Données manquantes et régression multinomiale

Analyse multidimensionnelle appliquée

Léo Belzile

HEC Montréal

Données manquantes

Plusieurs champs d'une base de donnée peuvent être manquants

- non-réponse
- valeurs erronées (erreur d'encodage)
- perte de suivi et censure
- plusieurs versions de formulaires (question optionnelles)

Pourquoi s'en préoccuper?

La plupart des procédures ne gèrent que les cas complets (toute observation avec des valeurs manquantes est éliminée).

Les données manquantes réduisent l'information disponible.

Sans traitement adéquat, les estimations seront biaisées.

Références

- van Buuren, S. (2018). Flexible imputation of missing data, CRC Press, 2e édition.
- Little, R. et D. Rubin (2019). Statistical Analysis with Missing Data, Wiley, 3e édition
- Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall / CRC.

Les valeurs manquantes dans un contexte de prédictions sont couvertes dans le cours MATH 60600.

Classification des données manquantes (MCAR)

Cas 1: Données manquantes de façon complètement aléatoire (missing completely at random)

La probabilité que la valeur soit manquante ne dépend ni de la valeur, ni de celles des autres variables.

Exemple: questionnaire trop long, la personne ne répond pas à tout (sans lien avec les questions posées).

Hypothèse souvent irréaliste en pratique.

Classification des données manquantes (MAR)

Cas 2: données manquantes de façon aléatoire (missing at random): la probabilité que la valeur soit manquante ne dépend pas de la valeur une fois qu'on a contrôlé pour les autres variables.

Exemple: les hommes sont plus susceptibles dans l'ensemble de divulguer leur âge que les femmes.

Classification des données manquantes (MNAR)

Cas 3: données manquantes de façon non-aléatoire (missing not at random): la probabilité que la mesure soit manquante dépend de la valeur elle-même, pas déterminable avec d'autres variables

Exemple: une personne transgenre ne répond pas à la question genre (si seulement deux choix, homme/femme) et aucune autre question ne se rattache au genre ou à l'identité sexuelle.

Type de données manquantes

Comment déterminer le type de données manquantes?

Par exemple, si une personne ne divulgue pas son salaire, est ce qu les données sont manquantes de manière aléatoire ou non aléatoire?

L'hypothèse pas testable, la réponse dépend du contexte et des variables auxiliaires disponibles.

Valeur logique

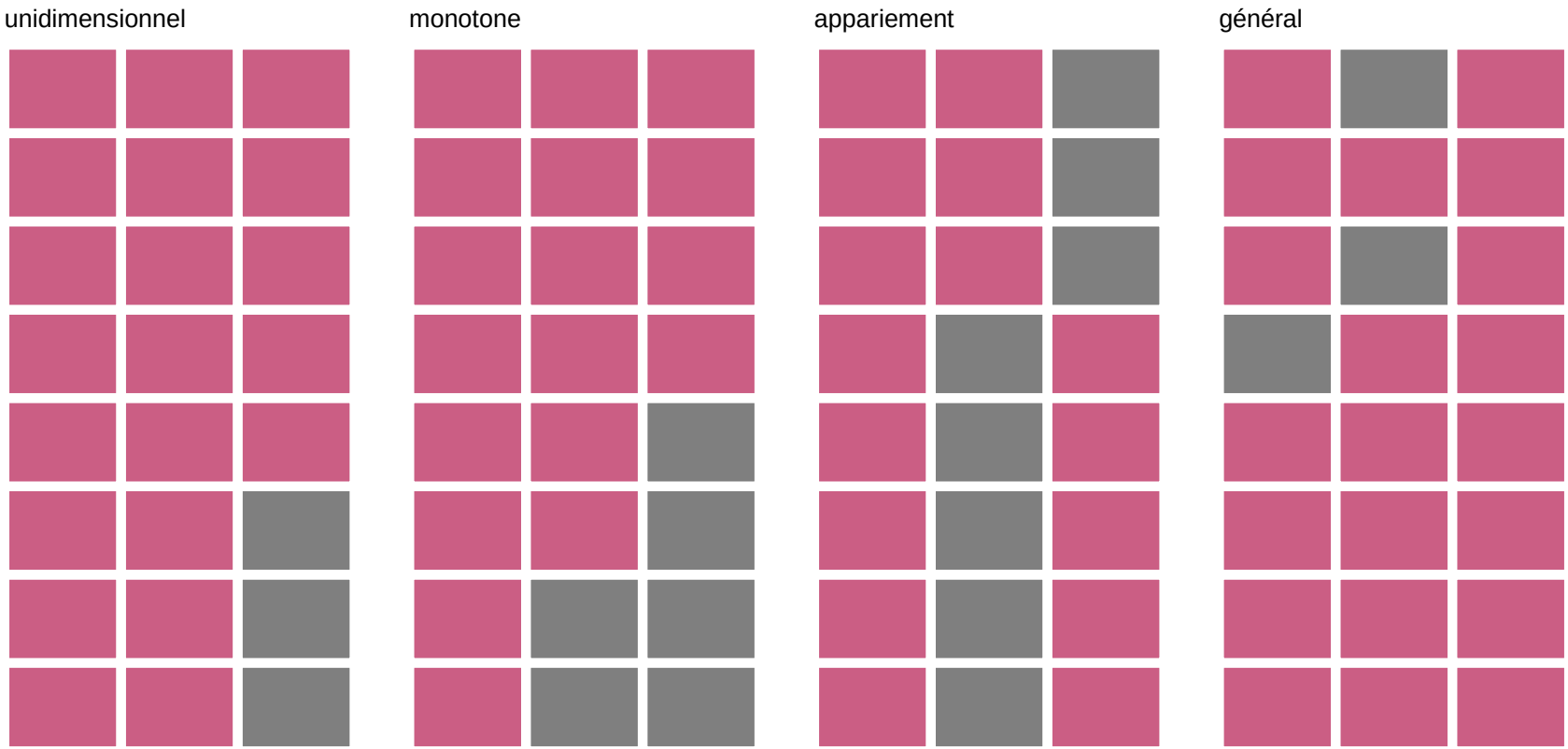
Les données manquantes ont souvent une valeur logique:

- un client qui n'a pas de carte de crédit a un solde de 0!

D'où l'importance des validations d'usage et du nettoyage préliminaire de la base de données.

Types de schémas de données manquantes

Matrice $n \times p$ (observations en lignes, variables en colonnes).



Les cases grises représentent des valeurs manquantes. Illustration adapté de la Figure 4.1 de van Buuren (2022)

Analyse des cas complet

Retirer les observations avec données manquantes pour conserver les cas complets.

- Valide uniquement pour complètement aléatoire.
- On perd de la précision en utilisant moins d'observations.

Méthode par défaut dans les logiciels.

Méthodes d'imputation

Imputation: remplacer les valeurs manquantes par une valeur judicieuse pour combler les trous.

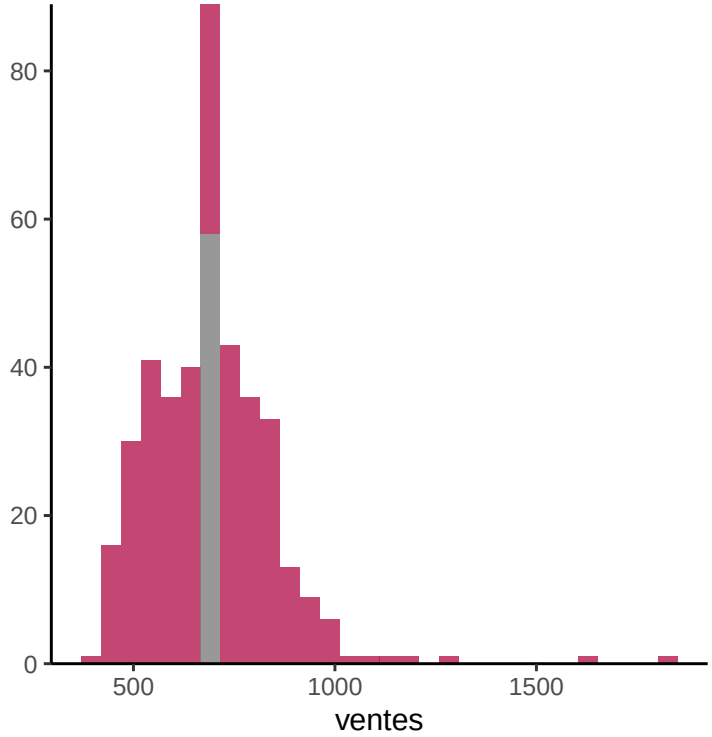
> Le concept d'imputation est à la fois séduisant et dangereux. (Dempster et Rubin, 1983)

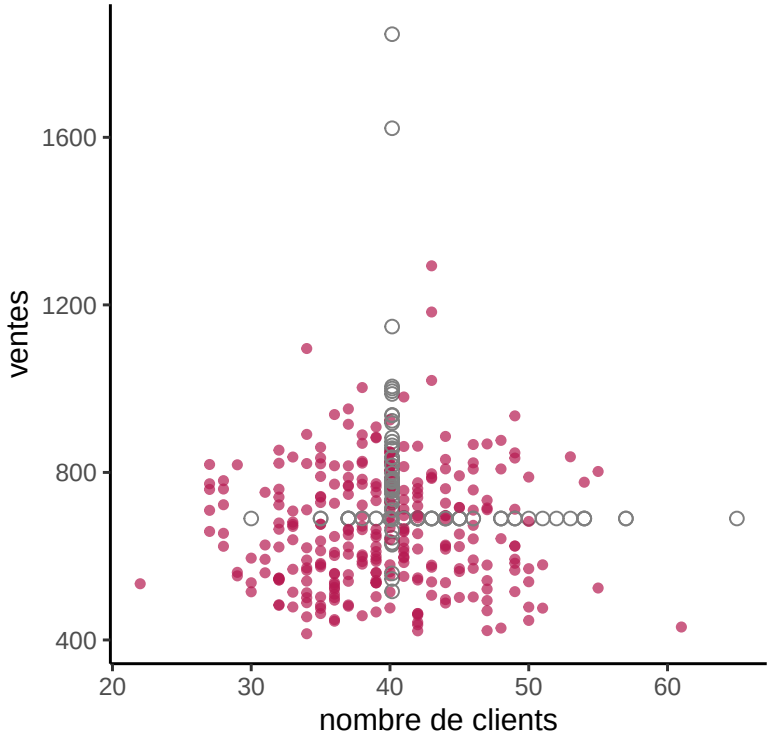
On distinguera

- l'imputation déterministe (par exemple, remplacer les valeurs manquantes par la moyenne) versus aléatoire
- l'imputation simple (une copie) versus multiple (plusieurs imputations)

Imputation par la moyenne

Dilution de la relation (corrélation) entre variables explicatives.
Réduction de la variabilité.





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 O/1 avec probabilité $(1 - \hat{p}_i, \hat{p}_i)$.

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

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

Faut-il toujours imputer?

Il faut utiliser son jugement.

Une observation imputée ne remplacera jamais une vraie observation.

- Si la proportion d'observations manquantes est petite (moins de 5%), on pourrait faire une analyse avec les cas complets (et valider au besoin en utilisant l'imputation multiple).
- Si la proportion de valeurs manquantes est 30% et que cette proportion baisse à 3% lorsque vous éliminez quelques variables peu importantes pour votre étude, alors procédez à leur élimination.

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.

Inspection des valeurs manquantes

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

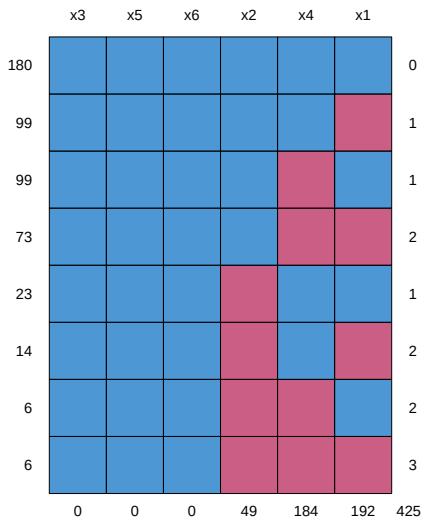
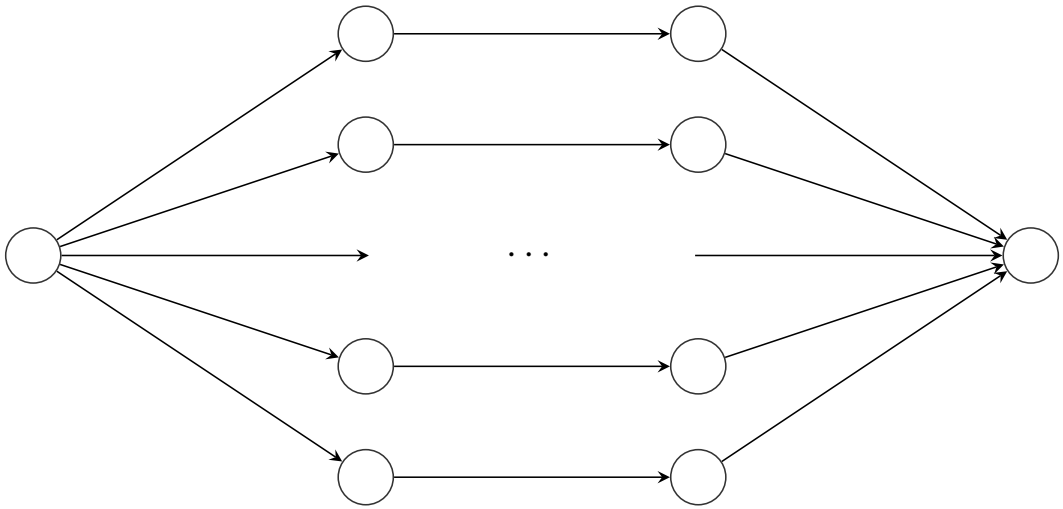


Figure 1: Les noms des variables sont indiquées au dessus, le nombre total de valeurs manquantes par variable en dessous, le nombre d'observations pour chaque configuration de valeurs manquantes à gauche et le nombre de variables avec des valeurs manquantes par configuration à droitef.

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

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

Estimations

L'estimation finale de θ, dénotée $\hat{\theta}$, est obtenue tout simplement en faisant la moyenne des estimations de tous les modèles.

Pour la variance de $\hat{\theta}$, on calcule la somme de

- la moyenne des K variances pour chaque imputation, $\hat{\sigma}_1^2, \dots, \hat{\sigma}_K^2$, appelée variance intra-groupe.
- la variance des estimations moyennes, $\hat{\theta}_1, \dots, \hat{\theta}_K$, appelée variance inter-groupe.

En combinant ces deux sources variabilités, on enfle la variance par rapport à l'imputation simple

Imputation multiple par équations chaînées (MICE)

Avec p variables $X_1, \dots, X_p$, 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 $j = 1, \dots, p$, 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`

```
# 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

- Les estimations avec les données complètes ont la plus grande incertitude, parce qu'elles utilisent uniquement $n = 180$ observations.
- L'imputation multiple donne $n = 500$ observations et prend correctement en compte l'incertitude.
- L'imputation simple sous-estime l'incertitude des coefficients.

Table 2: Estimation des coefficients et erreurs type pour quelques paramètres du modèle.

données complètes		imputation multiple		imputation simple	
coef.	err. type	coef.	err. type	coef.	err. type
-0.48	0.76	-0.90	0.58	-1.61	0.47
-0.38	0.77	-0.53	0.59	-0.95	0.46
-0.84	0.80	-0.76	0.55	-0.89	0.46
0.09	0.85	-0.58	0.64	-1.15	0.50
1.19	0.76	-0.03	0.44	-0.16	0.42

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.

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. Il y a trois modalités possible (rarement ou jamais, occasionnellement et toujours).

Analyse exploratoire

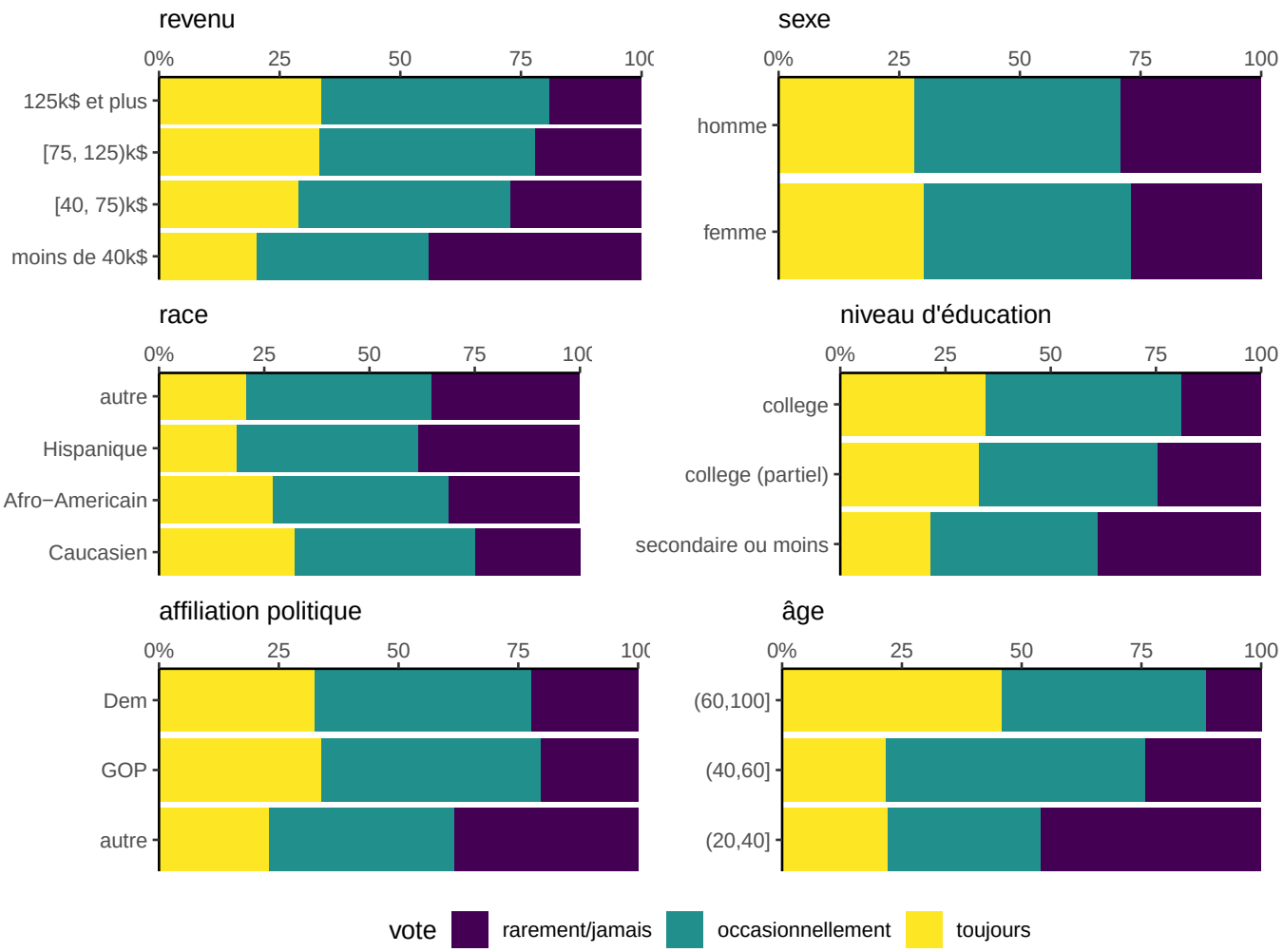
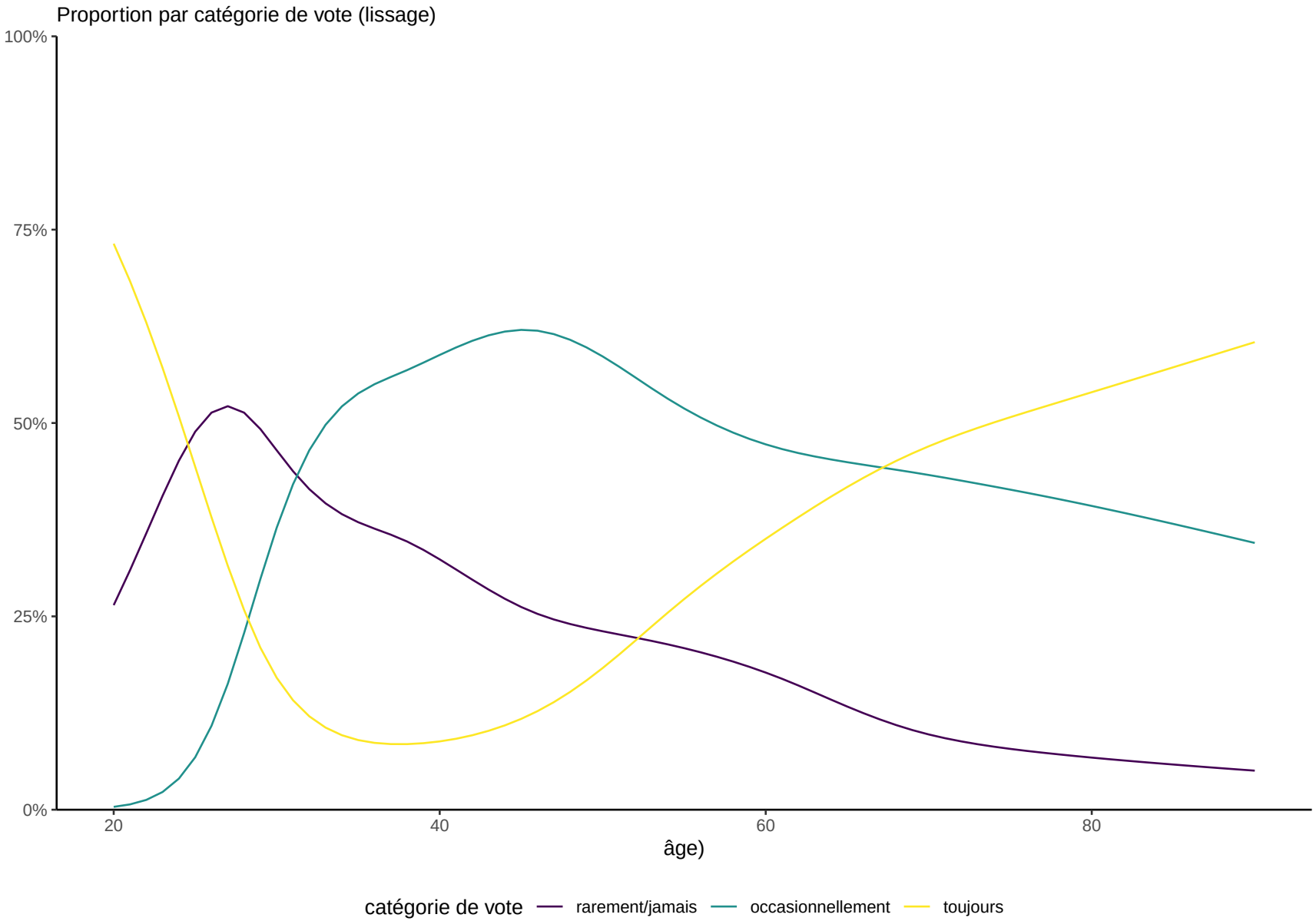


Figure 2: Proportion des modalités des variables sociodémographiques des données de participation électorale.

Analyse exploratoire



Analyse exploratoire (age)

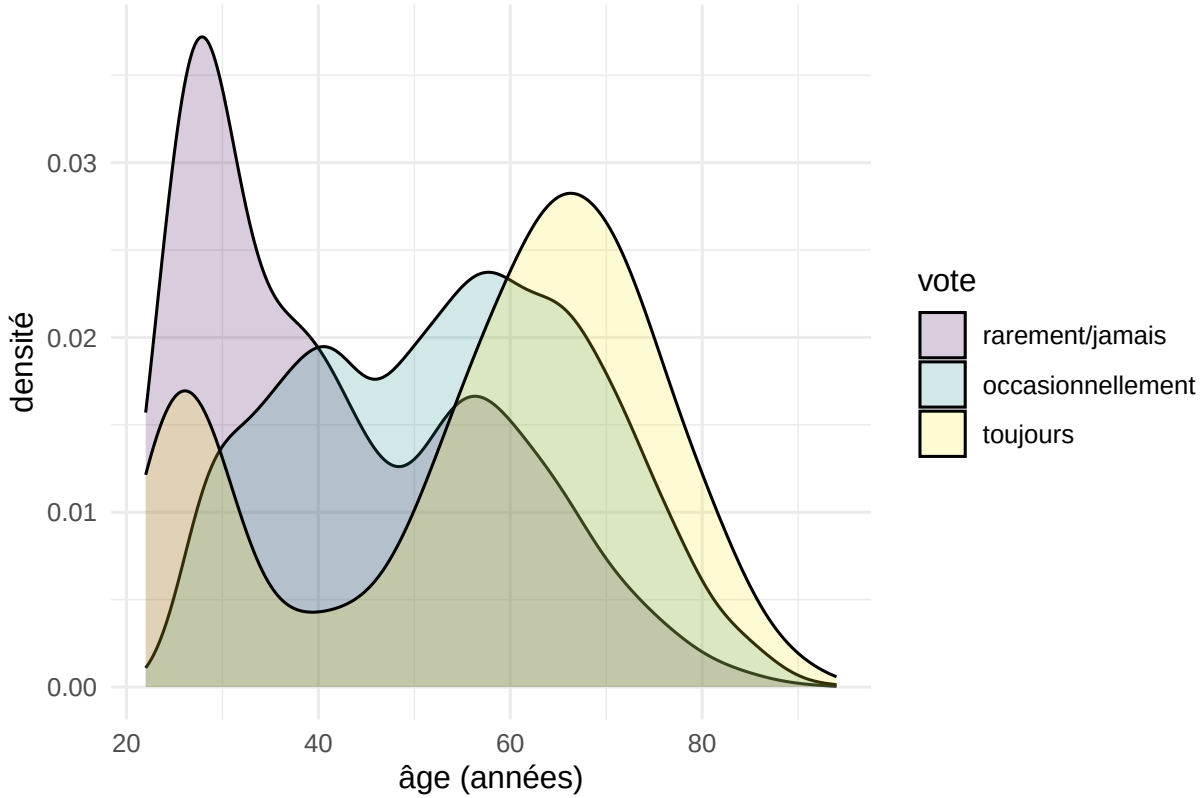


Figure 3: Fréquence de vote selon l'âge.

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

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,

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

- 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,

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

où η_{ij} est le prédicteur linéaire de l'individu i pour le log de la cote de $Y_i = j$ versus la référence $Y_i = 1$. On fixe $\eta_{i1} = 0$.

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.

```
multi_cst <- nnet::multinom(
  catvote ~ 1,
  weights = poids,
  subset = age > 30,
  data = vote,
  trace = FALSE)
head(predict(multi_cst, type = "probs"), n = 3)
```

```
  rarement/jamais occasionnellement  toujours
1       0.2282527          0.4910229 0.2807244
2       0.2282527          0.4910229 0.2807244
3       0.2282527          0.4910229 0.2807244
```

Tests de rapport de vraisemblance

On peut comparer des modèles emboîtés avec la fonction `anova`: ici, on compare le modèle complet au même modèle, moins la variable `sexe`.

ddl resid	deviance	ddl	stat	valeur-p
9670	8514.54			
9668	8504.74	2	9.8	0.01

La différence est significative à niveau 5%, l'effet de `sexe` est non null.

Prédictions

Pour un profil 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

Matrice de confusion

Table 4: Prédictions (lignes) versus observations (colonnes) pour le résultat du sondage pour les personnes âgées de plus de 30 ans.

	rare	occas.	toujours
rare	264	159	19
occas.	640	1820	843
toujours	76	421	605

Le taux de bonne classification est 2689/4847=0.55.

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...

Voir les notes de cours pour plus de détails.

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.