Interprétation des modèles de régression compositionnelle basées sur les ratio de paires de composantes

Lukas Dargel [1] & Christine Thomas-Agnan [2]

[1] *Toulouse School of Economics, France, christine.thomas@tse-fr.eu*
[2] *Toulouse School of Economics, France, lukas.dargel@tse-fr.eu*

Résumé. L'interprétation des modèles de régression comportant des vecteurs de parts aussi nommés compositions en tant que variable réponse et/où explicative a été abordée sous divers angles. Les premières approches de la littérature se font dans l'espace dit des coordonnées c'éest-à-dire après transformation des variables de composition par des transformations de type log-ratio. Etant donné que ces modèles sont non linéaires par rapport aux opérations classiques de l'espace $\mathbb{R}^D$, une autre approche a été proposée basée sur des incréments infinitésimaux et sur des dérivées au sens du simplexe. Cette dernière conduit à une interprétation à base d'élasticités ou de semi-élasticités. elles se font dans l'espace d'origine du vecteur de composition et sont indépendantes de toute transformation. Après un rappel sur ces deux points de vue, nous montrons que certaines fonctions des élasticités ou semi-élasticités sont constantes au travers de l'échantillon, ce qui en fait des paramètres naturels pour l'interprétation des modèles de régression compositionnelle. Ces paramètres sont liés à des variations relatives de ratio de deux composantes des variables réponse et/où explicatives. Nous proposons également des approximations de ces quantités pour un petit increment et montrons leur lien avec les paramètres naturels précédemment cités ce qui conduit à des interprétations libres de toute transformation et portant sur les variations exprimées dans l'espace d'origine. Nous utilisons un jeu de données sur les élections présidentielles françaises de 2022 pour illustrer chaque type d'interprétation.

Mots-clés. modèles de régression compositionnelle, ratios de paires, mesures d'impact, dérivées dans le simplexe

Abstract. The interpretation of regression models with compositional vectors as response and/or explanatory variables has been approached from different perspectives. The first approaches that appear in the literature are done in coordinate space after some log-ratio transformation of the compositional vectors. Considering the fact that these models are non-linear with respect to classical operations of the real space, another approach has been proposed based on infinitesimal increments or derivatives understood in a simplex sense, leading to elasticities or semi-elasticities interpretations in the original share space that have the advantage of being independent of any log-ratio transformations. After briefly reviewing these two points of view, we show that some functions of elasticities or semi-elasticities are constant throughout the sample observations, which makes them natural parameters for interpreting CoDa models. These parameters are linked to relative variations of pairwise share ratios of the response and/or of the explanatory variables. We derive approximations of share ratio variations and link them to these natural parameters leading to transformation-free

interpretations in the original share space. We use a real dataset on the French presidential election to illustrate each type of interpretation in detail.

Keywords. Compositional data regression, pairwise share ratio, impact measures, simplicial derivative.

1 Difficultés de l'interprétation d'un modèle de régression comportant des vecteurs de parts

L'interprétation des paramètres dans un modèle de régression est une étape essentielle pour comprendre l'impact marginal des changements d' une variable explicative sur la variable de réponse. Rappelons tout d'abord que, dans un modèle de régression linéaire classique expliquant la réponse Y par un ensemble de variables explicatives X, l'espérance conditionnelle $\mathbb{E}(Y \mid X)$ est une fonction linéaire de X. Par conséquent, pour toute variable explicative spécifique X_k, nous pouvons comprendre le paramètre β_{X_k} de X_k comme l'accroissement additif de $\mathbb{E}(Y \mid X_k)$ lorsque X_k augmente d'une unité (accroissement fini), toutes les autres variables explicatives restant fixées (ceteris paribus), ou alternativement comme la dérivée de $\mathbb{E}(Y \mid X_k)$ par rapport à X_k (accroissement infinitésimal). En économétrie, l'interprétation basée sur la dérivée est connue sous le nom d'effet marginal et les deux points de vue (accroissements finis et infinitésimaux) coïncident pour les modèles linéaires. Cependant, les modèles de régression CoDa impliquent des variables vecteurs de parts, également appelées variables compositionnelles du côté droit et/ou du côté gauche de l'équation de régression, ce qui implique qu'au moins certains paramètres ou variables du modèle sont à valeur dans un simplexe. Pour cette raison, les modèles CoDa ne sont pas linéaires pour la structure de l'espace vectoriel de l'espace réel, et c'est pourquoi les premières interprétations dans la littérature sont effectuées dans l'espace dit des coordonnées après une certaine transformation en log-ratio des vecteurs de parts. L'absence de linéarité du côté gauche (lorsque la réponse est compositionnelle) peut être résolue en adaptant la définition de l'espérance aux variables compositionnelles (voir par exemple [Pawlowsky-Glahn et al., 2015a]) comme cela a déjà été mentionné dans [Morais et Thomas-Agnan, 2021]. Comme nous le montrons dans ce texte, l'absence de linéarité du côté droit (lorsque l'explicative est compositionnelle) se traduit par le fait qu'on ne peut pas changer une composante d'un vecteur de parts en maintenant les autres composantes constantes: cette difficulté peut être résolue en considérant des accroissements linéaires des variables explicatives où la linéarité est comprise par rapport à la géométrie du simplexe introduite par Aitchison et définie par les opérations de perturbation et d'exponentiation.

2 Interprétations classiques et nouvelles d'un modèle de régression CoDa

Ce travail présente et illustre des interprétations classiques mais aussi nouvelles de l'impact des variables explicatives dans les modèles de régression comportant des vecteurs de parts. Les exemples de l'exposé seront centrés sur l'interprétation d'une régression expliquant les parts de vote pour différentes candidats ou groupes de candidats ainsi que le taux de participation lors du premier tour de l' élection présidentielle de 2022 en fonction de variables socio-économiques. L'utilisation de techniques de données compositionnelles pour analyser les données électorales est naturelle et plusieurs références peuvent être trouvées comme [Katz and King, 1999] et [Nguyen et al., 2022]. Après avoir rappelé les bases de la régression CoDa, notre premier objectif est de présenter une illustration complète des interprétations des élasticités/semi-élasticités de [Morais et al., 2018] et [Morais et Thomas-Agnan, 2021]. Nous discutons dans un deuxième temps des interprétations s'appuyant sur l'espace des coordonnées et de montrons l'avantage de l'interprétation dans le simplexe. Enfin, notre dernier objectif est de proposer une nouvelle interprétation basée sur les variations des ratios de paires de composantes dans l'espace du simplexe.

Les résultats mathématiques utilisés pour obtenir ces interprétations sont des approximations de parts et de ratios de parts dues à de petits accroissements linéaires d'un vecteur de parts explicatif le long d'un chemin linéaire dans le simplexe. Comprendre les chemins linéaires dans le simplexe est utile pour manipuler les variations linéaires des variables explicatives compositionnelles ainsi que pour interpréter les variations des vecteurs de parts de la variable réponse lorsque celle-ci est compositionnelle. Nous établissons une nouvelle formule de Taylor pour approximer une fonction d'un simplexe vers un autre simplexe le long d'un chemin linéaire dans une direction générale du simplexe.

Dans le cas de la réponse scalaire, nous montrons que toutes les approches sont équivalentes mais la nôtre permet de considérer des accroissements plus généraux des variables explicatives compositionnelles. Dans le cas de la variable réponse compositionelle, nous montrons que les variations infinitésimales des ratio de paires de parts sont indépendantes de l'individu dans l'échantillon, en faisant ainsi des paramètres essentiels pour l'interprétation de ces modèles. Par ailleurs ces interprétations sont plus faciles à comprendre pour les utilisateurs.

Tous les outils présentés sont disponibles dans le package R **CoDaImpact**, qui peut être utilisé en coordination avec le package R **compositions** [van den Boogaart et al.,2023]. Les illustrations sont disponibles dans la vignette

`\url{https://github.com/LukeCe/CoDaImpact}`.

Bibliographie

Aitchison, J., et Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. Biometrika, 71(2), 323-330.

Coenders, G., et Pawlowsky-Glahn, V. (2020). On interpretations of tests and effect sizes in regression models with a compositional predictor. SORT-Statistics and Operations Research Transactions, 201-220.

Dargel, L., et Thomas-Agnan, C. (2023). Pairwise share-ratio interpretations of compositional regression models. TSE working paper n. 23-1456, Juillet 2023, révisé le 20 septembre 2023.

Morais, J., et Thomas-Agnan, C. (2021). Impact of covariates in compositional models and simplicial derivatives. Austrian Journal of Statistics, 50(2), 1-15.

Muller I. et al. (2018). Interpretation of Compositional Regression with Application to Time Budget Analysis. Austrian Journal of Statistics, (2), 3-19.

van den Boogaart K.et al. (2021). Classical and robust regression analysis with compositional data. Mathematical geosciences, 53:823–858.

van den Boogaart, K., Tolosana-Delgado, R., et Bren, M. (2023). compositions: Compositional Data Analysis. R package version 2.0-5