Données manquantes et régression multinomiale

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

Données manquantes et inférence

- Les données manquantes réduisent la quantité d'information disponible et augmentent l'incertitude.
- On ne peut pas les ignorer (étude des cas complets) sans biaiser les interprétations et réduire la quantité d'information disponible.
- On considère des méthodes d'imputation qui remplacent les valeurs manquantes.

Imputation aléatoire

Considérons le cas d'une régression logistique pour une variable explicative binaire.

Plutôt que d'assigner à la classe la plus probable, une prédiction aléatoire simule une variable 0/1 avec probabilité $(1-\widehat{p}_i, \widehat{p}_i)$.

```
pred <- 0.3 #probabilité de succès
rbinom(n = 15, size = 1, prob = pred)
```

```
 [1] 0 1 1 0 0 0 0 1 1 1 1 0 1 0 1
```

Problèmes de l'imputation simple

On ne tient pas compte du fait que des valeurs ont été remplacées (on fait comme si c'était de vraies observations).

On sous-évalue encore une fois la variabilité des données

- les écarts-type des estimations sont trop petits.

Il est donc nécessaire d'examiner la configuration des valeurs manquantes avant de faire quoi que ce soit.

```
data(manquantes, package = 'hecmulti')
summary(manquantes)
# Pourcentage de valeurs manquantes
apply(manquantes, 2, function(x){mean(is.na(x))})
# Voir les configurations de valeurs manquantes
md.pattern(manquantes) # graphique diapo suivante
```

Table 1: Nombre et pourcentage de valeurs manquantes par variable.

	x1	x2	x3	x4	x5	x6	y
nombre	192	49	0	184	0	0	0
pourcentage	38.4	9.8	0	36.8	0	0	0

Configuration des valeurs manquantes

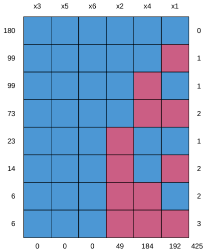
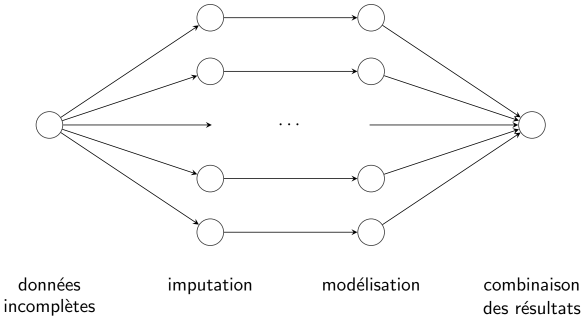


Figure 1: Configurations des valeurs manquantes pour `manquantes`.

Imputation multiple

Valides pour les données manquantes de manière aléatoire et complètement aléatoires (MAR et MCAR).

1. Procéder à plusieurs imputations aléatoires pour obtenir un échantillon complet (`mice`)
2. Ajuster le modèle d'intérêt avec chaque échantillon (`with`). 3. Combiner les résultats obtenus (`pool` et `summary`)



Combinaison des résultats

Considérons un seul paramètre θ (ex: coefficient d'une régression) et supposons qu'on procède à K imputations.

On estime les paramètres du modèle séparément pour chacun des K ensembles de données imputés, disons

- $\widehat{\theta}_k$ pour l'estimation du paramètre θ dans l'échantillon k et
- $\hat{\sigma}_k^2 = \mathsf{Va}(\widehat{\theta}_k)$ pour l'estimation de la variance de $\widehat{\theta}_k$.

Estimation du paramètre moyen

L'estimation finale de θ, dénotée $\widehat{\theta}$, est obtenue tout simplement en faisant la moyenne des estimations de tous les modèles, c'est-à-dire,

$$\widehat{\theta} = \frac{\widehat{\theta}_1 + \cdots + \widehat{\theta}_K}{K}.$$

Estimation des erreurs-types

Une estimation ajustée de la variance de $\hat{\theta}$ est

$$\mathsf{Va}(\hat{\theta}) = W + \frac{K+1}{K} B,$$

où

$$W = \frac{1}{K}\sum_{k=1}^{K} \hat{\sigma}_k^2 = \frac{\hat{\sigma}_1^2 + \cdots + \hat{\sigma}_K^2}{K},$$

$$B = \frac{1}{K-1}\sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta})^2.$$

- W est la moyenne des variances (variance intra-groupe) et
- B la variance des moyennes (variance inter-groupe).

Imputation multiple

Des formules analogues existent pour les degrés de liberté, les valeurs-p, etc. ainsi que pour la cas multidimensionnel (plusieurs paramètres).

Si on procédait à une seule imputation (même en ajoutant une part d'aléatoire pour essayer de reproduire la variabilité des données), on ne serait pas en mesure d'estimer la variance inter-groupe de l'estimateur.

On peut estimer la fraction de l'information manquante sur θ avec $(1+1/K)B/\mathsf{Va}(\widehat{\theta})$.

Imputation multiple par équations chaînées (MICE)

Avec p variables $X_1, \ldots, X_p$, spécifier un ensemble de modèles conditionnels pour chaque variable X_j en fonction de

- toutes les autres variables, X_{-j}
- les valeurs observées pour cette variable, $X_{j,\text{obs}}$

1. Initialisation: remplir les trous avec des données au hasard parmi $X_{j,\text{obs}}$ pour $X_{j,\text{man}}$
2. À l'itération t, pour chaque variable $j = 1, \ldots, p$, à tour de rôle:
   a) tirage aléatoire des paramètres $\phi_j^{(t)}$ du modèle pour $X_{j,\text{man}}$ conditionnel à $X_{-j}^{(t-1)}$ et $X_{j,\text{obs}}$
   b) échantillonnage de nouvelles observations $X_{j,\text{man}}^{(t)}$ du modèle avec paramètres $\phi_j^{(t)}$ conditionnel à $X_{-j}^{(t-1)}$ et $X_{j,\text{obs}}$
3. Répéter le cycle

Imputation multiple avec `mice`

```r
library(mice)
# Intensif en calcul, réduire "m" si nécessaire
impdata <- mice(
   data = manquantes,
   # argument "method" pour le modèle
   # dépend du type des variables, par ex.
   # régression logistique pour données binaires
   m = 50, # nombre d'imputations
   seed = 60602, # germe aléatoire
   printFlag = FALSE)
# Extraite une copie (m=1,..., 50) imputée
complete(data = impdata,
         action = 1) #no de la copie
```

Estimation et combinaison avec `mice`

```r
# ajuste le modèle avec les données imputées
adj_im <- with(
  data = impdata,
  expr = glm(y ~ x1 + x2 + x3 + x4 + x5 + x6,
             family = binomial))
# combinaison des résultats
fit <- pool(adj_im)
summary(fit)
```

Résultats

terme	estimation	erreur-type	stat	ddl	valeur-p
(Intercept)	-2.86	1.08	-2.65	327.28	0.009
x12	-0.90	0.58	-1.55	172.63	0.124
x13	-0.53	0.59	-0.90	164.98	0.372
x14	-0.76	0.55	-1.39	241.12	0.166
x15	-0.58	0.64	-0.91	176.76	0.365
x22	-0.03	0.44	-0.07	403.25	0.944
x23	-0.65	0.49	-1.34	353.29	0.181
x24	-2.60	0.65	-4.00	271.66	<1e-04
x25	-1.32	0.72	-1.84	283.49	0.066
x3	1.61	0.31	5.19	300.31	<1e-04
x4	2.56	0.42	6.14	185.05	<1e-04
x5	0.11	0.02	5.27	436.59	<1e-04
x62	-1.41	0.37	-3.83	408.17	1e-04
x63	-2.58	0.41	-6.24	399.50	<1e-04

Récapitulatif

- Les données manquantes réduisent la quantité d'information disponible et augmentent l'incertitude.
- On ne peut pas les ignorer (étude des cas complets) sans biaiser les interprétations et réduire la quantité d'information disponible.
- Pour bien capturer l'incertitude et ne pas modifier les relations entre variables, il faut utiliser une méthode aléatoire.
- Avec l'algorithme MICE, on utilise un modèle conditionnel pour chaque variable à tour de rôle

Récapitulatif

L'imputation multiple est préférée à l'imputation simple car elle permet d'estimer l'incertitude sous-jacente en raison des données manquantes.

- On procède à l'imputation plusieurs fois (avec un modèle conditionel, prédictions différentes chaque fois)
- on crée plusieurs copies
- ajuste le modèle sur chacune et
- combine les résultats

Traitement spécial pour erreurs-type, degrés de liberté, valeurs-p et intervalles de confiance.

Régression pour données multinomiale

On considère une variable réponse catégorielle avec $K \geq 2$ modalités.

Objectif: modéliser la probabilité de chaque catégorie de la variable réponse.

Soit la probabilité d'appartenir à la modalité k,

$$p_{ik} = \mathsf{Pr}(Y_i = k \mid \mathsf{X}_i), \qquad (k = 1, \ldots, K).$$

La somme des probabilités, $p_{i0} + \cdots + p_{iK}$, vaut 1.

Modèle multinomial logistique

Comme avec la régression logistique, on fixe une catégorie de référence (disons 1) et on modélise le log de la cote de chacune des autres catégories par rapport à cette référence,

$$\ln\left(\frac{p_{ij}}{p_{i1}}\right) = \eta_{ij} = \beta_{0j} + \beta_{1j}\mathsf{X}_{i1} + \cdots + \beta_{pj}\mathsf{X}_{ip}, \quad (j = 2, \ldots, K).$$

- Avec K modalités et p variables explicatives, on obtiendra $(K-1) \times (p+1)$ paramètres à estimer, en incluant l'ordonnée à l'origine.

L'interprétation des paramètres se fait comme en régression logistique sauf qu'il faut y aller équation par équation.

Modèle logistique multinomial et probabilités

On peut aussi exprimer le modèle en termes des probabilités,

$$
\begin{aligned}
p_{i1} &= \mathsf{Pr}(Y_i = 1 \mid \mathsf{X}_i) = \frac{1}{1+\mathsf{exp}(\eta_{i2}) + \cdots + \mathsf{exp}(\eta_{iK})}\\
p_{ik} &= \mathsf{Pr}(Y_i = k \mid \mathsf{X}_i) = \frac{\mathsf{exp}(\eta_{ik})}{1+\mathsf{exp}(\eta_{i2}) + \cdots + \mathsf{exp}(\eta_{iK})}, \qquad k = 2, \ldots, K.
\end{aligned}
$$

où η_{ij} est le prédicteur linéaire de l'individu i pour le log de la cote de $Y_i = j$ versus la référence $Y_i = 1$.

Taux de participation lors des élections américaines

Les données de cet exemple sont tirées d'un sondage Ipsos réalisé pour le site de nouvelles FiveThirtyEight.

La base de données `vote` contient 5837 observations avec les pondérations associées.

Nous allons modéliser l'intention de vote, `catvote` à l'aide d'une régression logistique multinomiale.

Analyse exploratoire

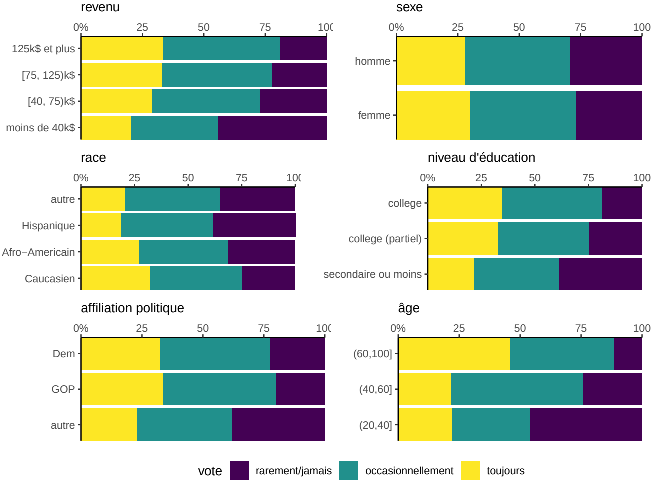


Figure 2: Proportion des modalités des variables sociodémographiques des données de participation électorale.

Analyse exploratoire (age)

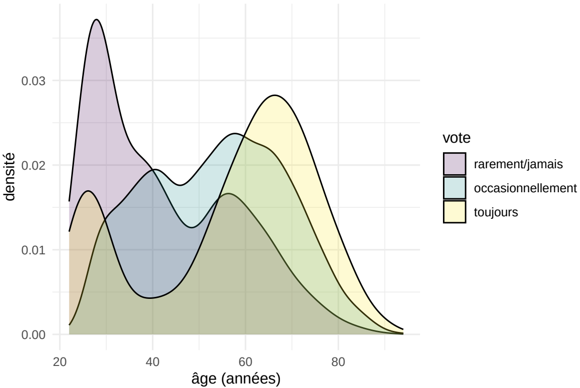


Figure 3: Fréquence de vote selon l'âge.

Notez le comportement des jeunes voteurs (bimodal). Ces personnes n'ont souvent eu qu'une seule occasion de voter...

Ajustement du modèle

La fonction `multinom` du paquet `nnet` ajuste le modèle multinomial logistique.

```
data(vote, package = "hecmulti")
levels(vote$catvote)
```

```
[1] "rarement/jamais"   "occasionnellement" "toujours"
```

```
# Modèle multinomial
multi1 <- nnet::multinom(
  catvote ~ age + sexe + race + revenu +
    educ + affiliation,
  data = vote,        # base de données
  subset = age > 30, # sous-ensemble des données
  weights = poids,    # poids de sondage
  trace = FALSE)      # infos sur convergence
```

```r
# Tableau résumé de l'ajustement
summary(multi1)
# Estimations des coefficients
coef(multi1)
# Intervalles de confiance (Wald)
confint(multi1)
# Critères d'information
AIC(multi1)
BIC(multi1)
# Prédiction: probabilité de chaque modalité
predict(multi1, type = "probs")
# Prédiction: classe la plus susceptible
predict(multi1, type = "class")
```

Comparaison de modèles emboîtés

Le modèle avec uniquement l'ordonnée à l'origine possède $K-1$ paramètres. Il retourne comme probabilité prédite la proportion empirique de chaque catégorie.

```
multi0 <- nnet::multinom(catvote ~ 1,
                         weights = poids,
             subset = age > 30,
                         data = vote,
                         trace = FALSE)
# Test de rapport de vraisemblance
anova(multi0, multi1)
```

Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
9692	9781.07				
9668	8504.74	1 vs 2	24	1276.33	0

Prédictions

Pour un profil $\mathbf{x}_i$ donné, on peut

- calculer chacun des $K-1$ prédicteurs linéaires $\widehat{\eta}_{i2}, \ldots, \widehat{\eta}_{iK}$.
- écrire $p_{ik} = p_{i1}\exp(\widehat{\eta}_{ik})$ (formule de la cote)
- substituer cette mesure dans l'équation $p_{i1} + \cdots + p_{iK} = 1$
- isoler la prédiction numérique pour p_{i1}.
- en déduire les probabilités de succès de chaque modalité de Y.

Exemple au tableau

Classification

La prédiction du modèle est une probabilité pour chacune des K modalités.

On peut toujours classifier les événements

- avec $K-1$ points de coupure...
- ou assigner à la modalité la plus probable

Avec les prédictions, on peut comparer les observations et les prédictions à l'aide d'une matrice de confusion $K \times K$.

- Le taux de bonne classification est toujours valide
- Il existe des extensions multidimensionnelles de l'aire sous la courbe

Commentaires

- Contrairement à la régression logistique, le nombre de paramètres augmente rapidement avec le nombre de variables explicatives, p.
- Il y a moins d'information pour estimer les paramètres qu'une régression linéaire: prévoir de plus grandes tailles d'échantillon.
- Attention aux modalités à faible fréquence et à la répartition des variables explicatives au sein des différentes modalités.

Données ordinales

Outre la régression multinomiale logistique, on peut également considérer la régression logistique cumulative à cotes proportionnelles.

- modèle plus parcimonieux que le modèle multinomial logistique,
- mais au prix de postulats supplémentaires...

En R, la variable réponse doit être de classe `ordered`, un facteur dont les niveaux sont ordonnés en ordre croissant.

```r
class(hecmulti::vote$catvote)
```

```
[1] "ordered" "factor"
```

Notation du modèle

- Soit $p_1, \ldots, p_K$ les probabilités associées aux événements $Y = 1, \ldots$
- On définit les points de coupure pour les K classes,

$$-\infty = \zeta_0 < \zeta_1 < \cdots < \zeta_K = \infty.$$

- Il y a $K - 1$ paramètres ζ à déduire pour identifier les probabilités puisque $p_1 + \cdots + p_K = 1$.

Formulation du modèle

Le modèle logistique à cote proportionnelle spécifie $K-1$ équations logistiques; pour $k=1,\ldots,K-1$,

$$\ln\left(\frac{\mathsf{Pr}(Y_i > k \mid \mathsf{X}_i)}{\mathsf{Pr}(Y_i \leq k \mid \mathsf{X}_i)}\right) = -\zeta_k + \beta_1 \mathsf{X}_{i1} + \cdots + \beta_p \mathsf{X}_{ip}.$$

- Les paramètres associés aux variables explicatives, $\beta_1, \ldots, \beta_p$ sont les mêmes pour chacune des log-cotes
- mais il y a une ordonnée à l'origine différente par rapport de cote, $-\zeta_k$.

Interprétation des paramètres

On considère la cote de $\mathsf{Pr}(Y_i > k \mid \mathsf{X}_i)$ versus $\mathsf{Pr}(Y_i \leq k \mid \mathsf{X}_i)$, qui mesure à quel point il est plus probable que Y_i prenne une valeur supérieure à k par rapport à une valeur inférieure ou égale à k, avec

Pour chaque augmentation d'une unité de X_j, cette cote est multipliée par $\exp(\beta_j)$, peu importe la valeur de Y (cote proportionnelle).

Interprétation des coefficients

Table 2: Tableau des estimations des coefficients du modèle pour réponses ordinales pour la régression logistique à cotes proportionnelles avec sexe.

effet	coefficient	erreur-type
sexe [homme]	-0.166	0.055
cst [rarement/jamais|occasionnellement]	-1.297	0.044
cst [occasionnellement|toujours]	0.865	0.04

Les hommes de plus de 30 ans sont moins susceptibles de voter fréquemment que les femmes.

La cote catégorie plus fréquente de vote (vs moins fréquente) pour les hommes est $\exp(-0.166) = 0.847$ fois celle des femmes, soit une diminution de 15.3% de la cote.

Pour simplifier, on utilise uniquement `sexe` comme variable explicative.

```
# with(vote, is.ordered(catvote))
multi2a <- MASS::polr(
  catvote ~ sexe,
  data = vote,
  subset = age > 30,
  weights = poids,
  method = "logistic",
  Hess = TRUE)
summary(multi2a)
```

Méthodes pour `polr`

```r
# IC pour beta_x (vraisemblance profilée)
confint(multi2a)
# On peut obtenir les intervalles de Wald
# avec confint.default (PAS RECOMMANDÉ)

# Critères d'information
AIC(multi2a); BIC(multi2a)
# Tableau des coefficients
# Coefficients (variables explicatives)
coef(multi2a)
# Négatif de l'ordonnée à l'origine:
multi2a$zeta
```

Coefficients pour l'ordonnée à l'origine

Si on écrit les équations pour la cote, on obtient

$$\begin{aligned}\frac{\Pr(Y = \texttt{rarement} \mid \texttt{sexe})}{\Pr(Y \geq \texttt{occasionnellement} \mid \texttt{sexe})} &= \exp(-0.166\texttt{sexe} + 1.297)\\ \frac{\Pr(Y \leq \texttt{occasionnellement} \mid \texttt{sexe})}{\Pr(Y = \texttt{toujours} \mid \texttt{sexe})} &= \exp(-0.166\texttt{sexe} - 0.865).\end{aligned}$$

Probabilités prédites

En terme de probabilité cumulée d'excéder k,

$$\Pr(Y_i > k \mid \mathsf{X}_i) = \operatorname{expit}(-\eta_k + \beta_1 \mathsf{X}_{i1} + \cdots + \beta_p \mathsf{X}_{ip}), \quad k = 1, \ldots, K-1.$$

En utilisant ces expressions, on peut obtenir la probabilité de chaque catégorie,

$$\Pr(Y_i = k \mid \mathsf{X}_i) = \Pr(Y_i > k \mid \mathsf{X}_i) - \Pr(Y_i > k-1 \mid \mathsf{X}_i).$$

Prédictions

- Soit $p_1 = \Pr(Y = \texttt{rarement/jamais} \mid \texttt{femme})$, etc.
- On a $\mathrm{expit}(\zeta_k)$ ($k = 0, \ldots, K$) qui donne $0, 0.215, 0.704$ et 1.
- Les différences donnent $\widehat{p}_1 = 0.215, \widehat{p}_2 = 0.489$ et $\widehat{p}_3 = 0.296$.
- Un rapide calcul numérique montre que c'est bien ce que retourne les prédictions.

```r
predict(multi2a,
          newdata = data.frame(sexe = factor("femme")),
          type = "probs")
```

Postulat de rapport de cote proportionnel

Une des hypothèses de ce modèle est que les effets des variables explicatives sont les mêmes pour chaque équation.

- $\mathscr{H}_0$: l'effet de chaque variable est le même pour les K logit du modèle .

Une très petite valeur-p (rejet de $\mathscr{H}_0$) pour ce test serait une indication que le modèle multinomial logistique serait préférable.

Test de rapport de vraisemblance

Ce test compare les deux modèles emboîtés, avec

- hypothèse nulle $\mathscr{H}_0$: modèle cumulatif à cotes proportionnelles, avec $p+K-1$ paramètres
- hypothèse alternative $\mathscr{H}_a$: modèle multinomial, avec $(K-1)\times(p+1)$ paramètres

```r
multi2b <- nnet::multinom(catvote ~ sexe,
  data = vote,  subset = age > 30,
  weights = poids, trace = FALSE)
# Valeur-p du test de rapport de vraisemblance
pchisq(q = deviance(multi2a) - deviance(multi2b),
       df = length(coef(multi2a)),
       lower.tail = FALSE)
```

```
[1] 0.7886889
```

Le modèle sous $\mathscr{H}_0$ semble être une simplification adéquate.

Test du rapport de vraisemblance

Si on ajuste plutôt le modèle avec uniquement `age`, la valeur-p est inférieure à 10^{-5}: le modèle cumulatif à cote proportionnelles ne serait pas une simplification adéquate.

On peut également effectuer des tests pour déterminer la significativité

- la significativité globale (ordonnée à l'origine vs modèle complet)
- l'effet d'une variable explicative (modèle complet, moins une variable)

Comparaison des prédictions

Prédictions pour le modèle avec uniquement `age` comme variable explicative.

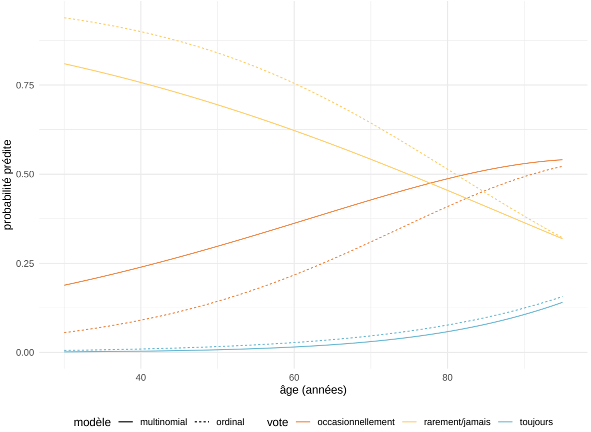


Figure 4: Probabilités prédites pour chaque modalité selon l'âge.

Récapitulatif

- La régression multinomiale logistique pour une variable catégorielle à K niveaux est une extension directe de la régression logistique pour données binaires
  - la somme des probabilités vaut 1.
  - il y a $K-1$ équations de cote en termes des variables explicatives,
  - donc le nombre de paramètres croît rapidement.

Récapitulatif

On met beaucoup l'accent sur l'interprétation des coefficients à l'échelle de la cote.

- rapports de cote = modèle multiplicatif: la cote de catégorie k vs référence est multipliée par $\exp(\beta_{jk})$ pour chaque augmentation de X_j d'une unité.
- les coefficients manquants (cote de $Y = k$ vs $Y = l$) peut être déduits par des manipulations algébriques.

Récapitulatif

Les outils usuels d'inférence pour les modèles estimés par maximum de vraisemblance sont applicables.

- intervalles de confiance (Wald ou vraisemblance profilée)
- tests de rapport de vraisemblance
- critères d'information

Récapitulatif

Côté classification, on va règle générale assigner à la classe la plus probable.

- il existe des équivalents multidimensionnels directs à ce qu'on a couvert (matrice de confusion, taux de bonne classification, gain, etc.)
- certains concepts (sensibilité, spécificité, fonction d'efficacité du récepteur) ne sont en revanche pas applicables ou n'ont pas d'équivalent.

Récapitulatif

Le modèle cumulatif à cote proportionnelle est une simplification du modèle multinomial pour des données ordinales.

- On suppose que l'effet des variables est le même pour la cote de la survie de chaque modalité.
- Moins de paramètres, mais postulat à vérifier (via test de rapport de vraisemblance).