

TD 3 — Régression multinomiale (R)

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

Pierre Beaucoral

Objectifs du TD

- Manipuler un **modèle logit multinomial nominal** (SUBTYPE).
- Manipuler un **modèle logit ordinal** (GRADE).
- Savoir :
 - préparer les données (codage en facteurs / variables ordinales) ;
 - estimer, simplifier et interpréter un modèle ;
 - tester :
 - * l’ajustement (Hosmer–Lemeshow multinomial) ;
 - * des interactions via test de rapport de vraisemblance (LR).

Données : *cancer.dta* (288 femmes avec cancer de l’endomètre).

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 **ordinale**.

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 **ordinale**, $Y = 0$ 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.
-

Packages utilisés

```
library(tidyverse)
library(janitor)
library(haven)
library(broom)
library(dplyr)
library(gt)
library(nnet)           # multinom (logit multinomial nominal)
library(generalhoslem)  # logitgof : test de Hosmer-Lemeshow multinomial
library(MASS)           # polr : logit ordinal

theme_set(theme_minimal())
```

Import des données et préparation

On suppose que le fichier `cancer.dta` se trouve dans le dossier `./data/`.

```
cancer_raw <- read_dta("./data/cancer.dta") |>
  clean_names()

glimpse(cancer_raw)
```

```
Rows: 288
Columns: 7
$ id      <dbl> 10009, 10025, 10038, 10042, 10049, 10113, 10131, 10160, 10164~
$ grade   <dbl+lbl> 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 2, 1, 2, 2, ~
$ race     <dbl+lbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ estrogen <dbl+lbl> 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, ~
$ subtype  <dbl+lbl> 1, 2, 1, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 2, 0, 0, ~
$ age      <dbl+lbl> 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, ~
$ smoking  <dbl+lbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

On dispose notamment des variables :

- grade (3 modalités ordonnées),
- race,
- estrogen,
- subtype,
- age,
- smoking.

Recodage des variables

On crée des facteurs explicites pour la régression, en choisissant des **références** cohérentes avec l'énoncé :

```
cancer <- cancer_raw |>
  mutate(
    # convertir les labels Stata en facteurs R
    grade_f      = as_factor(grade),
    subtype_f     = as_factor(subtype),
    race_f        = as_factor(race),
    estrogen_f    = as_factor(estrogen),
    age_f         = as_factor(age),
    smk_f         = as_factor(smoking),
```

```

# forcer l'ordre pour l'ordinal (adapter les noms à ce que tu vois)
grade_ord = fct_relevel(
  grade_f,
  "bien différencié",
  "moyennement différencié",
  "peu différencié"
)
)

cancer |>
  dplyr::select(grade, grade_ord, subtype, subtype_f,
    race_f, estrogen_f, age_f, smk_f) |>
  head()

```

```

# A tibble: 6 x 8
  grade      grade_ord subtype subtype_f race_f estrogen_f age_f smk_f
<dbl> <dbl> <fct>      <fct>      <fct> <fct>      <fct> <fct>
1 1 [moyennement diff~ moyennem~ 1 [ade~ adenosqu~ blanc~ never too~ 50-64 yes
2 0 [bien différencié] bien dif~ 2 [oth~ other    blanc~ took oest~ 50-64 no
3 1 [moyennement diff~ moyennem~ 1 [ade~ adenosqu~ blanc~ never too~ 65-79 no
4 0 [bien différencié] bien dif~ 0 [ade~ adenocar~ blanc~ never too~ 65-79 no
5 0 [bien différencié] bien dif~ 0 [ade~ adenocar~ blanc~ took oest~ 50-64 no
6 0 [bien différencié] bien dif~ 0 [ade~ adenocar~ blanc~ took oest~ 65-79 no

```

Modèle multinomial pour expliquer SUBTYPE

Variables explicatives : RACE, ESTROGEN, SMK, AGE.

Estimation du premier modèle (logit multinomial nominal)

```

mod_sub_full <- multinom(
  subtype_f ~ race_f + estrogen_f + smk_f + age_f,
  data = cancer
)

```

```
# weights: 18 (10 variable)
initial value 314.203115
iter 10 value 247.216796
final value 246.965190
converged
```

```
res_sub_ful <- tidy(mod_sub_full,
                    exponentiate = TRUE, # passe en OR
                    conf.int = TRUE)    # ajoute IC 95%

res_sub_ful
```

```
# A tibble: 10 x 8
```

	y.level <chr>	term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>	conf.low <dbl>	conf.high <dbl>
1	adenosquamous	(Inter~	0.169	0.447	-3.97	7.18e-5	0.0705	0.407
2	adenosquamous	race_f~	0.806	0.413	-0.521	6.02e-1	0.359	1.81
3	adenosquamous	estrog~	0.483	0.378	-1.93	5.39e-2	0.230	1.01
4	adenosquamous	smk_fy~	2.43	0.526	1.69	9.06e-2	0.869	6.82
5	adenosquamous	age_f6~	2.66	0.412	2.38	1.75e-2	1.19	5.96
6	other	(Inter~	0.282	0.378	-3.35	8.07e-4	0.134	0.591
7	other	race_f~	1.13	0.376	0.319	7.49e-1	0.539	2.36
8	other	estrog~	0.943	0.343	-0.171	8.64e-1	0.482	1.85
9	other	smk_fy~	0.166	1.05	-1.71	8.67e-2	0.0214	1.29
10	other	age_f6~	1.33	0.329	0.872	3.83e-1	0.699	2.54

```
summary(mod_sub_full)
```

Call:

```
multinom(formula = subtype_f ~ race_f + estrogen_f + smk_f +
          age_f, data = cancer)
```

Coefficients:

	(Intercept)	race_fnoire	estrogen_ftook	oestrogen	treatment
adenosquamous	-1.775326	-0.2151038			-0.72813726
other	-1.266281	0.1201888			-0.05867755
		smk_fyes	age_f65-79		
adenosquamous	0.889793	0.9780758			
other	-1.793171	0.2865677			

Std. Errors:

	(Intercept)	race_fnoire	estrogen_ftook	oestrogen	treatment
adenosquamous	0.4471631	0.4127306			0.3777940
other	0.3779403	0.3762200			0.3425219
	smk_fyes	age_f65-79			
adenosquamous	0.5257988	0.4117731			
other	1.0467384	0.3285697			

Residual Deviance: 493.9304

AIC: 513.9304

On obtient, pour chaque modalité $g \neq$ référence, une équation :

$$\log \frac{P(\text{SUBTYPE}=g)}{P(\text{SUBTYPE}=\text{adenocarcinomous})} = \beta g_0 + \beta g, \text{race} + \dots$$

À caractéristiques identiques (race, oestrogènes, âge), les **fumeuses** ont environ **2,4 fois plus de chances (odds)** d'avoir un cancer *adenosquamous* plutôt que le sous-type de référence, comparées aux **non fumeuses**.

Résumé et odds-ratios

```
res_sub_full <- tidy(mod_sub_full, exponentiate = TRUE, conf.int = TRUE)
res_sub_full
```

A tibble: 10 x 8

	y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	adenosquamous	(Inter~	0.169	0.447	-3.97	7.18e-5	0.0705	0.407
2	adenosquamous	race_f~	0.806	0.413	-0.521	6.02e-1	0.359	1.81
3	adenosquamous	estrog~	0.483	0.378	-1.93	5.39e-2	0.230	1.01
4	adenosquamous	smk_fy~	2.43	0.526	1.69	9.06e-2	0.869	6.82
5	adenosquamous	age_f6~	2.66	0.412	2.38	1.75e-2	1.19	5.96
6	other	(Inter~	0.282	0.378	-3.35	8.07e-4	0.134	0.591
7	other	race_f~	1.13	0.376	0.319	7.49e-1	0.539	2.36
8	other	estrog~	0.943	0.343	-0.171	8.64e-1	0.482	1.85
9	other	smk_fy~	0.166	1.05	-1.71	8.67e-2	0.0214	1.29
10	other	age_f6~	1.33	0.329	0.872	3.83e-1	0.699	2.54

Ici, $\text{estimate} = \text{odds-ratio}$, $\text{conf.low} / \text{conf.high} = \text{IC } 95 \%$.

Probabilités prédites

On génère les probabilités prédites pour chaque modalité de SUBTYPE :

```
phat_sub <- predict(mod_sub_full, type = "probs")
head(phat_sub)
```

	adenocarcinomous	adenosquamous	other
1	0.6852096	0.2826449	0.03214548
2	0.7420517	0.0607007	0.19724761
3	0.5476498	0.2467524	0.20559787
4	0.5476498	0.2467524	0.20559787
5	0.7420517	0.0607007	0.19724761
6	0.6363103	0.1384208	0.22526893

```
colMeans(phat_sub) # moyennes des proba par modalité
```

adenocarcinomous	adenosquamous	other
0.6433555	0.1573435	0.1993010

On recolle ces probabilités aux données :

```
# 1) Sous-échantillon sans NA sur les variables du modèle
cancer_complete <- cancer |>
  drop_na(subtype_f, race_f, estrogen_f, smk_f, age_f)

# 2) Modèle multinomial sur cancer_complete
mod_sub_full <- multinom(
  subtype_f ~ race_f + estrogen_f + smk_f + age_f,
  data = cancer_complete
)
```

```
# weights: 18 (10 variable)
initial value 314.203115
iter 10 value 247.216796
final value 246.965190
converged
```

```
# 3) Probabilités prédites (286 x 3)
phat_sub <- predict(mod_sub_full, type = "probs")
colMeans(phat_sub)
```

adenocarcinomous	adenosquamous	other
0.6433555	0.1573435	0.1993010

```
# 4) On transforme en tibble et on renomme
phat_sub_tbl <- as_tibble(phat_sub)
names(phat_sub_tbl) <- paste0("p_", levels(cancer_complete$subtype_f))
# ex : "p_adenocarcinomous" "p_adenosquamous" "p_other"
```

```
# 5) On colle aux données *complètes* (286 lignes)
cancer_sub <- bind_cols(cancer_complete, phat_sub_tbl)
```

```
cancer_sub |>
  dplyr::select(subtype_f, starts_with("p_")) |>
  slice(1:5)
```

```
# A tibble: 5 x 4
```

subtype_f	p_adenocarcinomous	p_adenosquamous	p_other
<fct>	<dbl>	<dbl>	<dbl>
1 adenosquamous	0.685	0.283	0.0321
2 other	0.742	0.0607	0.197
3 adenosquamous	0.548	0.247	0.206
4 adenocarcinomous	0.548	0.247	0.206
5 adenocarcinomous	0.742	0.0607	0.197

Test d'ajustement (Hosmer–Lemeshow multinomial)

On utilise `logitgof()` du package **generalhoslem**.

- `obs` : modalités de SUBTYPE sous forme numérique (1, 2, 3).
- `exp` : matrice de probabilités prédites.
- `g` : nombre de groupes (à ajuster si nécessaire).

```
y_num <- as.numeric(cancer_sub$subtype_f)
exp_mat <- as.matrix(phat_sub)

# Exemple avec 10 groupes
gof_10 <- logitgof(obs = y_num, exp = exp_mat, g = 10)
gof_10
```

Hosmer and Lemeshow test (multinomial model)

```
data: y_num, exp_mat
X-squared = 4.0727, df = 10, p-value = 0.944
```

Le test de Hosmer–Lemeshow (généralisé au multinomial et implémenté par `logitgof()` dans **generalhoslem**) compare, dans des groupes de probabilités prédites similaires, les effectifs observés de chaque catégorie de la variable dépendante aux effectifs attendus selon le modèle. La statistique de test est de type ². Une grande p-value indique que l'on ne détecte pas de mauvais ajustement global du modèle aux données ; une petite p-value suggère un manque d'adéquation (modèle mal calibré).

Si le test ne passe pas (classes attendues trop petites), on réduit le nombre de groupes :

```
library(purrr)

map_df(4:10, ~{
  out <- try(logitgof(y_num, exp_mat, g = .x), silent = TRUE)
  tibble(
    g      = .x,
    stat    = if (inherits(out, "try-error")) NA_real_ else out$statistic,
    p_value = if (inherits(out, "try-error")) NA_real_ else out$p.value
  )
})
```

```
# A tibble: 7 x 3
      g  stat p_value
<int> <dbl>   <dbl>
1     4  1.69  0.792
2     5  2.01  0.735
3     6 15.7   0.0474
4     7  2.63  0.955
5     8  2.63  0.955
6     9 16.5   0.170
7    10  4.07  0.944
```

Lecture :

- p-value **grande** → pas d'évidence de mauvais ajustement.
- p-value **petite** → modèle mal ajusté (au moins pour certains groupes).

En diminuant le nombre de groupes, le test Hosmer–Lemeshow devient applicable. Les p-values varient avec g, ce qui montre que le test est assez instable dans ce petit échantillon multinomial. Néanmoins, pour la plupart des partitions (g = 4, 5, 7, 8, 9, 10), on ne rejette pas l'hypothèse d'un bon ajustement (p-value > 5 %). On peut donc conclure qu'il **n'y a pas de signe clair de mauvais ajustement du modèle aux données**, tout en rappelant que ce test doit être interprété avec prudence.

Simplification du modèle (tests LR)

On cherche un modèle **plus parcimonieux** en retirant les variables non significatives.

Exemple : on teste si on peut retirer `smk_f` puis `age_f`

```
# Modèle sans SMK
mod_sub_nosmk <- multinom(
  subtype_f ~ race_f + estrogen_f + age_f,
  data = cancer
)
```

```
# weights: 15 (8 variable)
initial value 314.203115
iter 10 value 251.550761
final value 251.468001
converged
```

```
# Test LR : mod_sub_nosmk vs mod_sub_full
anova(mod_sub_nosmk, mod_sub_full)
```

Likelihood ratio tests of Multinomial Models

Response: subtype_f

	Model	Resid. df	Resid. Dev	Test	Df
1	race_f + estrogen_f + age_f	564	502.9360		
2	race_f + estrogen_f + smk_f + age_f	562	493.9304	1 vs 2	2

	LR stat.	Pr(Chi)
1		
2	9.005622	0.01107781

Test 1 : peut-on retirer smk_f ?

Modèles comparés :

1. Modèle réduit : `subtype_f ~ race_f + estrogen_f + age_f`
2. Modèle complet : `subtype_f ~ race_f + estrogen_f + smk_f + age_f`

Résultat du test LR :

- À 5 %, on **rejette** H “on peut enlever smk_f” : le **tabagisme (smk_f)** apporte une **information significative** pour expliquer le sous-type de cancer → on **garde smk_f**.

```
# Modèle sans AGE (à partir du modèle sans SMK par exemple)
mod_sub_noage <- multinom(
  subtype_f ~ race_f + estrogen_f + smk_f,
  data = cancer
)
```

```
# weights: 15 (8 variable)
initial value 314.203115
iter 10 value 250.492692
final value 250.192192
converged
```

```
anova(mod_sub_noage, mod_sub_full)
```

Likelihood ratio tests of Multinomial Models

Response: subtype_f

	Model	Resid. df	Resid. Dev	Test	Df
1	race_f + estrogen_f + smk_f	564	500.3844		
2	race_f + estrogen_f + smk_f + age_f	562	493.9304	1 vs 2	2
	LR stat.	Pr(Chi)			
1					
2	6.454004	0.03967628			

Interprétation :

- Si la p-value du test LR est $> 5\%$, on ne rejette pas H_0 : le modèle réduit n'est pas significativement pire \rightarrow on peut **retirer** la variable.
- On garde donc **smk_f** mais pas **age**

Les degrés de liberté du test de rapport de vraisemblance sont égaux au **nombre de paramètres supprimés** entre le modèle complet et le modèle réduit.

Dans un modèle multinomial, retirer une variable facteur à L modalités enlève $(J-1)(L-1)$ coefficients, donc $(J-1)(L-1)$ degrés de liberté. J catégories et L modalités.

Test meilleur modèle

On teste sur les deux variables restantes

```
# Modèle sans RACE (à partir du modèle sans SMK par exemple)
mod_sub_norace <- multinom(
  subtype_f ~ age_f + estrogen_f + smk_f,
  data = cancer
)
```

```
# weights:  15 (8 variable)
initial  value 314.203115
iter  10 value 247.380882
final   value 247.202541
converged
```

```
anova(mod_sub_norace, mod_sub_full)
```

Likelihood ratio tests of Multinomial Models

Response: subtype_f

	Model	Resid. df	Resid. Dev	Test	Df
1	age_f + estrogen_f + smk_f	564	494.4051		
2	race_f + estrogen_f + smk_f + age_f	562	493.9304	1 vs 2	2
	LR stat.	Pr(Chi)			
1					
2	0.4747033	0.7887139			

On retire la variable estrogen pour expliquer SUBTYPE :

```
# 1) Construire un sous-échantillon complet pour toutes les variables en jeu
cancer_lr <- cancer |>
  filter(
    !is.na(subtype_f),
    !is.na(age_f),
    !is.na(race_f),
    !is.na(smk_f),
    !is.na(estrogen_f)  # même si tu ne l'utilises pas dans tous les modèles
  )

# 2) Re-estimer les modèles sur CE MÊME jeu de données
mod_sub_full <- multinom(
```

```

    subtype_f ~ age_f + race_f + smk_f + estrogen_f,
    data = cancer_lr
)

```

```

# weights:  18 (10 variable)
initial  value 314.203115
iter   10 value 247.216796
final   value 246.965190
converged

```

```

mod_sub_noestro <- multinom(
  subtype_f ~ age_f + race_f + smk_f,
  data = cancer_lr
)

```

```

# weights:  15 (8 variable)
initial  value 314.203115
iter   10 value 249.150731
final   value 248.865000
converged

```

```

# 3) Maintenant, le test LR fonctionne
anova(mod_sub_noestro, mod_sub_full, test = "Chisq")

```

Likelihood ratio tests of Multinomial Models

Response: subtype_f

	Model	Resid. df	Resid. Dev	Test	Df
1	age_f + race_f + smk_f	564	497.7300		
2	age_f + race_f + smk_f + estrogen_f	562	493.9304	1 vs 2	2
	LR stat.	Pr(Chi)			
1					
2	3.799621	0.1495969			

La fonction se base uniquement sur les **résultats passés** (modèles déjà estimés) et sélectionne celui qui minimise l'AIC (ou le BIC).

Interprétation économique du modèle final

```
mod_sub_final <- multinom(  
  subtype_f ~ age_f + smk_f,  
  data = cancer  
)
```

```
# weights:  12 (6 variable)  
initial  value 316.400339  
iter   10 value 250.030104  
final   value 250.024195  
converged
```

```
res_sub <- tidy(mod_sub_final,  
  exponentiate = TRUE, # passe en OR  
  conf.int = TRUE)     # ajoute IC 95%  
  
res_sub
```

```
# A tibble: 6 x 8  
  y.level      term      estimate std.error statistic p.value conf.low conf.high  
  <chr>      <chr>      <dbl>    <dbl>    <dbl>   <dbl>   <dbl>    <dbl>  
1 adenosquamous (Interc~  0.111    0.374    -5.88  4.16e-9  0.0535  0.231  
2 adenosquamous age_f65~  2.67     0.409     2.40  1.63e-2  1.20    5.95  
3 adenosquamous smk_fyes  2.36     0.520     1.65  9.83e-2  0.853   6.54  
4 other         (Interc~  0.283    0.271    -4.66  3.10e-6  0.166   0.481  
5 other         age_f65~  1.30     0.328     0.812 4.17e-1  0.686   2.48  
6 other         smk_fyes  0.167    1.05     -1.71  8.69e-2  0.0214  1.30
```

Principe d'interprétation :

- Un odds-ratio > 1 pour une modalité donnée signifie que la variable augmente les **cotes** d'appartenir à ce type de cancer par rapport à la référence (**adenocarcinomous**), toutes choses égales par ailleurs.
- Un odds-ratio < 1 signifie au contraire une diminution des cotes.

Tableau de contingence des individus bien / mal classés

(a) Effectifs par type de tumeur observée

```
cancer_sub |>
  count(subtype_f)
```

```
# A tibble: 3 x 2
  subtype_f      n
  <fct>      <int>
1 adenocarcinoma 184
2 adenosquamous  45
3 other          57
```

(b) Classe prédite par « probabilité max » (règle du 1er choix)

On attribue à chaque individu la **classe prédite** correspondant à la probabilité la plus élevée :

```
phat_sub <- predict(mod_sub_final, type = "probs")
colMeans(phat_sub)
```

adenocarcinoma	adenosquamous	other
0.6458337	0.1562522	0.1979141

```
phat_sub_tbl <- as_tibble(phat_sub)
names(phat_sub_tbl) <- paste0("p_", levels(cancer_complete$subtype_f))
# ex : "p_adenocarcinoma" "p_adenosquamous" "p_other"

cancer <- bind_cols(cancer, phat_sub_tbl)

# vecteur des probas sous forme de matrice
probs_mat <- as.matrix(
  cancer[, c("p_adenocarcinoma", "p_adenosquamous", "p_other")]
)

# indices de la proba max par individu (1, 2 ou 3)
idx_max <- max.col(probs_mat)
```

```
# noms des modalités dans le bon ordre
lev <- c("adenocarcinomous", "adenosquamous", "other")

cancer <- cancer |>
  mutate(
    pred_subtype = factor(lev[idx_max],
                          levels = levels(subtype_f))
  )

cancer |>
  dplyr::select(subtype_f, pred_subtype, starts_with("p_")) |>
  slice(1:5)
```

```
# A tibble: 5 x 5
  subtype_f      pred_subtype p_adenocarcinomous p_adenosquamous p_other
<fct>          <fct>          <dbl>          <dbl>    <dbl>
1 adenosquamous adenocarcinomous 0.763          0.201  0.0360
2 other         adenocarcinomous 0.717          0.0798 0.203
3 adenosquamous adenocarcinomous 0.600          0.178  0.221
4 adenocarcinomous adenocarcinomous 0.600          0.178  0.221
5 adenocarcinomous adenocarcinomous 0.717          0.0798 0.203
```

```
tab_sub <- table(
  Observed = cancer$subtype_f,
  Predicted = cancer$pred_subtype
)
```

```
tab_sub
```

	Predicted		
Observed	adenocarcinomous	adenosquamous	other
adenocarcinomous	186	0	0
adenosquamous	45	0	0
other	57	0	0

```
prop_ok <- sum(diag(tab_sub)) / sum(tab_sub)
1-prop_ok
```

[1] 0.3541667

B. Modèle ordinal pour expliquer GRADE

On modélise le **stade de la tumeur** (bien / moyennement / peu différenciée) en fonction de :

- RACE, ESTROGEN, SUBTYPE, AGE, SMK.

Modèle de base (logit ordinal)

On utilise `polr()` (MASS) avec `grade_ord` comme variable ordinale.

```
mod_grade_base <- polr(
  grade_ord ~ race_f + estrogen_f + subtype_f + age_f + smk_f,
  data = cancer,
  Hess = TRUE
)

summary(mod_grade_base)
```

Call:

```
polr(formula = grade_ord ~ race_f + estrogen_f + subtype_f +
      age_f + smk_f, data = cancer, Hess = TRUE)
```

Coefficients:

	Value	Std. Error	t value
race_fnoire	0.59764	0.2790	2.14222
estrogen_ftook oestrogen treatment	-0.61411	0.2549	-2.40904
subtype_fadenosquamous	1.78110	0.3252	5.47653
subtype_fother	0.07823	0.2991	0.26153
age_f65-79	0.13110	0.2486	0.52742
smk_fyes	0.01358	0.3992	0.03401

Intercepts:

Value	Std. Error	t value
-------	------------	---------

bien différencié moyennement différencié	-0.0205	0.2890	-0.0708
moyennement différencié peu différencié	1.9682	0.3196	6.1589

Residual Deviance: 540.4501

AIC: 556.4501

(2 observations deleted due to missingness)

Les sorties donnent :

- Les coefficients β (effets sur le **score latent**),
- Les seuils (cutpoints) séparant les catégories de **GRADE**.

Test d'ajustement via interactions (LR tests)

Interaction ESTROGEN \times SUBTYPE

On ajoute l'interaction `estrogen_f * subtype_f` :

```
mod_grade_es <- polr(
  grade_ord ~ race_f + estrogen_f * subtype_f + age_f + smk_f,
  data = cancer,
  Hess = TRUE
)

anova(mod_grade_base, mod_grade_es)
```

Likelihood ratio tests of ordinal regression models

Response: grade_ord

	Model	Resid. df	Resid. Dev	Test
1	race_f + estrogen_f + subtype_f + age_f + smk_f	278	540.4501	
2	race_f + estrogen_f * subtype_f + age_f + smk_f	276	537.5181	1 vs 2
	Df LR stat. Pr(Chi)			
1				
2	2 2.932012 0.2308457			

- `anova()` réalise un **test LR** entre modèle sans interaction (`mod_grade_base`) et modèle avec interaction (`mod_grade_es`).
- On lit la **p-value** :

- si petite \rightarrow l'interaction améliore le modèle ;
- si grande \rightarrow on peut s'en passer.

Autres interactions possibles

On peut tester d'autres interactions pertinentes, par exemple :

- ESTROGEN \times AGE :

```
mod_grade_eage <- polr(
  grade_ord ~ race_f + estrogen_f * age_f + subtype_f + smk_f,
  data = cancer,
  Hess = TRUE
)

anova(mod_grade_base, mod_grade_eage)
```

Likelihood ratio tests of ordinal regression models

Response: grade_ord

	Model	Resid. df	Resid. Dev	Test
1	race_f + estrogen_f + subtype_f + age_f + smk_f	278	540.4501	
2	race_f + estrogen_f * age_f + subtype_f + smk_f	277	539.2085	1 vs 2

	Df	LR stat.	Pr(Chi)
1			
2	1	1.241621	0.2651588

-
- SUBTYPE \times AGE :

```
mod_grade_sage <- polr(
  grade_ord ~ race_f + estrogen_f + subtype_f * age_f + smk_f,
  data = cancer,
  Hess = TRUE
)

anova(mod_grade_base, mod_grade_sage)
```

Likelihood ratio tests of ordinal regression models

Response: grade_ord

	Model	Resid. df	Resid. Dev	Test
1	race_f + estrogen_f + subtype_f + age_f + smk_f	278	540.4501	
2	race_f + estrogen_f + subtype_f * age_f + smk_f	276	538.9157	1 vs 2
	Df LR stat.	Pr(Chi)		
1				
2	2	1.534397	0.4643119	

Pour chaque interaction :

- Si la p-value LR est **faible** → interaction importante → à garder ;
- Sinon → pas de gain significatif → on privilégie le modèle sans interaction.

Sélection de modèle

```
summary(mod_grade_base)
```

Call:

```
polr(formula = grade_ord ~ race_f + estrogen_f + subtype_f +  
      age_f + smk_f, data = cancer, Hess = TRUE)
```

Coefficients:

	Value	Std. Error	t value
race_fnoire	0.59764	0.2790	2.14222
estrogen_ftook oestrogen treatment	-0.61411	0.2549	-2.40904
subtype_fadenosquamous	1.78110	0.3252	5.47653
subtype_fother	0.07823	0.2991	0.26153
age_f65-79	0.13110	0.2486	0.52742
smk_fyes	0.01358	0.3992	0.03401

Intercepts:

	Value	Std. Error	t value
bien différencié moyennement différencié	-0.0205	0.2890	-0.0708
moyennement différencié peu différencié	1.9682	0.3196	6.1589

Residual Deviance: 540.4501

AIC: 556.4501
(2 observations deleted due to missingness)

Interprétation du modèle ordinal final

- **Race (noire vs non noire)**
 - Coefficient = 0,60 (t = 2,14) → **significatif**.
 - Les femmes noires ont des **cotes 1,8 fois plus élevées** d’avoir un grade plus mauvais (passer vers “moyennement/peu différencié”), toutes choses égales par ailleurs.
 - **Traitement aux œstrogènes (oui vs non)**
 - Coefficient = -0,61 (t = -2,41) → **significatif**.
 - Les femmes ayant reçu un traitement œstrogénique ont des **cotes divisées par 2** d’avoir un grade plus mauvais. → Le traitement est associé à des **grades un peu meilleurs**.
 - **Sous-type tumoral (réf. = adenocarcinomous)**
 - *adenosquamous* : coef = 1,78 (t = 5,48) → très significatif.
→ Cotes 6 fois plus élevées d’avoir un **grade plus défavorable** que l’adenocarcinome.
 - *other* : effet faible et **non significatif**.
 - **Âge (65–79 ans) et tabagisme (smk_fyes)**
 - Coefficients proches de 0, t très faibles → **pas d’association significative** avec le grade, une fois contrôlé pour race, sous-type et œstrogènes.
-

Conclusion TD3

- On a estimé :
 - un **logit multinomial nominal** pour SUBTYPE ;
 - un **logit ordinal** pour GRADE.
- On a :
 - testé l'ajustement du modèle nominal via un **Hosmer–Lemeshow multinomial** ;
 - utilisé des **tests de rapport de vraisemblance** pour :
 - * simplifier les modèles (variables non significatives),
 - * tester l'intérêt d'interactions dans le modèle ordinal ;
 - Choisi les spécifications finales.

À retenir :

- Pour les variables **nominales**, on compare chaque modalité à une **référence** via des odds-ratios.
- Pour les variables **ordinales**, le modèle logit/probit ordonné repose sur une **variable latente** et des **seuils**, avec une interprétation en termes de tendance vers des catégories plus élevées ou plus basses.