

TD3 – Modèles de régression multinomiale

Année universitaire 2025–2026 ■ Parcours Économie de la santé &
Développement durable

Pierre Beaucoral

Introduction

Un **modèle de régression multinomiale** est un modèle Logit ou Probit dans lequel la variable à expliquer Y est une variable qualitative à $k > 2$ modalités. Cette variable peut être qualitative **nominale** ou **ordinaire**.

Cas d'une variable expliquée nominale

Dans le cas d'une variable expliquée **nominale**, on prend n'importe quelle modalité comme modalité de référence (modalité 0), et on estime des *pseudo-côtes*, c'est-à-dire :

- $\frac{\Pr(Y = 1)}{\Pr(Y = 0)}$
- $\frac{\Pr(Y = 2)}{\Pr(Y = 0)}$
- etc.

Par exemple, dans le cas $k = 3$ modalités de Y , on a :

$$\Pr(Y = 0) + \Pr(Y = 1) + \Pr(Y = 2) = 1$$

MAIS $\Pr(Y = 0) + \Pr(Y = 1) < 1$ et $\Pr(Y = 0) + \Pr(Y = 2) < 1$

On estime alors les paramètres β_g tels que :

$$\ln \left(\frac{\Pr(Y = g)}{\Pr(Y = 0)} \right) = \beta_{g0} + \sum_{j=1}^p \beta_{gj} X_j$$

avec $g = 1, \dots, k - 1$.

On estime donc :

- $(k - 1)$ paramètres pour chaque variable explicative quantitative ;
- $(k - 1)(q - 1)$ paramètres pour une variable explicative qualitative à q modalités.

Cas d'une variable expliquée ordinale

Dans le cas d'une variable expliquée **ordonale**, $Y = 0$ ou 1 ou 2 , etc. représente une réponse graduée.

La résolution suppose l'existence d'une variable continue sous-jacente Y^* , et de $(k - 1)$ bornes c_j telles que :

- si $y_i^* < c_1$ alors $y_i = 1$
- si $c_{j-1} < y_i^* < c_j$ alors $y_i = j$
- si $y_i^* > c_{k-1}$ alors $y_i = k$

On a :

$$y_i^* = X_i B + \varepsilon_i$$

et on estime conjointement :

- les paramètres β_j correspondant à chaque variable explicative ;
- les seuils c_g ($g = 1, \dots, k - 1$).

On prédit alors l'appartenance de chaque individu à chaque classe par les formules :

$$\Pr(Y_i = 0) = \Phi(c_1 - X_i B)$$

$$\Pr(Y_i = g) = \Phi(c_g - X_i B) - \Phi(c_{g-1} - X_i B)$$

où Φ est :

- la fonction de répartition d'une loi gaussienne centrée réduite dans le cas du **modèle Probit multivarié** ;
- l'inverse de la fonction Logit dans le cas du **Logit multivarié**.

Présentation de l'étude et des données

Les données étudiées proviennent de *Hill et al.* (1995) et sont utilisées comme exemple dans l'ouvrage de Kleinbaum et Klein.

- 288 femmes avec un cancer de l'endomètre participent à l'étude.

Dictionnaire des variables

- **ID** : identifiant individuel.
- **GRADE** : variable ordinale indiquant le stade de la tumeur
 - 0 : bien différenciée
 - 1 : modérément différenciée
 - 2 : peu différenciée
- **RACE** : variable indicatrice à deux modalités
 - 1 : peau noire
 - 0 : peau blanche
- **ESTROGEN** : variable indicatrice à deux modalités
 - 1 : la femme a déjà pris des œstrogènes
 - 0 : sinon
- **SUBTYPE** : variable qualitative à trois modalités codant le sous-type de tissu cancéreux
 - 0 : Adénocarcinome
 - 1 : Adenosquamous
 - 2 : Autre
- **AGE** : âge recodé en deux classes
 - 0 : 50–64 ans
 - 1 : 65–79 ans
- **SMK** : variable binaire indiquant le statut tabagique au moment de l'étude
 - 1 : fumeuse
 - 0 : non-fumeuse

Références

- Hill, H.A., Coates, R.J., Austin, H., Correa, P., Robboy, S.J., Chen, V., Click, L.A., Barrett, R.J., Boyce, J.G., Kotz, H.L., and Harlan, L.C., *Racial differences in tumor grade among women with endometrial cancer*, Gynecol. Oncol. 56: 154–163, 1995.
 - David G. Kleinbaum, Mitchel Klein, *Logistic Regression – A Self-Learning Text*, Third Edition, Springer, 2010.
-

Import des données

Ouvrir R et importer les données (`cancer.dta` utiliser le package `haven`).

Modèle multinomial pour expliquer la variable SUBTYPE

Les variables explicatives sont : RACE, ESTROGEN, SMK et AGE.

1. Estimation du premier modèle

Appliquer un premier modèle de régression logit multinomiale prenant en compte les effets des quatre variables explicatives (commande R : `nnet`).

2. Sauvegarde des résultats

Sauvegarder les résultats du modèle ajusté.

3. Valeurs prédites et distribution

Générer les valeurs prédites.

Observer et expliquer la répartition de ces données (commande R : `predict`).

4. Test d'ajustement du modèle

Tester l'ajustement de ce modèle aux données (commande R : `generalhoslem`), en réduisant le nombre de groupes jusqu'à ce que le test soit applicable.

- Expliquer ce qui se passe.
- Le modèle est-il ajusté aux données ?

5. Simplification du modèle

Essayer de simplifier ce modèle, en se basant sur des tests de **rapport de vraisemblance** entre modèles emboîtés.

- Combien de degrés de liberté sont appliqués à chaque test ?
- Quel modèle est finalement choisi ?

6. Interprétation

Interpréter les résultats du **modèle final**.

7. Tableau de contingence des individus bien et mal classés

- Tabuler la variable `SUBTYPE` pour constater qu'il y a :
 - 186 adénocarcinomes
 - 45 adenosquames
 - 57 autres cas
- Tabuler les valeurs prédites dans `cancer_sub` et construire une nouvelle variable `pred_subtype` prenant la valeur 0 pour les 186 (environ) individus avec les plus grandes valeurs de `cancer_sub`.
- Établir le tableau de contingence des variables `subtype_f` et `pred_subtype`, et calculer la **proportion de cas mal prédits**.

B. Modèle multinomial ordonné pour expliquer la variable GRADE

Le stade de la tumeur dépend des variables précédentes mais aussi du type de cancer.

8. Modèle ordonné de base

Ajuster un modèle de régression multinomiale ordonnée, avec comme variables explicatives `RACE`, `ESTROGEN`, `SUBTYPE`, `AGE` et `SMK` (commande R : `polr()` (`MASS`)).

Attention : il faut bien utiliser la variable `grade_ord`.

9. Test d'ajustement via interactions (en R)

R, comme Stata, ne fournit pas de test d'ajustement global « clé en main » pour les modèles logit/probit ordonnés.

On va donc tester l'apport de certaines **interactions** en comparant des modèles **emboîtés** au moyen de **tests de rapport de vraisemblance** (Likelihood Ratio, LR).

On utilise pour cela la fonction `polr()` du package `MASS`, qui permet d'estimer un modèle logit ordinal.

1. Modèle de base (rappel de la question 8)

- Ajuster dans R un premier modèle de régression multinomiale ordonnée avec **GRADE** comme variable expliquée et les variables explicatives : **RACE**, **ESTROGEN**, **SUBTYPE**, **AGE** et **SMK**.
 - On utilisera la fonction `polr()` du package **MASS** (modèle noté par exemple `mod_base`).
2. **Ajout de l'interaction ESTROGEN × SUBTYPE**
- Ajuster un deuxième modèle ordinal contenant **tous les effets simples** et, en plus, l'effet de l'**interaction** entre **ESTROGEN** et **SUBTYPE**.
 - En R, on peut écrire cette interaction sous la forme **ESTROGEN * SUBTYPE**, qui inclut automatiquement les effets simples et le terme d'interaction.
 - Noter ce modèle, par exemple, `mod_int_ES`.
3. **Test de rapport de vraisemblance entre les deux modèles**
- Comparer `mod_base` et `mod_int_ES` à l'aide d'un **test de rapport de vraisemblance** (LR test) via la fonction `anova(mod_base, mod_int_ES)` dans R.
 - Interpréter :
 - la statistique de test (χ^2),
 - le nombre de degrés de liberté (lié au nombre de paramètres supplémentaires dans le modèle avec interaction),
 - la p-value.
 - Conclure : l'interaction **ESTROGEN × SUBTYPE** améliore-t-elle significativement le modèle ? Faut-il la conserver dans le modèle final ?
4. **Autres interactions possibles**
- Répéter la même démarche avec **une ou deux autres interactions** en effets simples, par exemple :
 - **ESTROGEN × AGE** ;
 - **SUBTYPE × AGE** ;
 - ou toute autre interaction jugée pertinente.
 - Pour chaque nouvelle interaction :
 1. Ajuster le modèle étendu (par exemple `mod_int_EAGE`, `mod_int_SAGE`, etc.) ;
 2. Comparer ce modèle au modèle de base `mod_base` au moyen d'un **test LR** via `anova()` ;
 3. Discuter de l'intérêt de conserver ou non l'interaction dans le modèle au vu de la p-value et, éventuellement, du critère **AIC**.
5. **Discussion**
- À partir de ces tests, proposer un modèle ordinal « raisonnable » :
 - suffisamment souple pour capter les effets importants ;

– mais pas trop complexe (principe de parcimonie).

- Discuter brièvement des limites de ce type de « test d’ajustement via interactions » pour juger de la qualité globale du modèle.

10. Sélection de modèle par AIC

En utilisant le critère **AIC**, rechercher un modèle plus simple permettant de prédire le stade de la tumeur selon son type.

11. Modèle final et interprétation

- Quel **modèle final** choisit-on ?
- Interpréter les résultats de ce modèle.