revenus après la première réforme ?

La réforme de Mac Sharry (réforme de 1992) consiste à remplacer le mécanisme d'aide au prix garantis par des aides directes à l'hectare et par tête de bétail. Cette réforme a pour objet de limiter les surproduction des produits laitiers, céréaliers, des oléagineux et des protéagineux. Après cette première réforme, les inégalités ont augmenté.

Après la réforme de 1992

L'approche économétrique résumée dans le tableau suivant fournit les valeurs estimées des contributions des régresseurs à l'inégalité totale.

Tableau 5 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (%) (année 1997)

$\hat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	15,7	7,5	0,2*	0,6*	15,9*
x_2	104,1	120,2	-0,00006*	0,01*	-9,1*
x_3	1,1	1,8	-0,00000002	0,00002	0,2
x_4	0,5	0,4	0,000000005	0,00001	0,4
x_5	50,9	73,8	0,0004*	0,04*	48,5*
x_6	-74,2	-92,6	-0,00006*	0,008*	-7,9*
x_7	41,9	37	88,2*	86,6*	36,7*
x_8	-40,7	-61,1	5,8*	2,9*	6,4*
x_9	10,7	1,67	0,03*	0,09*	2,4*
x_{10}	1,8	19	0,1*	0,3*	1,3*
x_{11}	-12,2	-7,7	-12,9*	-8,5*	-1,5*
Résidus	0,2	-0,07	18,3	17,8	6,3
Durbin-Watson	2,699	1,603	2,425	2,372	2,767
Test de White[17]	3,44	1,55	4,94	4,542	9,852

Coefficients significatifs à : *10%.

Durbin-Watson : Autocorrélation négative : [0 ;0,098] ; Absence d'autocorrélation : [0,497 ;3,503] ; Autocorrélation positive : [3,902 ;4].
Test de White : La statistique F est égale à $3,48$.

Vu les fortes corrélations entre les variables explicatives (3.2.3) , les régressions MCO et Gini surestiment la variable main d'œuvre. Les contributions de cette variable dépassent 100%.
Nous allons nous intéresser maintenant aux régressions PLS1, Gini1'-PLS1 et Gini2-PLS1. La statistique Q prend des valeurs autour de 0.6 et permet de retenir une seule composante significative. Les variables total cheptel et produits

laitiers accroissent les inégalités d'environ 85% dans le cas de la régression Gini2-PLS1. La variable produits laitiers accroit les inégalités d'environ 87% dans chacune des régression PLS1 et Gini1'-PLS1. Nous constatons une inversion de signe pour la régression Gini1'-PLS1 et ce pour les variables liées à la superficie agricole utilisée (superficie totale et superficie consacrée aux fourrages). Cette période correspond au système d'aides par hectare ou par tête de cheptel. Donc si l'exploitant possède plus d'hectares de terre, il va bénéficier de subventions, donc probablement, il y aura moins de disparités à ce niveau.

Comparativement à la période précédent la réforme de 1992, les grandes cultures (céréales, Oléagineux et Protéagineux) ont contribué à la hausse de l'inégalité totale uniquement de 10% environ (comparativement à 50% avant la réforme) dans le cas de la régression Gini2-PLS1. Ceci est dû à la baisse des prix garantis de ces produits. Les superficies de ces cultures sont aussi règlementées. En effet, les agriculteurs reçoivent des aides à l'hectare pour les cultures COP. Cependant, s'ils dépassent un certain seuil, les agriculteurs seront obligés de geler leurs terres, et ils reçoivent des aides. Ce mécanisme de la PAC a pour objet de limiter les productions des COP (céréales, oléagineux, protéagineux) et a favorisé la légère réduction de contributions des céréales à l'inégalité totale de revenu.

Les contributions des résidus à l'inégalité totale pour les deux modèles PLS1 et Gini1'-PLS1 sont assez importants (environ 18%), alors que celle estimée à l'aide de la régression Gini2-PLS1 n'est pas significative.

Le T^2 de Hotelling (3.2.3) montre qu'il n'existe pas d'outliers dans les trois modèles PLS1, Gini1'-PLS1 et Gini2-PLS1. Cependant, les covariances montrent la présence d'erreurs de mesure (endogénéité), ainsi, les corrélations entre les régresseurs x_k et le terme d'erreur ε ne sont pas nuls (les variables sont exogènes)(voir tableau 3.2.3 en Annexes).

Pour choisir le modèle de régression approprié, nous examinons les redondances qui représentent les parts de variances de la composante t_1 expliquées par les différentes variables y et x_k.

Tableau 6–Rd(y,t_1) et Rd(x_k,t_1) (année 1997)

Variables	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,78	0,77	0,60
x_1	0,32	0,33	0,48
x_2	0,37	0,45	0,77
x_3	0,22	0,22	0,16
x_4	0,008	0,001	0,002
x_5	0,79	0,81	0,77
x_6	0,36	0,41	0,69
x_7	0,99	0,98	0,67
x_8	0,24	0,33	0,61
x_9	0,07	0,12	0,44
x_{10}	0,11	0,17	0,54
x_{11}	0,23	0,31	0,51

Les coefficients de déterminations R^2 (ou bien Rd(y,t_1)) des modèles PLS1 et Gini1'-PLS1 sont plus élevés que celui du Gini2-PLS1. Or, les parts de variance de l'axe t_1 expliquées par les x_k pour le Gini2-PLS1 sont plus intéressantes que les modèles PLS1 et Gini1'-PLS1. Nous pouvons donc retenir la régression Gini2-PLS1, puisque ce modèle est très significatif de point de vue le cohérence des signes, l'absence d'auto-corrélations et l'absence d'hétéroscédasticité, en plus de la faible contribution du résidu.

Dans la période qui suit la réforme de Mac Sharry, les contributions des produits laitiers à l'accroissement de l'inégalité totale restent assez élevées. Les aides directes ont amélioré les revenus agricoles des zones de grandes cultures, (voir aussi Desriers et al.(2000)). Cependant, cette première réforme de la PAC ne s'est pas traduite par une diminution des disparités de revenu, [Butault et Lerouvillois (1999)]. Les grandes exploitations reçoivent des aides plus que proportionnelles à leur dimension, [Desriers et al.(2000)].

Tableau 7 – Valeurs estimées des contributions des x_k à l'inégalité totale (%)(1997)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
$\hat{s_k}$	y_1	y_2	y_1	y_2	y_1	y_2
x_1	26,9	6,1	23,8	41	31	5
x_2	-6,1	-1	-7,4	-2	-2,7	-1,5
x_3	-1,6	0,2	-2,4	1,1	-1,7	-2,3
x_4	0,1*	-0,3*	-1*	0,1***	0,4	-0,1
x_5	29,1	50,5	28,2	45,4	20,7	58,6
x_6	-15,5	-9,8	-16,2	-8,9	-15,6	-8,2
x_7	39,9	53,3	42,5	54,8	45,5	46,7
x_8	8,9	9,1	17.8	12	8.7	10.4
x_9	13.9	2,7	7,7	0,7	8,3	0,1
x_{10}	7	-4,4	6	0,6	9,5	-4,9
x_{11}	-13,4	-7,5	9,5	-10,5	-14,6	-5,6
Résidus	10,8	1,1	10,2	2,6	10,3	0,9
Durbin-Watson	0,09	0,0002	2,911	2,183	2,335	2,878
Test de White	0,9	0,08	10,7	14	11,6	1,3

Coefficients significatifs à : ***1%, *10%.

Durbin-Watson : Autocorrélation négative : [0 ;0,098] ; Absence d'autocorrélation : [0,497 ;3,503] ; Autocorrélation positive : [3,902 ;4].
Test de White : La statistique F est égale à $3,48$.

La statistique Q est de $0,38$; $0,39$; $0,35$ respectivement pour les modèles PLS2, Gini1'-PLS2 et Gini2-PLS2. elle permet de retenir la première composante. La variance expliquée par la première composante t_1 (Rd) est autour de 50% pour la variable dépendante y_1 (revenu d'activité agricole). Cette même composante explique autour de 64% de la variance totale pour la variable y_2 (revenu d'élevage).

Les statistiques VIP (3.2.3) des trois modèles PLS2, Gini1'-PLS2 et Gini2-PLS2 montrent que les variables taille du cheptel et produits laitiers sont importantes. Les variables main d'œuvre, taille du cheptel et produits laitiers accroissent l'inégalité des revenus d'activité végétale d'environ 90% dans le cas

de la régression Gini1'-PLS2. Ces variables ont des redondances assez élevées dans la régression Gini1'PLS2. Le T^2 de Hotelling (3.2.3) montre la présence d'outliers pour les deux modèles de régressions PLS2 et Gini2-PLS2, plus précisément au niveau de la variable y_2 qui contient des valeurs aberrantes (des revenus très élevés). Il existe aussi des erreurs de mesures pour ces modèles, (détectées à l'aide des tests d'exogénéité 3.2.3).

Dans le cas de la régression Gini2-PLS2, les variables accroissent les inégalités des revenus issus de la production animale de plus de 90%. Ces mêmes variables accroissent les inégalités des revenus d'activité végétale d'environ 65%. Nous constatons une inversion de signe pour les modèles PLS2 et Gini2-PLS2 pour les variables COP. Les rendements des COP (Céréales, Oléagineux et Protéagineux) accroissent les inégalités des revenus d'activités végétales d'environ 30% dans le cas de la régression Gini2-PLS2. Nous retenons ici le modèle de régression Gini1'-PLS2 vu l'absence d'auto-corrélations et d'hétéroscédasticité, en plus de la faible contribution du résidu à l'inégalité totale de y_2 (2.6%).

Tableau 8 –Rd(y,t_1) et Rd(x_k,t_1) (année 1997)

Variables	PLS2	Gini1'-PLS2	Gini2-PLS2
$y1$	0,427	0,427	0,376
$y2$	0,643	0,642	0,564
x_1	0,495	0,499	0,448
x_2	0,744	0,748	0,807
x_3	0,179	0,165	0,153
x_4	0,002	0,001	0,015
x_5	0,789	0,799	0,419
x_6	0,660	0,680	0,679
x_7	0,699	0,702	0,627
x_8	0,610	0,589	0,692
x_9	0,439	0,422	0,508
x_{10}	0,521	0,513	0,593
x_{11}	0,501	0,492	0,559

Pour cette période qui suit la réforme de Mac Sharry, nous pouvons déduire qu'avec les mêmes variables technico-économiques, les contributions à

l'inégalité totale des revenus sont différentes des contributions à l'inégalité horizontale (c'est à dire à chacune des inégalités $I(y_1)$ et $I(y_2)$). Les variables liées à l'élevage laitier restent les plus contributrices à la formation des inégalités.

Après la réforme de 2003

La réforme de 2003 (accord de Luxembourg) est essentiellement un "découplage des aides des quantités produites proposée par la Commission européenne. " Les aides sont sous forme de paiement unique à l'exploitation, c'est à dire versées à l'exploitant indépendamment de la production, [Bureau, (2007)].

Nous constatons un accroissement des inégalités au cours de la période suivant cette réforme.

Tableau 9 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (%) (année 2008)

$\hat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	-1,9	41,6	0,08*	0,7	3,6*
x_2	104,6	-159,4	-0,00002*	0,02*	48,3*
x_3	-0,2	4,3	0,000000001	0,00001*	0,6**
x_4	1,5	4,7	-0,000000008	0,00005**	0,5*
x_5	84,3	17,8	0,00006*	0,02*	16,8*
x_6	-107,8	29,1	-0,00001*	0,01*	-1,2*
x_7	48,2	92,5	94,7*	93,3*	16,9*
x_8	-122,6	70,1	2,8*	0,8*	8*
x_9	25,2	7,3	0,0001*	0,08*	6,4*
x_{10}	69,6	-8,3	0,1*	0,8*	-33,2*
x_{11}	-1,7	-1,3	-1,8*	-0,3*	9,2*
Résidus	0,7	1,2	3,9	4,3	23,7***
Durbin-Watson	1,922	2,243	2,142	2,141	1,511
Test de White	14,750	0,302	9,606	9,378	3,598

Coefficients significatifs à : ***1%, **5%, *10%.

Durbin-Watson : Autocorrélation négative : [0 ;0,098] ; Absence d'autocorrélation : [0,497 ;3,503] ; Autocorrélation positive : [3,902 ;4]

Test de White : La statistique F est égale à 2.83.

La statistique Q prend des valeurs autour de 0.8 dans la régression PLS1 et Gini1'-PLS1, elle est d'environ 0.6 pour la régression Gini2-PLS1. Nous retenons donc une seule composante significative pour les trois modèles.

Ce tableau montre des inversions de signes pour les modèles MCO, Gini, PLS1 et Gini2-PLS1. De plus des contributions de certaines variables qui dépassent 100% pour les deux régressions MCO et Gini. En plus de la multicolinéarité (3.2.3) et de l'endogénéité (3.2.3), le T^2 de Hotelling (3.2.3) indique la présence d'outliers (Slovaquie est une observation aberrante. Nous constatons une inversion de signe pour la contribution de la variable subventions sur activités à l'inégalité totale dans la régression Gini2-PLS1. Nous remarquons aussi une inversion de signe pour la régression PLS1 et ceci pour les variables liées aux superficies (totale, hors production, consacrée aux fourrages, arboriculture fruitière). Les régressions PLS1 et Gini1'-PLS1 montrent que les produits laitiers accroissent les inégalités de plus de 90 %. La contribution des résidus à l'inégalité totale n'est pas significative dans le cas des régressions PLS1 et Gini1'-PLS1. Cependant, la contribution du résidu pour le Gini2-PLS1 dépasse 20%.

Dans cette période suivant la réforme de 2003 (découplage des aides), la plupart des coefficients estimés à l'aide des régressions PLS1, Gini1'-PLS1 et Gini2-PLS1 sont significatifs au seuil 1%.

Le tableau suivant des redondances permettra de connaître les parts de variances expliquées par les différentes variables.

Tableau 10 –Rd(y,t_1) et Rd(x_k,t_1) (année 2008)

Variables	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,89	0,85	0,71
x_1	0,59	0,71	0,84
x_2	0,71	0,82	0,96
x_3	0,12	0,19	0,33
x_4	0,11	0,14	0,23
x_5	0,86	0,81	0,64
x_6	0,71	0,78	0,91
x_7	0,98	0,93	0,75
x_8	0,74	0,84	0,94
x_9	0,53	0,58	0,71
x_{10}	0,63	0,75	0,89
x_{11}	0,80	0,88	0,93

Les parts de variances expliquées par la première composante sont assez élevés pour les trois modèles , nous pouvons retenir le modèle Gini1'-PLS1. Cet algorithme montre que les produits laitiers augmentent les inégalités d'environ 93%, et que les subventions d'activités agricoles réduisent les inégalités de 0.3 %.

La présence de points aberrants au niveau de la variable dépendante y, la cohérence des signes des coefficients estimés, l'absence d'auto-corrélation et d'hétéroscédasticité, ainsi que la significativité des paramètres estimés, nous guident à retenir le modèle Gini1'-PLS1.

Malgré les prélèvements effectués sur les comptes des exploitations " qui reçoivent plus de 5000 € par an " pour supporter le second pilier de développement rural [Bureau, (2007)], les inégalités des revenus se sont accrues.

Tableau 11 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (%) (année 2008)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
$\hat{s_k}$	y_1	y_2	y_1	y_2	y_1	y_2
x_1	33,7	1,1	39,5	-3,4	16,8	15
x_2	-10,5**	-8	-18,6***	-5,1	-3,3	-8,6***
x_3	6,6*	5*	9 *	1*	3,4*	-0,2*
x_4	0,7*	0,9*	-0,8*	-0,9*	0,7*	1,4*
x_5	27,5*	68,1*	24,2*	70,9*	40,2*	71,9*
x_6	-32,6*	-16,9*	-13,5*	-9,5*	-24,3*	-23,5*
x_7	61,1	57,6	56,1	41,5	33,8	50,7
x_8	1,1	3	-13	3,2	19,2	10,2
x_9	23,4	-0,8	26,7	5,2	-1,4	-3,8
x_{10}	11,3	-2,2	-7,1	-9,6	15,2	1,4
x_{11}	-23,5	-45	7,6	5,3	-5,2	-1,6
Résidus	1,1	0,6	5,2	1,4	4,8	0,4
Durbin-Watson [18]	0,011	2,122	2,192	2,183	1,897	2,514
Test de White [19]	9,4	8,7	10,4	1,4	4,8	2,9

Coefficients significatifs à : ***1%, **5%, *10%.

La statistique Q est autour de $0,4$ pour les trois régressions, elle permet de retenir la première composante t_1. Dans cette période suivant la réforme 2003, concernée par le découplage des aides (subventions indépendantes des productions).

Dans le cas de la régression Gini2-PLS2, les variables céréales et protéagineux accroissent les inégalités des revenus d'activité végétale d'environ 35%. La variable taille du cheptel accroit les inégalités des revenus d'activités animale d'environ 70% dans le cas de la régression Gini2-PLS2. Nous retenons ici le modèle de régression Gini2-PLS2 vu l'absence d'auto-corrélation et d'hétéroscédasticité, pour les deux variables y_1 et y_2.

Les statistiques VIP (3.2.3) des trois modèles PLS2, Gini1'-PLS2 et Gini2-PLS2 montrent que les variables taille du cheptel et produits laitiers sont importantes. Les variables main d'œuvre, taille du cheptel et produits laitiers

accroissent l'inégalité des revenus d'activité végétale d'environ 90% dans le cas de la régression Gini1'-PLS2.

Le T^2 de Hotelling (3.2.3) montre la présence d'outliers pour les trois modèles de régressions. Il existe aussi des erreurs de mesures pour ces modèles. Ceci montre l'absence d'exogénéité des régresseurs. Nous remarquons des inversions de signe pour les deux modèles PLS2 et Gini1'-PLS2 pour les variables COP.

Tableau 12–Rd(y,t_1) et Rd(x_k,t_1) (2008)

Variables	PLS2	Gini1'-PLS2	Gini2-PLS2
$y1$	0,807	0,810	0,815
$y2$	0,477	0,497	0,511
x_1	0,838	0,823	0,813
x_2	0,952	0,945	0,938
x_3	0,346	0,319	0,318
x_4	0,238	0,229	0,229
x_5	0,664	0,685	0,697
x_6	0,909	0,908	0,902
x_7	0,768	0,787	0,796
x_8	0,939	0,933	0,930
x_9	0,703	0,710	0,707
x_{10}	0,887	0,872	0,865
x_{11}	0,931	0,931	0,929

La variance expliquée par la première composante t_1 est autour de 80% pour la variable dépendante y_1 (revenu d'activité végétale). Cette même composante explique autour de 50% de la variance totale pour la variable y_2(revenu d'activité animale). Les redondances des variables explicatives sur la première composante sont assez élevées pour les trois modèles.

Vu les cohérences de signes des paramètres estimés, l'absence d'auto-corrélation et d'hétéroscédasticité, la présence de valeurs aberrantes au niveau de la matrice des variables explicatives, nous retenons la régression Gini2-PLS2. Dans cette période suivant la réforme de 2003, l'activité d'élevage continue à accroître les inégalités. Les superficies des cultures fourragères réduisent les in-

égalités de chacune des rémunérations végétales et animales d'environ 25%. Les subventions accordés aux exploitants européens réduisent légèrement les inégalités. Pour cette période, nous pouvons déduire qu'avec les mêmes variables technico-économiques, les contributions à l'inégalité totale des revenus sont différentes des contributions à l'inégalité horizontale ($I(y_1)$ et $I(y_2)$). Les variables liées à l'élevage laitier (produits laitiers) restent les plus contributrices à la formation des inégalités des revenus (y, y_1 et y_2).

Conclusion

Les approches RISD Gini-PLS que nous proposons dans cette thèse sont des extensions des méthodes économétriques de Morduch et Sicular (2002). L'importance de ce travail réside dans le fait que les régressions Gini-PLS sont capables de résoudre simultanément les problèmes de données manquantes, de faible taille de l'échantillon, d'endogénéité, de multi-colinéarité et d'outliers. Nous avons eu des résultats assez fiables du point de vue de la robustesse des modèles. L'estimation des contributions des variables technico-économiques agricoles à l'inégalité totale des rémunérations nous a permis d'éclaircir l'impact de la PAC et de ses réformes sur les inégalités. Étant donné que la base de données RICA présente les problèmes déjà cités (outliers, multicolinéarité, etc.), les résultats des estimations à l'aide de régressions Gini-PLS (univariés et multivariés) ont permis de déterminer les variables qui contribuent à l'inégalité totale de revenus. Avant la première réforme de 1992, le mécanisme d'aides par les prix garantis a favorisé l'intensification des activités agricoles, en particulier l'élevage et les grandes cultures. Puisque tout ce qui est produit est vendu au prix garanti, ces activités intenses ont favorisé l'accroissement des inégalités. Après la réforme de Mac Sharry, les inégalités sont encore intenses. Les subventions à l'hectare se sont traduites par les fortes contributions des variables produits laitiers à l'accroissement du niveau d'inégalité totale des revenus d'activité. Les superficies (fourragères, totales) contribuent à la légère réduction de l'inégalité des revenus, grâce au système d'aide à l'hectare. Après la réforme de 2003, la coexistence de multicolinéarité, d'endogénéité et d'outliers a permis aux régressions Gini1'-PLS1 et Gini1'-PLS2 de prendre les premières places en terme de significativité des modèles et des parts de variances de l'axe t_1 expliquées par les différents régresseurs. Les résultats montrent que les produits laitiers accentuent les inégalités des rémunérations totales, et que les productions des céréales et des protéagineux accroissent les inégalités des revenus d'activité végétale.

Conclusion générale

Dans notre thèse, nous proposons de nouveaux modèles économétriques : les régressions "Gini-PLS". Notre conception consiste à combiner deux approches de régressions existantes (PLS et Gini), où chacune résorbe un problème bien déterminé, tel que, les valeurs aberrantes, la forte corrélation entre les régresseurs, les erreurs de mesure ou l'endogénéité, la faible taille de l'échantillon et les données manquantes. Ces difficultés sont souvent rencontrées en économétrie. Pour cette raison, les méthodes de détection s'avèrent importantes pour choisir le modèle approprié et donc aboutir à des estimations robustes.

En nous appuyant sur les propriétés des algorithmes PLS et Gini, nous avons pu construire quatre modèles de régressions Gini-PLS univariés (Gini-PLS1) et multivariés (Gini-PLS2) [20]. Ces régressions résultent de la création de nouveaux estimateurs qui tiennent compte simultanément des propriétés désirables des régressions PLS et Gini, en d'autres termes, s'affranchir des problèmes de multi-colinéarité et de valeurs aberrantes en laissant la base de données inchangée. Les modèles univariés consistent à régresser une variable dépendante sur une ou plusieurs variable(s) explicative(s), alors que les régressions multivariées s'intéressent à l'estimation d'un bloc de variables à expliquer en fonction d'une matrice de régresseurs. Nous distinguons les modèles Gini(1,1')-PLS fondés sur le concept rang des variables explicatives $(R(x_k))$ [21] ; et les modèles Gini2-PLS qui reprennent les mêmes étapes des régressions PLS(1 et 2) en employant des poids inspirés de la régression Gini

20. Nous avons aussi proposé deux régressions Gini1'-PLS (univariée et multivariée), dans le troisième chapitre, il s'agit d'une légère modification des algorithmes Gini1-PLS.

21. $R(x_k)$: est une matrice où la plus petite valeur de x_k prend la valeur 1 et la plus grande prend la valeur n (avec k est le nombre de variables explicatives et n la taille de l'échantillon.)

semi-paramétrique.

Les simulations Benchmark et de Monte Carlo que nous avons programmées sous Gauss, ont révélé la robustesse de nos modèles Gini-PLS, quelle que soit la forme de distribution des données. Nos modèles se distinguent des versions classiques PLS et Gini par leurs capacités à résoudre simultanément les difficultés omniprésentes dans la plupart des bases de données, à savoir, la multi-colinéarité, les outliers, les erreurs de mesure, l'endogénéité, la faible taille de l'échantillon et les données manquantes. Les simulations ont montré que les modèles Gini(1,1')-PLS sont capables de neutraliser les contaminations des données, notamment les valeurs aberrantes qui pourraient exister à la fois dans les matrices des variables dépendantes et explicatives ; et que les modèles de régressions Gini2-PLS sont robustes lorsque les outliers sont localisés dans la matrice des régresseurs. Ces régressions pourraient s'appliquer à plusieurs domaines.

Dans cette thèse, nous avons choisi l'exemple des inégalités des revenus des exploitations agricoles européennes. Le premier travail économétrique qui s'intéresse à l'étude des inégalités est la régression basée sur la décomposition en sources de revenus (RISD-MCO) de Morduch et Sicular (2002). Bien qu'il soit fondamental, nous pouvons lui reprocher l'usage des régressions MCO dans le cadre des inégalités des revenus, où on peut fréquemment rencontrer des difficultés liées à la présence de valeurs aberrantes, d'erreurs de mesures et de multi-colinéarité. Notre appui sur l'approche (RISD-MCO) nous a permis de l'étendre aux modèles PLS , Gini, et Gini-PLS.[22] Une fois les paramètres des régressions (MCO, PLS, Gini et Gini-PLS) estimés, nous les incorporons dans une mesure d'inégalité comme l'indice de Gini absolu[23].

Dans notre exemple d'application sur les données agrégées du Réseau d'Information Comptable Agricole européen (RICA : FADN) et dans le cadre de la Politique Agricole Commune (PAC), nous avons constaté la présence d'outliers et des fortes corrélations entre les variables explicatives. Lors de l'estimation des contributions des sources à l'inégalité totale des revenus agricoles euro-

22. Nous avons programmé les approches RISD-MCO, RISD-PLS, RISD-Gini et RISD-Gini-PLS sous GAUSS (versions 9.0 et 10).

23. Le Gini absolu est insensible au centrage des données.

péens, nous avons remarqué la concordance des simulations théoriques avec les résultats empiriques, pour les échantillons de faible ou de grande taille. En particulier, les régressions Gini1-PLS sont robustes en présence de fortes corrélations accompagnées d'outliers (au niveau des variables dépendantes et explicatives) et d'erreurs de mesures (ou endogénéité). Les régressions Gini2-PLS aboutissent à des résultats plus intéressants que les modèles PLS et Gini1-PLS en présence simultanée d'outliers dans la martice des régresseurs, d'endogénéité et de multi-colinéarité. Les tests d'hétéroscédaticité et d'auto-corrélation des régressions en question confirment aussi ces résultats.

Les objectifs de départ de la PAC étaient de garantir la sécurité alimentaire et d'améliorer les revenus des exploitants. Cependant, malgré les budgets importants et les réformes successives, l'inégalité des rémunérations agricoles entre les pays européens persiste, voire s'est légèrement accentuée. Ces inégalités se sont manifestées à travers les orientations technico-économiques des exploitations. Les résultats des estimations des inégalités des rémunérations agricoles montrent un meilleur ajustement des régressions Gini-PLS. Avant la première réforme, les systèmes d'aides aux prix garantis ont favorisé une surproduction considérable à la fin des années 1980, pour les produits laitiers et les cultures de plein champ (COP). Pour cette période, les modèles de régressions montrent que les productions laitières occupent la première place dans l'accroissement des inégalités de rémunérations liées à la production brute des exploitations. Les modèles multivariés montrent que les activités d'élevage (taille du cheptel et produits laitiers) et les grandes cultures, notamment les céréales, accroissent beaucoup les inégalités. Les modèles de régressions Gini2-PLS sont retenus pour cette période vu l'absence d'outliers et la présence de multi-colinéarité et d'endogénéité. Dans la période suivant la réforme de Mac Sharry, les statistiques T^2 de Hotelling indiquent la présence de valeurs aberrantes. En revenant aux données, ces outliers sont localisés au niveau des variables dépendantes. En plus de l'endogénéité et de la forte corrélation entre les variables explicatives, nous retenons les régressions Gini1'-PLS. Les variables liées à l'activité d'élevage accroissent les inégalités de plus de 50%. Nous constatons aussi que la politique d'aides directes à l'hectare a eu comme conséquence une légère réduction des inégalités. Après la réforme de la PAC

de 2003 (Accord de Luxembourg), les subventions via les aides découplées ont maintenu des niveaux d'inégalités élevés. Les activités d'élevage ont participé à l'accroissement des inégalités des revenus de plus de 50%. À travers les réformes successives, les instruments de la PAC ont favorisé l'intensification de l'élevage et des cultures COP. Ces activités sont fortement contributrices à l'accroissement des inégalités. Le système de subventions via les aides découplées a légèrement réduit les inégalités de rémunérations. Notons que les résultats des estimations sont robustes. Ils pourraient toutefois orienter les politiques économiques comme la PAC en vue de réduire les inégalités de revenus. Par exemple, une réduction des activités d'élevage intensif, un changement des orientations technico-économiques des exploitations. En effet, un système de redistribution des revenus aux petites exploitations pourrait être mis en place.

Les perspectives ouvertes par les régressions sont larges aussi bien sur le plan théorique que pratique. Les modèles économétriques Gini-PLS pourraient être élargis, de manière itérative, pour estimer les paramètres d'une régression (linéaire ou non linéaire) en maitrisant l'influence exercée par les outliers, la corrélation excessive des co-variables, l'endogénéité et les erreurs de mesures. Le changement d'espace engendré par la régression Gini-PLS oblige son utilisateur à utiliser les statistiques proches de celles des modèles d'analyse des données afin d'attester la qualité d'ajustement des données (statistiques VIP, redondances sur les variables explicatives, expliquées, etc.). L'algorithme multivarié est construit sur la même base, cependant il est valable uniquement lorsque plusieurs cibles (variables expliquées) sont envisagées (équivalent des modèles VAR). S'il n'y a pas de problème d'outliers et qu'il y a de l'endogénéité, et dans le cas où le vecteur rang $(R(x))$ n'est pas un instrument adéquat pour traiter l'endogénéité : il serait intéressant d'analyser les algorithmes de régression Gini-PLS-VI en essayant de prendre $R(z)$ au lieu de $(R(x))$ comme dans l'approche Gini-VI de Schechtman et Yitzhaki (2004).

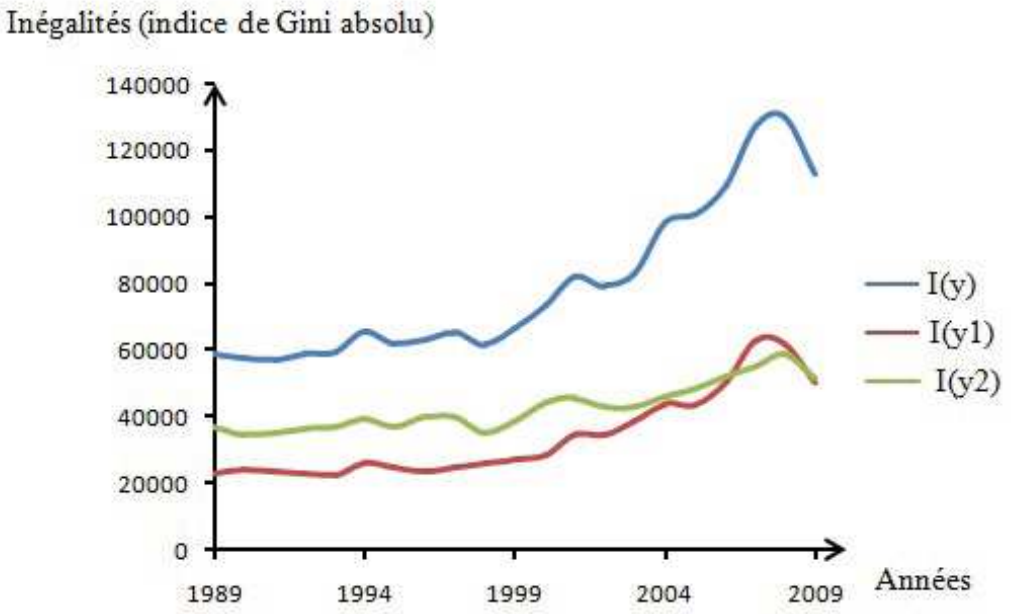
Il est à noter que les coefficients de pondération provenant des Gini2-PLS pourraient être construits avec la régression Gini paramétrique (à la place de la régression Gini semi-paramétrique). La régression Gini paramétrique est intéressante si la relation entre entre la variable dépendante et les variables explicatives est linéaire. Par conséquent, la régression Gini2-PLS1 pourrait

être élargie. Sur le plan pratique, plusieurs domaines pourraient bénéficier des résultats probants des régressions Gini-PLS. En particulier, dans le domaine de l'agronomie, le problème d'erreurs de mesures et d'observations aberrantes est souvent rencontré sur les parcelles expérimentales. Par conséquent, un modèle Gini-PLS pourrait donner des résultats intéressants. Dans le domaine de la finance, les valeurs aberrantes sont très fréquents. Ensuite, un algorithme de Gini-PLS pourrait être étudié pour la série temporelle tels que le rendement des actions. Dans le domaine de la mesure des inégalité, il serait possible de modéliser le revenu des personnes à l'aide des régressions Gini-PLS lorsque les revenus élevés (valeurs aberrantes) existent dans le dernier décile de la distribution. Ce qui aiderait à capturer les dimensions (variables) qui sont principalement responsables de l'inégalité du revenu total en évitant les biais provenant des corrélations entre les dimensions. Dans plusieurs domaines, la robustesse des estimations est très recherchée, de ce fait, les régressions Gini-PLS pourraient donner des résultats intéressants.

Annexes

Indices d'inégalités Gini absolu

Graphique1 : Indices d'inégalité : Gini absolu (Retour au texte3.2).



Matrice des corrélations des variables explicatives

x_k

Tableau 1 – Corrélations (1989)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,597	-0,294	-0,081	0,708	0,640	0,634	0,405	0,173	0,050	0,390
x_2		1	-0,351	-0,284	0,743	0,952	0,587	0,849	0,596	0,543	0,758
x_3			1	0,581	-0,584	-0,430	-0,578	-0,295	-0,192	-0,216	-0,191
x_4				1	-0,531	-0,292	-0,517	-0,304	-0,301	-0,239	-0,214
x_5					1	0,736	0,923	0,606	0,451	0,301	0,480
x_6						1	0,598	0,660	0,380	0,304	0,686
x_7							1	0,436	0,343	0,193	0,152
x_8								1	0,825	0,858	0,734
x_9									1	0,851	0,610
x_{10}										1	0,545
x_{11}											1

Tableau 2 – Corrélations (1990) Retour au texte 3.2.3

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,571	-0,381	-0,121	0,729	0,634	0,681	0,364	0,133	0,107	0,357
x_2		1	-0,316	-0,295	0,729	0,951	0,583	0,864	0,634	0,647	0,725
x_3			1	0,554	-0,566	-0,399	-0,566	-0,258	-0,202	-0,207	-0,155
x_4				1	-0,538	-0,301	-0,512	-0,301	-0,263	-0,258	-0,216
x_5					1	0,729	0,927	0,577	0,407	0,358	0,427
x_6						1	0,603	0,681	0,415	0,412	0,681
x_7							1	0,409	0,275	0,236	0,111
x_8								1	0,869	0,921	0,681
x_9									1	0,935	0,601
x_{10}										1	0,541
x_{11}											1

Tableau 3 – Corrélations (1991)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,531	-0,418	-0,125	0,719	0,603	0,689	0,326	0,228	0,178	0,420
x_2		1	-0,410	-0,253	0,731	0,952	0,581	0,857	0,749	0,733	0,861
x_3			1	0,604	-0,658	-0,486	-0,632	-0,312	-0,243	-0,238	-0,245
x_4				1	-0,529	-0,265	-0,505	-0,258	-0,224	-0,210	-0,168
x_5					1	0,734	0,926	0,567	0,480	0,413	0,522
x_6						1	0,606	0,670	0,546	0,512	0,808
x_7							1	0,391	0,323	0,277	0,255
x_8								1	0,927	0,961	0,779
x_9									1	0,933	0,758
x_{10}										1	0,665
x_{11}											1

Tableau 4 – Corrélations (1992)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,546	-0,388	-0,207	0,674	0,605	0,712	0,395	0,262	0,334	0,275
x_2		1	-0,261	-0,264	0,725	0,931	0,602	0,863	0,757	0,802	0,777
x_3			1	0,611	-0,402	-0,357	-0,500	-0,250	-0,160	-0,111	-0,084
x_4				1	-0,531	-0,304	-0,549	-0,255	-0,156	-0,107	-0,143
x_5					1	0,676	0,918	0,612	0,487	0,469	0,463
x_6						1	0,604	0,642	0,519	0,565	0,664
x_7							1	0,468	0,344	0,332	0,195
x_8								1	0,907	0,966	0,743
x_9									1	0,894	0,815
x_{10}										1	0,749
x_{11}											1

Tableau 5 – Corrélations (1993)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1 1	0,538	-0,378	0,109	0,724	0,608	0,696	0,329	0,288	0,322	0,459	
x_2		1	-0,219	0,555	0,694	0,929	0,575	0,829	0,793	0,848	0,822
x_3			1	0,174	-0,481	-0,311	-0,476	-0,236	-0,136	-0,041	-0,211
x_4				1	0,092	0,335	-0,051	0,668	0,742	0,770	0,646
x_5					1	0,659	0,922	0,601	0,502	0,457	0,573
x_6						1	0,578	0,585	0,557	0,616	0,642
x_7							1	0,450	0,313	0,322	0,337
x_8								1	0,917	0,943	0,895
x_9									1	0,922	0,858
x_{10}										1	0,908
x_{11}											1

Tableau 6 – Corrélations (1994)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,560	-0,342	0,163	0,707	0,632	0,682	0,358	0,259	0,283	0,454
x_2		1	-0,284	0,624	0,671	0,935	0,537	0,854	0,763	0,776	0,916
x_3			1	0,170	-0,567	-0,392	-0,578	-0,263	-0,097	-0,0153	-0,377
x_4				1	0,157	0,401	0,008	0,747	0,812	0,810	0,601
x_5					1	0,639	0,912	0,609	0,444	0,390	0,661
x_6						1	0,540	0,637	0,537	0,542	0,787
x_7							1	0,440	0,255	0,250	0,541
x_8								1	0,900	0,908	0,921
x_9									1	0,937	0,771
x_{10}										1	0,841
x_{11}											1

Tableau 7 – Corrélations (1995)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,486	-0,278	0,063	0,631	0,539	0,566	0,374	0,312	0,418	0,445
x_2		1	-0,205	0,574	0,657	0,929	0,565	0,865	0,781	0,737	0,652
x_3			1	0,095	-0,398	-0,246	-0,512	-0,226	-0,034	-0,148	-0,432
x_4				1	0,079	0,302	0,098	0,683	0,581	0,689	0,517
x_5					1	0,644	0,890	0,612	0,493	0,368	0,344
x_6						1	0,532	0,657	0,611	0,522	0,507
x_7							1	0,495	0,315	0,343	0,368
x_8								1	0,885	0,840	0,674
x_9									1	0,801	0,513
x_{10}										1	0,689
x_{11}											1

Tableau 8 – Corrélations (1996)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1 1	0,423	-0,320	-0,082	0,620	0,483	0,540	0,287	0,310	0,423	0,439	
x_2		1	-0,271	0,473	0,626	0,916	0,581	0,855	0,701	0,775	0,687
x_3			1	0,160	-0,414	-0,327	-0,521	-0,267	-0,030	-0,058	-0,422
x_4				1	-0,060	0,160	0,048	0,635	0,411	0,451	0,279
x_5					1	0,640	0,868	0,538	0,413	0,388	0,404
x_6						1	0,558	0,603	0,519	0,608	0,587
x_7							1	0,490	0,249	0,280	0,408
x_8								1	0,819	0,840	0,666
x_9									1	0,940	0,510
x_{10}										1	0,616
x_{11}											1

Tableau 9 – Corrélations (1997) Retour au texte 3.2.3

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,424	-0,254	-0,194	0,635	0,475	0,563	0,308	0,301	0,477	0,417
x_2		1	-0,232	0,351	0,636	0,921	0,599	0,839	0,685	0,752	0,751
x_3			1	0,318	-0,366	-0,305	-0,473	-0,211	-0,048	-0,036	-0,407
x_4				1	-0,153	0,077	-0,102	0,476	0,318	0,303	0,241
x_5					1	0,645	0,888	0,551	0,439	0,426	0,435
x_6						1	0,595	0,593	0,494	0,593	0,658
x_7							1	0,478	0,266	0,319	0,468
x_8								1	0,848	0,846	0,695
x_9									1	0,879	0,518
x_{10}										1	0,591
x_{11}											1

Tableau 10 – Corrélations (1998)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,466	-0,225	-0,210	0,630	0,528	0,579	0,325	0,268	0,464	0,406
x_2		1	-0,241	0,290	0,642	0,912	0,598	0,843	0,707	0,739	0,796
x_3			1	0,376	-0,371	-0,305	-0,488	-0,187	-0,013	-0,049	-0,403
x_4				1	-0,193	0,019	-0,160	0,399	0,333	0,253	0,209
x_5					1	0,656	0,892	0,546	0,396	0,426	0,420
x_6						1	0,587	0,590	0,490	0,572	0,684
x_7							1	0,476	0,252	0,351	0,454
x_8								1	0,886	0,877	0,719
x_9									1	0,871	0,547
x_{10}										1	0,622
x_{11}											1

Tableau 11 – Corrélations (1999)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,415	-0,167	-0,072	0,640	0,424	0,629	0,396	0,406	0,530	0,440
x_2		1	-0,328	0,504	0,627	0,902	0,605	0,856	0,771	0,740	0,802
x_3			1	0,065	-0,365	-0,389	-0,480	-0,242	-0,044	-0,115	-0,414
x_4				1	-0,027	0,213	0,031	0,545	0,430	0,396	0,406
x_5					1	0,613	0,890	0,585	0,407	0,468	0,468
x_6						1	0,575	0,597	0,603	0,533	0,674
x_7							1	0,503	0,278	0,416	0,479
x_8								1	0,858	0,905	0,743
x_9									1	0,875	0,587
x_{10}										1	0,641
x_{11}											1

Tableau 12 – Corrélations (2000)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,489	-0,162	-0,040	0,669	0,513	0,653	0,350	0,402	0,473	0,374
x_2		1	-0,343	0,508	0,664	0,892	0,608	0,854	0,808	0,680	0,786
x_3			1	0,026	-0,348	-0,380	-0,488	-0,260	-0,060	-0,086	-0,467
x_4				1	0,031	0,190	0,049	0,638	0,506	0,415	0,466
x_5					1	0,641	0,886	0,570	0,401	0,369	0,408
x_6						1	0,568	0,555	0,648	0,454	0,640
x_7							1	0,484	0,246	0,344	0,471
x_8								1	0,811	0,813	0,704
x_9									1	0,786	0,549
x_{10}										1 0,524	
x_{11}											1

Tableau 13 – Corrélations (2001)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,493	-0,141	0,147	0,702	0,511	0,696	0,350	0,646	0,499	0,363
x_2		1	-0,348	0,632	0,684	0,915	0,678	0,821	0,804	0,720	0,776
x_3			1	0,061	-0,380	-0,377	-0,522	-0,289	-0,110	-0,126	-0,524
x_4				1	0,139	0,446	0,159	0,578	0,500	0,436	0,426
x_5					1	0,670	0,904	0,597	0,613	0,423	0,498
x_6						1	0,652	0,554	0,706	0,534	0,665
x_7							1	0,553	0,578	0,431	0,603
x_8								1	0,719	0,831	0,665
x_9									1	0,731	0,439
x_{10}										1	0,558
x_{11}											1

Tableau 14 – Corrélations (2002)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,401	-0,135	0,002	0,636	0,484	0,679	0,238	0,509	0,385	0,250
x_2		1	-0,349	0,595	0,683	0,889	0,667	0,890	0,764	0,792	0,811
x_3			1	-0,048	-0,356	-0,361	-0,504	-0,288	-0,113	-0,186	-0,552
x_4				1	0,151	0,289	0,198	0,655	0,429	0,430	0,503
x_5					1	0,648	0,891	0,631	0,624	0,469	0,450
x_6						1	0,624	0,612	0,654	0,626	0,665
x_7							1	0,557	0,591	0,469	0,572
x_8								1	0,735	0,829	0,753
x_9									1	0,698	0,443
x_{10}										1	0,653
x_{11}											1

Tableau 15– Corrélations (2003)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,395	-0,084	0,040	0,648	0,467	0,698	0,277	0,543	0,407	0,258
x_2		1	-0,341	0,635	0,692	0,887	0,666	0,877	0,800	0,826	0,787
x_3			1	-0,065	-0,329	-0,368	-0,476	-0,257	-0,113	-0,191	-0,554
x_4				1	0,182	0,349	0,192	0,652	0,460	0,495	0,461
x_5					1	0,648	0,895	0,691	0,603	0,564	0,533
x_6						1	0,636	0,605	0,732	0,663	0,668
x_7							1	0,566	0,556	0,513	0,614
x_8								1	0,738	0,860	0,673
x_9									1	0,789	0,456
x_{10}										1	0,628
x_{11}											1

Tableau 16 – Corrélations (2004)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,955	0,456	0,859	0,685	0,910	0,736	0,914	0,778	0,953	0,548
x_2		1	0,372	0,880	0,766	0,976	0,809	0,959	0,874	0,954	0,708
x_3			1	0,318	0,135	0,271	0,082	0,426	0,268	0,466	-0,053
x_4				1	0,545	0,876	0,599	0,771	0,760	0,775	0,513
x_5					1	0,771	0,950	0,795	0,782	0,732	0,765
x_6						1	0,810	0,896	0,887	0,885	0,715
x_7							1	0,808	0,763	0,774	0,798
x_8								1	0,864	0,981	0,735
x_9									1	0,815	0,729
x_{10}										1	0,657
x_{11}											1

Tableau 17 – Corrélations (2005)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,957	0,574	0,631	0,651	0,912	0,746	0,899	0,644	0,943	0,768
x_2		1	0,495	0,744	0,734	0,975	0,814	0,956	0,768	0,955	0,877
x_3			1	0,239	0,213	0,389	0,220	0,519	0,282	0,597	0,237
x_4				1	0,469	0,736	0,553	0,681	0,616	0,626	0,704
x_5					1	0,743	0,946	0,784	0,722	0,702	0,804
x_6						1	0,814	0,891	0,794	0,890	0,870
x_7							1	0,827	0,713	0,783	0,868
x_8								1	0,784	0,971	0,881
x_9									1	0,742	0,765
x_{10}										1	0,828
x_{11}											1

Tableau 18 – Corrélations (2006)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,945	0,656	0,316	0,625	0,884	0,669	0,850	0,665	0,955	0,793
x_2		1	0,545	0,517	0,716	0,968	0,750	0,927	0,803	0,957	0,900
x_3			1	0,087	0,248	0,419	0,220	0,569	0,353	0,662	0,360
x_4				1	0,355	0,523	0,374	0,481	0,527	0,369	0,519
x_5					1	0,727	0,949	0,797	0,702	0,665	0,820
x_6						1	0,753	0,852	0,835	0,878	0,880
x_7							1	0,789	0,691	0,712	0,857
x_8								1	0,824	0,943	0,912
x_9									1	0,743	0,788
x_{10}										1	0,866
x_{11}											1

Tableau 19 – Corrélations (2007)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,944	0,761	0,223	0,633	0,891	0,714	0,893	0,797	0,959	0,864
x_2		1	0,689	0,438	0,721	0,974	0,788	0,960	0,911	0,964	0,946
x_3			1	0,194	0,364	0,601	0,385	0,673	0,554	0,760	0,557
x_4				1	0,325	0,452	0,338	0,448	0,537	0,298	0,442
x_5					1	0,734	0,954	0,778	0,694	0,672	0,799
x_6						1	0,788	0,897	0,935	0,897	0,928
x_7							1	0,817	0,723	0,750	0,858
x_8								1	0,872	0,959	0,943
x_9									1	0,821	0,862
x_{10}										1	0,903
x_{11}											1

Tableau 20 – Corrélations (2008) Retour au texte 3.2.3

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,945	0,676	0,393	0,595	0,893	0,701	0,903	0,677	0,967	0,888
x_2		1	0,616	0,524	0,687	0,973	0,779	0,958	0,808	0,965	0,951
x_3			1	0,415	0,262	0,531	0,284	0,611	0,426	0,676	0,517
x_4				1	0,229	0,515	0,292	0,461	0,479	0,428	0,462
x_5					1	0,708	0,944	0,733	0,668	0,624	0,755
x_6						1	0,787	0,894	0,836	0,895	0,921
x_7							1	0,793	0,686	0,718	0,837
x_8								1	0,806	0,969	0,946
x_9									1	0,723	0,768
x_{10}										1	0,923
x_{11}											1

Tableau 21 – Corrélations (2009)

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
x_1	1	0,943	0,723	0,111	0,576	0,888	0,696	0,840	0,358	0,957	0,912
x_2		1	0,670	0,281	0,659	0,973	0,780	0,912	0,530	0,958	0,965
x_3			1	0,245	0,261	0,589	0,311	0,586	0,171	0,704	0,607
x_4				1	0,040	0,290	0,087	0,187	0,293	0,145	0,196
x_5					1	0,684	0,940	0,765	0,566	0,607	0,706
x_6						1	0,769	0,849	0,606	0,887	0,935
x_7							1	0,805	0,517	0,720	0,803
x_8								1	0,573	0,938	0,906
x_9									1	0,421	0,491
x_{10}										1	0,934
x_{11}											1

Pourcentages des contributions des variables x_k à l'inégalité totale de y

Tableau 22 – Valeurs estimées des contributions ($\widehat{s_k}$) (%) (1989)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	50,8	50,8	0,07*	0,8*	11,2*
x_2	627	627	-0,00001*	0,01*	-17,7*
x_3	41,1	41,1	-0,000000006	0,0001	5,5
x_4	34,8	34,8	0,00000008***	0,0005*	8,2*
x_5	158	158	0,0001*	0,06*	43,2*
x_6	-405	-405	-0,00003*	0,01*	-3,7*
x_7	-150	-150	78*	77*	35,7*
x_8	-261	-261	7,8*	7,4*	4,2*
x_9	15,4	15	0,07*	0,5*	8,7*
x_{10}	13	13	0,2*	0,3*	0,9*
x_{11}	-24	-24	0,1*	0,07*	-1,7*
$\widehat{s^{\varepsilon}}$	0	0	13	12,9	5,3
Durbin-Watson	1,96	2,62	2,68	2,66	2,31
Test de White	6,55	1,97	3,37	3,40	5,62

Tableau 23 – Valeurs estimées des contributions ($\hat{s_k}$) (%) (1991)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	-15,5	-15,5	0,09*	1,1*	14,5*
x_2	-463	-463	-0,000005*	0,02*	-21,2*
x_3	-34	-34,6	-0,000000005*	0,0001	5,5
x_4	-16	-15,8	0,00000007	0,0006*	9,1*
x_5	-82,2	-82,2	0,0001***	0,06*	40,9*
x_6	241	241	-0,0000003*	0,01*	2,1*
x_7	262	262	82,1*	81*	35,4*
x_8	197	198	6,8*	6,2*	2,63*
x_9	0,02	0,03	0,04*	0,4*	8*
x_{10}	-16,8	-17	0,19*	0,6*	2,7
x_{11}	26,4	26	0,17*	0,13*	-1,18
$\widehat{s^\varepsilon}$	0	0	10,4	10,2	5,7

Tableau 24 – Valeurs estimées des contributions ($\hat{s_k}$) (%) (1993)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	-12,8	-12,8	0,14*	0,98*	17,9*
x_2	-499	-499	-0,00006*	0,01*	-12,8*
x_3	-39,23	-39,23	-0,00000002	0,0001	4,1*
x_4	0,23	0,26	0,000000002	-0,000003	-0,07*
x_5	-38,71	-39	0,0002*	0,05*	45,4*
x_6	278	279	-0,00006*	0,009*	-4,6*
x_7	200	201	82,8*	84,2*	36,1*
x_8	117	117	2,43*	0,4*	3*
x_9	58,19	58	0,01*	0,22*	1,7*
x_{10}	-17,63	-17,6	0,01*	0,17*	0,9*
x_{11}	52,79	52,8	2,98*	1,9*	1,4*
$\widehat{s^\varepsilon}$	0	0	11,51	11,8	6,7

Tableau 25 – Valeurs estimées des contributions ($\hat{s_k}$) (%) (1994)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	26,6	27	1,5*	0,65*	17,87*
x_2	231	231	-0,0009*	0,01*	-22,2*
x_3	18	18,3	-0,0000002	0,0001	6,5
x_4	-2,88	-2,9	0,0000007***	0,00004	0,11
x_5	46,74	46,7	0,002*	0,04*	47,7*
x_6	-150	-150,5	-0,0009*	0,007*	-3,8*
x_7	26,5	26,5	99*	88,3*	35,6*
x_8	-44,6	-44,6	3,7*	0,87*	3,9*
x_9	-13	-13,7	0,2*	0,1*	3,8*
x_{10}	21,2	21,3	0,38*	0,13*	1,9*
x_{11}	-59,4	-59,5	-48,2*	-2,5	-0,14*
$\widehat{s^\varepsilon}$	0	0	8,8	12,32	8,5

Tableau 26 – Valeurs estimées des contributions ($\hat{s_k}$) (%) (1995)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	25	26,8	0,24*	0,7*	17,3*
x_2	59,9	112,7	-0,00005*	0,01*	-21*
x_3	-0,98	-5,54	-0,00000002	0,00004***	2,17
x_4	-2,24	-2,73	-0,00000005	0,00007**	0,19
x_5	31,7	59,49	0,0004*	0,04*	48,56*
x_6	-53,3	-82,6	-0,00003*	0,008*	-2,08*
x_7	50,6	32,2	88,2*	87,4*	39,4*
x_8	-10,5	-27,6	5,19*	2,5*	4,3*
x_9	147	-2,28	0,01*	0,12*	1,9*
x_{10}	-6,86	-4,06	0,02*	0,21*	3,33*
x_{11}	-0,10	-7,66	-8,9*	-5,35*	1,3*
$\widehat{s^\varepsilon}$	1,17	1,2	15,2	14,34	5,3

Tableau 27 – Valeurs estimées des contributions ($\hat{s_k}$) (%)(1996)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	23,75	21,88	0,28*	0,7*	17,1*
x_2	85,07	105,6	-0,00006*	0,01*	-12*
x_3	-0,43	0,17	-0,00000002	0,00006***	2,3
x_4	-1,93	-2,03	-0,00000002	-0,00001***	-0,44**
x_5	40,2	49,4	0,0005*	0,05*	47,1*
x_6	-63,46	-77,58	-0,00005*	0,008*	-8,2*
x_7	39,33	35,9	81,39*	80,6*	37,2*
x_8	-15,96	-27,3	6,4*	2,9*	6,4*
x_9	16,91	11,4	0,02*	0,12*	4,3*
x_{10}	-12,68	-6,39	0,09*	0,25*	0,52*
x_{11}	-11,29	-11	-9,77*	-5,4*	-0,39*
$\widehat{s^\varepsilon}$	0,4	-0,15	21,4	20,63	6

Tableau 28 – Valeurs estimées des contributions ($\hat{s_k}$) (%) (1998)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	18,55	17,3	0,25*	0,67*	15,88*
x_2	65,34	105,4	-0,00006*	0,01*	-8,8*
x_3	-1,72	-1,7	-0,00000002	0,000029	0,34
x_4	160	2,26	0,00000001	0,000057***	1,15
x_5	40,28	48,9	0,0004*	0,04*	47,96*
x_6	-55,16	-81,59	-0,00005*	0,008*	-8,43*
x_7	53,93	50,1	90,88*	89,14*	40,48*
x_8	-29,4	-51,3	4,64*	2,82*	5,8*
x_9	14,94	15,15	0,01*	0,03*	0,14*
x_{10}	1,15	5,4	0,14*	0,33*	2,36*
x_{11}	-9,493071	-9	-11,66*	-7,9	-1,68*
$\widehat{s^\varepsilon}$	-0,14	-0,9	15,7	14,8	4,89

Tableau 29 – Valeurs estimées des contributions ($\hat{s_k}$) (%) (1999)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	14,45	15,69	0,26*	0,61*	16,66*
x_2	39,63	64,6	-0,00009*	0,01*	-10*
x_3	-0,87	-1,34	-0,00000001	0,000022	0,88
x_4	-1,3	-1,8	-0,00000005*	0,000007**	-0,49
x_5	34,96	39,7	0,0005*	0,04*	43,19*
x_6	-36,7	-47,5	-0,00006*	0,006*	-2,1*
x_7	60,3	53,6	92,4*	90,19*	39,4*
x_8	-7,23	-16,14	8,53*	3,52*	4,66*
x_9	17,2	13,2	1,54*	0,06*	0,4*
x_{10}	-10,8	-9,2	0,2*	0,47*	1,7*
x_{11}	-9,7	-10,4	-14,3*	-7,35*	-0,9*
$\widehat{s^{\varepsilon}}$	0,07	-3,9	12,9	12,42	7,2

Tableau 30 – Valeurs estimées des contributions ($\hat{s_k}$) (%)(2000)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	6,86	8,55	0,12*	0,6*	16,6*
x_2	15,7	-0,09054103	-0,00003*	0,01*	-15*
x_3	-1,3	-0,02647024	-0,00000001	0,00002***	0,59
x_4	-0,44	-0,00757590	-0,00000001	0,00003**	-0,08
x_5	45,6	0,28569747	0,0003*	0,04*	47,8*
x_6	-30	-0,15244988	-0,0003*	0,006*	-1,9*
x_7	64,9	0,76204090	92,9*	91,2*	40,3*
x_8	-99,5	0,11269872	4,4*	3,08*	4,4*
x_9	17,2	0,18696051	0,006*	0,05*	0,16*
x_{10}	-3,6	-0,09447082	0,04*	0,2*	2*
x_{11}	-5,4	-0,06659500	-12,14*	-9,1*	-1,3*
$\widehat{s^{\varepsilon}}$	0,5	0,00510552	14,15	1,3	6,3

Tableau 31 – Valeurs estimées des contributions ($\hat{s_k}$) (%)(2001)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	0,55	5,2	0,15*	0,44*	19,45*
x_2	14,17	-6,2	-0,0001*	0,005*	-21,8*
x_3	0,7	-1,3	-0,000000008*	0,000015*	1,37
x_4	0,8	-0,54	-0,0001*	0,000030*	-0,6**
x_5	64,7	44	0,0002*	0,02*	48,4*
x_6	-43,46	-25	-0,00005*	0,003*	-9,9*
x_7	49,19	57	0,97*	0,98*	43,9*
x_8	-17,9	3,3	0,34*	-0,64*	4,4*
x_9	34	3,5	0,01	0,09*	11,7*
x_{10}	-1,9	-10,6	-0,02	0,03*	-1*
x_{11}	-1,8	-1,7	-19,7*	-0,17*	-0,13*
$\widehat{s^{\varepsilon}}$	0,7	0,6	18,4	1,7	4,2

Tableau 32 – Valeurs estimées des contributions ($\hat{s_k}$) (%)(2002)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	4,95	25,3	0,08*	0,34*	16,14*
x_2	12,24	274,4	-0,00006*	0,006*	-22,8*
x_3	-0,81	1,18	-0,000000006	0,00001	0,9
x_4	-1,2	-14,19	-0,00000006**	0,00006***	0,08
x_5	40,5	14,17	0,0001*	0,02*	45,8*
x_6	-28,5	-130,4	-0,00003*	0,003*	-6,2*
x_7	53,45	41	99,2*	97,4*	42,5*
x_8	-11,7	-69	0,2*	-0,3*	4,5*
x_9	33,5	7	0,01*	0,1*	17,02*
x_{10}	-0,32	-18,9	-	0,07*	-0,96*
x_{11}	-0,1	-30,7	0,02*	-12,1*	-0,27*
$\widehat{s^{\varepsilon}}$	0,9	0,04	-14,3*	14,3	3,2

Tableau 33 – Valeurs estimées des contributions ($\hat{s_k}$) (%) (2003)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	38,47	28,64	0,11*	0,41*	17,24*
x_2	46,35	-218	-0,00008*	0,006*	-15,62*
x_3	-0,004	-11,7	-0,000000009	0,00001	47,35
x_4	-17,3	-6,55	-0,00000007**	0,00006**	-0,2***
x_5	53,07	-182,54	0,0001*	0,02*	48,8*
x_6	-46,6	113	-0,00003*	0,003*	-7,14*
x_7	53,15	193,5	0,97*	0,98*	41,7*
x_8	-25,9	260,2	4,6*	1,4*	4,44*
x_9	29	19,14	0,01*	0,12*	9,2*
x_{10}	-7,3	-94,3	-0,05*	0,07*	-3,33
x_{11}	-5,9	-2	-18,8*	-15,7*	-0,2*
$\widehat{s^\varepsilon}$	1,2	0,9	13,53	13,17	4,6

Tableau 34 – Valeurs estimées des contributions ($\hat{s_k}$) (%) (2004)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	3,8	28,6	-0,22*	0,14*	17,2*
x_2	46	-218,5	-0,0003*	0,01*	-15,6*
x_3	-0,004	-11,7	0,0000000019	0,000002	0,4
x_4	-1,7	-6,5	-0,000001*	0,0004*	-0,2*
x_5	53	-182,6	0,0002*	0,01*	48,8*
x_6	-46	113,3	-0,0001*	0,007*	7,1*
x_7	53,1	193,5	92,12*	96*	41,7*
x_8	-25	260,2	3,36*	-1,95*	4*
x_9	29	19,1	0,03*	0,11*	9,2*
x_{10}	7,3	-94,5	-0,16*	0,23*	-3,3*
x_{11}	-5	-2	-15,2*	-4,39*	-0,2*
$\widehat{s^\varepsilon}$	1	0,9	10	9	4,6

Tableau 35 – Valeurs estimées des contributions ($\widehat{s_k}$) (%)(2005)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	30	31	-0,4*	-0,1*	31*
x_2	-127	-138	-0,0002*	0,01*	-38*
x_3	1,4	6	0,000000002	0,000003	6,3
x_4	2,2	5,9	-0,0000002*	0,0001*	5,9*
x_5	3,8	-2,8	0,0001*	0,01*	-2,8*
x_6	56	18,5	-0,00008*	0,05*	18,5*
x_7	73,31	128	102	0,97*	28*
x_8	18,12	78,7	0,08*	-2,3*	78*
x_9	15,42	20,5	0,00001*	0,05*	20,5*
x_{10}	43,4	-41,9	-0,4*	-0,09*	-41,9
x_{11}	-6	-6,5	-14*	-8,8*	-6,5
$\widehat{s^{\varepsilon}}$	11	-0,08	7	7,8	-0,8

Tableau 36 – Valeurs estimées des contributions ($\widehat{s_k}$) (%)(2006)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	9,82	41,81	-0,51*	-0,18*	2,59
x_2	61,34	-289	-0,0002*	0,10*	47,44
x_3	1,42	8,4	0,000000001**	0,000003	0,33
x_4	0,48	11,7	-0,0000001*	0,000067**	1,63
x_5	69,67	-55,58	00,0001*	11*	23,07
x_6	-67,729550	95,82	-0,0007*	0,004*	-4,47
x_7	60,23	182	87*	88*	22,55
x_8	-35,33	108	-2,14*	-13*	10,41
x_9	14,51	0,04679673	-0,001*	0,35*	7,10
x_{10}	-07,92	0,01888180	-0,8*	-0,5*	-46,82
x_{11}	-6,893821	-0,08952520	-13,5*	-10	13,01
$\widehat{s^{\varepsilon}}$	0,3	-0,01370999	6,17	6,3	0,23

Tableau 37 – Valeurs estimées des contributions ($\hat{s_k}$) (%) (2007)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	18,613	18,91	0,05*	0,61*	-5,48*
x_2	77,62	-77,694850	-0,00004*	0,02*	52
x_3	3,47	7,16	0,000000001	0,00002***	0,37
x_4	-0,34	4,7	-0,00000001	0,0001**	0,69
x_5	46,96	23,88	0,00006*	0,023*	30,57*
x_6	-83,25	-18,51	-0,00001*	0,01*	-5,63
x_7	67,94	100,3	98,23*	96,71*	31,82
x_8	-21,1	12,21	2,38*	-0,40*	8,06
x_9	22,71	22,82	-0,00009*	0,11	-13,09
x_{10}	-30,36	9,83	-0,18*	0,68	-6,79
x_{11}	-2,21	-3,52	-3,37*	-1,12	6,7
$\widehat{s^\varepsilon}$	-0,01	-0,001	2,88	3,3	11,6

Tableau 38 – Valeurs estimées des contributions ($\hat{s_k}$) (%) (2009)

$\widehat{s_k}$	MCO	Gini	PLS1	Gini1'-PLS1	Gini2-PLS1
x_1	3,21	34,1	-0,05	0,47	3,24
x_2	90,5	-218	-0,0001	0,01	71,29
x_3	1,59	6,61	0,0000000004	0,00001	10,7
x_4	0,14	2,91	-0,000000008	0,000007	0,08
x_5	101	34,04	0,00008	0,02	35,98
x_6	-96,7	37,94	-0,00003	0,008	-25,32
x_7	37,76	103,6	103,2	97,3	35,84
x_8	-76,24	36,3	0,13	1,02	15,75
x_9	11,29	17,14	0,01	0,05	11,45
x_{10}	29,34	50,85	-0,44	0,06	-21,21
x_{11}	-2,94	-4,01	-8,8	-14,5	-44,37
$\widehat{s^\varepsilon}$	1,06	-1,5	6	6,04	16,1

Pourcentages des contributions des variables x_k aux inégalités totales de y_1 et y_2

Tableau 39 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (%) (1989)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	27,73	15,55	15,96	12,93	28,83	14,05
x_2	-12,20	-15,76	-4,91	-7,82	-12,05	-4,91
x_3	-8,3	-8,52	-0,93	-2,76	-8,83	-1,99
x_4	10,53	4,79	8,55	11,28	12,14	8,51
x_5	25,78	33,78	35,39	36,41	26,38	41,33
x_6	-11,65	-8,34	-8,13	-5,45	-11,66	-9,11
x_7	29,44	33,75	43,11	40,33	27,27	42,88
x_8	-8,02	2,90	0,18	1,24	-5,62	2,25
x_9	38,53	41,96	8,11	13,04	34,73	5,07
x_{10}	2,62	-8,37	0,26	-3,11	3,69	0,54
x_{11}	4,46	3,59	-3,82	-3,01	3,95	-3,66
$rsidu$	1,07	4,66	6,23	6,91	1,16	5,01
Durbin-Watson	0,019	0,019	2,485	1,469	2,80	2,02
Test de White	6,944	5,487	8,96	6,29	10,11	5

Tableau 40 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (%) (1991)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	27	18	16	14	32	18
x_2	-17	-25	-6	-11	-14	-2
x_3	-12	-9	-1	-4	-13	-4
x_4	11	0	9	9	13	8
x_5	31	45	40	45	24	39
x_6	-18	-18	-13	-10	-15	-9
x_7	34	38	46	45	30	45
x_8	-4	-12	2	-1	-3	4
x_9	30	36	7	12	28	3
x_{10}	6	4	2	0	8	3
x_{11}	10	16	-5	-4	6	-8
$rsidu$	4	6	4	3	5	4
Durbin-Watson	0,01	0,011	2,10	2,52	2,70	2,42
Test de White	13,2	4,4	11,9	4,3	16,9	5,12

Tableau 41 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (%) (1992)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	497	447	26,36	8,70	47,93	14,88
x_2	-17,48	-3,65	-17,4	-13,91	18,25	-3,04
x_3	-5,46	-2,69	-4,9	-1,45	-6,48	-3,28
x_4	9,4	6,52	-1,51	6,77	10,14	7,97
x_5	12,51	52,72	37,71	50,27	18,52	41,39
x_6	-18,7	-11,09	-10,56	-11,43	-18,49	-8,650849
x_7	38,258	45,88	42,11	48,64	38,35	40,38
x_8	-5,244	4,6	-1	-0,457	-10,29	4,78
x_9	22,20	4,50	16,20	9,12	20,96	2,78
x_{10}	12,12	12,15	3,31	-0,076	9,48	1,38
x_{11}	1,5	-3,7	-0,11	1,25	6,71	-4,85
$rsidu$	1,09	2,25	9,90	2,55	1,66	6,250289

Tableau 42 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 1993)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,46753584	0,26518597	0,08590477	0,07758148	0,44366652	0,11587022
x_2	-0,19033292	-0,16089172	-0,03773726	-0,08113372	-0,20300956	-0,02895787
x_3	-0,03128882	-0,05526077	-0,02510530	-0,00231542	-0,03175972	-0,04631160
x_4	-0,02667495	-0,03166164	0,00307754	0,00183550	-0,05427981	0,00553637
x_5	0,25590496	0,37832127	0,50145706	0,50297227	0,27145457	0,46439715
x_6	-0,21865187	-0,14722792	-0,11561845	-0,12980843	-0,21901444	-0,10266331
x_7	0,37120681	0,42668946	0,49515724	0,52993430	0,35904610	0,47981168
x_8	-0,04549967	-0,02826463	0,07568038	-0,01490186	-0,11373714	0,07139394
x_9	0,22738851	0,10449162	0,02409935	0,02258338	0,34867213	0,03243699
x_{10}	0,10357966	0,07468712	-0,00617112	0,01612720	0,10923255	0,00501653
x_{11}	0,05410341	0,09633684	-0,00787445	0,07057384	0,06543704	-0,00686081
$rsidu$	0,03272903	0,07759440	0,00713024	0,00655145	0,02429177	0,01033073

Tableau 43 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 1994)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,40266958	0,36093896	0,08646271	0,04087724	0,40783392	0,10270916
x_2	-0,08418610	-0,17039227	-0,05075928	-0,07187283	-0,08459365	-0,04829523
x_3	-0,02171340	-0,04380528	-0,01346446	-0,00634809	-0,02727172	-0,01963623
x_4	-0,05700557	-0,01947754	-0,00208581	-0,00589097	-0,05591511	-0,00647577
x_5	0,17988341	0,37324568	0,53509479	0,59500800	0,24441711	0,48814681
x_6	-0,11595553	-0,11869676	-0,11373991	-0,03637722	-0,11894735	-0,11360716
x_7	0,50667560	0,40920279	0,46509849	0,45402605	0,45148395	0,46712593
x_8	-0,02479040	-0,00765127	0,06500668	0,01779299	-0,00376979	0,07668566
x_9	0,35370105	0,29606733	0,01224653	0,08819098	0,33978645	0,01558115
x_{10}	0,20190759	0,03539415	-0,00362042	0,00998834	0,16827266	0,01219128
x_{11}	-0,36196211	-0,15367467	0,00700336	-0,10968453	-0,34163546	0,00771546
$rsidu$	0,02077589	0,03884888	0,01275733	0,02429001	0,02033899	0,01785893

Tableau 44 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 1995)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,29687298	0,22516768	0,06327808	0,06898593	0,28212657	0,05031847
x_2	-0,10242223	-0,11110865	-0,03692052	-0,05797516	-0,09474597	-0,03880223
x_3	-0,00796972	-0,01081024	0,00477753	0,00378887	-0,00840593	0,00787029
x_4	-0,00022544	0,00647977	-0,00128140	-0,00531552	-0,01411963	-0,00672058
x_5	0,31898222	0,41198561	0,50745564	0,49033167	0,32159844	0,51652910
x_6	-0,15742308	-0,14030666	-0,10286738	-0,09124594	-0,15571755	-0,10758056
x_7	0,38344769	0,40393424	0,50012858	0,50263521	0,36514260	0,47439631
x_8	0,04258097	0,02077578	0,07397368	0,04675934	0,07608129	0,09430954
x_9	0,19636336	0,10145649	0,02988551	0,04374022	0,17096688	0,02182505
x_{10}	0,06591931	0,07584134	-0,01221345	0,02187604	0,09619736	0,01434185
x_{11}	-0,06231305	-0,03586200	-0,04101130	-0,04546156	-0,06230353	-0,04417917
$rsidu$	0,02618699	0,05244663	0,01479502	0,02188091	0,02317947	0,01769192

Tableau 45– Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 1996)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,26346525	0,22298722	0,08973787	0,06170996	0,25154193	0,06910422
x_2	-0,09028917	-0,10767308	-0,02194026	-0,03211732	-0,07576016	-0,01943949
x_3	-0,01375647	-0,01442948	0,00286836	0,02342196	-0,01040265	-0,01655205
x_4	-0,00006381	-0,01519457	0,00017727	-0,00018012	-0,00750886	-0,00021011
x_5	0,29646237	0,28830350	0,50882008	0,47676044	0,21109398	0,56602546
x_6	-0,15804611	-0,15092681	-0,09407564	-0,06521628	-0,16541812	-0,08202582
x_7	0,37215773	0,39075662	0,47912065	0,47952782	0,43458419	0,43782553
x_8	0,07796686	0,08274404	0,08272719	0,05996402	0,08955557	0,10047032
x_9	0,22266054	0,30232046	0,02059117	0,05500329	0,24356408	0,00571646
x_{10}	0,07382165	0,03278799	-0,02924482	-0,01624043	0,09249804	-0,02039070
x_{11}	-0,10480887	-0,09305135	-0,06165301	-0,08536120	-0,12326505	-0,05756285
$rsidu$	0,06043003	0,06137546	0,02287115	0,04272787	0,05951706	0,01703904

Tableau 46 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 1998)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,30290408	0,25311707	0,05370274	0,03742975	0,32274707	0,04444320
x_2	-0,06380473	-0,04269517	-0,01135263	-0,00146473	-0,04109899	-0,01875406
x_3	0,00259097	0,00401504	0,01430671	0,01660420	0,00348729	0,00389059
x_4	0,00001229	0,00003853	-0,00304855	0,00448556	0,00000145	-0,00157291
x_5	0,26392914	0,28006177	0,46429773	0,42043805	0,30316667	0,51192230
x_6	-0,15640016	-0,17966135	-0,09339788	-0,09537944	-0,12150119	-0,09449440
x_7	0,39722121	0,39525417	0,55049274	0,55351647	0,35830966	0,51638259
x_8	0,05998059	0,12354379	0,07707597	0,10700067	0,05587147	0,09317197
x_9	0,19856479	0,10293560	0,02148691	-0,00665953	0,13250466	0,00334952
x_{10}	0,02810445	0,06158821	-0,03325746	0,01243072	0,01157065	-0,02626612
x_{11}	-0,08576856	-0,05125905	-0,05338317	-0,07584633	-0,08610298	-0,04108766
$rsidu$	0,05266593	0,05306140	0,01307691	0,02744461	0,06104424	0,00901496

Tableau 47 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 1999)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,24772989	0,37637350	0,05630825	0,06062255	0,25320853	0,05342624
x_2	-0,03853951	0,09527548	0,00065148	0,01328807	-0,00419135	-0,00481544
x_3	-0,00402058	-0,02303047	0,00070982	0,02117746	-0,01566432	0,00124089
x_4	0,00164125	-0,00395032	0,00132030	0,00016598	0,00441556	0,00095290
x_5	0,20067752	0,11696036	0,51511017	0,41455820	0,25841942	0,53102934
x_6	-0,16017302	-0,24856259	-0,08478170	-0,10771255	-0,10728886	-0,09515648
x_7	0,53642838	0,55798102	0,49856313	0,56329029	0,49442764	0,48261314
x_8	0,01783817	0,17267495	0,08322106	0,09514127	-0,00902754	0,09296144
x_9	0,37351362	0,32221090	-0,02433471	0,01660266	0,35522061	-0,02689993
x_{10}	-0,04954163	-0,33919669	-0,01972621	-0,03110148	-0,10557860	-0,01563795
x_{11}	-0,19981777	-0,07967501	-0,05088963	-0,08212812	-0,20955955	-0,04076783
$rsidu$	0,07426369	0,05293887	0,02384804	0,03609566	0,08561846	0,02105369

Tableau 48 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 2000)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,27589930	0,24444103	0,07177066	0,01738353	0,31904391	0,05295344
x_2	-0,11429069	-0,19489158	-0,01387946	-0,03202350	-0,12277846	-0,01151057
x_3	0,00004040	-0,00052141	0,00810909	0,01326356	-0,00020894	0,00451453
x_4	0,00588004	0,00873174	0,00621539	0,00435684	0,01277168	0,00457412
x_5	0,22574148	0,27242452	0,50233297	0,50444091	0,30533964	0,51787957
x_6	-0,22815917	-0,22952231	-0,09473878	-0,06656588	-0,22974368	-0,09483117
x_7	0,54936181	0,58780699	0,48971961	0,53299758	0,44973879	0,49719995
x_8	-0,04119686	-0,11706509	0,10594739	0,07435085	-0,02028317	0,10526274
x_9	0,47278187	0,52329574	-0,01155830	-0,00115301	0,43880149	-0,00474901
x_{10}	-0,06353719	-0,04192269	-0,02357467	0,00654299	-0,09561472	-0,02650342
x_{11}	-0,13747865	-0,09349188	-0,05393749	-0,07904436	-0,11625811	-0,05664927
$rsidu$	0,05495766	0,04071494	0,01359360	0,02545050	0,05919157	0,01185911

Tableau 49 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 2001)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,31074065	0,13460325	0,03710013	-0,00484343	0,36746777	0,06720981
x_2	-0,14913509	-0,14359209	-0,02777195	-0,04159050	-0,10036045	-0,00719370
x_3	0,00522290	-0,01375510	0,00153972	0,01028223	-0,00442419	-0,03829839
x_4	0,00843509	-0,02312773	-0,00219883	0,00059324	0,00590699	-0,00336964
x_5	0,16093968	0,12059033	0,59030954	0,59598952	0,26280200	0,59200284
x_6	-0,27491004	-0,26791169	-0,14131344	-0,12794003	-0,17259341	-0,09841374
x_7	0,48845447	0,61013718	0,43359226	0,48973558	0,37064194	0,42921160
x_8	0,01472379	-0,00443605	0,21181226	0,15049417	-0,01115841	0,15703653
x_9	0,65181616	0,84036934	-0,00334754	0,00643726	0,50612432	-0,03094154
x_{10}	-0,14835369	-0,22566770	-0,06310253	-0,02366385	-0,16866092	-0,08025079
x_{11}	-0,12209109	-0,05313836	-0,04539575	-0,07438567	-0,14620773	-0,00309528
$rsidu$	0,05415717	0,02592864	0,00877614	0,01889148	0,09046210	0,01610228

Tableau 50 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 2002)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,20032862	0,14001591	0,03949047	-0,02466259	0,21161561	0,04181644
x_2	-0,11900586	-0,09773796	-0,02184654	-0,05506582	-0,10185000	-0,03071756
x_3	-0,00077460	-0,00854716	0,01639573	0,01523065	0,00028745	0,01464301
x_4	-0,01101044	-0,02827790	0,01550286	0,00710028	-0,00759136	0,00729351
x_5	0,12237841	0,09184926	0,57505294	0,67678759	0,16253454	0,57516133
x_6	-0,18745787	-0,13203192	-0,13460752	-0,10192709	-0,15877325	-0,15688684
x_7	0,40706700	0,42658373	0,42709783	0,38081758	0,34762481	0,42765843
x_8	-0,03433980	0,01406696	0,14711234	0,13128979	-0,01086993	0,16529189
x_9	0,65443220	0,71435518	0,03650368	0,06051034	0,63596033	0,02794644
x_{10}	-0,03442156	-0,13740969	-0,07946850	-0,05552066	-0,07241218	-0,05828023
x_{11}	-0,06818636	-0,06147099	-0,03274247	-0,04922993	-0,09045783	-0,02317538
$rsidu$	0,07099025	0,07860459	0,01150917	0,01466986	0,08393180	0,00924895

Tableau 51 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 2003)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,22112288	0,07646198	0,05668635	-0,01765281	0,18049802	0,03129185
x_2	-0,08140674	-0,15658813	-0,01999595	-0,03569538	-0,06627253	-0,00770052
x_3	-0,00003160	0,00017030	0,01172525	0,02451684	-0,00006160	-0,00063308
x_4	0,01144979	-0,00154067	0,00831523	-0,00315785	0,00028504	0,00399683
x_5	0,18914262	0,25515781	0,54431817	0,61401325	0,26234837	0,53870420
x_6	-0,19628263	-0,17645735	-0,15054233	-0,11974170	-0,16850441	-0,15421336
x_7	0,46484097	0,50566148	0,44458804	0,42776839	0,40554588	0,47599406
x_8	0,00464059	-0,02186145	0,18560107	0,17490793	0,06582773	0,18368600
x_9	0,51704054	0,67446522	-0,01580596	0,03612610	0,49023581	0,01368235
x_{10}	-0,13681438	-0,19880665	-0,06073956	-0,04232106	-0,14414357	-0,09045004
x_{11}	-0,08459045	-0,03111124	-0,01382375	-0,07367154	-0,11776369	0,00114634
$rsidu$	0,09088844	0,07444872	0,00967345	0,01490783	0,09200495	0,00449537

Tableau 52 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 2004)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,13159484	0,20559721	-0,04562162	-0,07957671	0,12124977	-0,03113127
x_2	-0,04552073	-0,12213822	-0,06283320	-0,09409690	-0,04499382	-0,05493060
x_3	0,05391658	0,10518481	-0,00248467	-0,00329697	0,03927684	-0,00180613
x_4	-0,09938989	-0,04324606	0,03464637	0,08488919	-0,08673636	0,00302257
x_5	0,36431179	0,48079945	0,83681404	0,71856260	0,44359684	0,81153445
x_6	-0,10826357	-0,11786720	-0,12876649	-0,31794044	-0,16697530	-0,09454359
x_7	0,55951515	0,43347206	0,48873998	0,57061547	0,44632034	0,48707356
x_8	0,00741307	-0,06479013	0,03429809	0,09204556	0,11074476	0,02228665
x_9	0,22852796	0,28349524	-0,07881099	0,08317586	0,19529643	-0,06011908
x_{10}	0,05350436	-0,15069022	-0,07818977	-0,10010637	0,07592303	-0,06710439
x_{11}	-0,22899761	-0,06171177	-0,01530232	0,02095700	-0,21930368	-0,03766196
$rsidu$	0,08338805	0,05189483	0,01751059	0,02477169	0,08560117	0,02337979

Tableau 53 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 2005)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,13249395	0,16975448	-0,04751851	-0,13891375	0,11853517	-0,05818799
x_2	-0,07402030	-0,13220941	-0,07452629	-0,01536959	-0,08186597	-0,08776331
x_3	0,08221480	0,09983647	0,00352164	0,00370418	0,07699482	0,00482859
x_4	0,02869621	-0,01787785	0,03509674	0,00227270	0,03846501	0,03584943
x_5	0,44600131	0,42805224	0,75135240	0,64986348	0,47143952	0,73813155
x_6	-0,17866322	-0,12409344	-0,16471165	-0,20115770	-0,21196010	-0,16453736
x_7	0,52606846	0,39110102	0,51709837	0,47268124	0,46220985	0,51645089
x_8	-0,04219892	-0,02476276	0,08744324	0,17210850	0,03490707	0,06669093
x_9	0,28174423	0,32351777	-0,04934772	-0,00067125	0,27911432	-0,02544556
x_{10}	-0,03929725	-0,16945520	-0,06101243	-0,04385330	-0,03096797	-0,05731700
x_{11}	-0,22999513	-0,02114070	-0,01181097	0,07956883	-0,22339733	0,01414926
$rsidu$	0,06695587	0,07727738	0,01441519	0,01976667	0,06652561	0,01715057

Tableau 54 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 2006)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,12535323	0,23622923	-0,05629861	-0,05120030	0,18860193	-0,05446380
x_2	-0,07518463	-0,19938372	-0,07149727	-0,05130682	-0,03740362	-0,08396148
x_3	0,12451887	0,11901445	0,00794672	0,00540672	0,10315918	0,01340411
x_4	-0,01529352	-0,01875860	0,01785955	0,00237267	-0,00852361	0,01838983
x_5	0,42970818	0,45540785	0,67631270	0,67018600	0,46166143	0,62933080
x_6	-0,08681638	-0,09641739	-0,14320367	-0,20119293	-0,12197547	-0,13992601
x_7	0,74163900	0,63746704	0,48440823	0,43243433	0,73734907	0,54518258
x_8	-0,20619768	-0,18281824	0,15682683	0,17197328	-0,17710114	0,08511600
x_9	0,18695442	0,24662803	-0,07532658	-0,00006057	0,15611249	-0,02091332
x_{10}	-0,08937661	-0,17873773	-0,04639208	-0,06596866	-0,10645621	-0,06444428
x_{11}	-0,18346556	-0,06660068	0,04722441	0,08321290	-0,25089912	0,06462733
$rsidu$	0,04816067	0,04796976	0,00213976	0,00414338	0,05547506	0,00765823

Tableau 55 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 2007)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,27957688	0,11261316	0,00882228	-0,04979476	0,23714068	-0,01175758
x_2	-0,06928892	-0,07596325	-0,07720134	-0,02192154	-0,08598783	-0,09679487
x_3	0,11676235	0,15707057	0,01496435	0,02162953	0,09807846	0,01535823
x_4	-0,00334031	-0,03286385	0,01028873	-0,00836515	-0,01672178	-0,00015002
x_5	0,32011936	0,43612539	0,65934811	0,72916330	0,30782783	0,64651894
x_6	-0,25426629	-0,09459909	-0,14826217	-0,15517663	-0,35249585	-0,19824411
x_7	0,62473546	0,37303576	0,59020301	0,44512270	0,56911945	0,57758649
x_8	-0,01705429	0,01360874	0,04721006	0,10652568	0,15903407	0,10176247
x_9	0,30910422	0,21275194	0,00036883	-0,01028193	0,36264505	0,04858567
x_{10}	-0,12287189	-0,16234841	-0,06674104	-0,12102630	-0,09188222	-0,06485516
x_{11}	-0,21288452	0,01508258	-0,03901373	0,07198725	-0,19947102	-0,01679854
$rsidu$	0,02940795	0,04548648	0,00001289	-0,00786215	0,01271317	-0,00121151

Tableau 56 – Valeurs estimées des contributions des régresseurs à l'inégalité totale (comparées à 1) (année 2009)

	PLS2		Gini1'-PLS2		Gini2-PLS2	
	y_1	y_2	y_1	y_2	y_1	y_2
x_1	0,15531209	0,19685104	-0,02069631	0,07141189	0,08582102	-0,04550672
x_2	-0,07720959	-0,12506567	-0,07434672	-0,09602635	-0,08847804	-0,08851580
x_3	0,13818505	0,11926772	0,00912777	0,00539449	0,13255938	0,01394004
x_4	-0,01649665	-0,01745986	0,00192485	0,00125899	-0,01472903	0,00069272
x_5	0,45932467	0,45779785	0,72096706	0,66215324	0,48921445	0,72639657
x_6	-0,14880408	-0,21089449	-0,12546476	-0,24048754	-0,16825222	-0,13479027
x_7	0,54096757	0,43636541	0,51751703	0,50315752	0,41777136	0,49116280
x_8	-0,07379901	0,02642371	0,08423137	0,05691330	0,05013925	0,11430520
x_9	0,09301524	0,15119861	-0,04302705	0,07938543	0,07909872	-0,03764027
x_{10}	-0,02517745	-0,07023670	-0,05241525	-0,09146684	-0,01867614	-0,05205053
x_{11}	-0,10150968	-0,01464040	-0,02964609	0,02584983	-0,02198688	0,00275441
$rsidu$	0,05619185	0,05039279	0,01182810	0,02245604	0,05751813	0,00925187

Statistiques VIP (Variable Importance in the Projection

Tableau 57 – VIP (1989)

	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,048	1,002	0,001	0,885	0,821	0,741
x_2	0,001	0,923	0,047	0,690	0,715	0,728
x_3	0,001	0,774	3,203	0,568	0,564	0,616
x_4	0,001	0,586	0,854	0,697	0,463	0,794
x_5	0,002	1,558	0,058	1,225	1,267	1,192
x_6	0,001	0,912	0,062	0,589	0,709	0,644
x_7	3,300	1,604	0,001	1,230	1,296	1,142
x_8	0,320	0,940	0,001	0,697	0,752	0,720
x_9	0,025	1,152	0,003	0,747	0,971	0,724
x_10	0,037	0,398	0,001	0,458	0,312	0,468
x_11	0,027	0,029	0,001	0,395	0,047	0,474

Tableau 58 – VIP (1990) Retour au texte 3.2.3

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,054	1,125	0,001	0,948	0,904	0,788
x_2	0,001	0,920	0,052	0,705	0,707	0,730
x_3	0,001	0,713	3,163	0,590	0,525	0,612
x_4	0,001	0,619	0,991	0,692	0,488	0,798
x_5	0,002	1,504	0,061	1,236	1,202	1,203
x_6	0,001	0,910	0,069	0,621	0,701	0,653
x_7	3,301	1,571	0,001	1,239	1,245	1,139
x_8	0,315	0,795	0,001	0,667	0,620	0,698
x_9	0,020	1,141	0,003	0,607	0,929	0,626
x_{10}	0,041	0,706	0,001	0,502	0,557	0,529
x_{11}	0,030	0,084	0,001	0,335	0,091	0,421

Tableau 59 – VIP (1991)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,050	1,159	0,002	0,922	0,894	0,764
x_2	0,001	0,867	0,046	0,652	0,641	0,706
x_3	0,001	0,745	3,255	0,638	0,539	0,660
x_4	0,001	0,426	0,629	0,657	0,325	0,747
x_5	0,002	1,489	0,057	1,195	1,142	1,158
x_6	0,001	0,905	0,063	0,572	0,671	0,634
x_7	3,302	1,558	0,001	1,202	1,190	1,090
x_8	0,301	0,765	0,001	0,616	0,568	0,662
x_9	0,020	1,247	0,003	0,612	0,963	0,635
x_{10}	0,040	0,711	0,001	0,504	0,535	0,526
x_{11}	0,043	0,246	0,001	0,421	0,199	0,492

Tableau 60 – VIP (1992)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,051	1,178	0,002	1,023	0,937	0,789
x_2	0,001	0,820	0,060	0,617	0,613	0,735
x_3	0,001	0,739	3,145	0,493	0,530	0,503
x_4	0,001	0,649	1,044	0,015	0,498	0,735
x_5	0,002	1,487	0,074	1,260	1,170	1,145
x_6	0,001	0,879	0,086	0,521	0,662	0,632
x_7	3,302	1,580	0,001	1,272	1,239	1,102
x_8	0,299	0,922	0,001	0,612	0,708	0,710
x_9	0,020	1,017	0,003	0,521	0,815	0,619
x_{10}	0,017	0,744	0,002	0,454	0,579	0,556
x_{11}	0,039	0,238	0,001	0,551	0,203	0,465

Tableau 61 – VIP (1993)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,051	1,215	0,003	1,023	0,933	0,870
x_2	0,001	0,771	0,070	0,617	0,565	0,706
x_3	0,001	0,676	3,301	0,493	0,484	0,465
x_4	0,001	0,196	0,274	0,015	0,170	0,011
x_5	0,003	1,541	0,095	1,260	1,181	1,141
x_6	0,001	0,868	0,109	0,521	0,640	0,635
x_7	3,304	1,620	0,001	1,272	1,239	1,078
x_8	0,235	0,888	0,001	0,612	0,659	0,657
x_9	0,013	1,034	0,006	0,521	0,795	0,564
x_{10}	0,016	0,697	0,003	0,454	0,518	0,504
x_{11}	0,158	0,581	0,001	0,551	0,430	0,614

Tableau 62 – VIP (1994)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,047	1,206	0,002	0,996	0,897	0,821
x_2	0,001	0,755	0,039	0,603	0,533	0,700
x_3	0,001	0,776	3,308	0,561	0,546	0,485
x_4	0,001	0,287	0,223	0,082	0,194	0,132
x_5	0,002	1,609	0,061	1,250	1,190	1,127
x_6	0,001	0,848	0,062	0,520	0,608	0,630
x_7	3,301	1,602	0,001	1,274	1,185	1,061
x_8	0,248	0,881	0,001	0,604	0,626	0,653
x_9	0,010	0,807	0,004	0,470	0,602	0,506
x_{10}	0,016	0,678	0,001	0,401	0,481	0,463
x_{11}	0,195	0,726	0,001	0,575	0,510	0,676

Tableau 63 – VIP (1995)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,050	1,164	0,005	0,942	0,876	0,818
x_2	0,001	0,845	0,123	0,658	0,621	0,705
x_3	0,001	0,283	3,258	0,407	0,180	0,486
x_4	0,001	0,266	0,544	0,060	0,168	0,160
x_5	0,003	1,687	0,177	1,322	1,289	1,129
x_6	0,001	0,951	0,205	0,577	0,706	0,610
x_7	3,297	1,626	0,001	1,290	1,233	1,080
x_8	0,317	0,907	0,001	0,698	0,662	0,736
x_9	0,011	0,855	0,013	0,600	0,625	0,562
x_{10}	0,020	0,744	0,004	0,461	0,529	0,551
x_{11}	0,155	0,531	0,001	0,280	0,357	0,501

Tableau 64 – VIP (1996)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,046	1,164	0,003	0,944	0,890	0,816
x_2	0,001	0,872	0,081	0,645	0,637	0,707
x_3	0,001	0,442	3,308	0,430	0,267	0,344
x_4	0,001	0,086	0,127	0,006	0,102	0,110
x_5	0,003	1,691	0,118	1,300	1,304	1,091
x_6	0,001	0,947	0,134	0,540	0,691	0,584
x_7	3,296	1,604	0,001	1,253	1,221	0,995
x_8	0,313	0,951	0,001	0,686	0,707	0,738
x_9	0,010	0,831	0,008	0,587	0,632	0,630
x_{10}	0,023	0,697	0,002	0,507	0,506	0,602
x_{11}	0,182	0,526	0,001	0,328	0,355	0,544

Tableau 65 – VIP (1997) Retour au texte 3.2.3

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,046	1,157	0,005	0,947	0,901	0,797
x_2	0,001	0,868	0,122	0,662	0,650	0,724
x_3	0,001	0,255	3,091	0,342	0,137	0,280
x_4	0,001	0,374	1,162	0,163	0,320	0,001
x_5	0,003	1,705	0,181	1,309	1,345	1,107
x_6	0,001	1,006	0,215	0,563	0,756	0,584
x_7	3,299	1,619	0,001	1,268	1,262	1,035
x_8	0,283	0,970	0,001	0,687	0,737	0,776
x_9	0,012	0,580	0,007	0,580	0,419	0,650
x_{10}	0,031	0,782	0,003	0,550	0,575	0,629
x_{11}	0,187	0,523	0,001	0,345	0,360	0,550

Tableau 66 – VIP (1998)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,045	1,181	0,004	0,968	0,920	0,861
x_2	0,001	0,824	0,091	0,674	0,629	0,746
x_3	0,001	0,316	3,182	0,365	0,223	0,351
x_4	0,001	0,387	0,908	0,256	0,342	0,167
x_5	0,003	1,707	0,135	1,300	1,372	1,143
x_6	0,001	0,928	0,158	0,590	0,713	0,658
x_7	3,302	1,696	0,001	1,284	1,355	1,058
x_8	0,248	1,031	0,001	0,674	0,798	0,725
x_9	0,008	0,302	0,004	0,528	0,190	0,585
x_{10}	0,033	0,782	0,002	0,557	0,578	0,596
x_{11}	0,183	0,509	0,001	0,367	0,366	0,553

Tableau 67 – VIP (1999)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,045	1,127	0,004	1,005	0,896	0,907
x_2	0,001	0,838	0,090	0,667	0,622	0,748
x_3	0,001	0,308	3,309	0,337	0,190	0,258
x_4	0,001	0,039	0,070	0,041	0,018	0,109
x_5	0,002	1,639	0,129	1,291	1,318	1,128
x_6	0,001	0,846	0,145	0,540	0,632	0,590
x_7	3,297	1,655	0,001	1,282	1,311	1,072
x_8	0,274	1,026	0,001	0,724	0,781	0,785
x_9	0,007	0,618	0,008	0,538	0,446	0,598
x_{10}	0,037	0,993	0,003	0,639	0,747	0,656
x_{11}	0,224	0,608	0,001	0,444	0,434	0,601

Tableau 68 – VIP (2000)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,045	1,182	0,004	1,032	0,959	0,873
x_2	0,001	0,843	0,098	0,709	0,614	0,787
x_3	0,001	0,285	3,277	0,315	0,142	0,283
x_4	0,001	0,233	0,452	0,039	0,154	0,201
x_5	0,002	1,691	0,139	1,317	1,349	1,129
x_6	0,001	0,915	0,170	0,568	0,684	0,613
x_7	3,299	1,710	0,001	1,284	1,330	1,059
x_8	0,271	0,948	0,001	0,708	0,713	0,790
x_9	0,007	0,577	0,008	0,509	0,436	0,593
x_{10}	0,025	0,806	0,003	0,541	0,594	0,574
x_{11}	0,199	0,507	0,001	0,361	0,342	0,543

Tableau 69 – VIP (2001)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,051	1,247	0,003	0,985	0,896	0,812
x_2	0,001	0,792	0,066	0,616	0,580	0,707
x_3	0,001	0,375	3,306	0,360	0,277	0,391
x_4	0,001	0,205	0,218	0,067	0,154	0,190
x_5	0,002	1,565	0,098	1,197	1,159	1,059
x_6	0,001	0,801	0,108	0,514	0,583	0,590
x_7	3,298	1,595	0,001	1,162	1,158	1,017
x_8	0,264	0,879	0,001	0,653	0,647	0,720
x_9	0,014	1,248	0,011	0,815	0,881	0,714
x_{10}	0,028	0,483	0,001	0,443	0,341	0,506
x_{11}	0,223	0,659	0,001	0,397	0,481	0,575

Tableau 70 – VIP (2002)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,045	1,070	0,004	0,930	0,789	0,826
x_2	0,001	0,846	0,084	0,673	0,635	0,737
x_3	0,001	0,335	3,275	0,343	0,236	0,376
x_4	0,001	0,347	0,480	0,171	0,252	0,238
x_5	0,002	1,570	0,123	1,195	1,207	1,089
x_6	0,001	0,848	0,143	0,553	0,635	0,593
x_7	3,295	1,578	0,001	1,190	1,180	1,057
x_8	0,271	0,943	0,001	0,677	0,718	0,758
x_9	0,016	1,193	0,012	0,922	0,878	0,820
x_{10}	0,032	0,571	0,002	0,498	0,416	0,600
x_{11}	0,256	0,770	0,001	0,435	0,576	0,591

Tableau 71 – VIP (2003)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,046	1,140	0,005	0,964	0,844	0,845
x_2	0,001	0,781	0,109	0,679	0,606	0,714
x_3	0,001	0,256	3,259	0,311	0,196	0,306
x_4	0,001	0,300	0,546	0,175	0,230	0,242
x_5	0,002	1,578	0,176	1,201	1,221	1,123
x_6	0,001	0,822	0,194	0,558	0,629	0,540
x_7	3,291	1,591	0,001	1,194	1,198	1,101
x_8	0,305	0,982	0,001	0,733	0,760	0,802
x_9	0,015	1,231	0,016	0,786	0,905	0,702
x_{10}	0,042	0,548	0,002	0,568	0,413	0,632
x_{11}	0,270	0,673	0,001	0,483	0,517	0,617

Tableau 72 – VIP (2004)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,231	0,865	0,001	0,753	0,673	0,629
x_2	0,004	0,989	0,071	0,800	0,798	0,773
x_3	0,001	0,205	3,271	0,243	0,171	0,122
x_4	0,001	0,516	0,498	0,539	0,390	0,468
x_5	0,002	1,378	0,160	1,100	1,153	1,185
x_6	0,002	0,959	0,147	0,763	0,775	0,746
x_7	3,188	1,347	0,001	1,077	1,115	1,146
x_8	0,732	1,032	0,001	0,867	0,841	0,848
x_9	0,018	1,183	0,016	0,826	0,978	0,812
x_{10}	0,197	0,881	0,001	0,807	0,702	0,736
x_{11}	0,458	1,083	0,001	0,774	0,880	0,981

Tableau 73 – VIP (2005)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,199	0,847	0,001	0,740	0,711	0,640
x_2	0,003	0,985	0,079	0,787	0,794	0,768
x_3	0,001	0,193	3,125	0,324	0,170	0,232
x_4	0,001	0,580	1,079	0,491	0,436	0,512
x_5	0,002	1,376	0,186	1,106	1,153	1,176
x_6	0,001	0,959	0,165	0,753	0,779	0,737
x_7	3,180	1,357	0,001	1,080	1,127	1,130
x_8	0,612	1,043	0,001	0,869	0,850	0,877
x_9	0,013	1,079	0,020	0,784	0,895	0,797
x_{10}	0,166	0,856	0,001	0,788	0,681	0,741
x_{11}	0,667	1,139	0,001	0,840	0,930	0,940

Tableau 74 – VIP (2006)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,159	0,883	0,002	0,697	0,707	0,618
x_2	0,003	0,975	0,086	0,744	0,765	0,730
x_3	0,001	0,193	3,063	0,345	0,179	0,274
x_4	0,001	0,465	1,240	0,314	0,344	0,369
x_5	0,002	1,418	0,199	1,141	1,135	1,161
x_6	0,001	0,941	0,177	0,713	0,741	0,706
x_7	3,167	1,417	0,001	1,129	1,126	1,153
x_8	0,603	1,051	0,001	0,867	0,833	0,877
x_9	0,010	0,969	0,022	0,725	0,763	0,758
x_{10}	0,189	0,871	0,001	0,728	0,681	0,692
x_{11}	0,738	1,145	0,000	0,870	0,904	0,954

Tableau 75 – VIP (2007)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,165	0,828	0,001	0,792	0,695	0,641
x_2	0,003	1,029	0,042	0,835	0,860	0,792
x_3	0,001	0,447	3,258	0,519	0,382	0,409
x_4	0,001	0,452	0,609	0,342	0,361	0,446
x_5	0,002	1,374	0,089	1,091	1,164	1,226
x_6	0,001	0,962	0,083	0,806	0,806	0,774
x_7	3,001	1,358	0,001	1,092	1,143	1,215
x_8	1,040	1,044	0,001	0,906	0,878	0,914
x_9	0,015	1,030	0,008	0,782	0,857	0,756
x_{10}	0,259	0,847	0,001	0,810	0,703	0,747
x_{11}	0,905	1,151	0,001	0,901	0,968	0,970

Tableau 76 – VIP (2008) Retour au texte 3.2.3

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,162	0,920	0,001	0,802	0,768	0,642
x_2	0,003	1,004	0,064	0,838	0,841	0,799
x_3	0,001	0,306	3,211	0,438	0,268	0,329
x_4	0,001	0,313	0,804	0,293	0,247	0,278
x_5	0,002	1,392	0,144	1,090	1,170	1,264
x_6	0,001	1,008	0,137	0,808	0,841	0,787
x_7	3,002	1,378	0,001	1,093	1,151	1,229
x_8	0,963	1,031	0,001	0,900	0,867	0,898
x_9	0,011	0,991	0,016	0,754	0,828	0,783
x_{10}	0,368	0,862	0,001	0,815	0,721	0,718
x_{11}	0,949	1,146	0,001	0,897	0,959	0,955

Tableau 77 – VIP (2009)

VIP	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
x_1	0,141	0,971	0,002	0,741	0,771	0,620
x_2	0,003	0,980	0,073	0,759	0,770	0,741
x_3	0,001	0,285	3,286	0,391	0,228	0,308
x_4	0,001	0,122	0,378	0,003	0,087	0,140
x_5	0,002	1,445	0,174	1,174	1,151	1,236
x_6	0,001	0,992	0,155	0,738	0,780	0,742
x_7	3,086	1,420	0,001	1,152	1,125	1,199
x_8	0,550	1,029	0,001	0,895	0,813	0,922
x_9	0,008	0,965	0,019	0,575	0,764	0,661
x_{10}	0,242	0,782	0,001	0,744	0,609	0,685
x_{11}	1,047	1,146	0,001	0,820	0,902	0,835

Redondances

Tableau 78 –Rd(y,t_1) et Rd(x_k,t_1) (1989)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,836	0,851	0,672
x_1	0,4046	0,421	0,442
x_2	0,3566	0,423	0,774
x_3	0,3345	0,335	0,366
x_4	0,2691	0,271	0,299
x_5	0,8574	0,885	0,865
x_6	0,3654	0,408	0,68
x_7	0,9997	0,989	0,7
x_8	0,2021	0,274	0,657
x_9	0,1258	0,175	0,434
x_{10}	0,0425	0,077	0,306
x_{11}	0,0267	0,051	0,381

Tableau 79 –Rd(y,t_1) et Rd(x_k,t_1) (1990)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,847	0,861	0,654
x_1	0,465	0,478	0,447
x_2	0,352	0,421	0,787
x_3	0,321	0,321	0,343
x_4	0,264	0,267	0,293
x_5	0,867	0,893	0,831
x_6	0,373	0,419	0,69
x_7	0,999	0,989	0,664
x_8	0,179	0,249	0,674
x_9	0,083	0,129	0,45
x_{10}	0,062	0,106	0,426
x_{11}	0,015	0,033	0,349

Tableau 80–Rd(y,t_1) et Rd(x_k,t_1) (1991)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,805	0,819	0,598
x_1	0,476	0,488	0,454
x_2	0,349	0,419	0,821
x_3	0,401	0,404	0,402
x_4	0,255	0,255	0,227
x_5	0,864	0,892	0,83
x_6	0,376	0,425	0,729
x_7	0,999	0,989	0,646
x_8	0,163	0,23	0,668
x_9	0,112	0,167	0,551
x_{10}	0,084	0,135	0,493
x_{11}	0,071	0,11	0,535

Tableau 81–Rd(y,t_1) et Rd(x_k,t_1) (1992)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,867	0,615	0,867
x_1	0,509	0,501	0,509
x_2	0,374	0,8	0,374
x_3	0,25	0,276	0,25
x_4	0,301	0,269	0,3
x_5	0,847	0,798	0,847
x_6	0,372	0,679	0,372
x_7	0,999	0,694	0,999
x_8	0,23	0,688	0,23
x_9	0,126	0,516	0,126
x_{10}	0,119	0,522	0,119
x_{11}	0,042	0,413	0,042

Tableau 82 –Rd(y,t_1) et Rd(x_k,t_1) (1992)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,839	0,844	0,577
x_1	0,487	0,503	0,533
x_2	0,341	0,416	0,803
x_3	0,227	0,227	0,233
x_4	0,001	0,001	0,144
x_5	0,855	0,888	0,816
x_6	0,34	0,387	0,673
x_7	0,999	0,988	0,649
x_8	0,212	0,292	0,667
x_9	0,105	0,164	0,552
x_{10}	0,111	0,175	0,555
x_{11}	0,122	0,188	0,634

Tableau 83 –Rd(y,t_1) et Rd(x_k,t_1) (1993)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,846	0,839	0,554
x_1	0,467	0,481	0,512
x_2	0,3	0,386	0,818
x_3	0,335	0,332	0,268
x_4	0,001	0,011	0,232
x_5	0,838	0,87	0,785
x_6	0,3	0,359	0,692
x_7	0,999	0,985	0,624
x_8	0,206	0,294	0,709
x_9	0,072	0,126	0,51
x_{10}	0,069	0,126	0,497
x_{11}	0,306	0,401	0,797

Tableau 84 –Rd(y,t_1) et Rd(x_k,t_1) (1994)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,801	0,792	0,559
x_1	0,324	0,353	0,469
x_2	0,334	0,415	0,831
x_3	0,263	0,268	0,154
x_4	0,012	0,034	0,196
x_5	0,796	0,812	0,728
x_6	0,293	0,344	0,683
x_7	0,999	0,983	0,614
x_8	0,262	0,352	0,747
x_9	0,109	0,167	0,57
x_{10}	0,128	0,195	0,546
x_{11}	0,147	0,216	0,448

Tableau 85 –Rd(y,t_1) et Rd(x_k,t_1) (1995)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,777	0,772	0,587
x_1	0,294	0,314	0,454
x_2	0,353	0,436	0,791
x_3	0,273	0,279	0,212
x_4	0,003	0,014	0,055
x_5	0,757	0,771	0,733
x_6	0,321	0,372	0,686
x_7	0,999	0,983	0,632
x_8	0,257	0,347	0,652
x_9	0,07	0,119	0,478
x_{10}	0,088	0,146	0,547
x_{11}	0,175	0,246	0,485

Tableau 86 –Rd(y,t_1) et Rd(x_k,t_1) (1996)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,809	0,804	0,644
x_1	0,337	0,355	0,505
x_2	0,37	0,448	0,764
x_3	0,239	0,238	0,181
x_4	0,023	0,011	0,001
x_5	0,797	0,808	0,781
x_6	0,353	0,401	0,692
x_7	0,999	0,987	0,71
x_8	0,238	0,319	0,597
x_9	0,069	0,115	0,388
x_{10}	0,131	0,191	0,512
x_{11}	0,218	0,286	0,513

Tableau 87 –Rd(y,t_1) et Rd(x_k,t_1) (1997)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,809	0,804	0,644
x_1	0,337	0,355	0,505
x_2	0,37	0,448	0,764
x_3	0,239	0,238	0,181
x_4	0,023	0,011	0,001
x_5	0,797	0,808	0,781
x_6	0,353	0,401	0,692
x_7	0,999	0,987	0,71
x_8	0,238	0,319	0,597
x_9	0,069	0,115	0,388
x_{10}	0,131	0,191	0,512
x_{11}	0,218	0,286	0,513

Tableau 88 –Rd(y,t_1) et Rd(x_k,t_1) (1998)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,837	0,832	0,63
x_1	0,398	0,421	0,492
x_2	0,381	0,465	0,793
x_3	0,232	0,235	0,166
x_4	0,001	0,01	0,072
x_5	0,795	0,808	0,723
x_6	0,34	0,389	0,634
x_7	0,999	0,984	0,671
x_8	0,267	0,358	0,722
x_9	0,085	0,139	0,53
x_{10}	0,184	0,258	0,631
x_{11}	0,244	0,323	0,579

Tableau 89 –Rd(y,t_1) et Rd(x_k,t_1) (1999)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,826	0,817	0,612
x_1	0,428	0,441	0,522
x_2	0,385	0,46	0,836
x_3	0,241	0,246	0,17
x_4	0,003	0,014	0,105
x_5	0,787	0,796	0,735
x_6	0,331	0,371	0,657
x_7	0,999	0,988	0,677
x_8	0,249	0,325	0,665
x_9	0,067	0,108	0,504
x_{10}	0,126	0,176	0,488
x_{11}	0,235	0,301	0,509

Tableau 90 –Rd(y,t_1) et Rd(x_k,t_1) (2000)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,826	0,817	0,612
x_1	0,428	0,441	0,522
x_2	0,385	0,46	0,836
x_3	0,241	0,246	0,17
x_4	0,003	0,014	0,105
x_5	0,787	0,796	0,735
x_6	0,331	0,371	0,657
x_7	0,999	0,988	0,677
x_8	0,249	0,325	0,665
x_9	0,067	0,108	0,504
x_{10}	0,126	0,176	0,488
x_{11}	0,235	0,301	0,509

Tableau 91 –Rd(y,t_1) et Rd(x_k,t_1) (2001)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,749	0,72	0,563
x_1	0,482	0,481	0,557
x_2	0,473	0,544	0,828
x_3	0,275	0,281	0,175
x_4	0,028	0,049	0,184
x_5	0,818	0,817	0,763
x_6	0,432	0,47	0,697
x_7	0,999	0,988	0,767
x_8	0,319	0,395	0,622
x_9	0,34	0,382	0,704
x_{10}	0,195	0,252	0,514
x_{11}	0,378	0,451	0,519

Tableau 92–Rd(y,t_1) et Rd(x_k,t_1) (2002)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,798	0,763	0,599
x_1	0,455	0,43	0,418
x_2	0,462	0,555	0,847
x_3	0,258	0,271	0,178
x_4	0,044	0,076	0,188
x_5	0,796	0,799	0,75
x_6	0,399	0,451	0,687
x_7	0,999	0,982	0,755
x_8	0,327	0,423	0,69
x_9	0,357	0,402	0,669
x_{10}	0,232	0,305	0,574
x_{11}	0,345	0,441	0,557

Tableau 93 –Rd(y,t_1) et Rd(x_k,t_1) (2003)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,8	0,772	0,595
x_1	0,481	0,457	0,448
x_2	0,462	0,546	0,834
x_3	0,23	0,238	0,146
x_4	0,042	0,071	0,202
x_5	0,805	0,818	0,772
x_6	0,414	0,46	0,696
x_7	0,999	0,985	0,751
x_8	0,337	0,423	0,688
x_9	0,317	0,367	0,7
x_{10}	0,277	0,349	0,651
x_{11}	0,395	0,472	0,527

Tableau 94 –Rd(y,t_1) et Rd(x_k,t_1) (2004)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,901	0,873	0,714
x_1	0,573	0,646	0,848
x_2	0,69	0,773	0,958
x_3	0,01	0,023	0,102
x_4	0,384	0,447	0,682
x_5	0,903	0,886	0,75
x_6	0,685	0,751	0,921
x_7	0,997	0,973	0,792
x_8	0,694	0,787	0,933
x_9	0,612	0,677	0,831
x_{10}	0,638	0,728	0,892
x_{11}	0,659	0,709	0,617

Tableau 95 –Rd(y,t_1) et Rd(x_k,t_1) (2005)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,893	0,859	0,709
x_1	0,586	0,668	0,828
x_2	0,698	0,787	0,949
x_3	0,057	0,087	0,188
x_4	0,33	0,387	0,521
x_5	0,891	0,861	0,722
x_6	0,692	0,765	0,913
x_7	0,997	0,973	0,814
x_8	0,721	0,811	0,934
x_9	0,531	0,579	0,703
x_{10}	0,648	0,741	0,89
x_{11}	0,787	0,853	0,856

Tableau 96 –Rd(y,t_1) et Rd(x_k,t_1) (2006)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,884	0,831	0,614
x_1	0,489	0,594	0,801
x_2	0,613	0,727	0,937
x_3	0,063	0,11	0,245
x_4	0,155	0,184	0,258
x_5	0,902	0,878	0,723
x_6	0,609	0,7	0,891
x_7	0,996	0,961	0,758
x_8	0,675	0,786	0,921
x_9	0,512	0,582	0,73
x_{10}	0,558	0,68	0,874
x_{11}	0,781	0,868	0,901

Tableau 97 –Rd(y,t_1) et Rd(x_k,t_1) (2007)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,909	0,871	0,719
x_1	0,596	0,691	0,841
x_2	0,718	0,815	0,964
x_3	0,202	0,268	0,43
x_4	0,136	0,153	0,201
x_5	0,893	0,845	0,686
x_6	0,702	0,78	0,924
x_7	0,987	0,943	0,766
x_8	0,766	0,857	0,944
x_9	0,601	0,675	0,836
x_{10}	0,661	0,763	0,895
x_{11}	0,822	0,897	0,933

Tableau 98 –Rd(y,t_1) et Rd(x_k,t_1) (2008)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,895	0,854	0,708
x_1	0,592	0,703	0,849
x_2	0,714	0,821	0,961
x_3	0,128	0,192	0,338
x_4	0,114	0,147	0,232
x_5	0,866	0,804	0,645
x_6	0,707	0,788	0,917
x_7	0,984	0,931	0,754
x_8	0,74	0,845	0,941
x_9	0,534	0,586	0,705
x_{10}	0,631	0,754	0,896
x_{11}	0,801	0,889	0,934

Tableau 99 –Rd(y,t_1) et Rd(x_k,t_1) (2009)

Rd	PLS1	Gini1'-PLS1	Gini2-PLS1
y	0,844	0,79	0,622
x_1	0,568	0,678	0,819
x_2	0,667	0,78	0,937
x_3	0,139	0,208	0,359
x_4	0,011	0,017	0,046
x_5	0,862	0,805	0,656
x_6	0,669	0,758	0,908
x_7	0,989	0,943	0,766
x_8	0,724	0,817	0,9
x_9	0,281	0,292	0,387
x_{10}	0,609	0,731	0,87
x_{11}	0,736	0,84	0,923

Tableau 100 –Rd(y,t_1) et Rd(x_k,t_1) (1989)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,523	0,532	0,523
y_2	0,645	0,661	0,645
x_1	0,453	0,462	0,453
x_2	0,772	0,771	0,772
x_3	0,359	0,352	0,359
x_4	0,304	0,28	0,304
x_5	0,869	0,875	0,869
x_6	0,677	0,675	0,677
x_7	0,684	0,711	0,684
x_8	0,66	0,652	0,66
x_9	0,432	0,431	0,432
x_{10}	0,307	0,298	0,307
x_{11}	0,407	0,375	0,407

Tableau 101 –Rd(y,t_1) et Rd(x_k,t_1) (1990)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,504	0,511	0,489
y_2	0,658	0,647	0,631
x_1	0,491	0,476	0,457
x_2	0,771	0,782	0,781
x_3	0,356	0,335	0,355
x_4	0,306	0,277	0,321
x_5	0,866	0,849	0,847
x_6	0,7	0,691	0,701
x_7	0,693	0,685	0,664
x_8	0,624	0,657	0,648
x_9	0,391	0,43	0,416
x_{10}	0,367	0,404	0,393
x_{11}	0,346	0,339	0,368

Tableau 102 –Rd(y,t_1) et Rd(x_k,t_1) (1991)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,435	0,439	0,416
y_2	0,647	0,623	0,617
x_1	0,484	0,48	0,45
x_2	0,78	0,808	0,798
x_3	0,444	0,404	0,439
x_4	0,271	0,223	0,279
x_5	0,87	0,848	0,85
x_6	0,718	0,724	0,727
x_7	0,692	0,671	0,661
x_8	0,604	0,645	0,632
x_9	0,482	0,527	0,509
x_{10}	0,427	0,469	0,454
x_{11}	0,496	0,517	0,521

Tableau 103 –Rd(y,t_1) et Rd(x_k,t_1) (1992)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,484	0,472	0,446
y_2	0,663	0,637	0,617
x_1	0,529	0,519	0,487
x_2	0,776	0,8	0,811
x_3	0,268	0,254	0,255
x_4	0,286	0,254	0,279
x_5	0,835	0,814	0,808
x_6	0,671	0,674	0,689
x_7	0,732	0,707	0,683
x_8	0,646	0,682	0,689
x_9	0,471	0,509	0,517
x_{10}	0,48	0,519	0,524
x_{11}	0,387	0,408	0,437

Tableau 104 –Rd(y,t_1) et Rd(x_k,t_1) (1993)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,405	0,393	0,365
y_2	0,666	0,651	0,611
x_1	0,577	0,552	0,537
x_2	0,778	0,793	0,822
x_3	0,223	0,219	0,202
x_4	0,128	0,134	0,163
x_5	0,845	0,835	0,806
x_6	0,66	0,666	0,682
x_7	0,687	0,673	0,633
x_8	0,629	0,649	0,678
x_9	0,508	0,532	0,561
x_{10}	0,519	0,538	0,578
x_{11}	0,606	0,617	0,656

Tableau 105 –Rd(y,t_1) et Rd(x_k,t_1) (1994)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,422	0,403	0,375
y_2	0,644	0,607	0,578
x_1	0,558	0,531	0,515
x_2	0,771	0,801	0,825
x_3	0,292	0,27	0,259
x_4	0,177	0,215	0,227
x_5	0,834	0,805	0,786
x_6	0,674	0,681	0,701
x_7	0,691	0,65	0,626
x_8	0,644	0,688	0,706
x_9	0,436	0,487	0,5
x_{10}	0,426	0,475	0,494
x_{11}	0,75	0,779	0,804

Tableau 106 –Rd(y,t_1) et Rd(x_k,t_1) (1995)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,432	0,418	0,375
y_2	0,646	0,609	0,579
x_1	0,504	0,491	0,477
x_2	0,768	0,804	0,805
x_3	0,194	0,16	0,196
x_4	0,144	0,175	0,186
x_5	0,788	0,759	0,727
x_6	0,649	0,667	0,659
x_7	0,686	0,649	0,632
x_8	0,686	0,719	0,735
x_9	0,499	0,537	0,536
x_{10}	0,482	0,518	0,54
x_{11}	0,416	0,43	0,48

Tableau 107 –Rd(y,t_1) et Rd(x_k,t_1) (1996)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,44	0,435	0,389
y_2	0,621	0,607	0,538
x_1	0,472	0,472	0,441
x_2	0,769	0,769	0,812
x_3	0,206	0,206	0,197
x_4	0,047	0,047	0,081
x_5	0,761	0,761	0,691
x_6	0,672	0,672	0,673
x_7	0,658	0,658	0,598
x_8	0,628	0,628	0,699
x_9	0,456	0,456	0,509
x_{10}	0,522	0,522	0,587
x_{11}	0,466	0,466	0,522

Tableau 108 –Rd(y,t_1) et Rd(x_k,t_1) (1997)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,425	0,43	0,377
y_2	0,672	0,628	0,609
x_1	0,511	0,519	0,478
x_2	0,752	0,739	0,805
x_3	0,184	0,182	0,17
x_4	0,001	0,002	0,001
x_5	0,784	0,805	0,73
x_6	0,672	0,678	0,696
x_7	0,711	0,735	0,654
x_8	0,603	0,571	0,662
x_9	0,4	0,362	0,454
x_{10}	0,516	0,486	0,567
x_{11}	0,506	0,49	0,562

Tableau 109 –Rd(y,t_1) et Rd(x_k,t_1) (1998)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,425	0,43	0,377
y_2	0,672	0,692	0,609
x_1	0,511	0,519	0,478
x_2	0,752	0,739	0,805
x_3	0,184	0,182	0,17
x_4	0,001	0,002	0,001
x_5	0,784	0,805	0,73
x_6	0,672	0,678	0,696
x_7	0,711	0,735	0,654
x_8	0,603	0,57	0,662
x_9	0,401	0,362	0,454
x_{10}	0,516	0,486	0,567
x_{11}	0,506	0,49	0,562

Tableau 110 –Rd(y,t_1) et Rd(x_k,t_1) (1999)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,465	0,467	0,421
y_2	0,644	0,648	0,583
x_1	0,511	0,522	0,485
x_2	0,769	0,759	0,811
x_3	0,163	0,178	0,161
x_4	0,06	0,055	0,087
x_5	0,753	0,755	0,699
x_6	0,622	0,619	0,64
x_7	0,7	0,706	0,646
x_8	0,694	0,682	0,738
x_9	0,499	0,49	0,547
x_{10}	0,606	0,594	0,641
x_{11}	0,559	0,56	0,607

Tableau 111 –Rd(y,t_1) et Rd(x_k,t_1) (2000)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,466	0,471	0,411
y_2	0,629	0,627	0,554
x_1	0,551	0,546	0,498
x_2	0,804	0,811	0,862
x_3	0,181	0,162	0,171
x_4	0,083	0,092	0,128
x_5	0,765	0,765	0,698
x_6	0,644	0,643	0,662
x_7	0,709	0,705	0,64
x_8	0,629	0,639	0,699
x_9	0,468	0,475	0,536
x_{10}	0,458	0,466	0,507
x_{11}	0,489	0,485	0,549

Tableau 112 –Rd(y,t_1) et Rd(x_k,t_1) (2001)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,358	0,348	0,302
y_2	0,658	0,646	0,612
x_1	0,574	0,562	0,524
x_2	0,799	0,817	0,845
x_3	0,188	0,18	0,195
x_4	0,159	0,176	0,195
x_5	0,787	0,778	0,745
x_6	0,674	0,688	0,698
x_7	0,792	0,783	0,758
x_8	0,608	0,615	0,651
x_9	0,677	0,686	0,677
x_{10}	0,502	0,501	0,53
x_{11}	0,51	0,52	0,56

Tableau 113 –Rd(y,t_1) et Rd(x_k,t_1) (2002)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,406	0,385	0,359
y_2	0,679	0,668	0,642
x_1	0,448	0,422	0,407
x_2	0,818	0,84	0,854
x_3	0,183	0,18	0,191
x_4	0,165	0,184	0,192
x_5	0,771	0,763	0,738
x_6	0,671	0,679	0,684
x_7	0,779	0,767	0,747
x_8	0,663	0,685	0,704
x_9	0,66	0,658	0,654
x_{10}	0,558	0,563	0,592
x_{11}	0,533	0,555	0,579

Tableau 114 –Rd(y,t_1) et Rd(x_k,t_1) (2003)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,381	0,381	0,347
y_2	0,693	0,689	0,672
x_1	0,457	0,447	0,421
x_2	0,818	0,828	0,845
x_3	0,162	0,152	0,167
x_4	0,188	0,198	0,213
x_5	0,789	0,787	0,768
x_6	0,683	0,689	0,687
x_7	0,775	0,767	0,752
x_8	0,678	0,686	0,709
x_9	0,664	0,676	0,663
x_{10}	0,64	0,638	0,663
x_{11}	0,537	0,535	0,569

Tableau 115–Rd(y,t_1) et Rd(x_k,t_1) (2004)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,701	0,706	0,704
y_2	0,524	0,544	0,551
x_1	0,836	0,814	0,808
x_2	0,949	0,937	0,933
x_3	0,101	0,088	0,082
x_4	0,663	0,644	0,64
x_5	0,771	0,789	0,793
x_6	0,91	0,904	0,902
x_7	0,811	0,826	0,834
x_8	0,93	0,922	0,919
x_9	0,825	0,832	0,825
x_{10}	0,886	0,87	0,866
x_{11}	0,628	0,651	0,662

Tableau 116 –Rd(y,t_1) et Rd(x_k,t_1) (2005)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,684	0,687	0,687
y_2	0,512	0,527	0,527
x_1	0,82	0,803	0,805
x_2	0,941	0,932	0,933
x_3	0,196	0,174	0,18
x_4	0,515	0,512	0,516
x_5	0,738	0,753	0,751
x_6	0,902	0,901	0,899
x_7	0,825	0,838	0,837
x_8	0,934	0,926	0,929
x_9	0,696	0,707	0,701
x_{10}	0,887	0,872	0,875
x_{11}	0,855	0,865	0,865

Tableau 117–Rd(y,t_1) et Rd(x_k,t_1) (2006)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,626	0,623	0,626
y_2	0,536	0,541	0,545
x_1	0,782	0,771	0,768
x_2	0,921	0,918	0,915
x_3	0,246	0,227	0,232
x_4	0,254	0,263	0,264
x_5	0,75	0,756	0,758
x_6	0,872	0,876	0,871
x_7	0,779	0,787	0,787
x_8	0,923	0,917	0,92
x_9	0,726	0,731	0,731
x_{10}	0,862	0,851	0,851
x_{11}	0,902	0,907	0,907

Tableau 118–Rd(y,t_1) et Rd(x_k,t_1) (2007)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,791	0,793	0,797
y_2	0,527	0,543	0,553
x_1	0,834	0,818	0,81
x_2	0,956	0,949	0,944
x_3	0,436	0,412	0,409
x_4	0,201	0,207	0,212
x_5	0,704	0,72	0,728
x_6	0,913	0,913	0,907
x_7	0,781	0,795	0,802
x_8	0,944	0,94	0,94
x_9	0,822	0,826	0,821
x_{10}	0,891	0,876	0,872
x_{11}	0,931	0,934	0,933

Tableau 119 –Rd(y,t_1) et Rd(x_k,t_1) (année 2008)

Variables	PLS2	Gini1'-PLS2	Gini2-PLS2
$y1$	0,807	0,810	0,815
$y2$	0,477	0,497	0,511
x_1	0,838	0,823	0,813
x_2	0,952	0,945	0,938
x_3	0,346	0,319	0,318
x_4	0,238	0,229	0,229
x_5	0,664	0,685	0,697
x_6	0,909	0,908	0,902
x_7	0,768	0,787	0,796
x_8	0,939	0,933	0,930
x_9	0,703	0,710	0,707
x_{10}	0,887	0,872	0,865
x_{11}	0,931	0,931	0,929

Tableau 120 –Rd(y,t_1) et Rd(x_k,t_1) (2009)

Rd	PLS2	Gini1'-PLS2	Gini2-PLS2
y_1	0,673	0,667	0,674
y_2	0,46	0,462	0,473
x_1	0,807	0,789	0,784
x_2	0,918	0,914	0,91
x_3	0,356	0,334	0,337
x_4	0,037	0,043	0,045
x_5	0,693	0,7	0,708
x_6	0,885	0,891	0,884
x_7	0,796	0,801	0,807
x_8	0,905	0,898	0,902
x_9	0,367	0,397	0,389
x_{10}	0,861	0,844	0,843
x_{11}	0,914	0,909	0,906

Exogénéité des variables explicatives x_k

Tableau 121 – Endogénéité des variables explicatives

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_10	x_11
1989	−	−	−	−	−	−	+	+	−	−	+
1990	−	−	−	−	−	−	+	+	−	−	+
1991	−	−	−	−	−	−	+	+	−	−	+
1992	−	−	−	−	−	−	+	+	−	−	+
1993	−	−	−	−	−	−	+	+	−	−	+
1994	−	−	−	−	−	−	+	+	−	−	+
1995	−	−	−	−	−	−	+	+	−	+	+
1996	−	−	−	−	−	−	+	+	−	−	+
1997	−	−	−	−	−	−	+	+	−	−	+
1998	−	−	−	−	−	−	+	+	−	−	+
1999	−	−	−	−	−	−	+	+	−	−	+
2000	+	−	−	−	−	+	+	+	−	−	+
2001	+	−	−	−	−	+	+	+	−	−	+
2002	+	−	−	−	−	+	+	+	−	−	+
2003	−	−	−	−	−	−	+	+	+	+	+
2004	+	−	−	−	+	−	+	+	+	+	+
2005	−	−	−	−	−	−	+	+	+	+	+
2006	−	−	−	−	−	−	+	+	−	+	+
2007	+	−	−	−	−	−	+	+	−	−	+
2008	+	−	−	−	−	−	+	+	−	−	+
2009	+	−	−	−	−	−	+	+	−	+	+

+ : Variables endogènes, − : Variables exogènes,

P-value du T^2 de Hotelling

Table 122 – p-Value du T^2 de Hotelling (1989)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,7364	0,7566	0,8081	0,7547	0,8001	0,7681
Danemark	0,9384	0,8641	0,6384	0,6746	0,6686	0,6441
Allemagne	0,8613	0,8463	0,9229	0,9307	0,9356	0,9343
Grèce	0,2579	0,2478	0,2884	0,3208	0,2937	0,337
Espagne	0,3133	0,3072	0,1901	0,2095	0,2175	0,2077
France	0,8771	0,9561	0,6716	0,6624	0,6941	0,6497
Irlande	0,8167	0,7655	0,9444	0,9529	0,9121	0,9922
Italie	0,3247	0,3189	0,3705	0,3831	0,3794	0,3883
Luxembourg	0,1746	0,1995	0,5174	0,5924	0,5393	0,6061
Pays-Bas	0,1475	0,182	0,6176	0,6371	0,5991	0,6812
Portugal	0,2824	0,2742	0,2436	0,2552	0,2814	0,2382
Royaume-Uni	0,2574	0,199	0,042	0,0486	0,049	0,0486

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 123– p-Value du T^2 de Hotelling (1990) Retour au texte 3.2.3

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,7253	0,7511	0,8326	0,7523	0,8116	0,774
Danemark	0,9697	0,8972	0,6306	0,7417	0,6826	0,6965
Allemagne	0,8884	0,8776	0,961	0,9484	0,9654	0,9533
Grèce	0,26186	0,2516	0,3026	0,3245	0,3064	0,3404
Espagne	0,3208	0,3122	0,1927	0,2018	0,2171	0,2041
France	0,8962	0,976	0,5826	0,6537	0,6172	0,6312
Irlande	0,7923	0,7491	0,9216	0,974	0,9025	0,9943
Italie	0,3286	0,3218	0,3529	0,3719	0,3608	0,3803
Luxembourg	0,1588	0,1837	0,5271	0,5492	0,5377	0,5776
Pays-Bas	0,1537	0,1865	0,668	0,6241	0,6376	0,6735
Portugal	0,2841	0,2778	0,2559	0,2684	0,2936	0,2477
Royaume-Uni	0,2601	0,1998	0,0408	0,0494	0,0483	0,0484

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 124– p-Value du T^2 de Hotelling (1991)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,7544	0,7511	0,7767	0,7693	0,8263	0,7987
Danemark	0,9714	0,8972	0,9133	0,8023	0,7723	0,7612
Allemagne	0,8651	0,8776	0,8553	0,9324	0,9564	0,9413
Grèce	0,2646	0,2518	0,2556	0,3398	0,3106	0,3551
Espagne	0,3123	0,3122	0,3032	0,1976	0,2209	0,2037
France	0,8998	0,9761	0,9852	0,6458	0,6006	0,6254
Irlande	0,7903	0,7491	0,7481	0,9646	0,8901	0,9932
Italie	0,3353	0,3218	0,328	0,3649	0,3522	0,3705
Luxembourg	0,1515	0,1837	0,1741	0,4844	0,4967	0,5058
Pays-Bas	0,148	0,1866	0,1804	0,6377	0,6651	0,6972
Portugal	0,2869	0,2778	0,2794	0,2707	0,3184	0,2543
Royaume-Uni	0,2809	0,1998	0,2147	0,0505	0,0455	0,048

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 125– p-Value du T^2 de Hotelling (1992)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,7674	0,7873	0,7947	0,7317	0,7832	0,7761
Danemark	0,7232	0,6685	0,5613	0,6177	0,5981	0,5826
Allemagne	0,8395	0,827	0,9393	0,948	0,949	0,9705
Grèce	0,2463	0,2381	0,2928	0,305	0,2943	0,3186
Espagne	0,2998	0,2941	0,1901	0,2212	0,2312	0,2345
France	0,9001	0,9999	0,5483	0,6191	0,5748	0,5727
Irlande	0,7848	0,7375	0,8637	0,8744	0,837	0,9112
Italie	0,3201	0,3138	0,3422	0,3552	0,3504	0,3585
Luxembourg	0,2144	0,2397	0,4974	0,5132	0,5196	0,5447
Pays-Bas	0,1372	0,1747	0,6851	0,6142	0,6601	0,7101
Portugal	0,2685	0,2601	0,2455	0,2435	0,269	0,2348
Royaume-Uni	0,2812	0,2194	0,0487	0,0602	0,0562	0,0548

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 126 – p-Value du T^2 de Hotelling (1993)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,7446	0,7549	0,7628	0,7126	0,7428	0,7691
Danemark	0,6891	0,6234	0,5487	0,6109	0,5859	0,5864
Allemagne	0,8526	0,8259	0,9244	0,8967	0,9284	0,9018
Grèce	0,2479	0,2393	0,2749	0,2736	0,2823	0,2788
Espagne	0,296	0,2903	0,2039	0,2358	0,2403	0,2587
France	0,9023	0,9976	0,5534	0,6184	0,5926	0,5732
Irlande	0,7659	0,718	0,7609	0,7403	0,7496	0,7498
Italie	0,3225	0,3107	0,3201	0,3215	0,3331	0,3247
Luxembourg	0,186	0,2186	0,5746	0,5855	0,58	0,6224
Pays-Bas	0,1432	0,1824	0,6922	0,6186	0,6522	0,7251
Portugal	0,2686	0,2567	0,3104	0,3367	0,319	0,3406
Royaume-Uni	0,3265	0,2518	0,0424	0,0529	0,0517	0,0461

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 127 – p-Value du T^2 de Hotelling (1994)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,7659	0,7843	0,8009	0,7247	0,7759	0,7944
Danemark	0,6593	0,5974	0,5331	0,5992	0,5632	0,5717
Allemagne	0,8051	0,7837	0,9004	0,8854	0,8995	0,8954
Grèce	0,2404	0,2301	0,2464	0,2575	0,2582	0,261
Espagne	0,2966	0,2963	0,2528	0,2566	0,2731	0,2921
France	0,9475	0,9506	0,512	0,602	0,5506	0,5451
Irlande	0,7627	0,7136	0,7273	0,7381	0,7238	0,7371
Italie	0,312	0,2991	0,2832	0,2951	0,2978	0,3006
Luxembourg	0,2085	0,23	0,5636	0,5216	0,5652	0,5587
Pays-Bas	0,1298	0,1786	0,7423	0,6403	0,7006	0,7672
Portugal	0,2668	0,254	0,3305	0,3338	0,3462	0,3381
Royaume-Uni	0,3583	0,2703	0,0423	0,0581	0,0532	0,0494

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 128 – p-Value du T^2 de Hotelling (1995)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,7609	0,7963	0,8289	0,7357	0,7817	0,8206
Danemark	0,5828	0,5291	0,5489	0,5675	0,5733	0,5554
Allemagne	0,5081	0,4599	0,4667	0,476	0,4723	0,4451
Grèce	0,1956	0,1829	0,1914	0,207	0,1983	0,208
Espagne	0,2577	0,2474	0,2645	0,2231	0,2752	0,2263
France	0,9619	0,9419	0,4345	0,524	0,4831	0,485
Irlande	0,7223	0,6622	0,6537	0,6852	0,6553	0,6729
Italie	0,2743	0,2616	0,2601	0,2707	0,2708	0,2662
Luxembourg	0,1851	0,2137	0,5558	0,5099	0,5438	0,5516
Pays-Bas	0,103	0,1511	0,7291	0,5938	0,6557	0,7462
Autriche	0,5087	0,5482	0,7228	0,7362	0,7312	0,7783
Portugal	0,2193	0,2043	0,2769	0,279	0,2858	0,2814
Finlande	0,8508	0,9168	0,8426	0,8426	0,837	0,9291
Suède	0,6419	0,6774	0,9505	0,9443	0,9393	0,9864
Royaume-Uni	0,2989	0,2182	0,0217	0,033	0,0286	0,0301

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 129 – p-Value du T^2 de Hotelling (1996)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,815	0,848	0,815	0,767	0,771	0,867
Danemark	0,616	0,561	0,612	0,614	0,629	0,599
Allemagne	0,526	0,489	0,571	0,549	0,58	0,542
Grèce	0,205	0,189	0,201	0,208	0,208	0,205
Espagne	0,265	0,258	0,223	0,221	0,248	0,262
France	0,932	0,949	0,393	0,45	0,434	0,382
Irlande	0,735	0,671	0,69	0,674	0,685	0,639
Italie	0,273	0,255	0,243	0,249	0,254	0,249
Luxembourg	0,163	0,183	0,495	0,505	0,497	0,551
Pays-Bas	0,106	0,159	0,668	0,587	0,607	0,773
Autriche	0,536	0,562	0,733	0,745	0,734	0,765
Portugal	0,229	0,208	0,231	0,247	0,242	0,253
Finlande	0,718	0,78	0,765	0,777	0,757	0,838
Suède	0,459	0,453	0,731	0,735	0,769	0,705
Royaume-Uni	0,354	0,269	0,029	0,04	0,035	0,033

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 130– p-Value du T^2 de Hotelling (1997) Retour au texte 3.2.3

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,798	0,84	0,789	0,77	0,744	0,886
Danemark	0,589	0,538	0,629	0,599	0,65	0,561
Allemagne	0,565	0,526	0,583	0,579	0,597	0,569
Grèce	0,205	0,191	0,211	0,222	0,222	0,215
Espagne	0,262	0,257	0,227	0,223	0,247	0,269
France	0,926	0,962	0,438	0,466	0,484	0,402
Irlande	0,714	0,656	0,7	0,675	0,703	0,637
Italie	0,276	0,26	0,256	0,259	0,27	0,256
Luxembourg	0,162	0,182	0,454	0,48	0,455	0,525
Pays-Bas	0,106	0,159	0,667	0,616	0,607	0,809
Autriche	0,551	0,569	0,723	0,744	0,725	0,743
Portugal	0,23	0,212	0,217	0,24	0,225	0,251
Finlande	0,75	0,809	0,747	0,773	0,742	0,831
Suède	0,577	0,559	0,751	0,768	0,789	0,705
Royaume-Uni	0,285	0,217	0,027	0,036	0,033	0,031

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 131 – p-Value du T^2 de Hotelling (1998)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,811	0,849	0,759	0,765	0,722	0,85
Danemark	0,574	0,53	0,63	0,608	0,648	0,592
Allemagne	0,585	0,543	0,534	0,554	0,554	0,55
Grèce	0,208	0,196	0,21	0,227	0,223	0,225
Espagne	0,262	0,258	0,204	0,207	0,219	0,23
France	0,939	0,949	0,47	0,46	0,515	0,415
Irlande	0,7	0,646	0,688	0,675	0,696	0,663
Italie	0,278	0,262	0,253	0,265	0,269	0,263
Luxembourg	0,129	0,143	0,327	0,368	0,33	0,393
Pays-Bas	0,107	0,155	0,619	0,606	0,568	0,761
Autriche	0,553	0,568	0,747	0,768	0,749	0,774
Portugal	0,237	0,217	0,199	0,218	0,208	0,224
Finlande	0,775	0,815	0,736	0,753	0,737	0,798
Suède	0,615	0,598	0,797	0,811	0,827	0,765
Royaume-Uni	0,334	0,262	0,038	0,046	0,047	0,039

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 132 – p-Value du T^2 de Hotelling (1999)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,757	0,787	0,781	0,723	0,735	0,803
Danemark	0,606	0,563	0,624	0,654	0,641	0,641
Allemagne	0,336	0,296	0,259	0,282	0,277	0,281
Grèce	0,209	0,196	0,196	0,212	0,205	0,207
Espagne	0,259	0,252	0,247	0,24	0,269	0,273
France	0,899	0,992	0,44	0,492	0,485	0,448
Irlande	0,631	0,583	0,576	0,592	0,585	0,568
Italie	0,276	0,263	0,241	0,252	0,253	0,25
Luxembourg	0,193	0,213	0,441	0,44	0,437	0,478
Pays-Bas	0,099	0,153	0,753	0,668	0,688	0,823
Autriche	0,557	0,571	0,691	0,716	0,695	0,713
Portugal	0,23	0,211	0,221	0,23	0,228	0,243
Finlande	0,838	0,89	0,784	0,802	0,779	0,849
Suède	0,702	0,697	0,811	0,851	0,848	0,791
Royaume-Uni	0,36	0,277	0,044	0,055	0,054	0,047

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 133 – p-Value du T^2 de Hotelling (2000)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,755	0,796	0,82	0,759	0,763	0,874
Danemark	0,629	0,58	0,613	0,639	0,63	0,609
Allemagne	0,323	0,295	0,293	0,313	0,309	0,316
Grèce	0,206	0,195	0,187	0,209	0,196	0,205
Espagne	0,258	0,253	0,265	0,255	0,292	0,284
France	0,876	0,967	0,463	0,511	0,504	0,461
Irlande	0,654	0,608	0,604	0,62	0,602	0,6
Italie	0,267	0,254	0,217	0,231	0,226	0,234
Luxembourg	0,179	0,194	0,41	0,411	0,416	0,452
Pays-Bas	0,104	0,15	0,709	0,638	0,646	0,818
Autriche	0,572	0,577	0,664	0,693	0,671	0,685
Portugal	0,227	0,21	0,223	0,226	0,234	0,24
Finlande	0,867	0,93	0,852	0,855	0,837	0,925
Suède	0,682	0,667	0,824	0,864	0,861	0,793
Royaume-Uni	0,392	0,316	0,043	0,054	0,053	0,045

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 134 – p-Value du T^2 de Hotelling (2001)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,803	0,843	0,904	0,849	0,872	0,927
Danemark	0,669	0,626	0,722	0,72	0,717	0,698
Allemagne	0,395	0,363	0,369	0,36	0,384	0,359
Grèce	0,205	0,196	0,219	0,236	0,229	0,231
Espagne	0,256	0,251	0,267	0,261	0,279	0,267
France	0,836	0,926	0,588	0,623	0,624	0,607
Irlande	0,652	0,605	0,591	0,607	0,6	0,614
Italie	0,272	0,259	0,243	0,253	0,251	0,251
Luxembourg	0,159	0,176	0,475	0,469	0,471	0,468
Pays-Bas	0,161	0,23	0,569	0,541	0,564	0,681
Autriche	0,566	0,581	0,661	0,682	0,668	0,678
Portugal	0,255	0,24	0,294	0,295	0,307	0,305
Finlande	0,846	0,93	0,806	0,806	0,822	0,873
Suède	0,72	0,71	0,889	0,918	0,894	0,852
Royaume-Uni	0,183	0,147	0,022	0,031	0,028	0,028

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 135 – p-Value du T^2 de Hotelling (2002)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,728	0,792	0,935	0,888	0,907	0,951
Danemark	0,582	0,535	0,544	0,564	0,541	0,555
Allemagne	0,386	0,37	0,425	0,419	0,438	0,409
Grèce	0,197	0,187	0,204	0,224	0,214	0,223
Espagne	0,246	0,24	0,268	0,264	0,283	0,266
France	0,866	0,958	0,652	0,686	0,68	0,655
Irlande	0,615	0,572	0,554	0,564	0,565	0,569
Italie	0,279	0,269	0,257	0,273	0,27	0,271
Luxembourg	0,214	0,231	0,51	0,518	0,511	0,518
Pays-Bas	0,144	0,226	0,615	0,563	0,616	0,68
Autriche	0,524	0,536	0,598	0,623	0,604	0,627
Portugal	0,252	0,232	0,29	0,3	0,302	0,3
Finlande	0,789	0,895	0,805	0,783	0,814	0,844
Suède	0,578	0,55	0,658	0,704	0,671	0,659
Royaume-Uni	0,225	0,169	0,024	0,031	0,03	0,03

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 136 – p-Value du T^2 de Hotelling (2003)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,736	0,784	0,853	0,817	0,828	0,868
Danemark	0,55	0,495	0,527	0,53	0,52	0,506
Allemagne	0,388	0,372	0,434	0,414	0,443	0,407
Grèce	0,196	0,187	0,212	0,222	0,217	0,218
Espagne	0,249	0,246	0,288	0,277	0,3	0,282
France	0,837	0,924	0,711	0,75	0,742	0,724
Irlande	0,621	0,578	0,542	0,562	0,56	0,559
Italie	0,299	0,289	0,29	0,299	0,299	0,295
Luxembourg	0,22	0,234	0,501	0,484	0,491	0,482
Pays-Bas	0,135	0,205	0,621	0,623	0,63	0,711
Autriche	0,524	0,53	0,559	0,583	0,566	0,583
Portugal	0,25	0,234	0,299	0,306	0,309	0,307
Finlande	0,795	0,869	0,748	0,77	0,767	0,817
Suède	0,66	0,642	0,758	0,772	0,756	0,736
Royaume-Uni	0,212	0,163	0,02	0,029	0,027	0,028

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 137 – p-Value du T^2 de Hotelling (2004)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,844	0,934	0,922	0,968	0,975	0,979
Chypre	0,361	0,369	0,452	0,448	0,437	0,431
République Tchèque	0,091	0,073	0,101	0,101	0,102	0,103
Danemark	0,674	0,689	0,867	0,837	0,816	0,8
Allemagne	0,402	0,442	0,737	0,72	0,701	0,688
Grèce	0,327	0,348	0,471	0,457	0,452	0,442
Espagne	0,383	0,408	0,574	0,589	0,563	0,548
Estonie	0,946	0,997	0,995	0,995	0,982	0,984
France	0,961	0,975	0,88	0,883	0,861	0,868
Hongrie	0,482	0,556	0,723	0,732	0,705	0,693
Irlande	0,724	0,701	0,687	0,689	0,702	0,71
Italie	0,425	0,436	0,518	0,52	0,507	0,498
Lituanie	0,528	0,54	0,672	0,655	0,641	0,636
Luxembourg	0,36	0,437	0,87	0,836	0,8	0,774
Lettonie	0,556	0,564	0,712	0,701	0,678	0,677
Malte	0,476	0,484	0,521	0,518	0,515	0,516
Pays-Bas	0,216	0,406	0,996	0,959	0,938	0,936
Autriche	0,656	0,684	0,683	0,681	0,686	0,69
Pologne	0,411	0,417	0,496	0,487	0,477	0,47
Portugal	0,389	0,397	0,537	0,538	0,514	0,507
Finlande	0,888	0,979	0,82	0,831	0,851	0,879
Suède	0,685	0,712	0,936	0,932	0,912	0,894
Slovaquie	0,007	0,004	0,0004	0,001	0,001	0,001
Slovénie	0,427	0,432	0,492	0,483	0,477	0,475
Royaume-Uni	0,36	0,352	0,305	0,317	0,277	0,285

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 138 – p-Value du T^2 de Hotelling (2005)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,845	0,933	0,914	0,958	0,959	0,963
Chypre	0,344	0,359	0,433	0,434	0,426	0,426
République Tchèque	0,098	0,085	0,094	0,093	0,101	0,1
Danemark	0,664	0,693	0,893	0,861	0,845	0,836
Allemagne	0,42	0,474	0,701	0,686	0,68	0,669
Grèce	0,332	0,358	0,447	0,437	0,436	0,43
Espagne	0,386	0,409	0,55	0,586	0,549	0,562
Estonie	0,97	0,983	0,99	0,991	0,989	0,983
France	0,951	0,998	0,893	0,893	0,885	0,886
Hongrie	0,469	0,538	0,694	0,711	0,679	0,688
Irlande	0,719	0,7	0,675	0,665	0,686	0,677
Italie	0,431	0,448	0,51	0,517	0,505	0,505
Lituanie	0,521	0,538	0,66	0,645	0,64	0,638
Luxembourg	0,326	0,419	0,861	0,848	0,814	0,819
Lettonie	0,552	0,566	0,713	0,712	0,695	0,702
Malte	0,499	0,501	0,504	0,498	0,502	0,497
Pays-Bas	0,232	0,418	0,984	0,949	0,922	0,93
Autriche	0,675	0,702	0,678	0,671	0,678	0,674
Pologne	0,417	0,427	0,485	0,478	0,474	0,469
Portugal	0,392	0,405	0,548	0,568	0,541	0,554
Finlande	0,938	0,986	0,842	0,843	0,859	0,865
Suède	0,752	0,762	0,875	0,882	0,86	0,857
Slovaquie	0,006	0,003	0,001	0,001	0,001	0,001
Slovénie	0,427	0,438	0,475	0,463	0,466	0,459
Royaume-Uni	0,337	0,337	0,225	0,243	0,217	0,228

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 139 – p-Value du T^2 de Hotelling (2006)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,827	0,907	0,937	0,99	0,977	0,984
Chypre	0,32	0,335	0,415	0,416	0,408	0,408
République Tchèque	0,123	0,091	0,085	0,084	0,095	0,091
Danemark	0,407	0,45	0,702	0,655	0,646	0,635
Allemagne	0,337	0,386	0,663	0,649	0,642	0,637
Grèce	0,305	0,333	0,433	0,422	0,422	0,418
Espagne	0,355	0,38	0,532	0,557	0,532	0,545
Estonie	0,995	0,955	0,971	0,992	0,996	0,991
France	0,941	0,997	0,896	0,888	0,886	0,878
Hongrie	0,437	0,511	0,663	0,678	0,648	0,659
Irlande	0,722	0,695	0,665	0,659	0,673	0,663
Italie	0,399	0,417	0,494	0,498	0,488	0,49
Lituanie	0,535	0,536	0,604	0,586	0,59	0,583
Luxembourg	0,298	0,393	0,818	0,798	0,778	0,782
Lettonie	0,543	0,56	0,679	0,662	0,666	0,662
Malte	0,487	0,489	0,497	0,493	0,493	0,487
Pays-Bas	0,181	0,359	0,953	0,898	0,894	0,896
Autriche	0,656	0,676	0,652	0,645	0,65	0,646
Pologne	0,393	0,408	0,473	0,464	0,46	0,456
Portugal	0,375	0,385	0,544	0,554	0,544	0,552
Finlande	0,882	0,952	0,832	0,831	0,848	0,852
Suède	0,627	0,678	0,841	0,842	0,814	0,814
Slovaquie	0,015	0,007	0,001	0,001	0,001	0,001
Slovénie	0,383	0,39	0,434	0,42	0,423	0,415
Royaume-Uni	0,277	0,285	0,211	0,215	0,2	0,2

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 140 – p-Value du T^2 de Hotelling (2007)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,752	0,846	0,987	0,97	0,964	0,955
Bulgarie	0,408	0,437	0,526	0,519	0,508	0,5
Chypre	0,378	0,401	0,485	0,487	0,477	0,471
République Tchèque	0,117	0,099	0,126	0,123	0,136	0,134
Danemark	0,421	0,478	0,736	0,701	0,675	0,647
Allemagne	0,377	0,451	0,759	0,743	0,728	0,712
Grèce	0,364	0,4	0,479	0,472	0,466	0,458
Espagne	0,419	0,452	0,633	0,659	0,634	0,636
Estonie	0,942	0,971	0,892	0,91	0,912	0,917
France	0,996	0,976	0,917	0,917	0,909	0,906
Hongrie	0,565	0,634	0,752	0,769	0,739	0,737
Irlande	0,744	0,723	0,704	0,698	0,713	0,708
Italie	0,464	0,486	0,561	0,566	0,554	0,548
Lituanie	0,596	0,614	0,691	0,675	0,674	0,668
Luxembourg	0,363	0,475	0,857	0,849	0,819	0,813
Lettonie	0,598	0,62	0,738	0,733	0,729	0,734
Malte	0,529	0,543	0,558	0,558	0,555	0,549
Pays-Bas	0,258	0,46	0,962	0,918	0,896	0,884
Autriche	0,702	0,716	0,681	0,676	0,678	0,673
Pologne	0,447	0,471	0,536	0,533	0,523	0,515
Portugal	0,431	0,448	0,633	0,655	0,639	0,655
Roumanie	0,386	0,413	0,485	0,474	0,466	0,456
Finlande	0,939	0,998	0,848	0,843	0,861	0,87
Suède	0,657	0,688	0,814	0,829	0,797	0,786
Slovaquie	0,003	0,001	0,001	0,0002	0,0003	0,0003
Slovénie	0,447	0,464	0,497	0,49	0,487	0,478
Royaume-Uni	0,314	0,347	0,283	0,3	0,274	0,274

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 141– p-Value du T^2 de Hotelling (2008) Retour au texte 3.2.3

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,766	0,868	0,987	0,977	0,965	0,946
Bulgarie	0,403	0,445	0,543	0,532	0,521	0,512
Chypre	0,377	0,406	0,472	0,47	0,462	0,457
République Tchèque	0,163	0,123	0,131	0,138	0,144	0,143
Danemark	0,419	0,489	0,74	0,703	0,681	0,652
Allemagne	0,359	0,434	0,728	0,713	0,7	0,685
Grèce	0,361	0,4	0,5	0,49	0,487	0,477
Espagne	0,442	0,481	0,68	0,738	0,684	0,698
Estonie	0,862	0,92	0,948	0,961	0,959	0,966
France	0,996	0,98	0,925	0,916	0,911	0,904
Hongrie	0,567	0,648	0,762	0,77	0,74	0,739
Irlande	0,746	0,723	0,703	0,693	0,713	0,709
Italie	0,454	0,477	0,538	0,541	0,53	0,526
Lituanie	0,597	0,619	0,71	0,697	0,694	0,684
Luxembourg	0,375	0,5	0,879	0,872	0,834	0,823
Lettonie	0,603	0,629	0,742	0,741	0,729	0,725
Malte	0,521	0,539	0,562	0,556	0,557	0,551
Pays-Bas	0,225	0,426	0,923	0,884	0,851	0,828
Autriche	0,694	0,707	0,676	0,668	0,672	0,667
Pologne	0,448	0,476	0,539	0,531	0,525	0,517
Portugal	0,431	0,45	0,6	0,631	0,598	0,604
Roumanie	0,388	0,418	0,49	0,474	0,471	0,459
Finlande	0,928	0,972	0,827	0,827	0,836	0,838
Suède	0,618	0,665	0,797	0,802	0,774	0,771
Slovaquie	0,003	0,001	0,002	0,001	0,0002	0,002
Slovénie	0,435	0,457	0,49	0,479	0,48	0,471
Royaume-Uni	0,352	0,406	0,3	0,306	0,278	0,276

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 142– p-Value du T^2 de Hotelling (2009)

Observations (Pays)	PLS1	Gini1'-PLS1	Gini2-PLS1	PLS2	Gini1'-PLS2	Gini2-PLS2
Belgique	0,724	0,816	0,964	0,891	0,911	0,897
Bulgarie	0,391	0,435	0,538	0,529	0,517	0,511
Chypre	0,366	0,39	0,469	0,464	0,458	0,45
République Tchèque	0,141	0,11	0,106	0,104	0,115	0,111
Danemark	0,429	0,481	0,701	0,623	0,63	0,594
Allemagne	0,331	0,395	0,68	0,634	0,646	0,626
Grèce	0,353	0,388	0,486	0,467	0,471	0,458
Espagne	0,429	0,462	0,616	0,645	0,612	0,634
Estonie	0,839	0,898	0,917	0,951	0,926	0,923
France	0,984	0,983	0,912	0,903	0,899	0,89
Hongrie	0,525	0,592	0,715	0,731	0,699	0,705
Irlande	0,754	0,727	0,695	0,703	0,707	0,698
Italie	0,448	0,47	0,527	0,532	0,521	0,52
Lituanie	0,577	0,597	0,676	0,647	0,659	0,652
Luxembourg	0,303	0,408	0,807	0,76	0,758	0,753
Lettonie	0,594	0,616	0,709	0,68	0,695	0,7
Malte	0,511	0,511	0,536	0,539	0,531	0,522
Pays-Bas	0,18	0,354	0,876	0,794	0,796	0,788
Autriche	0,694	0,706	0,661	0,661	0,657	0,65
Pologne	0,436	0,462	0,523	0,52	0,51	0,502
Portugal	0,426	0,442	0,54	0,528	0,533	0,547
Roumanie	0,377	0,403	0,462	0,454	0,447	0,437
Finlande	0,953	0,995	0,825	0,793	0,825	0,82
Suède	0,652	0,712	0,796	0,83	0,776	0,786
Slovaquie	0,005	0,002	0,0003	0,001	0,001	0,001
Slovénie	0,432	0,454	0,477	0,474	0,468	0,458
Royaume-Uni	0,346	0,38	0,202	0,238	0,191	0,202

Lorsque p-Value du T^2 de Hotelling $0,01$, il n'y a pas d'outliers,

Tableau 143 – Taille de l'échantillon en fonction des années

Années	Taille de l'échantillon
1989-1994	12
1995-2003	15
2004-2006	25
2007-2009	27

Tableau 144 – Statistiques Durbin-Watson

n=12	0	AN	0,098	?	0,497	AAC	3,503	?	3,902	AP	4
n=15	0	AN	0,098	?	0,497	AAC	3,503	?	3,902	AP	4
n=25	0	AN	0,470	?	1,298	AAC	2,702	?	3,530	AP	4
n=27	0	AN	0,544	?	1,400	AAC	2,600	?	3,456	AP	4

AN : Auto-corrélation négative,
AAC : Absence d'auto-corrélation,
AP : Auto-corrélation positive,

Tableau 145 – Test de White au seuil 5%

Taille de l'échantillon	F(0,95, n-K,K-1)
$n = 12$	4,96
$n = 15$	3,48
$n = 25$	2,86
$n = 27$	2,83

Bibliographie

[1] Arias O., Hallock K.F., Sosa-Escudero W. (2001), Individual Heterogeneity in the Returns to Schooling : Instrumental Variables Quantile Regression Using Twins Data, *Empirical Economics*, Vol. 26, No. 1, pp. 7-40.

[2] Atkinson, A.B., (1970), On the measurement of inequality,*Journal of Economic Theory*, Vol. 2, pp. 244-263.

[3] Basmann, R.L. and D.J. Slottje, (1987), A new index of income inequality : The B measure, *Economics Letters*, Vol. 24, pp. 384-389.

[4] Basmann, R.L. and D.J. Slottje, (1988), Some new weighted geometric mean measures of inequality and their relations to several well-known measures, *Working paper*, No. 8711 (Department of Economics, Southern Methodist University, Dallas, TX).

[5] Basmann R.L., Hayes K.J., SlottjeD.J., Johnson J.D., (1990), Ageneral functional form for approximating the Lorenz curve, *Journal of Econometrics*, Vol. 43, pp.77-90.

[6] Boisso D., Hayes K., Hirschberg J., and Silber J., (1994), Occupational segragation in the multidimensional case decomposition and tests of significance, *Journal of Econometrics*, No. 61, pp. 161-171.

[7] Bry X., Trottier C., Verron T., Mortier F., (2013), Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm, *Journal of Multivariate Analysis*, Vol. 119, pp. 47-60

[8] Bureau J.C., (2007), La Politique Agricole Commune, *Paris : La Découverte*, Vol. 1, pp. 1-121.

[9] Butault J.P., Lerouvillois P. (1999). *La réforme de la PAC et l'inégalité des revenus agricoles dans l'Union européenne : les premiers effets*, E conomie et Statistique, No. 329-330, pp 73-86.

[10] Capeye,Cellule de veille et de prospectivesur la politique agricole commune (2014), Histoire de la PAC, pp. 1-6, https://www.supagro.fr/capeye/histoire-de-la-pac/, 06 Août 2014.

[11] Chameni Nembua C., (2009), Inequality factor decomposition under uniform additions property with applications to Cameroonian rural data, *Munich Personal RePEc Archive*, MRPRA Paper, No. 31250, pp. 1-27.

[12] Chantreuil, F. and Trannoy A. (1999), Inequality Decomposition Values : The Trade-off Between Marginality and Consistency, DP 9924 THEMA.

[13] Choi S.W.(2009), The effect of outliers on regression analysis : regime type and foreign direct investment, *Quarterly journal of political science*, 4, 153-165.

[14] Chung D., Keles S., (2010), Sparce Partial Least Squares Classification for High Dimentional Data, *Statistical Applications in Genetics and Molecular Biology*, Vol.9, Issue 1, Article 17, pp. 1-30.

[15] Cowell F.A., Fiorio C.V., (2011), Inequality decompositions : a reconciliation, *Journal of Enonomic Inequality*, Vol. 9, pp. 509-528.

[16] De Bourmont M., (2012), La résolution d'un problème de multicolinéarité au sein des "études portant sur les déterminants d'une publication volontaire d'informations : proposition d'un algorithme de décision simplifié basé sur les indicateurs de Belsley, Kuh et Welsch (1980), 33^{me} *Congrès de l'AFC Comptabilités et innovation, Grenoble : France.*

[17] Desriers M., Burgue M. Sparhubert J. (2000) Le montant des aides directes de la PAC reste très lié à la taille des exploitations, Agreste cahiers numéro 3.

[18] Dixon W.J.(1950), Analysis of extreme values, *The annals of mathemetical statistics*, 2(4), 488-506.

[19] Durbin, J. (1954). Errors in variables, *Review of the International Statistical Institute*, 22, 23–32.

[20] FADN. Agriculture and rural development Farm accounting data network [en ligne] (page consultée le 13/05/2014). Disponible sur : http://ec.europa.eu/agriculture/ricaprod/database/database_en.cfm

[21] Fei J.C.H., Ranis G. and Kuo S.W.Y., (1978), Growth and the Family Distribution of Income by Factor Components, *Quarterly Journal of Economics*, Vol. 92, pp. 17-53.

[22] Fields G.S., Yoo G. (2000), Falling labour income inequality in Korea's economic growth : pattern and underlying causes, *Review of income and wealth*, Vol. 46, pp. 139-159.

[23] Furusjö E., Svenson A., Rahmberg M., Andersson M., (2006), The importance of outlier detection and training set selection for reliable environmental QSAR predictions, *Chemosphere* ,Vol. 63, pp. 99-108.

[24] Gini, C., (1912), Variabilita e Mutabilita, Bologna.

[25] Gini, C., (1914), Di una misura della concentratione independente della distribuzione dl carattere. *Atti del R.Instituto Veneto de Scinze, Lettere ed Arti, tomo LXXIII*, parte II, p. 1203-1258.

[26] Gneiting T., (2012), making and evaluating point forecasts, Journal of American Association.

[27] Greene W. (2005), *Économétrie*. Pearson Education. Édition française dirigée par Shlachter D., Azomahou T., Couderc N, Monjon S., Nguyen Van P., 5ème édition. 946 p.

[28] Hawkins D. M., Bradu D. and Kass,G. V., (1984), Location of several outliers in multiple regression data using elemental sets*Technometrics*, Vol. 26,pp. 197-208.

[29] John, G.H. (1995). Robust Decision Tree : Removing Outliers from Databases, *KDD-95 Proceeding*, 174-179.

[30] Kakwani, N. (1980a), On A Class of Poverty Measures, *Econometrica*, Vol. 48, pp. 437-446.

[31] Kakwani, N., (1980b), Income inequality and poverty *Oxford University Press, Oxford*.

[32] Knorr E., Ng R., (1998), Algorithms for mining distance-based outliers in large datasets, *Proc of the VLDB Conference, New York USA*, pp. 392-403.

[33] Lerman, R. and Yitzhaki S. (1985), Income Inequalities Effects by Income Source : A New Approach and Applications to United States, *Review of Economics and Statistics*, 67, 151–156.

[34] Lerman R. and Yitzhaki S. (1989a), Income Sources and Income Inequality : Measurement from Three US Income Surveys, *Journal of Economic and Social Measurement*, 15(2), pp. 167-179.

[35] Lerman, R.I., and Yitzhaki, S. (1989b), Improving the accuracy of estimates of Gini coefficients, *Journal of Econometrics*, Vol.42, No.1, pp. 43-47.

[36] Ministère de l'agriculture, de l'agroalimentaire et de la forêt(La politique agricole commune : les 50 ans de la PAC) 2015.

[37] Morduch J. et Sicular T. (2002), Rethinking Inequality Decomposition, with evidence from Rural China, *Economic Journal*, Vol. 112, No.476, pp. 93-106.

[38] Mussard S., Pi-Alperin M.N., Seyte F. et Terraza M. (2005), Extension of Dagum's Gini Decompsoition, *Colloque international en mémoire de Gini et Lorenz, Sienne, Italie*.

[39] Mussard S., (2007), La décomposition des mesures d'inégalité en sources de revenu : méthodes et applications, *L'actualité économique*, Vol. 83, No. 3, pp. 415-445.

[40] Mussard S., Souissi-Benrejab F., (2015a), Gini-PLS Regressions *Cahier de Recherche/Working Paper*, GRÉDI, Université de Sherbrooke, No. 15-01. http://gredi.recherche.usherbrooke.ca/wpapers/GREDI-1501.pdf.

[41] Mussard S., Souissi-Benrejab F., (2015b), Gini-PLS Regressions *Documents de recherche :*, LAMETA, Université Montpellier1, No.2015-03, http://www.lameta.univ-montp1.fr/Documents/DR2015-03.pdf.

[42] Oaxaca R.L.,(1973), Male-female wage differences in urban labour market, *International Economic Review*, No. 14, pp. 693-709.

[43] Olkin, I., Yitzhaki, S. (1992). Gini regression analysis, *International Statistical Review*, 602, 185–196.

[44] Planchon, V. (2005). Traitement des valeurs aberrantes : concepts actuels et tendances générales, *Biotechnologie Agronomie Société et Environnement*, 91, 185-196.

[45] Ramsawamy S., Rastogi R., Shim K., (2000), Efficient algorithm for mining outliers from a large data sets, *ACM SIGMOD RECORD*, pp. 427-438.

[46] Rao V.M., (1969),Two Decompositions of Concentration Ratio *Journal of the Royal Statistical Society, Series A (General)*, Vol. 132, No. 3, pp.418-425.

[47] Robin J.M., (2000), Endogénéité et variables instrumentales dans les sciences sociales, *INSEE Méthodes*, pp. 217-276.

[48] Rousseew P.J., Leroy A.M., (2003), Robust regression and outlier detection, Wiley Interscience.

[49] Russolillo G., (2012), Non-metric Partial Least Squares, *Electronic Journal of Statistics*, Vol. 6, pp. 1641-1669.

[50] Schechtman, E., and Yitzhaki, S. (1987), A measure of association based on Gini's mean difference, *Communications in Statistics, Theory and Methods*, Vol. 16, pp. 207-231.

[51] Schechtman, E., Yitzhaki, S., Pudalov, T. (2011). Gini's multiple regressions : Two approaches and their interaction. *Metron*, LXIX(1), 65-97.

[52] Shorrocks A.F., (1980), The Class of Additively Decomposable Inequality Measures,*Econometrica*, Vol. 48, pp. 613-625.

[53] Shorrocks A.F.,(1982), Inequality Decomposition by Factor Components, *Econometrica*, Vol. 50, pp. 193-211.

[54] Shorrocks, A. F., (1983), The Impact of Income Components on the Distribution of Family Income, *Quarterly Journal of Economics*, Vol. 98, pp. 311-326.

[55] Shorrocks, A. F., (1999), Decomposition procedure for distributional analysis : a unified framework based on the Shapley value, *Unpublished manuscript, Department of Economics, University Essex.*

[56] Silber, J., (1993), Inequality decomposition by income source : a note, *Review of Economics and Statistics*, No. 75, Vol. 3, pp. 545-547.

[57] Tenenhaus, M. (1998). *La régression PLS théorie et pratique.* Éditions TECHNIP. 254 p.

[58] Tenenhaus M., and V. Esposito Vinzi (2005), PLS regression, PLS path modeling and generalized procrustean analysis : a combined approach for PLS regression, PLS path modeling and generalized multiblock analysis, *Journal of Chemometrics*, 19, 145-153.

[59] Tenenhaus, M., Tenanhaus A. (2009) A criterion based PLS approach to Structural Equation Modeling, *in third workshop on PLS Developments, ESSEC Business School of Paris*, May 14^{th} 2009.

[60] Theil, H., (1967), Economics and information theory (North-Holland, Amsterdam).

[61] Tracy, N.D., Young,J.C., and Mason, R.L.(1992), Multivariate control chart for individual observations, *Journal of Quality Technology*, Vol.24, pp.88-95.

[62] Tsay, R.S., (1988), Outliers, level shifts and variance changes in time series, *Journal of forecasting*, Vol.7, pp.1-20.

[63] Tukey J.W., (1977), Exploratory data analysis, Addison Welsley.

[64] Vasyechko O.A., Benlagha N., Grun-Rehomme M., (2005), Comparaison de méthodes de détection des valeurs extrêmes : application en statistique d'entreprise.

[65] Wan G., (2004), Accounting for income inequality in rural China : a regression-based approach, *Journal of Comparative Economics*, No. 32, pp. 348-363 .

[66] Wold, S., Albano, C., Dunn III, W.J., Esbensen, K., Hellberg, S., Johansson, E., Sjöström, H. (1983a). *Pattern Recognition : Finding and Using Regularities in multivariate Data in Proc. UFOST conf.* Food research and Data Analysis, Martens J.(Ed.), Applied Science Publications, London.

[67] Wold, S., Martens, H., Wold, H. (1983b). *The multivariate calibration problem in chemistry solved by the PLS method*. in Proc.Conf. Matrix Pencils Rnhe A. Kagstroem B.(Eds). Lecture Notes in Mathematics, Springer verlag, Heidelberg, 286-293.

[68] Yitzhaki S. (2003), Gini's Mean difference : a superior measure of variability for non-normal distributions, *Metron*, LXI(2), 285-316.

[69] Yitzhaki, S., Schechtman, E. (2004). The Gini instrumental variable, or the ′double instrumental variable′ estimator, *Metron*, 52(3), 287–313.

[70] Yitzhaki, S., Itzhaki, R., Pudalov, T. (2011). A nonparametric item characteristic curve using the Gini's mean difference approach. http://ssrn.com

[71] Yitzhaki, S., Schechtman, E. (2013). *The Gini Methodology : A primer on statistical methodology*. Springer series in statistics. 548 p.

[72] Yitzhaki, S. and P. Lambert (2013), The Relationship between the Absolute Deviation from a Quantile and Gini's Mean Difference, *Metron*, 71, 97–104.

Table des matières

Table des matières

Introduction générale **1**

1 Les modèles de régression PLS et Gini **11**

1.1 Problématiques : valeurs aberrantes, erreurs de mesure et multi-colinéarité 14

1.1.1 Valeurs aberrantes : “Outliers ” 14

1.1.2 Problème d’erreurs de mesure (endogénéité) 16

1.1.3 Problème de multi-colinéarité 17

1.2 Régression PLS1 (univariée) 19

1.2.1 Les étapes de la régression PLS1 19

1.2.2 Validation croisée 23

1.3 Régression PLS2 (multivariée) 24

1.3.1 Étapes de la régression PLS2 25

1.3.2 Validation croisée 28

1.4 Covariance Gini et régression Gini 29

1.4.1 Régression Gini semi-paramétrique 32

1.4.2 Approche par minimisation (ou régression Gini paramétrique) 37

2 Construction des régressions Gini-PLS **41**

2.1 Les régressions univariées Gini-PLS1 44

2.1.1 La régression Gini1-PLS1 44

2.1.2 La régression Gini2-PLS1 46

2.1.3 Propriétés et aides à interprétations 54

2.2 Les régressions multivariées Gini-PLS2 60

2.2.1 La régression Gini1-PLS2 60

2.2.2 L’algorithme Gini2-PLS2 62

2.2.3 Propriétés et aides à interprétations 65

2.3 Simulations 70

3 Étude des inégalités des revenus agricoles européens **79**

3.1 Modèles économétriques pour l’étude des inégalités 82

3.1.1 Shorrocks (1982) 83

3.1.2 Morduch et Sicular (2002) (Approche RISD-MCO) 84

3.1.3 Une nouvelle approche RISD-Gini-PLS 87

3.2 Analyse des disparités des revenus agricoles européens 94

3.2.1 Les enjeux de la Politique Agricole Commune (PAC) 94

3.2.2 Présentation de la base de données 96

3.2.3 Résultats 97

Conclusion générale **117**

Annexes **123**

Table des matières **209**

VU et PERMIS D'IMPRIMER



A Montpellier, le

Le Président de l'Université de Montpellier

Philippe Augé

Doctorat de l'Université Montpellier
Faculté des Sciences économiques — Arrêté du 7 Août 2006

Les régressions Gini-PLS :
une application aux inégalités des revenus agricoles européens

Résumé

Dans cette thèse, nous introduisons des modèles de régression "`Gini-PLS`". Les algorithmes proposés combinent les propriétés des estimateurs relatifs aux régressions Gini et PLS Les quatre modèles construits dans cette thèse permettent de résoudre `simultanément` les problèmes : de valeurs extrêmes ("outliers"), de multi-colinéarité, de faible taille de l'échantillon, de données manquantes, d'erreurs de mesure et d'endogénéité. En présence des problèmes cités, les modèles uni-variés (Gini-PLS1) sont robustes pour estimer une variable dépendante en fonction d'une ou plusieurs variables explicatives ; tandis que les modèles multi-variés (Gini-PLS2) servent à estimer une matrice de variables dépendantes en fonction d'une matrice de variables explicatives.
Notre application dans le cadre de la thèse concerne l'estimation de contributions des variables technico-économiques aux inégalités des rémunérations pour les pays européens adhérents à la Politique Agricole Commune.
Nous proposons deux approches de régressions basées sur les modèles Gini-PLS (RISD-Gini-PLS) pour estimer les contributions des variables technico-économiques (sources de revenus, superficies, main d'œuvre, etc.) aux inégalités des revenus agricoles pour les pays de l'union européenne avant et après les réformes de Mac Sharry et de l'accord de Luxembourg.

The Gini-PLS regressions :
an application to the European agricultural income inequalities

Abstract

In this thesis we propose "`Gini-PLS`" regressions. The proposed algorithms combine the properties of the estimators related to the Gini and PLS regressions. The four models built in this thesis solve `simultaneously` the problems of : extreme values (outliers), multicollinearity, small sample, missing data, measurement errors, and endogeneity. in presence of these problems, the univariate models (Gini-PLS1) are robust to estimate a dependent variable with one or more explanatory variables. While, the multivariate models (Gini-PLS2) are used to estimate a matrix of dependent variables with a matrix of explanatory variables.
Our application in this thesis is the estimation of the contributions of technico-economic variables to the whole inequality of farm's income for European countries acceding to the Common Agricultural Policy.
We also propose Gini-PLS regressions approaches based on income source decomposition (RISD-Gini-PLS) to estimate the contributions of techno-economic variables (income sources, areas, labor, etc.) to the income inequalies of productions (total output crops and output livestock) for european countries

Discipline : Sciences Économiques (section $n°05$).

MOTS CLÉS : Régressions Gini-PLS, problèmes économétriques (*Multicolinéarité, endogénéité, erreurs de mesure, outliers, faible taille de l'échantillon et données manquantes*), Approches RISD (*Approches de régressions basées sur la décomposition en sources de revenus*), Politique Agricole Commune, Revenus agricoles européens.

ADRESSE : LAMETA - Laboratoire Montpelliérain d'Économie Théorique et Appliquée
Université Montpellier - Faculté des Sciences Économiques
Avenue Raymond Dugrand - Site de Richter - C.S. 79606
34960 Montpellier Cedex 2