ÉVALUATION DE LA CONTRIBUTION MARGINALE DES FACTEURS DE RISQUE : OPÉRATIONNALISER LA VALEUR DE SHAPLEY

Par Adjoua Ahonzo

(8890417)

Mémoire présenté au Département de science économique

de l'Université d'Ottawa

pour l'obtention du diplôme de Maîtrise

Directeur du mémoire : Professeure Victoria Barham

ECO 6999

Ottawa, Ontario

Août 2017

Remerciements

Tout d'abord, je tiens à remercier mon directeur de mémoire Victoria Barham pour sa disponibilité, sa patience, ses encouragements et ses remarques pertinentes.

Je souhaiterais également remercier mes parents, mon adorable famille ainsi que Jean-Baptiste Tondji et Sacha Nandlall.

Abstract

This study computes for the first time the Shapley's value of Pongou and Tondji (2017) in order to estimate the marginal contribution of diabetes' risk factors. The study uses US survey data for 2015 and 2016 and the values obtained are compared with those of the probit model. The results of Shapley's value suggest that hypertension and obesity are predominant in the risk of having diabetes while those in the probit model consider high blood pressure, obesity, hypercholesterolemia and physical inactivity as major correlates in the risk of developing diabetes.

Résumé

Cette étude implémente pour la première fois la valeur de Shapley de Pongou et Tondji (2017) afin d'estimer la contribution marginale des facteurs de risque du diabète. Elle utilise les données d'enquête des États-Unis pour la période de 2015 et 2016 et les valeurs obtenues sont comparées avec celles du modèle probit. Les résultats de la valeur de Shapley suggèrent que l'hypertension artérielle et l'obésité sont prédominants dans le risque d'avoir le diabète tandis que ceux du modèle probit préconisent de considérer l'hypertension artérielle, l'obésité, l'hypercholestérolémie et l'inactivité physique comme corrélations majeures du risque de contracter le diabète.

INTRODUCTION

Aujourd'hui, l'évolution des maladies chroniques encore appelées maladies non transmissibles (MNT) représente un vrai problème de santé publique. Ces maladies sont responsables d'environ 63% des décès dans le monde, selon l'Organisation mondiale de la santé (OMS, 2017). Mais tout particulièrement quatre d'entres elles, à savoir les maladies cardiovasculaires, les affections respiratoires chroniques, les cancers et le diabète sont à l'origine de 35 millions de décès. Bien qu'il ne soit pas encore possible de les guérir, elles peuvent pour la plupart être évitées si l'on adopte les mesures nécessaires afin d'éliminer les facteurs de risque.

Souvent dénommée comme un « tueur silencieux », le diabète est l'une de ces quatre pathologies qui progresse le plus avec désormais 8,5% de la population adulte atteinte. (OMS, 2016). Au delà de son évolution fulgurante, le diabète est surtout redouté à cause de ses nombreuses complications et conséquences qui affectent non seulement la situation socioéconomique mais dégradent principalement la santé des individus. Cependant, la mise en place de politiques capables de freiner le fléau exige au préalable une identification claire des facteurs de risque du diabète mais surtout une détermination de la contribution de chaque facteur.

Eu égard à cette problématique, le diabète continue de susciter un grand intérêt dans le domaine de la recherche scientifique et de nombreuses études ont été menées (par exemple, Montgomery et al. 2015 ; Forouhi et al. 2014 ; Hegazi et al. 2015) dans le but de déterminer premièrement différents facteurs de risque propres au diabète. Ces études effectuées sur différentes populations ont permis d'identifier un ensemble de causes majeurs, reliées généralement à l'environnement, à la condition physique, à l'héritage génétique ou encore au genre des individus.

Pourtant, encore peu de travaux se sont penchés spécifiquement sur la détermination de l'incidence des facteurs de risque pour le diabète. Murad et al. (2014), visaient dans leur étude à évaluer les facteurs de risque communs associés au diabète de type 2, chez les patients de la ville de Djeddah. Les analyses univariées et multivariées montrent que les facteurs de risque les plus influents sont : être un homme âgé de plus de 40 ans, avoir un

faible niveau d'éducation, appartenir à la classe moyenne et fumer. Ces derniers sont fortement associés et jouent un grand rôle dans le risque de développer le diabète. D'Souza et al. (2013), procèdent à une évaluation des risques permettant de dépister le diabète chez les adultes Omanais. Se basant sur la méthode du test de khi-deux évaluer le score de risque du diabète (obtenu selon les calculs du modèle finlandais), les résultats montrent de 10% des adultes sont forment susceptibles de contracter le diabète au vu des facteurs de risque qu'ils présentent.

Toutefois, les données disponibles pour les études préconisent de considérer le diabète comme une variable discrète pouvant prendre deux valeurs. Afin de tenir compte de cette contrainte, les modèles probit sont appliqués toujours dans le but d'évaluer les facteurs de risque du diabète. Bhargava (2003), analyse l'importance des facteurs de risque du diabète et des maladies coronariennes chez les descendants de Framingham. Le poids, le tabagisme ou encore la pression artérielle sont des facteurs significatifs qui contribuent au diabète, l'augmentation d'une unité du facteur relatif au poids et à la taille accroit à lui seul d'environ 15% le risque de développer le diabète. De même Dickerson et al. (2011), s'intéressent au risque de développer le diabète au début de l'âge adulte en tenant comte des facteurs de risque héréditaires et comportementaux. Les résultats pour leur modèle de probit ordonné indiquent que les individus qui n'appartiennent pas à la race blanche non-hispanique étaient les plus exposés au risque du diabète tout comme ceux dont certains membres de la famille souffrent déjà du diabète.

Seulement, devant le peu de méthodes proposées pour l'estimation de la contribution marginale des variables binaires, d'autres alternatives sont constamment évaluées et récemment certains chercheurs se sont intéressés au développement de nouveaux outils tirés de la théorie des jeux afin contourner certaines difficultés en micro économétrie. En particulier, il y a ces derniers temps un intérêt pour l'utilisation de la valeur de Shapley, proposée au départ par Lloyd Shapley (1953) inspirée des travaux de Von Neumann et Morgenstern.

Tandis que le calcul classique de la valeur de Shapley suppose que la contribution de chaque joueur est déterministe, il est évident que de nombreuses situations économiques sont caractérisées par un contexte d'incertitude quant à la productivité des différents

facteurs de production. Ainsi, Pongou et Tondji (2017) apportent une autre contribution méthodologique à l'estimation de l'impact direct de caractéristiques tel que le niveau de compétences, le nombres d'heures travaillées ou les facteurs de risque d'une maladie par la méthode de Shapley. Ils construisent un modèle dans lequel chaque offre d'input est incertaine. Ils parviennent à énoncer une série d'axiomes leur permettant de calculer une première valeur dénommée valeur à priori de Shapley lorsque la probabilité de distribution est connue. Ensuite, après avoir observé le résultat a priori, ils déterminent la valeur Bayésienne de Shapley, définie comme valeur a postériori et terminent par l'application de ce modèle à l'analyse des jeux non-coopératifs.

Toutefois, leur article reste très théorique, uniquement testé sur de très petits échantillons. Il n'existe à ce jour à notre connaissance, aucune étude utilisant la méthodologie proposée pour estimer la contribution des facteurs de risque à une maladie. De ce fait, il apparait opportun de développer une méthodologie pour l'application cette approche à l'analyse des données.

Notre étude visera ainsi à faire une extension des travaux de Pongou et Tondji (2017) appliquée au cas du diabète. Nous incluons six facteurs de risque dans notre modèle et maintenons leur hypothèse en accord avec Shapley (1953) selon laquelle : si un profil ne dispose d'aucun facteur de risque alors sa contribution marginale doit être aussi nulle au sein de la coalition.

Nos facteurs de risque seront assimilés aux « joueurs », la « valeur du jeu » à la probabilité espérée des contributions marginale et la valeur de Shapley est la participation du facteur de risque spécifique à la coalition. En appliquant la méthode de Shapley dans notre étude sur le diabète, notre objectif est d'estimer la contribution marginale de chaque facteur de risque.

Cette étude permettra une première application de la méthodologie de Pongou et Tondji (2017) et une comparaison des valeurs de Shapley obtenues avec les résultats d'une estimation par le modèle Probit proposé par Bourbonnais (2015). De ce fait, nous optons délibérément pour un modèle probit relativement simple, ne contrôlant rien d'autre que l'homoscédasticité ; notre contribution pourtant essentiellement sur l'opérationnalisation de la valeur de Shapley.

La première partie de notre travail consiste en une revue de la littérature sur les différentes méthodes d'estimation pour évaluer la contribution marginale des facteurs à risque dans la probabilité de contracter une maladie. Dans la suite de la section nous introduisons tout d'abord, l'estimation par le modèle Probit et la méthode de Shapley ainsi que les données utilisées. Puis, nous présentons les résultats de nos estimations. Enfin la dernière partie s'articule principalement autour d'une discussion et extensions possible de notre travail.

1. REVUE DE LITTÉRATURE

Dans cette section nous faisons un bref survol de la littérature des travaux ayant utilisé la méthode probit et la valeur de Shapley afin d'obtenir les effets marginaux.

La question de la détermination de l'effet marginal a toujours été une préoccupation économie particulièrement en microéconomie. En effet, déterminer la variation de l'effet des variables après un choc s'avère utile dans la plupart des études. Jebeli et al. (2014) utilisent un modèle probit afin d'estimer l'effet marginal des facteurs socioéconomiques dans la demande de médicaments spécialisés chez 70 ménages iraniens dont seulement 280 patients atteints de thalassémie, hémophilie, insuffisance rénale chronique et de sclérose en plaques sont retenus. Ils trouvent que la relation qui existe entre la couverture en assurance maladie et la demande de médicaments n'est pas significative. Cela est sans doute dû à la gravité de maladie qui obligent les patients à acquérir les médicaments sans vraiment se soucier de leur coût. La taille du foyer augmente faiblement la demande de médicaments, par contre le genre, le niveau d'éducation ou le fait de travailler dans le secteur privé réduisent la demande des médicaments. Cette étude présente comment le modèle probit peut aider à la détermination des contributions marginales.

En outre, Costa-Font et Gil (2003) apportent une autre contribution à l'estimation d'effet marginal avec le modèle probit, en appliquant dans leurs travaux, un modèle de probit indépendant afin de tenir compte des problèmes d'endogénéité. Les données proviennent des enquêtes nationales de 1999 sur les personnes en situation de handicap et les conditions

générales de santé en Espagne. Après avoir retiré les individus de moins de 16 ans, ils obtiennent un échantillon de 54 159 pour lesquels ils parviennent aux résultats que l'obésité accroit la probabilité de contracter toutes les maladies chroniques (diabète, hypertension, maladies du cœur et taux de cholestérol élevé). Le genre est un facteur qui affecte le taux de prévalence de ces maladies chroniques. Et les femmes sont plus exposées que les hommes, de même concernant la probabilité d'être obèse. Le facteur de risque obésité augmente respectivement de 43%, 47%, 20% et 15% les probabilités d'avoir le diabète, l'hypertension, un taux de cholestérol élevé et les maladies du cœur.

Toutefois, hormis cette solution que confère le modèle probit son utilisation impose des hypothèses sur la normalité des erreurs et l'homoscédaticité. Le premier risque avec cette approche, est que les inférences ne sont pas robuste sous toutes les hypothèses. De plus, son estimateur obtenu par le maximum de vraisemblance n'est pas convergent en présence d'hétéroscédasticité, hétérogénéité non mesurée, variables manquantes ou si la distribution ne suit pas une loi normale et nécessite une solution par la méthode d'itération rendant les calculs lourds et compliqués. Mais encore, les coefficients obtenus par la régression ne correspondent pas aux effets marginaux et doivent être ré-estimés pour pouvoir être interprétés. De surcroît, la mesure traditionnelle du R^2 utilisée pour évaluer le pouvoir prédictif du modèle ne peut pas être exploitée car la variable dépendante est binaire. Cette situation conduit à des valeurs anormalement faibles du R^2 ne reflétant pas forcément que les facteurs explicatifs ne sont pas déterminants.

La valeur de Shapley offre une option différente qui permet de résoudre ce problème. De part les axiomes qui la compose, elle attribue une valeur unique pour chacun des facteurs choisis en calculant toutes les combinaisons possibles de ce facteur dans le modèle. Stan Lipovetcsky (2006), démontre que la valeur de Shapley d'un facteur « a » représente la part qu'explique ce facteur « a » dans le modèle. De manière analogue, la somme des valeurs de Shapley correspond au R^2 utilisé en économétrie. La méthode de la valeur de Shapley offre ainsi une alternative au pouvoir prédictif du R^2 qui ne peut pas être employé si le modèle n'est pas correctement spécifié, ne possède pas de constante ou encore devant les cas de multicolinéarité imparfaite qui réduisent l'impact de certains facteurs.

Un autre attrait aussi important de cette méthode est son efficacité, en ce sens que l'utilisation de cette la valeur de Shapley ne nécessite aucune spécification de fonction particulière. Cette méthode a ainsi la capacité d'appréhender des interactions complexes et non linéaires. De plus, les hypothèses traditionnelles des régressions économétriques ne sont pas utiles sous cette approche. De ce fait, aucune violation de celles-ci ne pourrait limiter la validité de nos résultats. Enfin, les valeurs de Shapley calculées s'interprètent directement en terme d'effets marginaux et représentent la contribution marginale spécifique de chaque facteur de production dans le résultat obtenu.

Un autre attrait aussi important de cette méthode est son efficacité, en ce sens que l'utilisation de cette la valeur de Shapley ne nécessite aucune spécification de fonction particulière. Cette méthode a ainsi la capacité d'appréhender des interactions complexes et non linéaires. De plus, les hypothèses traditionnelles des régressions économétriques ne sont pas utiles sous cette approche. De ce fait, aucune violation de celles-ci ne pourrait limiter la validité de nos résultats. Enfin, les valeurs de Shapley calculées s'interprètent directement en terme d'effets marginaux et représentent la contribution marginale spécifique de chaque facteur de production dans le résultat obtenu.

Un des domaines ou la valeur de Shapley est déjà utilisée pour évaluer le risque de chaque facteur et le champ de la finance. Tarashev et al. (2015) déterminent l'attribution du risque systémique en utilisant la valeur de Shapley comme solution à l'utilité transférable dans le système bancaire. La méthodologie de Shapley impose que la mesure du risque assigne une quantité de risque à élément de la coalition. Leur conclusion, leur permet de formuler un théorème qui stipule que la contribution du risque systémique augmente plus vite que sa taille.

Tol (2011) propose une application de la valeur de Shapley à une étude portant sur l'éducation. Il remplace les coalitions par les écoles existantes déjà et la valeur de Shapley sera la contribution moyenne d'un étudiant à son établissement. Les données utilisées sont celles de onze écoles de commerce irlandaises. La performance est mesurée par le nombre de publications et de citations dans la littérature. Il sépare la valeur de Shapley en deux partie une ordinale et l'autre cardinale. Les résultats montrent que la valeur de marché des étudiants augmente avec le nombre de publications (cardinale). Aussi, au niveau de valeur

ordinale, les étudiants de certaines écoles moins bien classées doivent publier plus ou être plus cités pour tendre vers la même valeur de marché de leurs pairs. Ces travaux ouvrent la voie à des applications plus riches et variées.

2. COMPARAISON DU MODÈLE PROBIT ET DU MODÈLE DE LA VALEUR DE SHAPLEY : APPLICATION À L'ANALYSE DES FACTEURS DE RISQUE POUR LE DIABÈTE

Dans cette section, nous utilisons la méthode probit ainsi que la valeur de Shapley afin d'estimer la contribution des différents facteurs de risque à la probabilité qu'un individu ait le diabète. Dans un premier temps, nous expliquons le modèle probit en détail ; nous utiliserons par la suite STATA pour effectuer nos estimations probit. Ensuite, nous énonçons de manière détaillée le calcul de la valeur de Shapley, que nous allons coder à l'aide de MATLAB afin d'obtenir des estimés. Nous appliquons ces deux procédures sur la même base de données, et comparons alors les résultats obtenus selon les deux procédures.

2.1 Procédure d'estimation

2.1.1 Modèle Probit

Nous utilisons le modèle présenté par Bourbonnais (2015) que nous adaptons plus spécifiquement à notre étude.

Dans un modèle de choix binaire, nous cherchons à modéliser une alternative (y_i = 0 ou 1) et donc à estimer la probabilité P_i associée à l'événement ($y_i = 1$).

Les variables latentes sont une première réponse aux problèmes liés à l'utilisation des MCO, dans un modèle dont la variable à expliquer est dichotomique. Cette variable latente est une variable continue non observable et représentative du phénomène étudié, par exemple considérons dans notre étude la « *propension à avoir le diabète* » comme variable

latente, seulement nous ne pouvons observer que l'événement *« l'individu a le diabète ou non »*.

De ce fait cette variable latente nous permet de poser deux conditions exprimées sous la forme suivante :

1) L'individu *i* est porteur de la maladie si sa propension à avoir le diabète est $y_i = 1$ soit $y_i^* > 0$
2) y_i^* est une fonction linéaire des x_i, $y_i^* = a_0 + a_1 x_i + \varepsilon_i$. (1)

La variable à expliquer binaire y_i est alors définie par le modèle de décision suivant :

$$\begin{cases} y_i = 1 \text{ si } y_i^* > 0 \text{ , } L'individu\ i\ a\ le\ diabète \\ y_i = 0 \text{ si } y_i^* \leq 0\ L'individu\ i\ n'a\ pas\ le\ diabète \end{cases} \quad \text{où } y_i^* = ax_i + \varepsilon_i \quad (2)$$

Avec :

y_i^* : Variable endogène représentant l'état de santé de l'individu ;

x_i : Vecteur des variables exogènes comprenant les facteurs de risque spécifiques de la maladie (High_Blood, High_BMI, Smoker, High_cholesterol, No_Physical_Activity, Sex);

α : Vecteur des coefficients des variables exogènes ;

ε_i : Terme d'erreur

Intuitivement la règle de décision consiste à supposer simplement que la proportion des (y_i= 1) est élevée pour $a_0 + a_1 x_i + \varepsilon_i > 0$.

Soit P_i la probabilité que $y_i^* > 0$.

$$P_i = Prob\ (y_i = \ 1) = Prob\ (y_i^* > 0) = Prob\ (a_0 + a_1 x_i + \varepsilon_i > 0)$$

$$= Prob\ (\ \varepsilon_i > -(a_0 + a_1 x_i\ )) \qquad (3)$$

Si la distribution de ε_i est centrée par rapport à la moyenne, nous avons l'équivalence :

$$Prob\ (\ \varepsilon_i > -(a_0 + a_1 x_i\ )) = \ Prob\ (\ \varepsilon_i < a_0 + a_1 x_i\ )$$

Soit $P_i = Prob\ (y_i = \ 1) = \ Prob\ (\ \varepsilon_i < a_0 + a_1 x_i\ )$

La probabilité P_i dépend ainsi de la distribution du terme de l'erreur ε_i du modèle de décision. Dans le modèle Probit, la fonction de répartition de l'erreur ε_i suit une loi normale et est donnée par :

$$P_i = F(x_i a) = \ \phi(x_i a) \ = \int_{-\infty}^{a_0 + a_1 x_i} \frac{1}{\sqrt{2\pi}} e^{-t^{2/2} dt} \qquad (4)$$

Il s'agit d'une loi normale centrée et réduite $N\,(0,1)$.

L'estimation des paramètres du modèle est effectuée à l'aide des algorithmes de maximisation de la fonction de log-vraisemblance.

La vraisemblance s'écrit :

$$L(\alpha) = \prod_{i=n}^{N} F(x_i a)^{y_i} \ (\ 1 - F(x_i a))^{1-y_i} \ . \qquad (5)$$

D'où la log-vraisemblance est définie par :

$$\log L(\alpha) = \sum_{i=1}^{N} y_i \log F(x_i a) + \sum_{i=1}^{N} (1 - y_i) \log\big(1 - F(x_i a)\big)$$

$$= \sum_{i:y_i = 1} \log F(x_i a) + \sum_{i:y_i = 0} \log\big(1 - F(x_i a)\big). \qquad (6)$$

Cependant contrairement aux modèles linéaires estimés par la méthode des MCO, les coefficients obtenus à l'aide du modèle Probit ne peuvent pas être directement interprétés économiquement car ne correspondant pas encore à une propension marginale. La seule interprétation initiale possible est basée sur le signe des coefficients qui indique soit une relation positive ou négative montrant comment cette variable explicative impacte la probabilité P_i. En vue d'atteindre notre objectif celui étant d'évaluer la contribution marginale de chaque facteur de risque, nous devons déterminer la valeur des effets marginaux[1] de notre estimation Probit. Cette estimation sous STATA_14 s'effectue simplement à l'aide de la commande *« margins »* est une post-estimation c'est-à-dire devant s'effectuer après la régression.

3.1.2 Valeur de Shapley

Dans cette section nous appliquons la formule de Pongou et Tondji (2017), de manière à obtenir les valeurs de Shapley de nos facteurs de risque.

Nous considérons un environnement de production incertain modélisé par un tuple tel que :

$P = (N, (T_i)_{i \in N}, f, \pi)$ (1)

Où

$N = \{1, 2, 3, 4, 5, 6\}$, représente l'ensemble des six facteurs de risque de choisi ;

[1] De combien varie la probabilité P_i lorsqu'il y'a un changement d'une unité dans un facteur de risque spécifique x_i les autres étant supposées constantes ?

$T_i = \{0,1\}$, représente l'ensemble des valeurs numériques que peut prendre le facteur *i*. En effet, chacun des facteurs de risque étant une variable binaire dans notre étude (avoir ce facteur de risque ou non), il n'existe donc que deux valeurs numériques (0 et 1) possibles ;

f = est la fonction de production qui relie un profil de facteur de risque par exemple, la probabilité d'avoir le diabète ;

$x = (x_1 \ , \dots , x_n) \ \in T = \prod_{i=1}^{n} T_i$ au résultat d'un nombre réel $f(x)$;

Dans notre travail, n =6 donc $x = (x_1 \ , \dots , x_6)$, définit l'univers des combinaisons possibles pour les six facteurs, Nous obtenons donc l'univers des combinaisons possibles par le calcul de $2^6 = 64$; il existe en tout 64 profils de répartition des facteurs de risque tous différents.

$\pi = (\pi_i)_{i \epsilon N}$, est un profil de probabilité de distribution des facteurs de risque.

Pour chaque profil de risque *i* nous avons :

$$\pi_i \colon T_i \to [0, \quad 1]$$

$$x_i \mapsto \pi_i \ avec \sum_{x_i \in T_i} \pi_i \ (x_i) = 1.$$

Nous posons cette hypothèse que $f(0,0,0,0,0,0) = 0$, ce qui signifie que pour le premier des 64 profils obtenus, ne comportant aucun des facteurs de risque, nous restreignons la probabilité d'avoir le diabète à 0. Il n'a pour ainsi dire aucune répercussion dans l'explication de la maladie. Cette hypothèse est fondée sur les travaux de Shapley (1953), où le profil nul n'a alors aucune contribution et donc ne reçoit aucune valeur lors de la répartition du résultat.

$\pi(x)$, est la probabilité de réalisation d'un profil de facteur de risque $x \in T$.

Nous posons que les facteurs de risque sont indépendamment distribués et la probabilité de réalisation d'un profil de facteur de risque $x \in T$ est donnée par la formule suivante :

$\pi(x) = \prod_{i \in N} \pi_i \ (x_i)$.

Avant de présenter notre formule de calcul il est nécessaire de définir les principaux éléments qui la compose.

Définition 1

Soit $P = (N, (T_i)_{i \epsilon N}, f, \pi)$, l'environnement de production incertain et deux facteurs *i* et *j* qui sont dits symétriques de manière probabiliste si : (1) $\pi_i = \pi_j$; et (2) pour tout profil de facteurs x tel que $x_i = 0$ ou $x_j = 0$, $f(x) = f(\tau_{ij}(x))$, où $\tau_{ij}(x) = x - x_i e_i + x_j e_i - x_j e_j + x_i e_j$, avec $e_k = (0, \dots, 0, 1, 0, \dots .0), k \in \{i, j\}$, étant un vecteur ligne à unité *n*-composants, dont le $k^{ième}$composant prend la valeur un (1) et tous les autres composants la valeur zéro (0).

Deux facteurs de risque sont symétriques de manière probabiliste s'ils ont la même probabilité de distribution et si même lorsqu'ils sont permutés leur niveau de distribution ne change pas.

Définition 2

Une procédure d'évaluation est une fonction ϕ qui relie tout environnement de production $P = (N, (T_i)_{i \epsilon N}, f, \pi)$, au vecteur $\left(\emptyset_i(P)\right)_{i \in N}$, où $\emptyset_i(P)$ est un nombre réel représentant la valeur du facteur de risque *i*.

Par soucis de simplification nous remplaçons pour la suite de notre travail $\emptyset_i(P)$ par $\emptyset_i(f)$.

Pongou et Tondji (2017) construisent l'algorithme pour la valeur de Shapley sur les trois axiomes suivants que nous adaptons à notre étude.

Axiome 1 (*Symétrie probabiliste*)

Une procédure d'évaluation est symétrique de manière probabiliste si pour tout environnement de production $P = (N, (T_i)_{i\epsilon N}, f, \pi)$ et tout facteur de risque *i* et *j* symétriques de manière probabiliste, on a $\emptyset_i(f) = \emptyset_j(f)$.

Axiome 2 (*Efficience probabiliste*)

Une procédure d'attribution de valeurs ϕ est efficiente de manière probabiliste si pour tout environnement de production $P = (N, (T_i)_{i\epsilon N}, f, \pi)$, on a $\sum_{i\epsilon} \emptyset_i(f) = \sum_{x\epsilon} \pi(x) f(x)$.

Axiome 3 (*Principe de marginalité*)

Une procédure d'évaluation ϕ satisfait au principe d'axiome de marginalité si pour tout environnement de production $P = (N, (T_i)_{i\epsilon N}, f, \pi)$ et $P' = (N, (T_i)_{i\epsilon N}, g, \pi)$, tous les facteurs de risque $i \in N$ et tout $x, a \in T$ on a tel que $a \lhd_0^i x$, $[f(a + x_i e_i) - f(a) \geq g(a + x_i e_i) - g(a)] \Longrightarrow [\emptyset_i(f) \geq \emptyset_i(g)]$.

Où

- $a, x \in T, [a \trianglelefteq x]$ si et seulement si $[a_i \neq x_i \Longrightarrow a_i = 0]$;
- $[a \lhd x\ ]$ si et seulement si $[a \trianglelefteq x\ et\ a\ \neq x\ ]$;
- $i \in N\ et\ a, x \in T, a \lhd_0^i\ x$ signifie $a \lhd x$ et $a_i = 0$.

Cette écriture mathématique, nous dit que a est aussi un profil de facteurs de risque qui est considéré ici comme un sous-groupe du profil x dont les éléments sont inclus dans l'ensemble des éléments du profil x. Ce qui revient à dire de manière spécifique pour notre travail, que tous les éléments de a sont soient égale à 0 ou à 1, mais jamais supérieurs à ces deux valeurs. Ce nouveau profil de risque défini a, est essentiel car il permet d'obtenir la contribution marginale. Autrement dit, ce profil peut être considéré comme le profil de base $(a_i = 0)$, avant le profil généré par le « jeu ».

En outre, la spécification de ces trois fondements, correspond en premier lieu au fait que l'axiome de symétrie probabiliste est vérifié lorsque la probabilité interchangeable des facteurs de risque, possède la même valeur a priori. Quant à l'efficience probabiliste, elle précise que le résultat espéré de notre probabilité d'avoir le diabète doit être totalement partagée entre les différents facteurs de risque. Nous utiliserons cet axiome comme vérification de la véracité de nos résultats. D'après sa définition nous devons obtenir l'égalité entre la somme des valeurs de Shapley de nos six facteurs de risque et la somme de l'espérance mathématique d'être atteint du diabète. Pour terminer, le principe de marginalité prévoit que s'il y a un changement dans la technologie qui accroit la productivité marginale d'un facteur de risque, alors la procédure d'attribution des valeurs doit également accorder une valeur supplémentaire à ce facteur de risque sous cette nouvelle technologie.

Conséquemment, Pongou et Tondji (2017) démontrent dans leur théorème de la valeur de Shapley qu'il existe une et seule procédure d'évaluation qui satisfait simultanément les trois axiomes de symétrie probabiliste, d'efficience probabiliste et du principe de marginalité. Cette procédure d'évaluation est définie pour tout environnement de production incertain $P = (N, (T_i)_{i\in N}, f, \pi)$ et pour tout facteur de risque $i \in N$ par :

$$ASV_i(P) = \sum_{x\in T} \pi(x) \left\{ \sum_{a \lhd_0^i x} \frac{(|a|)!\,(|x| - |a| - 1)!}{(|x|)!} [f(a + x_i e_i) - f(a)] \right\}$$

Ce processus d'évaluation représente la valeur a priori de Shapley

Avec :

- $|x| = |\{i \in N : x_i > 0\}|$, le nombre de profil de facteurs de risque actif, à savoir différent du profil ne comportant pas de facteurs à risque pris en valeur absolue donc supposé strictement positif ;

- $(|x|)!$, ensemble des permutations possibles des différents profils de facteurs de risque ;

- $a + x_i e_i$, le nouveau profil de facteurs de risque lorsque le « jeu » est réalisé ;

- $[f(a + x_i e_i) - f(a)]$, la contribution marginale du facteur indexé i ;

- $\sum_{a \lhd_0^i x} \frac{(|a|)!(|x|-|a|-1)!}{(|x|)!}[f(a + x_i e_i) - f(a)]$, la contribution marginale espérée du facteur de risque i dans l'univers de l'ensemble des profils T.

La création du code permettant l'exécution de cette formule a nécessité deux étapes détaillées en annexes (fichier 1 et fichier 2). Premièrement, dans le fichier de base servant de support à notre formule, il a fallu créer nos variables ; par la suite, définir les probabilités des évènements « avoir le facteur de risque *i*, mais ne pas être atteint de diabète » et son évènement contraire. Puis, répertorier tous les différents profils et les classer selon leurs similitudes. La phase finale de cette première étape consistait à définir la probabilité des personnes atteintes de diabète dans chaque différent profil.

La seconde partie a consisté à définir chaque paramètre dans le théorème, ensuite procéder par emboitement en créant tous les éléments intérieurs des sommes pour les réunir en dernier lieu sous la même formule, les valeurs s'obtiennent après avoir exécuté le premier fichier et celui de la formule enfin.

2.2 Base de données

2.2.1 Données

La présente étude exploite une base de données sur le diabète construite à partir d'une enquête téléphonique[2] auprès des individus résidants dans les cinquante états des États Unis d'Amérique, et dans deux de ses territoires non incorporés (Guam et Porto Rico). Ces données collectées par le Système de surveillance des facteurs de risques[3]

[2] Collecte de données obtenue par ligne téléphonique fixe et mobile
[3] Behavioral Risk Factor Surveillance System (BRFSS) « en anglais »

comportementaux pour le compte des Centres pour le contrôle et la prévision de la maladie[4] des États-Unis, regroupent les informations pour les années 2015 et 2016. Ce système initié en 1984, est le pionner des systèmes nationaux d'enquêtes téléphoniques pour le champ de la santé. Débutant seulement avec quelques états, il contient aujourd'hui les informations des résidents de tous les États-Unis, sur les comportements à risque, les maladies chroniques et l'utilisation des services préventifs. Si bien qu'il représente le plus grand système de sondage sur la santé en continu dans le monde et est un outil optimal pour élaborer des politiques sanitaires. Les données de notre étude comprennent des renseignements sur la situation socioéconomique, démographique, et la santé des ménages. L'analyse de la cohorte des 441 456 individus dont le nombre passe après exclusion des données manquantes à 324 481, se fera avec MATLAB_2017 pour déterminer les valeurs de Shapley et STATA_14 pour l'estimation de notre modèle économétrique.

3.2.2 Variables

Nous définissons notre variable dépendante « diabète » qui représente le fait d'avoir la maladie ou non. Cette maladie chronique selon l'OMS (2017) est observée lorsque le taux de glycémie à jeun est supérieur à 1,26g/L ou supérieur à 2g/L à tout moment de la journée. Nous incluons dans cette notion les trois types de diabète qui sont : le diabète de type 1, le diabète de type 2 et le diabète gestationnel.

✓ *Variables indépendantes*

a) Hypertension artérielle (High_blood)

D'après les travaux d'Ohlson et al. (1988), qui ont trouvé que ce facteur était significatif dans l'explication du diabète chez les hommes Suédois nés en 1913. Nous définissons cette variable comme un facteur de risque de la maladie et représentant tous les individus ayant répondu oui à la question « *Un professionnel de la santé vous a t il déjà signifié que vous aviez une tension artérielle élevée*[5]*?* » avec les mêmes niveaux de seuil que ceux choisi

[4] Centers for Disease Control and Prevention (CDC) « en anglais »

[5] Tension artérielle systolique d'un adulte ≥ à 140 mm Hg et/ou tension artérielle diastolique ≥ à 90 mm Hg (OMS, 2017)

pour la variable diabète. Nous espérons comme dans l'étude précitée un signe positif et signification de la part du « *High Blood* ».

b) Indice de masse corporelle (IMC) élevé (High_BMI)

Cet outil de mesure est un indice anthropométrique aussi appelé « indice de Quételet », inventé par le mathématicien belge Adolphe Quételet et permet de classifier les individus en intervalles standards (maigreur, indice normal, surpoids, obésité) selon deux critères : la masse corporelle et la taille. Nous considérons dans la construction de cette variable les individus en surpoids et obèses c'est à dire ceux dont le IMC ≥ 25(, en accord avec les résultats trouvés par Chan et al. (1994) qui déterminent ce facteur de risque comme déterminant pour les hommes professionnels de la santé aux États-Unis. Comparés à ceux dont l'IMC < 23, ces personnes avaient au moins 2,2 plus de chance de contracter le diabète. Nous attendons pour notre étude à une contribution marginale significative et importante de la part de ce paramètre.

c) Tabagisme (Smoker)

Manson et al. (2000), dans leurs travaux réussissent à prouver le tabagisme est un facteur augmentant le risque de contracter le diabète. Ils établissent quatre sous-groupes pour cette variable : fumeur journalier ≥ 20 cigarettes/jours, fumeur journalier < 20 cigarettes/jours, anciens fumeurs et hommes n'ayant jamais fumé. Les résultats des trois premières catégories, sont significatifs comparés aux individus n'ayant jamais fumé[6]. Nous dissocions dans notre population les fumeurs actuels et les anciens fumeurs d'avec ceux qui n'ont jamais fumé.

d) Hypercholestérolémie (High_cholesterol)

Le taux de cholestérol sanguin permet d'obtenir quatre types d'information sur premièrement le niveau total de cholestérol ensuite sur les lipoprotéines de faible densité[7], lipoprotéines de haute densité[8] et le triglycéride (NIH, 2015). Selon l'Association

[6] Sont également assimilé comme n'ayant jamais fumé les hommes ayant abandonné la cigarette depuis au moins 10 ans.
[7] Low density lipoprotein (LDL) en anglais
[8] High density lipoprotein (HDL) en anglais

canadienne de diabète (ACD, 2014), un haut niveau du taux de cholestérol (≥ 240 mg/DL), est répertorié comme un facteur de risque du diabète de type 2. Nous considérons donc cette variable dans notre étude afin de déterminer son apport marginal à la maladie.

e) Inactivité physique (No_Physical_activity)

D'après l'OMS (2017) la non-pratique d'activité physique et un excès pondéral occasionne le diabète de type 2. Nous retirons l'information de cet élément à la question : « Durant le mois dernier, excepté votre occupation professionnelle, avez-vous participé à un type d'activité physique tel que la course, la gymnastique, le golf, le jardinage ou la marche ?», de notre base de données qui sera utilisé comme variable explicative.

f) Sexe (Sex)

Selon Jenum et al. (2005) dans leur étude sur facteurs de risques de la maladie chez les occidentaux, montrent qu'être un homme accroit le risque d'avoir le diabète. Nous étudierons donc dans notre travail l'apport de cette variable.

3. RESULTATS

3.1 Statistiques descriptives

Les résultats de l'analyse préliminaire de nos données consignés dans le tableau 2, nous indique que nous avons en moyenne 14% de la population qui souffrent de diabète. Environ 41 % des personnes ont une tension artérielle supérieur à 140 mm Hg. De plus, un grand nombre approximativement 66% sont obèses ou en surpoids. Le pourcentage de personnes ayant déjà fumé au étant présentement fumeur est de 43,4%. Un peu plus de 42% de cette population a un niveau de « mauvais » cholestérol élevé. Il existe de nombreuses personnes, 73% qui pratiquent une activité physique, c'est une population plutôt active. Enfin, nous remarquons que la plupart des individus sont de sexe féminin soit 57%.

3.2 Modèle Probit

Les résultats de l'estimation de la régression obtenus avec le modèle Probit sont consignés dans le tableau 3. L'hétéroscédasticité est contrôlée avec la commande « *robust* » préférée à la méthode du « *cluster »* car même si les individus sont porteur facteur de risque ou pas, ils font face à différentes réalités, et ne sont pas sujet aux mêmes variables inobservées. Toutefois, les résultats présentés ne peuvent pas être interprétés pour l'instant, en termes d'effet marginal. En effet, nous ne pouvons pas interpréter ces valeurs obtenues relativement à leur magnitude, mais plutôt nous pouvons discuter du signe des coefficients et vérifier leur significativité.

De ce fait, nous observons que l'hypothèse nulle de la non-significativité des paramètres peut être rejetée pour tous les paramètres et on conclut que tous les facteurs de risque sont significatifs ($p<0,05$) au seuil de 5%. De plus, l'analyse du signe des coefficients estimés révèle qu'il existe une relation positive entre l'hypertension artérielle et la probabilité d'avoir le diabète. Il en est de même pour fait d'être en surpoids ou encore de fumer. Avoir un haut niveau de cholestérol dans le sang, tout comme ne pas pratiquer d'activité physique évoluent dans le même sens que la prévalence du diabète. Par contre, être de sexe masculin évolue négativement avec la probabilité d'avoir le diabète.

Les effets marginaux de notre modèle sont présentés dans le tableau 4 et 5. Les variables étant toutes significatives au seuil de 5%, nous pouvons dire qu'elles aident à prévoir la probabilité d'avoir le diabète. Grâce aux effet marginaux calculés, il est désormais possible de tenir compte de la valeur en magnitude de nos coefficients estimés. Nous avons privilégié dans notre travail la valeur de l'effet marginal moyen[9], à la place celle donnée par les moyennes de l'échantillon de tous les facteurs de risque ou en fonction d'un individu dans la population pris comme référence.

Les deux dernières méthodes ont été écartées parque que respectivement la moyenne d'une variable explicative binaire a très rarement une interprétation utile et une analyse basé l'individu de référence n'est pas le but de notre étude. L'effet marginal moyen est aussi

[9] Donné par la commande *« margins, dydx (*) et mfx »*

une méthode de calcul recommandée par Wooldridge (2002) permettant de déterminer l'effet partiel pour l'individu moyen dans l'échantillon.

La valeur de l'effet marginal est obtenue suite à l'estimation de notre modèle et est donnée par la commande *« mfx »*.

Tout d'abord, nous observons que la variation d'une unité dans notre variable muette relative à l'hypertension artérielle, augmente la probabilité d'avoir le diabète environ 13%. Ce qui veut dire que pour un individu qui n'avait initialement pas l'hypertension artérielle, développer ce facteur de risque (passer à la valeur 1), augmente sa probabilité d'être atteint du diabète par 13%. De même, pour une personne qui était auparavant considérée comme comme maigre ou de poids normal, le fait de passer en surpoids ou être obèse, accroit de 8% sa probabilité d'être diabétique. En appliquant le même raisonnement analogue, devenir fumeur, être hypercholestérolémie, et ne pas pratiquer d'activité physique augmentent respectivement d'environ 1,5%, 8% et 5% la probabilité d'avoir le diabète. Enfin au niveau de l'estimation de résultat nous remarquons que la variable sexe est le seul facteur de risque en termes de contribution marginale, qui associé avec une réduction du risque d'avoir le diabète d'une valeur de 0,69%. Ce résultat pourrait provenir, de la présence de femmes ayant déjà développé du diabète gestationnel par le passé, lorsqu'on sait que selon l'OMS (2016), ce facteur augmente le risque d'être atteint de diabète.

Afin de valider la qualité d'ajustement de notre modèle probit, nous estimons le pourcentage de probabilité qu'il prédit. Les résultats sont donnés par le tableau 6.

On observe que la moyenne estimée avec le modèle probit, 15, 26% s'écarte légèrement de la vraie moyenne de 13, 81% de personnes ayant le diabète dans notre population.

3.3 Valeur de Shapley

Le tableau 7, comprend l'ensemble des valeurs de Shapley appliquées à notre étude sur le diabète. Ces valeurs constituent la contribution marginale de chaque facteur de risque à la probabilité d'avoir le diabète estimé à 15,71%. Les contributions marginales pour les facteurs de risque avoir une pression sanguine élevée (5,1%), et une valeur de BMI

supérieure à 25 (5,5%) sont quelque peu similaires. Elles représentent la part qu'ajoute chacun de ces facteurs à probabilité totale d'être diabétique. L'hypertension artérielle serait responsable de 5.1 points dans la maladie tout comme l'obésité ou le surpoids qui apporte 5.5 points. De même, fumer, ou encore avoir un taux de cholestérol élevé ajoute respectivement d'environ 1% et 3% dans l'apparition du diabète. Ne pas pratiquer d'activité physique représente un peu moins de 2% dans le risque de développer le diabète. Pour terminer, être un homme réduit de 0,1% la prévalence du diabète. Ces valeurs de Shapley sont toutes différentes les unes des autres, ce qui sous-tend qu'il n'y a pas de facteur symétrique dans notre étude et que chacun facteur augmente ou réduit de manière unique la probabilité d'être atteint de diabète.

3.4 Comparaison des deux méthodes

Les résultats du modèle probit et de la méthode de Shapley indiquent que les coefficients des facteurs de risque, conservent tous uniformément leur signe (négatif ou positif) suivant les deux estimations. Cette première similarité, démontre le sens unique de variation des covariables relativement à la variable dépendante. Les facteurs de risques responsables de l'augmentation ou diminution du risque de diabète maintiennent leurs propriétés dans les deux estimations. Toutefois, la magnitude des effets n'est pas la même. Premièrement au niveau de la régression économétrique, le modèle probit affiche une variation moins importante des effets marginaux, il n'existe pas de grandes disparités entre les valeurs. L'ordre de contribution décroissant des effets marginaux est le suivant : l'hypertension artérielle, le surpoids ou obésité, l'hypercholestérolémie, l'inactivité physique, le tabagisme et enfin le genre. En second lieu, la méthode de Shapley, présente des valeurs plus dispersées avec deux facteurs majeurs : l'hypertension artérielle et le surpoids ou obésité qui représentent les deux tiers des contributions au développement du diabète. Les quatre autres facteurs de risque se partageant ainsi le tiers restant dans la probabilité d'avoir le diabète. Cependant, la méthode de Shapley comparativement au modèle probit, ne possède aucun outil permettant d'attribuer un « poids » relatif à chaque facteur pour évaluer sa significativité. Elle tient arbitrairement compte des facteurs que le chercheur estime

déterminants, et tente d'assigner pour chacun d'eux leur contribution à la réalisation de l'évènement.

CONCLUSION

Cette étude a servi de cadre « test » à l'implémentation de la valeur de Shapley telle que proposée par Pongou et Tondji (2017). Elle permet tout d'abord, une critique de l'économétrie moderne face à l'introduction d'une nouvelle approche d'estimation de la contribution marginale. L'application de la méthode de Shapley et du modèle probit devant le problème de détermination de l'effet marginal des facteurs de risque dans le cadre du diabète conduit à des résultats différents.

Les estimés obtenus de la méthode de Shapley suggèrent une corrélation de l'hypertension artérielle et du surpoids ou obésité dans la probabilité d'être atteint de diabète. Tandis que les résultats de l'estimation probit indiquent que les contributions marginales des facteurs de risque sont sensiblement rangées dans le même ordre de grandeur. On peut supposer que l'estimation probit présente de tels résultats à cause de la multicolinéarité qui existe être les facteurs de risque. En effet, être obèse peut être associé à de l'hypertension artérielle ou même l'hypercholestérolémie tout comme, la non pratique d'activité physique reliée au surpoids ou à l'obésité.

Pour conclure, la différence des résultats de ces deux méthodes laissent supposer certaines insuffisances méthodologiques. C'est le cas de mentionner la significativité des facteurs de risque comme une des faiblesses de la méthode de Shapley pouvant induire des estimés sans réel impact économique. De plus, une des hypothèses fortes de la méthode de Shapley est la contribution du joueur nul égalisé à zéro. Par exemple, un individu qui ne participe pas à une tâche donnée, nous comprenons aisément que sa production à ce travail est nulle. Cependant, pour notre étude avec les facteurs de risque, un individu ne possédant aucun facteurs choisis, peut tout de même être diabétique. Dans ce cas, un forçage d'exclusion s'avère nécessaire afin de ne pas tenir compte des personnes concernées. Par conséquent,

l’introduction cette nouvelle méthode aux études économiques ne nécessite-t-elle pas des prérequis structurels afin d’accroitre son efficacité et sa portée ?

BIBLIOGRAPHIE

1. ACD (2014). “Guidelines diabetes” Association canadienne du diabète.
2. Bhargava Alok, (2003). “A longitudinal analysis of the risk factors for diabetes and coronary heart disease in the Framingham Offspring Study.” *Population Health Metrics*, 1(3), 1-10.
3. CDC (2017), Récupéré de : www.cdc.gov/brfss/about/index.htm consulté le 28 Juillet 2017.
4. Costa-Font, J. et Gil J., (2005) “Obesity and the incidence of chronic diseases in Spain: A seemingly unrelated Probit approach.” *Economics and Human Biology*, 3 (2005), 188-214.
5. Chan, J. M., Rimm, E. B., Colditz, G. A., Stampfer, M. J., et Willett, W. C “(1994). Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men.” *Diabetes care*, 17(9), 961-969.
6. Dickerson, J.B., Smith, L. M., Sosa E., Mckyer L.E., et Ory, M.G. (2011) “Perceived risk of developing diabetes in early adulthood: Beliefs about inherited and behavioral risk factors across the life course.” *Journal of Health Psychology*, 17 (2), 285-296.
7. D’Souza S.M., Amirtharaj, A., Venkatesaperumal, R., Isac, C., et Maroof, S. (2013) “Risk-assessment score for screening diabetes mellitus among Omani adults.” *SAGE Open Medicine*, 0 (0), 1-8
8. Forouhi, G.N. et Wareham, N.J. (2014) “Epidemiology of diabetes”. *Medicine*, 42 (12) 698-702.
9. Frisch Ragnar (1933) “Editor’s Note.” *Econometrica*, 1(1), 1-4.

10. Greene William (2011). *Économétrie 7ème édition*. Paris : Pearson Éducation France.

11. Hegazi, R., El-Gamal M., Abdel-Hady, N., et Hamdy, O. (2015) "Epidemiology and Risk factors for Type 2 Diabetes in Egypt." Annals of Global Health, 81(6) 814-820.

12. Jebeli, S.H., Barouni, M., Orojloo H.P. et Mehraban S. (2014) "Estimating the Marginal Effect of Socioeconomic Factors on the Demand of Specialty Drugs." *Global Journal of Health Sciences*, 7 (2).

13. Jenum, A.K., Holme, I., Graff-Iversen, S., et Birkeland, K. (2005) "Ethnicity and sex are strong determinants of diabetes in an urban Western society: implications for prevention." *Diabetologia*, 48(3), 435-439.

14. Manson, J.E., Ajani, U.A., Liu, S., Nathan, D.M. et Hennekens, C.H. (2000) "A prospective study of cigarette smoking and the incidence of diabetes mellitus among US male physicians." *The American Journal of Medecine,*109 (7), 538-542.

15. Montgomery, M., et Ewell, P. (2015) "The presence of Risk factors for Type 2 Diabetes Mellitus in Underserved." *Nursing Clinics of North America*, 50(2015), 585-594.

16. Murad, A. M., Abdulmageed, S.S., Iftikhar R. et Sagga, B.K. (2014) "Assessment of the Common Risk Factors associated with type 2 Diabetes Mellitus in Jeddah." *International Journal of Endocrinology*, (2014), 1-10

17. Ohlson, L-O., Larsson, B., Bjorntorp, P., Eriksson, H., Svärdsudd, K., Welin, L., Tibblin, G. et Wilhelmsen, L. (1988) "Risk factors for type 2 (non-insulin-

dependent) diabetes mellitus: thirteen and one-half years of follow-up of the participants in a study of Swedish men born in 1913." *Diabetologia* 31 (1988) 798-805.

18. OMS (2016) "*Rapport mondial sur le diabète*." Genève, Suisse: Organisation mondiale de la Santé.

19. OMS (2017), Récupré de www.who.int/diabetes/action_online/basics/fr/index2.html Consulté le 28 Juillet 2017.

20. Pongou, R., et Tondji, J-A., (2017) "Valuing inputs under supply uncertainty: The Bayesian Shapley value." *Working Paper,* University of Ottawa.

21. Stan Lipovetsky (2006), "Entropy Criterion in logistic regression and Shapley value of predictors". *Journal of Modern Applied Statistical Methods*, 5(1), 95-106.

22. Shapley, L. S. (1953). "A value for n-person games." *Annals of Mathematical Studies*, H. Kuhn and A. Tucker editors 28, 307–317.

23. Stock, J.H. et Watson, M.W. (2015). *Introduction to econometrics third edition*. Edinburgh: Pearson Education Limited, chp 1. pp 47.

24. Tarashev, N., Tsatsaronis, K. et Borio, C. (2015) "Risk Attribution Using the Shapley Value: Methodology and Policy Applications." Re*view of Finance,* (2016) 1189-1213.

25. Tol Richard S.J. (2011) "Shapley values for assessing research production and impact of schools and scholars." *Scicentometrics*, 90 (2012), 763-780.

26. Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press, 2002.

ANNEXES

Tableau 1: Codification des variables

Facteur ou Maladie	Nom de la variable	Code
Diabète de type 2	Diabetes	Oui diabete3=1 Pas de diabete3=0
Hypertension artérielle	High_Blood	Oui bphigh4=1 non bphigh4=0
IMC élevé	High_BMI	Oui _bmi5cat=1 non _bmi5cat=0
Tabagisme	Smoker	Oui_smoker3=1 non _smoker3=0
Hypercholestérolémie	High_cholesterol	Oui toldhi2=1 non toldhi2=0
Inactivité physique	No_Physical_activity	Oui exerany2=0 non exerany2=1
Sexe	Sex	Homme sex=1 femme sex=0

Source : *Réalisé par l'auteur à partir des données du CDC (2015)*

Tableau 2: Statistiques descriptives

Variables	Fréquence	Moyenne	Écart-Types	Pourcentage
Diabete = 0	379 794	0,1381	0, 3450	86, 19
Diabete = 1	60 864			13, 81
High Blood=0	258 630	0,4123	0, 4922	58,77
High Blood=1	181 459			41, 23
High_BMI=0	138 130	0,6589	0, 4740	34, 10
High_BMI=1	266 928			65, 90
Smoker=0	239 608	0,4341	0, 4956	56.58
Smoker=1	183 858			43, 42
High_cholesterol=0	218 771	0,4223	0, 4939	57, 76
High_cholesterol=1	159 970			42, 24
No_Physical_activity=0	296 020	0,2663	0, 4420	73, 37
No_Physical_activity=1	107 444			26, 63
Homme sex=1	186 938	0,4234	0, 4941	42, 35
Femme sex=0	254 518			57, 65

Source: *Réalisé par l'auteur à partir des données du CDC (2015)*

Tableau 3: Régression Probit

Variables	Coefficients estimés	Écart-types	P-Value
High_Blood	0.6198	0.0061	0.000**
High_BMI	0.4349	0.0071	0.000**
Smoker	0.0738	0.0058	0.000**
High_Cholesterol	0.3763	0.0059	0.000**
No_Physical_activity	0.2485	0.0062	0.000**
Sex	-0.0333	0.0058	0.000**
Constante	-1.9785	0.0079	0.000**
Nombre d'observations : 324 481			

Source: *Réalisé par l'auteur à partir des données du CDC (2015)*

**** : significatif au seuil de 1%* *** : significatif au seuil de 5%*

Tableau 4: Effets marginaux à la population moyenne du modèle Probit

Variables	Coefficients estimés	Écart-types	P-Value
High_Blood	0.1325	0.0013	0.000**
High_BMI	0.0837	0.0012	0.000**
Smoker	0.0154	0.0012	0.000**
High_Cholesterol	0.0804	0.0013	0.000**
No_Physical_activity	0.0558	0.0014	0.000**
Sex	-0.0069	0.0012	0.000**
Nombre d'observations : 324 481			

Source: *Réalisé par l'auteur à partir des données du CDC (2015)*

**** : significatif au seuil de 1%* *** : significatif au seuil de 5%*

Tableau 5: Effets marginaux du modèle Probit

Variables	Coefficients estimés	Écart-types	P-Value
High_Blood	0.1311	0.0010	0.000**
High_BMI	0.0809	0.0011	0.000**
Smoker	0.0151	0.0012	0.000**
High_Cholesterol	0.0789	0.0013	0.000**
No_Physical_activity	0.0541	0.0014	0.000**
Sex	-0.0067	0.0012	0.000**

Source: *Réalisé par l'auteur à partir des données du CDC (2015)*

**** : significatif au seuil de 1%* *** : significatif au seuil de 5%*

Tableau 6: Probabilité prédite

Variable	Observations	Moyenne	Écart-type	Minimum	Maximum
Diabète	440 658	0.1381	0.3450	0	1
PProbit	324 830	0.1526	0.1129	0.0221	0.4110

Source: *Réalisé par l'auteur à partir des données du CDC (2015)*

Tableau 7 : Estimation par la méthode de Shapley

Variables	Valeur de Shapley
High_Blood	0.0514
High_BMI	0.0551
Smoker	0.0093
High_Cholesterol	0.0292
No_Physical_activity	0.0131
Sex	-0.0009
Total	0.1571
Nombre d'observations : 324 481	

Source: *Réalisé par l'auteur à partir des données du CDC (2015)*

CODE MATLAB

- ✓ Étape 1 : comprenant la spécification des variables et calculs initiaux, servant de base au fichier de la formule.

FICHIER N°1

```
% Read data
D = csvread('base.csv', 1, 0);

column_names = {'Diabetes', 'High_blood', 'High_BMI', 'Smoker',
'High_Cholesterol', 'No_Physical_Activity', 'Sex'};

%% Diabetes
i = 1;

% Keep only unambiguous answers to diabetes question (1 and 2= yes,
3and 4 = no)
v = (D(:, i) == 1) | (D(:, i) == 2)| (D(:, i) == 3)| (D(:, i) == 4);
D = D(v, :);

% Change 3 and 4 to 0 for diabetes
v = (D(:, i) == 3)|(D(:, i) == 4);
D(v, i) = 0;

%% High_blood
i = 2;

% Filter by blood pressure (1 and 2 = high, 3 and 4 = not high)
v = (D(:, i) == 1) | (D(:, i) == 2)| (D(:, i) == 3)| (D(:, i) == 4);
D = D(v, :);

% Change 3 and 4 to 0 for blood pressure
v = (D(:, i) == 3)| (D(:, i) == 4);
D(v, i) = 0;

%% High_BMI
i = 3;

% Filter by BMI (1 and 2 = not high, 3 and 4 = high)
v = (D(:, i) == 1) | (D(:, i) == 2) | (D(:, i) == 3) | (D(:, i) == 4);
D = D(v, :);

% Change 1 and 2 to 0 for BMI
v = (D(:, i) == 1) | (D(:, i) == 2);
D(v, i) = 0;

% Change 3 and 4 to 1 for BMI
v = (D(:, i) == 3) | (D(:, i) == 4);
D(v, i) = 1;

%% Smoker
```

```
i = 4;

% Filter by smoker (1 and 2 and 3 = yes, 4 = no)
v = (D(:, i) == 1) | (D(:, i) == 2) | (D(:, i) == 3) | (D(:, i) == 4);
D = D(v, :);

% Change 2 and 3 to 1 for smoker
v = (D(:, i) == 2) | (D(:, i) == 3);
D(v, i) = 1;

% Change 4 to 0 for smoker
v = (D(:, i) == 4);
D(v, i) = 0;

%% High_cholesterol
i = 5;

% Filter by cholesterol (1 = yes, 2 = no)
v = (D(:, i) == 1) | (D(:, i) == 2 );
D = D(v, :);

% Change 3 to 0 for cholesterol
v = (D(:, i) == 2);
D(v, i) = 0;

%% No_Physical_activity
i = 6;

% Filter by physical activity (2= yes, 1 = no)
v = (D(:, i) == 1) | (D(:, i) == 2 );
D = D(v, :);

% Change 1 to 0 for physical activity
v = (D(:, i) == 1);
D(v, i) = 0;

% Change 2 to 1 for physical activity
v = (D(:, i) == 2);
D(v, i) = 1;

%% Sex
i = 7;

% Filter by Sex (1 = male, 2 = female)
v = (D(:, i) == 1) | (D(:, i) == 2);
D = D(v, :);

% Change 2 to 0 for sex
v = (D(:, i) == 2);
D(v, i) = 0;


%% Significance
% Determine incremental probabilities, contingency tables, and apply
Fisher
```

```
% test
prob_no = zeros(i - 1, 1);
prob_yes = zeros(i - 1, 1);

C = zeros(2, 2, i - 1);
fisher_h = zeros(i - 1, 1);
fisher_p = zeros(i - 1, 1);

for k = 2:i
    v_no = (D(:, k) == 0);
    num_no = length(v_no);
    v_diab = D(v_no, 1);
    num_diab = sum(v_diab(:));
    prob_no(k - 1) = num_diab / num_no;
    num_no_diab = num_no - num_diab;
    C(2, 1, k - 1) = num_diab;
    C(2, 2, k - 1) = num_no_diab;

    v_yes = (D(:, k) == 1);
    num_yes = length(v_yes);
    v_diab = D(v_yes, 1);
    num_diab = sum(v_diab(:));
    prob_yes(k - 1) = num_diab / num_yes;
    num_no_diab = num_no - num_diab;
    C(1, 1, k - 1) = num_diab;
    C(1, 2, k - 1) = num_no_diab;

    [fisher_h(k - 1), fisher_p(k - 1)] = fishertest(C(:, :, k - 1));
end

prob_incremental = prob_yes - prob_no;

%% Determine incidence of each profile
profile_incidence = zeros(2^(i-1), 1);
profile_num_diab = zeros(2^(i-1), 1);

[C_inc, ia, ic] = unique(D(:, 2:i), 'rows');

for k = 1:(2^(i-1))
    v_ind = find(ic == k);
    profile_incidence(k) = numel(v_ind);
    v_diab = D(v_ind, 1);
    profile_num_diab(k) = sum(v_diab(:));
end

profile_diab_prob = profile_num_diab ./ profile_incidence;
profile_prob = profile_incidence / sum(profile_incidence(:));
N = i - 1;
%% Save contingency tables and incidence
save('shapley_params.mat', 'profile_diab_prob', 'profile_prob', 'N');
csvwrite('results/Fisher Test Results.csv', [fisher_h, fisher_p]);
for k = 2:i
    csvwrite(['results/Contingency Table - ' column_names{k} '.csv'],
C(:, :, k - 1));
end
```

- ✓ Étape 2 : comprenant la formule de Shapley à exécuter après le fichier N°1 de base.

FICHIER N°2

```
%% Parameters

% Load number of risk factors (N), profile diabetes probabilities
% (for production function) and incidence (for probability
distribution)
% See file Shapley_base.m
load('shapley_params.mat');

%% Code

% Force the incidence of the null profile (all zeros) to zero
profile_diab_prob(1) = 0;

% Create universe T
T_num = 2^N;
T = zeros(T_num, N);
for k = 1:T_num
    T(k, :) = fliplr(de2bi(k - 1, N));
end

ASV = zeros(N, 1);
ASV_inner = zeros(N, T_num);
% Loop over all indices i to get each ASV
for i = 1:N

    for ix = 1:T_num

        % Get the current value of x
        x = T(ix, :);

        % Calculate |x| (number of non-zero elements in x)
        num_x = sum(x);

        % Get x_i * e_i
        e_i = zeros(1, N);
        e_i(i) = 1;
        xiei = x .* e_i;

        % Zero out the index i
        x(i) = 0;

        % Multiply x with the universe T to get allowable values of a
        x_mat = repmat(x, T_num, 1);

        A_values = T .* x_mat;
        A_values = unique(A_values, 'rows');

        % If there is a row equal to x itself, remove it
```

```
        x = T(ix, :);
        x_ind = find(ismember(A_values, x, 'rows'));
        A_values(x_ind, :) = [];

        % Calculate the inner ASV by looping over all values of a
        for ia = 1:size(A_values, 1)
            a = A_values(ia, :);

            % Calculate |a| (number of non-zero elements in a)
            num_a = sum(a);

            % Get a + x_i * e_i
            a_plus_xiei = a + xiei;

            % Calculate the term in the inner sum
            bin_coeff = factorial(num_a) * factorial(num_x - num_a - 1)
...
                / factorial(num_x);

            f_gain = f(a_plus_xiei, profile_diab_prob) - f(a,
profile_diab_prob);
            ASV_inner(i, ix) = ASV_inner(i, ix) + bin_coeff * f_gain;
        end

        % Multiply this inner sum by the probability and add the result
        % to the outer sum, which is the ASV
        ASV(i) = ASV(i) + pi(x, profile_prob) * ASV_inner(i, ix);

    end
end

% Check sums
disp(['Sum of ASVs: ' num2str(sum(ASV(:)))])
v = profile_diab_prob(:) .* profile_prob(:);
disp(['Sum of pi * f: ' num2str(sum(v(:)))])

return

%% Subfunctions

% Production function
function f_val = f(x, profile_diab_prob)
    profile_index = bi2de(flip(x)) + 1;
    f_val = profile_diab_prob(profile_index);
%f_val = x' .* prob_incremental;
%    f_val = sum(f_val(:));
    %f_val = 100 * f_val / sum(profile_diab_prob(:));
end

% Probability distribution
function pi_val = pi(x, profile_prob)
    profile_index = bi2de(flip(x)) + 1;
    pi_val = profile_prob(profile_index);
end
```