

BURKINA FASO

Unité-Progrès-Justice

**MINISTRE DE L'ECONOMIE, DE LA FINANCE ET DE LA
PROSPECTIVE**

**INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA
DEMOGRAPHIQUE (INSD)**






ISSP-UJKZ



**ECONOMETRIE DES VARIABLES
QUALITATIVES**

**THEME : PROBIT MULTINOMIAL DE
CHOIX**

Rédigé par :

 **CISSE Oumarou**
 **TONGO Lanyiwè Lazare**
 **SAWADOGO Yacouba**

Enseignant :

Dr. Israël SAWADOGO
Ingénieur Statistique Economiste (ISE)

Année académique 2023-2024

Table des matières

Introduction	3
I. Définition.....	3
II. Spécification du modèle	4
1. Fonction d'utilité	4
2. Probabilité de Choix de l'alternative	5
III. Estimation du modèle	6
1. Estimation par la méthode du maximum de vraisemblance	6
2. Simulation du maximum de vraisemblance (SML)	7
3. Méthode GHK : Simulation de la Probabilité de Choix	7
IV. Validation du modèle.....	10
1. Vérification des Coefficients et de leur Significativité :.....	10
2. Adéquation Globale du Modèle	11
3. Analyse des Résidus	13
V. Avantages et limites.....	15
1. Avantages.....	15
2. Limites	15
VI. Domaines d'applications	16
Conclusion.....	17
Bibliographie	18

Introduction

Dans un monde où les individus sont confrontés à une multitude de choix quotidiens, comprendre les mécanismes sous-jacents à leurs décisions est crucial pour divers domaines tels que l'économie, la sociologie, la psychologie et le marketing. Le modèle probit multinomial de choix émerge comme un outil statistique puissant pour démystifier ces processus de choix complexes.

Ce modèle offre une perspective analytique sophistiquée sur les décisions discrètes impliquant plusieurs alternatives. Contrairement aux approches binaires plus simples, le modèle probit multinomial permet de capturer la diversité des choix humains en tenant compte de trois ou plus de catégories distinctes. En essence, il s'agit de déchiffrer les motivations et les préférences qui guident les individus dans leurs prises de décision.

Au cœur du modèle probit multinomial réside une modélisation probabiliste robuste. Il suppose que les individus évaluent chaque alternative en fonction d'une combinaison linéaire de variables explicatives, auxquelles s'ajoute un terme d'erreur qui suit une distribution normale (probit). Ce terme d'erreur capture l'incertitude et les influences latentes qui peuvent affecter le choix final.

Ce rapport vise à explorer en profondeur le modèle probit multinomial de choix, depuis sa formulation théorique jusqu'à son application pratique. Nous examinerons comment spécifier et estimer ce modèle, ainsi que les considérations importantes lors de son interprétation. Des exemples concrets illustreront son utilisation dans différents contextes, mettant en lumière son potentiel à éclairer les décisions humaines complexes.

Dans ce voyage à travers le modèle probit multinomial, nous découvrirons comment il peut non seulement enrichir notre compréhension des comportements de choix, mais aussi nous permettre de prendre des décisions éclairées dans un monde de multiples possibilités.

I. Définition

Le modèle probit multinomial de choix est une méthode statistique utilisée pour modéliser les choix discrets lorsque les individus ont plus de deux options. Il est basé sur le modèle probit, qui est utilisé pour modéliser les résultats binaires (par exemple, succès ou échec).

Le modèle probit multinomial de choix est une méthode économétrique avancée utilisée pour modéliser le comportement de choix des individus parmi plusieurs alternatives. Il se distingue du modèle logit multinomial par l'utilisation d'une distribution normale pour les erreurs aléatoires, ce qui permet de capturer des corrélations complexes entre les alternatives; et aussi par la non vérification de l'hypothèse d'indépendance des alternatives non pertinentes (IIA) qui stipule que l'ajout ou la suppression d'une alternative ne doit pas modifier le choix de l'individu.

II. Spécification du modèle

Le modèle probit multinomial de choix suppose que chaque individu a une fonction d'utilité pour chaque option de choix. La fonction d'utilité est une fonction mathématique qui représente la préférence de l'individu pour une option par rapport aux autres. La probabilité qu'un individu choisisse une option particulière est donnée par la fonction de distribution cumulative probit, qui est une fonction mathématique qui convertit les valeurs de la fonction d'utilité en probabilités. Elle est donnée par :

$$P_{ij} = P(U_{ij} > U_{ik}) \quad \forall \quad k \neq j$$

avec P_{ij} la probabilité qu'un individu i choisisse l'alternative j .

1. Fonction d'utilité

Dans le modèle probit multinomial de choix, l'utilité U_{ij} d'une alternative j pour un individu i est représentée par une fonction qui dépend de caractéristiques individuelles et de caractéristiques des alternatives :

$$U_{ij} = \beta_j V_{ij} + \varepsilon_{ij}$$

- V_{ij} est la partie déterministe de l'utilité, qui dépend de variables observées
- β_j est le vecteur des coefficients à estimer
- ε_{ij} est la partie aléatoire de l'utilité, qui capture les facteurs non observés

2. Probabilité de Choix de l'alternative

a. En utilisant la fonction d'utilité latente

L'individu i choisit l'alternative j qui maximise son utilité latente U_{ij} . Cela signifie que l'individu i choisit l'alternative j si et seulement si l'utilité latente U_{ij} est supérieure à l'utilité latente de toutes les autres alternatives k . En d'autres termes : $U_{ij} > U_{ip} \quad \forall p \neq j$

Ainsi la probabilité que l'individu i choisisse l'alternative j est donnée par :

$$P(y_i = j | x_i) = P(U_{ij} > U_{ip} \quad \forall p \neq j)$$

On peut voir que la probabilité que l'individu i choisisse l'alternative j dépend de la différence d'utilité entre les alternatives. Pour simplifier la modélisation et l'estimation des probabilités de choix nous travaillons avec la différence d'utilité entre les alternatives. En faisant cela, nous passons de l'utilité absolue à l'utilité relative.

Pour cela nous définissons une différence d'utilité entre l'alternative j et une alternative de référence $j=0$.

Soit \tilde{U}_{ij} la différence entre l'utilité latente U_{ij} et l'utilité de référence U_{i0}

$$\tilde{U}_i(j, 0) = U_{ij} - U_{i0} = (X_i \beta_j + \varepsilon_{ij}) - (X_i \beta_0 + \varepsilon_{i0}) = X_i (\beta_j - \beta_0) + (\varepsilon_{ij} - \varepsilon_{i0})$$

Comme les termes d'erreur ε_{ij} suivent une distribution normale multivariée avec une moyenne nulle et une matrice de covariance Σ alors $\varepsilon_{ij} - \varepsilon_{i0}$ suit également une distribution normale multivariée avec une matrice de covariance modifiée que nous appellerons Σ' .

- ✓ Si $\tilde{U}_i(j, 0) > 0$ alors l'individu i perçoit l'alternative j comme ayant une utilité plus grande que l'alternative de référence.
- ✓ Si $\tilde{U}_i(j, 0) < 0$ alors l'individu i perçoit l'alternative j comme ayant une utilité moindre que l'alternative de référence.
- ✓ $(\beta_j - \beta_0)$ Mesure l'impact de la variable explicative. Si $(\beta_j - \beta_0) > 0$ alors la variable explicative a un effet plus positif sur l'utilité perçue de l'alternative j par rapport à l'alternative de référence.

b. En utilisant la fonction de répartition

En utilisant la fonction de répartition on a :

$P(y_i = j | x_i) = P(U_{ij} > U_{ip} \ \forall p \neq j) = \phi(X_i(\beta_j - \beta_0) + \Sigma')$ avec ϕ la fonction de répartition.

On démontre que pour toutes les alternatives j , les probabilités de choix peuvent être exprimées sous forme intégrale. La probabilité que l'individu i choisisse l'alternative j peut être calculée comme suit :

$$P(y_i = j | x_i) = \int_{-\infty}^{X_i \beta_j} \phi(\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{j-1} / \Sigma) d\varepsilon$$

$\phi(\cdot)$ est la fonction de répartition cumulative de la loi normale multivariée.

$X_i \beta_j$ est la limite d'intégration correspondant à l'alternative j .

III. Estimation du modèle

1. Estimation par la méthode du maximum de vraisemblance

Les paramètres sont estimés par la méthode du maximum de vraisemblance (MLE).

Cette méthode consiste à maximiser la fonction de vraisemblance pour obtenir les estimations des paramètres :

La fonction de vraisemblance est donnée par : $L(\beta) = \prod_{i=1}^N P(Y_i=j | X_i)$

$$L(\beta) = \prod_{i=1}^N \prod_{j=1}^J P(Y_i = j)^{d_{ij}}$$

Où $d_{ij}=1$ si l'individu i choisit l'alternative j et 0 sinon.

Pour faciliter les calculs nous travaillons avec la log-vraisemblance

$$\text{Log} L(\beta) = \sum_{i=1}^N \sum_{j=1}^J d_{ij} \text{Log} P(Y_i=j)$$

Pour maximiser la log-vraisemblance, nous utilisons des méthodes d'optimisation. Pour cela, nous dériverons cette fonction par rapport aux paramètres β_j et trouverons les points où les dérivées sont nulles. Ce qui n'est pas toujours évident à cause de la complexité des intégrales d'où le recours à la simulation de vraisemblance.

2. Simulation du maximum de vraisemblance (SML)

Dans le contexte du modèle probit multinomial, recourir à la simulation du maximum de vraisemblance (Simulated Maximum Likelihood, SML) devient nécessaire en raison de la complexité du calcul des intégrales. En effet l'estimation des paramètres du modèle probit multinomial implique le calcul de probabilités de choix qui nécessite l'intégration sur des distributions normales multivariées. Ces intégrales sont souvent de haute dimension et n'ont pas de solution analytique simple, rendant le calcul direct des fonctions de vraisemblance pratiquement impossible.

Alors la SML est utilisée pour surmonter ces difficultés en approximant les intégrales difficiles à calculer par des méthodes de simulation. Pour ce faire, la SML utilise des techniques de simulation pour générer des échantillons aléatoires à partir de la distribution des erreurs. Ces échantillons sont ensuite utilisés pour approximer les intégrales requises dans le calcul de la vraisemblance.

L'une des méthodes les plus utilisées est celle de Geweke-Hajivassiliou-Keane (GHK).

3. Méthode GHK : Simulation de la Probabilité de Choix

La méthode GHK est une technique utilisée pour simuler les probabilités de choix dans un modèle probit multivarié. Elle est particulièrement utile lorsque l'intégration analytique est complexe ou impraticable. Voici les étapes clés de cette méthode :

- *Contexte et Modèle Latent*

Le modèle latent est :

$$y_i^* = X_i\beta + \epsilon$$

Où $\epsilon \sim N(0, \Sigma)$.

Nous utilisons une factorisation de Cholesky $\Sigma=CC'$ pour réécrire : ϵ avec $\eta_i \sim N(0, I)$

- *Région de Troncature*

La région de troncature A_j pour chaque variable y_j est définie par :

$$A_j = \begin{cases} [-\infty, 0] & \text{si } y_j = 0 \\ [0, +\infty] & \text{si } y_j = 1 \end{cases}$$

- *Simulation par Échantillonnage d'Importance*

On échantillonne les variables latentes en utilisant des bornes de troncature redéfinies pour chaque étape. Pour des limites $[a, b]$, nous avons :

$$\frac{a - x_1\beta_1}{c_{11}} < \eta_1 < \frac{b - x_1\beta_1}{c_{11}}$$

$$\frac{a - (x_2\beta_2 + c_{21}\eta_1)}{c_{22}} < \eta_2 < \frac{b - (x_2\beta_2 + c_{21}\eta_1)}{c_{22}}$$

$$\vdots$$

$$\vdots$$

$$\frac{a - (x_j\beta_j + \sum_{k=1}^{j-1} c_{jk}\eta_k)}{c_{jj}} < \eta_j < \frac{b - (x_j\beta_j + \sum_{k=1}^{j-1} c_{jk}\eta_k)}{c_{jj}}$$

Maintenant, il suffit de tirer de manière itérative de la distribution normale univariée tronquée avec les bornes données ci-dessus. Cela peut être fait en utilisant la méthode de l'inverse de la fonction de répartition cumulative (CDF) et en notant que la distribution normale tronquée est donnée par :

$$u = \frac{\Phi\left(\frac{x-u}{\sigma}\right) - \Phi\left(\frac{a-u}{\sigma}\right)}{\Phi\left(\frac{b-u}{\sigma}\right) - \Phi\left(\frac{a-u}{\sigma}\right)}$$

où u sera un nombre entre 0 et 1 car ce qui précède est une CDF. Cela suggère que pour générer des tirages aléatoires de la distribution tronquée, il faut résoudre pour x en utilisant :

$$x = \sigma F^{-1}(u * (F(\beta) - F(\alpha)) + F(\alpha)) + u$$

Où $\alpha = \frac{a-u}{\sigma}$ et $\beta = \frac{b-u}{\sigma}$ et F est la CDF normale standard. Avec de tels tirages, on peut reconstruire les y_i^* par son équation simplifiée en utilisant la factorisation de Cholesky. Ces tirages seront conditionnels sur les tirages précédents et en utilisant les propriétés des normales, le produit des PDF conditionnelles sera la distribution jointe des y_i^*

$$q(y_i^* | X_1\beta, \Sigma) = q(y_1^* | X_1\beta, \Sigma) q(y_2^* | y_1^*, X_1\beta, \Sigma) \dots q(y_j^* | y_1^*, \dots, y_{j-1}^*, X_1\beta, \Sigma)$$

où $q(\cdot)$ est la distribution normale multivariée.

Parce que y_i^* conditionné à y_k , $k < j$ est restreint à l'ensemble A par la configuration utilisant la factorisation de Cholesky, nous savons que $q(\cdot)$ est une normale multivariée tronquée. La fonction de distribution d'une loi normale tronquée est :

$$\frac{\phi\left(\frac{x - \mu}{\sigma}\right)}{\sigma\left(\Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)\right)}$$

Ainsi, y_j^* a pour distribution :

$$\begin{aligned} q(y_i^* | X_1\beta, \Sigma) &= \frac{\frac{1}{c_{11}} \Phi_1\left(\frac{y_j^* - x_1\beta}{c_{11}}\right)}{\sigma\left(\Phi\left(\frac{b - z_1\beta}{\sigma}\right) - \Phi\left(\frac{a - z_1\beta}{\sigma}\right)\right)} \times \dots \\ &\times \frac{\frac{1}{c_{jj}} \Phi_j\left(\frac{y_j^* - (x_j\beta + c_{j1}\eta_1 + c_{j2}\eta_2 + \dots + c_{jJ-1}\eta_{J-1})}{c_{jj}}\right)}{\sigma\left(\Phi_j\left(\frac{b - (x_j\beta + c_{j1}\eta_1 + c_{j2}\eta_2 + \dots + c_{jJ-1}\eta_{J-1})}{c_{jj}}\right) - \Phi\left(\frac{a - (x_j\beta + c_{j1}\eta_1 + c_{j2}\eta_2 + \dots + c_{jJ-1}\eta_{J-1})}{c_{jj}}\right)\right)} \\ &= \frac{\prod_{j=1}^J \frac{1}{c_{jj}} \Phi_j\left(\frac{y_j^* - \sum_{k=1}^{k < j} c_{jk}\eta_k}{c_{jj}}\right)}{\prod_{j=1}^J \Phi_j\left(\frac{b - \sum_{k=1}^{k < j} c_{jk}\eta_k}{c_{jj}}\right) - \Phi\left(\frac{a - \sum_{k=1}^{k < j} c_{jk}\eta_k}{c_{jj}}\right)} \end{aligned}$$

Où ϕ_j est la densité de probabilité normale standard pour le choix j.

Puisque $y_{j|\{y_{k < j}^*\}}^* \sim N(X_i\beta + \sum_{k=1}^{k < j} c_{jk}\eta_k, c_{jj}^2)$, la standardisation ci-dessus rend chaque terme de moyenne 0 et de variance 1.

Laissez le dénominateur $\prod_{j=1}^J \Phi_j\left(\frac{b - \sum_{k=1}^{k < j} c_{jk}\eta_k}{c_{jj}}\right) - \Phi\left(\frac{a - \sum_{k=1}^{k < j} c_{jk}\eta_k}{c_{jj}}\right) = \prod_{j=1}^J l_{jj}$ et le numérateur $\prod_{j=1}^J \frac{1}{c_{jj}} \Phi_j\left(\frac{y_j^* - \sum_{k=1}^{k < j} c_{jk}\eta_k}{c_{jj}}\right) = f_N(y_i^* | X_i\beta, \Sigma)$ avec $f_N(\cdot)$ la densité de probabilité multivariée normale.

Revenons à l'objectif initial, pour évaluer :

$$P(y_i = j | x_i\beta, \Sigma) = \int_{A_j} f_N(y_i^* | x_i\beta, \Sigma) dy_i^*$$

En utilisant l'échantillonnage d'importance, nous pouvons évaluer cette intégrale :

$$P(y_i = j | x_i\beta, \Sigma) = \int_{A_j} f_N(y_i^* | x_i\beta, \Sigma) dy_i^*$$

$$\begin{aligned}
&= \int_{A_j} \frac{f_N(y_i^*/x_i\beta, \Sigma)}{q(y_i^*|X_1\beta, \Sigma)} q(y_i^* | X_1\beta, \Sigma) dy_i^* \\
&= \int_{A_j} \frac{f_N(y_i^*/x_i\beta, \Sigma)}{\frac{f_N(y_i^*/x_i\beta, \Sigma)}{\prod_{j=1}^J l_{jj}}} q(y_i^* | X_1\beta, \Sigma) dy_i^* \\
&= Eq(\prod_{j=1}^J l_{jj})
\end{aligned}$$

Cela est bien approximé par : $\frac{1}{S} \sum_{s=1}^S \prod_{j=1}^J l_{jj}$

- *Tirage de la Distribution Normale Tronquée*

Les tirages de la distribution normale tronquée se font via l'inverse de la fonction de répartition (CDF inverse).

- *Calcul des Probabilités*

Les tirages successifs η_i permettent de reconstruire y_i^* :

$$y_i^* = X_i\beta + C\eta_i$$

La densité jointe des y_i^* est donnée par : $q(y_i^*|X_i\beta, \Sigma)$.

Le facteur de normalisation est :

$$\prod_{j=1}^J (\Phi_j \left(\frac{b - \sum_{k=1}^{K < J} c_{jk} \eta_k}{c_{jj}} \right) - \Phi \left(\frac{a - \sum_{k=1}^{K < J} c_{jk} \eta_k}{c_{jj}} \right)) = \prod_{j=1}^J l_{jj} \text{ et le numérateur}$$

Après l'estimation des paramètres du modèle il faut voir si le modèle utilisé s'adapte parfaitement aux données. Pour cela plusieurs critères sont à vérifier.

IV. Validation du modèle

1. Vérification des Coefficients et de leur Significativité :

Après l'estimation des coefficients il faut vérifier les p-values des coefficients pour déterminer leur significativité statistique. Pour cela on compare les p-valeurs avec un seuil de signification prédéterminé (0.05 ou 5%). Si la p-valeur est inférieure à ce seuil, le coefficient est considéré comme significatif. Sinon le coefficient n'est pas significatif.

2. Adéquation Globale du Modèle

Après la vérification de la significativité des coefficients, il faut effectuer des tests d'adéquations pour évaluer la qualité et la pertinence du modèle. Les principaux tests d'adéquation utilisés dans le cadre des modèles probit multinomial sont : le Test du Rapport de Vraisemblance, les Critères d'Information (AIC, BIC), le pseudo R² de Cragg et Uhler,

a. Le Test du Rapport de Vraisemblance

Le Test du Rapport de Vraisemblance compare deux modèles pour déterminer si l'ajout de paramètres supplémentaires améliore significativement l'ajustement du modèle. Plus précisément on compare un modèle nul ou restreint à un modèle complet pour voir si les variables ajoutées ont un effet significatif sur le modèle ajusté. Encore une fois pour simplifier les calculs on travaille avec la log-vraisemblance.

Pour faire la comparaison la statistique calculée est la suivante :

$$\lambda = -2(\ln(L_R) - \ln(L_C))$$

L_R est la Log – vraisemblance du modèle restreint.

L_C Log-vraisemblance du modèle complet

La statistique λ suit une distribution chi-carré(χ^2) avec des degrés de liberté égaux à la différence entre le nombre de paramètres des deux modèles.

On compare λ à la valeur critique de la distribution χ^2 pour le niveau de signification choisi (généralement 0.05 ou 5%) et les degrés de liberté correspondant

- Si λ est supérieur à la valeur critique de χ^2 , on rejette l'hypothèse nulle selon laquelle le modèle restreint est suffisant, ce qui signifie que le modèle complet apporte une amélioration significative.
- Si λ est inférieur à la valeur critique de χ^2 , on ne peut rejeter l'hypothèse nulle, ce qui signifie que les paramètres supplémentaires dans le modèle complet n'apportent pas une amélioration significative.

b. Les Critères d'Information (AIC, BIC)

Les critères AIC et BIC permettent d'évaluer combien un modèle s'ajuste aux données. Ils prennent en compte la log-vraisemblance du modèle et ajoutent une pénalité pour le nombre de paramètres estimés, afin de prévenir le surajustement (overfitting).

Plus basse d'AIC ou de BIC est considérée comme offrant le meilleur compromis entre qualité de l'ajustement et complexité du modèle.

➤ Akaike Information Criterion (AIC)

L'AIC est défini comme suit :

$$AIC = -2 \ln(L) + 2k$$

L est la valeur de la log-vraisemblance du modèle estimé et k est le nombre de paramètres du modèle.

➤ Bayesian Information Criterion (BIC)

$$BIC = -2 \ln(L) + k \ln(n)$$

L est la valeur de la log-vraisemblance du modèle estimé.

k est le nombre de paramètres du modèle.

n est le nombre d'observations.

c. le pseudo R² de Cragg et Uhler

Le pseudo R^2 de Cragg et Uhler, également connu sous le nom de R^2 de Nagelkerke, est une mesure d'ajustement pour les modèles de régression non linéaires. Il représente pour les modèles de régression non linéaires ce que le R^2 traditionnel représente pour la régression linéaire ainsi fournit une interprétation similaire au R^2 des modèles linéaires, indiquant la proportion de variance expliquée par le modèle.

La formule du pseudo R^2 de Cragg et Uhler est la suivante :

$$R^2_{Cragg-Uhler} = \frac{1 - \left(\frac{\text{Log} - \text{Likelihood}_{\text{Modèle}}}{\text{Log} - \text{Likelihood}_{\text{Nul}}} \right)^{\frac{2}{N}}}{1 - (\text{Log} - \text{Likelihood}_{\text{Nul}})^{\frac{2}{N}}}$$

Où :

- $\text{Log} - \text{Likelihood}_{\text{Modèle}}$ est la log-vraisemblance du modèle estimé.
- $\text{Log} - \text{Likelihood}_{\text{Nul}}$ est la log-vraisemblance du modèle nul (modèle sans variables explicatives)
- N est le nombre d'observations

Si le pseudo R^2 de Cragg et Uhler est proche de 1, cela indique un bon ajustement du modèle, signifiant que le modèle explique bien la variance des données.

Si le pseudo R^2 est proche de 0, cela indique un faible ajustement, signifiant que le modèle n'explique pas bien la variance des données.

3. Analyse des Résidus

Pour que notre modèle soit validé, les résidus doivent respecter certains critères qui sont :

a. Normalité des Résidus

Les résidus du modèle multinomial doivent suivre une distribution normale pour que les tests statistiques soient valides. Pour vérifier cela, on peut voir le graphique des résidus. Ce graphique doit présenter une distribution approximativement normale pour qu'on puisse dire que les résidus suivent une distribution normale.

On peut aussi utiliser le test de Shapiro-Wilk pour vérifier formellement la normalité des résidus.

b. Homoscédasticité des Résidus

Les résidus doivent présenter une variance constante à travers toutes les catégories de la variable dépendante. On peut aussi vérifier cette hypothèse graphiquement en comparant le graphique des résidus et le graphique de variables explicatives. On vérifie si la dispersion des résidus doit être constante à travers différentes valeurs des variables explicatives.

On peut aussi utiliser des tests comme le test de Breusch-Pagan pour détecter formellement l'hétéroscédasticité.

c. Indépendance des Résidus

Les résidus doivent être indépendants les uns des autres. Comme les autres propriétés des résidus, cette propriété peut être vérifiée graphiquement en vérifiant l'absence de séquences dans le graphique des résidus. Il peut être aussi vérifié formellement en utilisant le test de Durbin-Watson qui permet de vérifier l'autocorrélation des résidus. Les résidus sont indépendants s'ils ne sont pas corrélés.

d. Linéarité

La relation entre les variables explicatives et la variable dépendante doit être linéaire pour chaque catégorie de la variable dépendante. Cette propriété peut être vérifiée graphiquement à travers le graphique des résidus et variables explicatives. Il faut vérifier que les résidus ne présentent pas de schémas clairs lorsqu'ils sont tracés par rapport aux variables explicatives.

e. Absence de Multicolinéarité

Les variables explicatives ne doivent pas être fortement corrélées entre elles.

Variance Inflation Factor (VIF) : Vérifier les valeurs du VIF pour chaque variable explicative afin de détecter la multicolinéarité

$$VIF(X_i) = \frac{1}{1 - R_i^2}$$

Où R_i^2 est le coefficient de détermination de la régression de X_i sur les autres variables explicatives.

Un VIF supérieur à 10 (ou parfois 5) est généralement considéré comme indiquant une multicollinéarité problématique. Un VIF élevé indique que la variance du coefficient de régression de la variable explicative correspondante est fortement augmentée en raison de la corrélation avec d'autres variables explicatives.

V. Avantages et limites

1. Avantages

Le modèle probit multinomial de choix présente plusieurs avantages par rapport à d'autres méthodes de modélisation des choix discrets, tels que :

- La modélisation d'un nombre quelconque d'options de choix en raison de sa flexibilité, de sa capacité à capturer les corrélations entre options, de son approche réaliste pour des choix complexes, de son potentiel pour une analyse détaillée des préférences et de sa capacité à prendre en compte les effets substitutifs et complémentaires entre les différentes options.
- Il ne suppose pas que les erreurs des individus sont indépendantes et identiquement distribuées au vu de sa capacité à capturer les corrélations entre les alternatives, à représenter plus fidèlement les processus choix complexes, à éviter les limitations de l'hypothèse IIA et à fournir des prédictions plus réalistes et robustes dans les applications pratiques.
- L'utilisation du modèle probit multinomial de choix pour modéliser des données complexes telles que des données longitudinales ou des données en grappes offre des avantages en terme de gestion des dépendances intra-groupes et temporelles, de flexibilité dans la spécification des structures de corrélation, de robustesse aux violations des hypothèses classiques.
- Flexibilité : il est capable de capturer des structures de dépendance plus riches. Elle (flexibilité) permet d'adapter le modèle à différentes structures de données, de capturer la corrélation entre les alternatives, d'être robuste face aux violations des hypothèses strictes et d'intégrer un grand nombre de variables explicatives.

2. Limites

Le modèle probit multinomial de choix présente également quelques limites :

- Il peut être difficile à interpréter à cause de la non-linéarité de la fonction de régression et des effets marginaux complexes. Une compréhension approfondie des principes statistiques et économétriques est souvent nécessaire pour une interprétation précise et pertinente des résultats.

- Il peut être sensible aux valeurs aberrantes en raison de son estimation par maximum de vraisemblance et de sa fonction de lien non-linéaire. Les valeurs aberrantes peuvent influencer négativement les estimations des paramètres du modèle, entraînant des résultats peu fiables et une interprétation moins précise des effets des variables explicatives sur les choix catégoriques.
- Il peut être coûteux à calculer en raison de la complexité des intégrales multidimensionnelles nécessaires pour les probabilités de choix, des algorithmes d'optimisation non-linéaire requis pour l'estimation des paramètres, de l'augmentation exponentielle de la complexité avec le nombre de catégories de choix et de variables explicatives et des méthodes numériques intensives souvent employées pour contourner les difficultés analytiques.

VI. Domaines d'applications

Le modèle probit multinomial de choix est utilisé dans divers domaines où il est essentiel de comprendre les comportements de choix entre plusieurs alternatives. Voici quelques domaines d'application clés :

- Economie des transports : Il analyse les facteurs influençant le choix entre différents modes de transports (voitures, train, bus, vélo, etc.). Également, il intervient dans l'évaluation des impacts des nouvelles infrastructures de transport sur les choix des usagers.
- Marketing et études de marché : il aide dans l'identification des segments de consommateurs en fonction de leur comportement de choix et préférences.
- Le choix d'une activité de loisirs
- Politique : Le choix d'un vote politique
- Finance et investissements à travers le choix d'investissement : Étude des décisions des investisseurs entre différents types d'actifs (actions, obligations, immobilier, etc.) et le comportement financier : Analyse des préférences pour divers produits financiers et des facteurs influençant ces choix.
- Etc...

Le modèle probit multinomial de choix est donc extrêmement polyvalent et peut être appliqué dans toute situation nécessitant l'analyse des décisions entre plusieurs alternatives. Il

permet de capturer les complexités et les corrélations dans les choix des individus, offrant des insights précieux pour la prise de décision, la politique et la stratégie dans divers domaines.

Exemple d'application

Soit un individu qui a le choix entre trois modes de transport pour se rendre au travail : la voiture, le bus et le train. Le modèle probit multinomial de choix peut être utilisé pour modéliser la probabilité que l'individu choisisse chaque mode de transport en fonction de ses caractéristiques individuelles, telles que son revenu, son lieu de résidence et ses préférences en matière de transport.

Conclusion

En conclusion, le modèle probit multinomial de choix est un outil statistique puissant pour analyser les décisions discrètes impliquant plusieurs alternatives. Ce modèle, en capturant les corrélations complexes entre les alternatives grâce à une distribution normale multivariée des erreurs, offre une flexibilité et une précision accrues par rapport aux autres modèles de choix discrets.

L'estimation des paramètres du modèle nécessite des techniques avancées comme le maximum de vraisemblance simulé, et la validation passe par des tests de significativité des coefficients et une analyse approfondie des résidus. Malgré ces défis, le modèle prouve son utilité dans des domaines variés comme l'économie, le marketing et les sciences sociales.

En somme, le modèle probit multinomial enrichit notre compréhension des comportements de choix et fournit un cadre robuste pour des décisions éclairées dans un monde complexe et diversifié.

Bibliographie

https://www.academia.edu/5916521/Th%C3%A9orie_et_Application_du_PROBIT_MULTINOMIAL_DE_CHOIX_sous_STATA_Eviews_et_R

https://wikimedia.org/api/rest_v1/media/math/render/svg/f9c8e7fa7de2ac8ed41b0dbfc913a7d2789c23db