Régression logistique: prédictions et données multinomiales

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

automne 2022

Sélection de variables en régression logistique

- On applique les mêmes principes que précédemment.
- Notre mesure d'ajustement (gain, taux de bonne classification, log-vraisemblance) peut différer selon l'objectif.
- Les modèles de régression logistique sont plus coûteux à estimer.
- Pour la classification, le point de coupure est à déterminer.

Fonctions **R** pour la sélection de modèles

- `glmbb::glmbb` permet une recherche exhaustive de tous les sous-modèles à au plus une certaine distance (`cutoff`) du modèle avec le plus petit critère d'information (`criterion`).
- `step` permet de faire une recherche séquentielle avec un critère d'information.
- `glmulti::glmulti` permet une recherche exhaustive (`method = "h"`) ou par le biais d'un algorithme génétique (`method = "g"`).
- `glmnet::glmnet` permet d'ajuster le modèle avec pénalité LASSO.

Voir le code en ligne.

Objectif du ciblage marketing

Déterminer si le revenu prévu justifie l'envoi du catalogue

$$\mathsf{E}(\mathsf{ymontant}_i) = \mathsf{E}(\mathsf{ymontant}_i \mid \mathsf{yachat}_i = 1)\,\mathsf{Pr}(\mathsf{yachat}_i = 1).$$

On peut combiner un modèle de régression logistique avec la régression linéaire (ajustés simultanément avec un modèle Heckit).

Ou simplement ignorer le montant d'achat et envoyer un catalogue si la probabilité d'achat excède notre point de coupure optimal.

Stratégie de référence

- Parmi les 100K clients, 23 179 auraient acheté si on leur avait envoyé le catalogue
- Ces clients auraient généré des revenus de 1 601 212$.
- Si on enlève le coût des envois (100 000 $\times$ 10$), la stratégie de référence permet un revenu net de 601 212$.

Stratégie d'ajustement

En résumé, la procédure numérique à réaliser est la suivante:

- Choisir les variables à essayer (termes quadratiques, interactions, etc.)
- Choisir l'algorithme ou la méthode de sélection du modèle.
- Construire un catalogue de modèles: pour chacun, calculer les prédictions par validation croisée.
- Calculer le point de coupure optimal pour chaque modèle selon la fonction de gain moyen.
- Sélectionner le modèle qui **maximise le gain**.

Prédiction et envoi

- Prédire les 100 000 observations de l'échantillon test.
- Envoyer un catalogue si la probabilité d'achat excède le point de coupure.
- Calculer le revenu résultant:
  - zéro si on n'envoie pas de catalogue
  - -10 si la personne n'achète pas
  - -10 plus l'achat si la personne achète.

En pratique, on ne pourrait pas *a priori* connaître le revenu résultant de cette stratégie.

Conclusion

Si on avait fait une bête recherche séquentielle et qu'on avait pris le modèle avec le plus petit BIC (8 variables explicatives), on aurait dégagé des revenus de 978 226$.

C'est une énorme amélioration, de plus de 56%, par rapport à la stratégie de référence.

Récapitulatif

- Les principes de sélection de variable couverts précédemment s'appliquent toujours (recherche exhaustive, séquentielle et LASSO).
- On peut aussi calculer les critères d'information puisque le modèle est ajusté par maximum de vraisemblance.
- Attention au surajustement! Suspect si les probabilités estimées sont près de 0/1 (vérifier la calibration).
- Deux étapes: sélectionner le modèle (variables) et le point de coupure.
- D'autres modèles que la régression logistique (arbres de classification, etc.) sont envisageables pour la classification.

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,

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

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,

$$p_{i1} = \Pr(Y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + \exp(\eta_{i2}) + \cdots + \exp(\eta_{iK})}$$

$$p_{ik} = \Pr(Y_i = k \mid \mathbf{x}_i) = \frac{\exp(\eta_{ik})}{1 + \exp(\eta_{i2}) + \cdots + \exp(\eta_{iK})}, \qquad k = 2, \ldots, K.$$

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

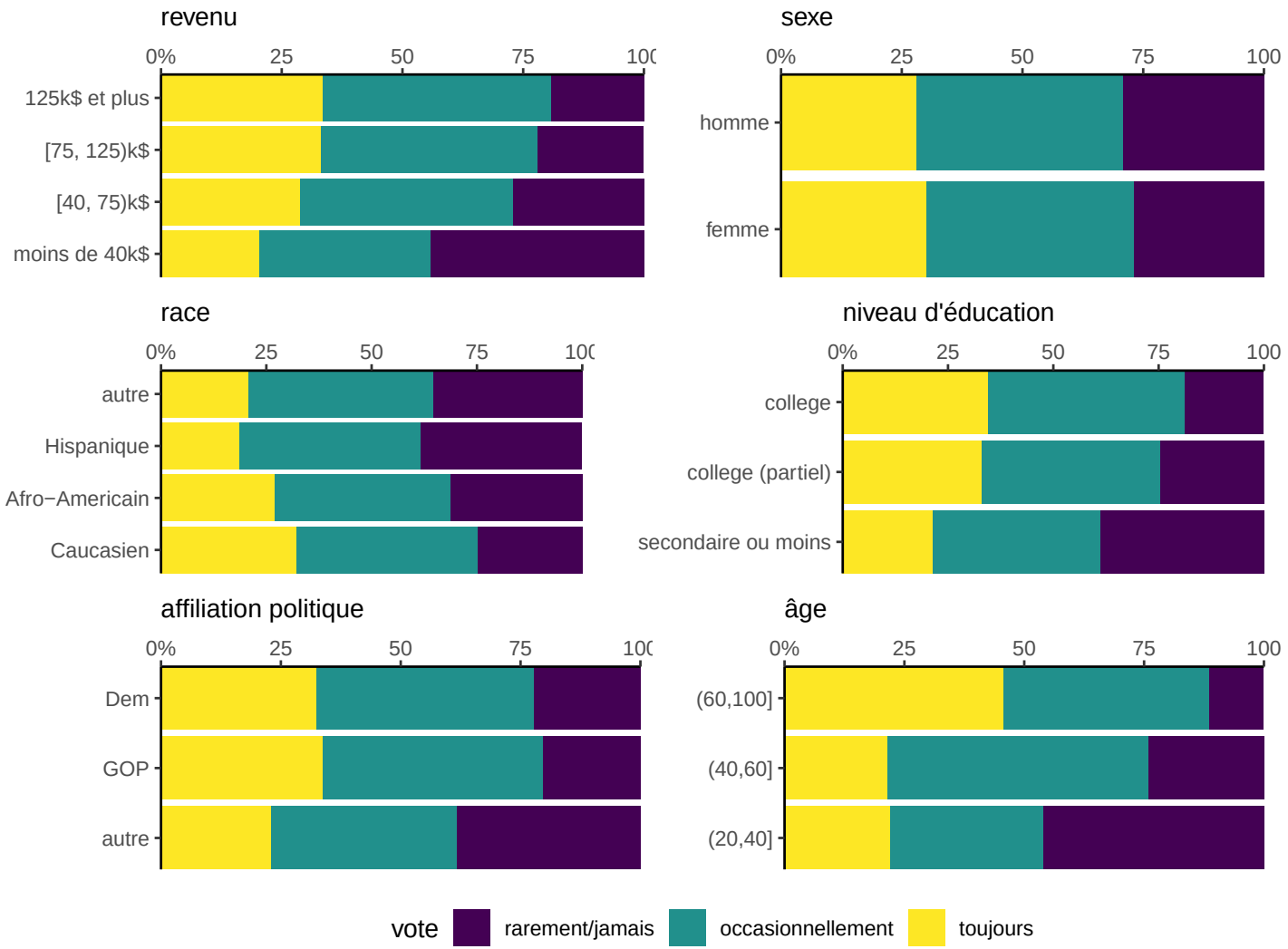

Figure 1: Proportion des modalités des variables sociodémographiques des données de participation électorale.

Analyse exploratoire (age)

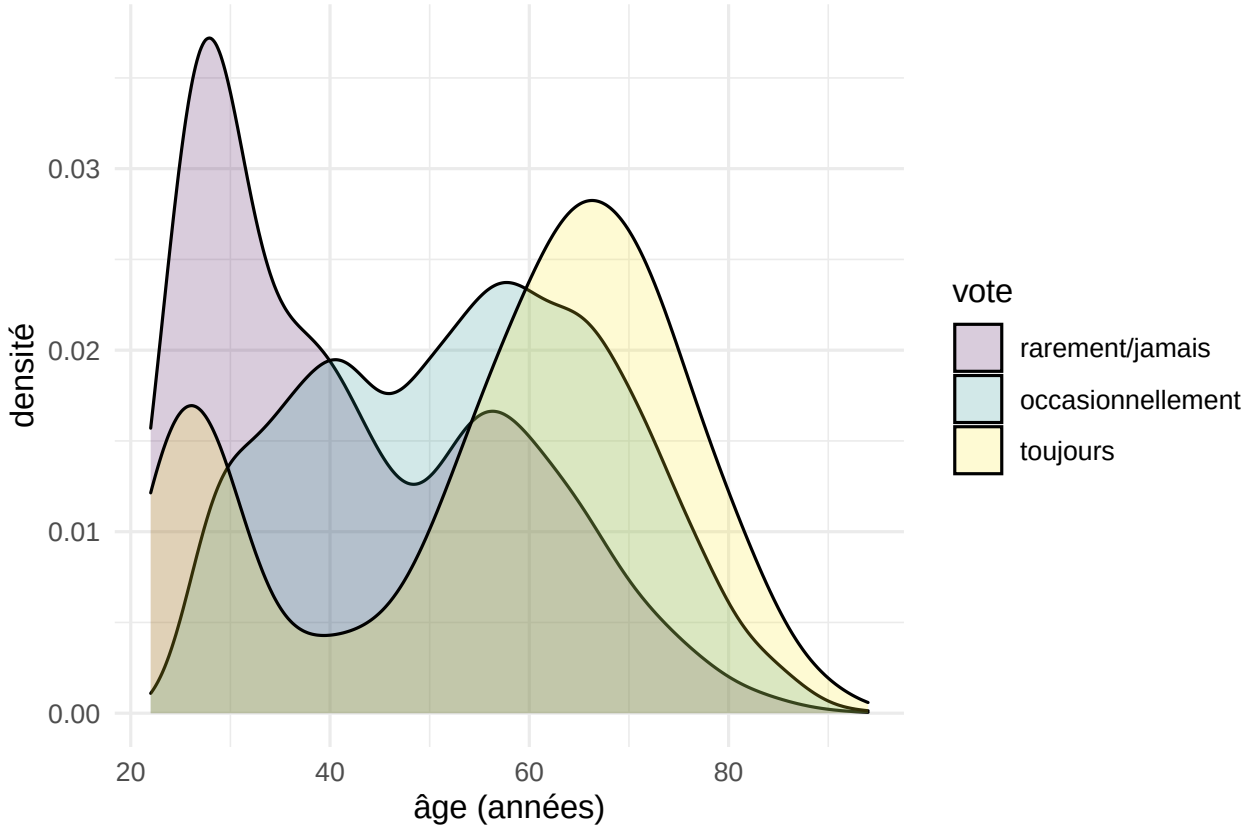

Figure 2: Fréquence de vote selon l'âge.

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
```

Méthodes **R** pour l'analyse

```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 $\hat{\eta}_{i2}, \dots, \hat{\eta}_{iK}$.
- écrire $p_{ik} = p_{i1} \exp(\hat{\eta}_{ik})$ (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.

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

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

Notation du modèle

Soit $p_1, \dots, p_K$ les probabilités associées aux événements $Y = 1, \dots, Y = K$. 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 $k = 1, \ldots, K-1$,

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

- 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 $\Pr(Y_i > k \mid \mathbf{x}_i)$ versus $\Pr(Y_i \leq k \mid \mathbf{x}_i)$, 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	0.041

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.

Ajustement dans **R**

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

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

Probabilités prédites

En terme de probabilité cumulée d'excéder k,

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

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

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

Prédictions

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

```
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

```
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.

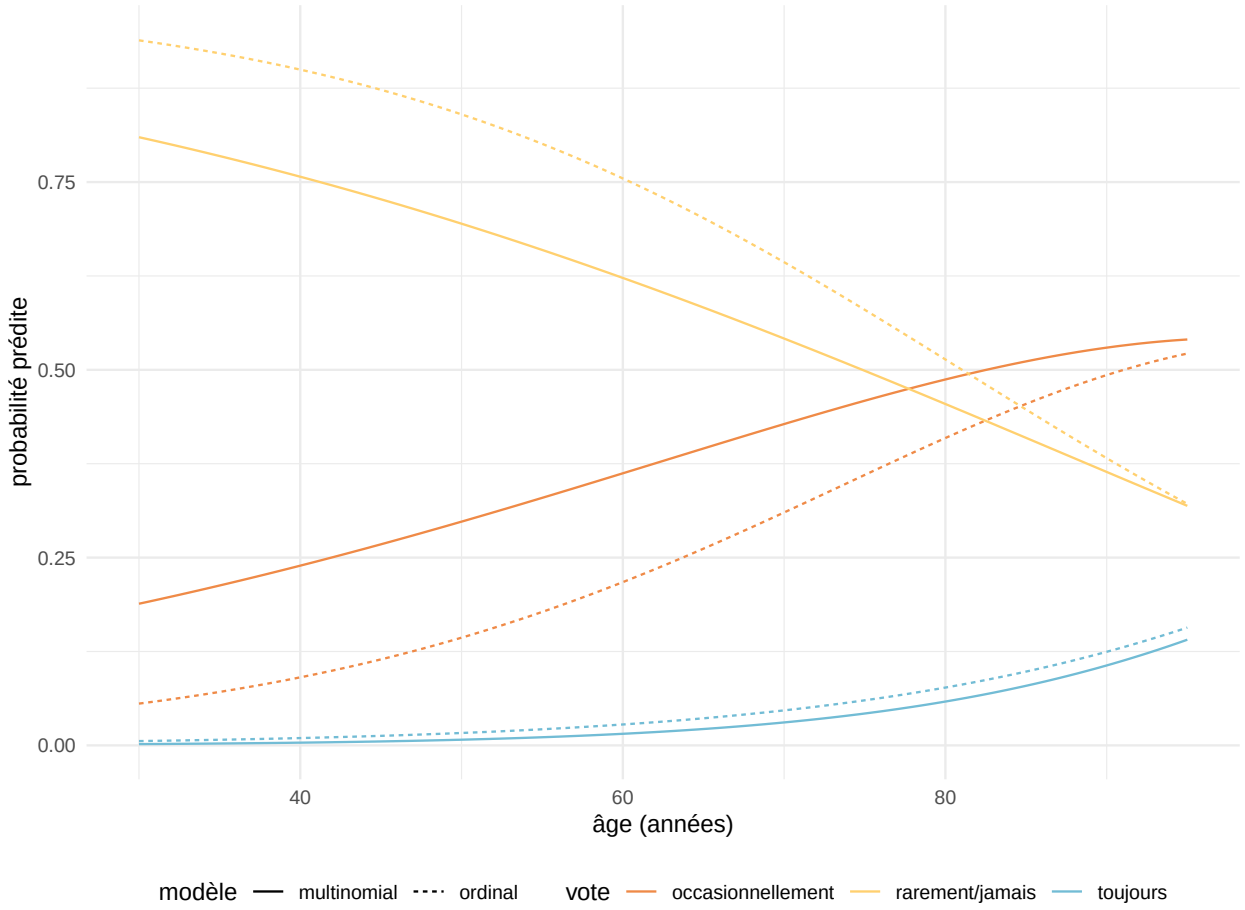

Figure 3: Probabilités prédites pour chaque modalité selon l'âge.

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).