

1 Préparation de données

Étapes du nettoyage de données

Cette liste n'est pas séquentielle, il est surtout important de *tout* le faire.

- **Comprendre** la structure des données ;
 - › dimensions, types de variables, str et summary.
- **Visualiser** les données ;
 - › head, summary et graphiques exploratoires.
- **Mettre en forme (format)** les données ;
 - › Chaque ligne est une observation ;
 - › Chaque colonne est une variable ;
 - › Supprimer les doublons ;
- Vérifier et corriger les **types** de variables ;
 - › booléens, entiers, numériques, facteurs, chaînes de caractères, dates, etc.
- **Manipuler** les **chaînes** de caractères ;
 - › Corriger les typos ;
 - › Changer la casse avec tolower ;
 - › Extraire des informations avec les expressions régulières.
- Identifier les **données aberrantes** ;
 - › Mettre NA et gérer plus tard.
- **Détecter** les **erreurs** flagrantes ou les changements structurels dans les données ;
 - › Prendre compte des réformes ;
 - › Constater les structures importantes.
- **Augmenter** les données à l'aide d'autres sources ;
 - › optionnel.

2 Données manquantes

Le chapitre utilise la **mise en contexte** suivante :

- › Il y a une réclamation pour un accident d'auto en Ontario ;
- › Le contrat d'assurance couvre les frais médicaux ;
- › On désire calculer la probabilité de paiement (variable réponse) en fonction de :
 1. La gravité de l'accident (variable explicative) ;
3 niveaux : mineur-majeur-catastrophique ;
 2. La souffrance du réclamant ;
Échelle de 1 (peu) à 5 (beaucoup) ;

Problèmes de modélisation :

- › Comment analyser les données malgré les valeurs manquantes ?
- › Quels enjeux ou problèmes devrait-on considérer dans la modélisation ?

Terminologie

Notation

Y_{ij} : Valeur de la variable explicative j pour l'observation i où $j \in \{1, \dots, p\}$ et $i \in \{1, \dots, n\}$;

$\mathbf{Y}_{n \times p}$: Matrice contenant les données **complètes** ;

\mathbf{Y} est partitionné en deux, $\mathbf{Y} = \{\mathbf{Y}_{obs}, \mathbf{Y}_{mis}\}$

\mathbf{Y}_{obs} : matrice avec les données ayant toutes les valeurs observées ;

\mathbf{Y}_{mis} : matrice avec les données comportant des valeurs manquantes ;

$\mathbf{R}_{n \times p}$: **Matrice de réponse** des variables indicatrices $R_{ij} = \mathbf{1}_{\{Y_{ij} \text{ observé}\}}$;

θ : **Paramètre de nuisance**

Exemple de notation

$$\mathbf{Y} = \begin{bmatrix} 2 & 3 \\ 8 & 6 \\ 3 & 12 \end{bmatrix} \quad \mathbf{Y}_{obs} = \begin{bmatrix} 2 & . \\ 8 & 6 \\ . & . \end{bmatrix} \quad \mathbf{Y}_{mis} = \begin{bmatrix} . & 3 \\ . & . \\ 3 & 12 \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$$

Mécanisme de non-réponse

La distribution de \mathbf{R} est le *mécanisme de non-réponse* ;

Types de données manquantes :

1. **MCAR** : Missing Completely at Random ;

- › Le patron de non-réponse (pattern of missing values) est indépendant des données \mathbf{Y} ;
- › Il s'ensuit que la probabilité de réponse $f(R|\mathbf{Y}, \theta)$ ne dépend pas des données complètes \mathbf{Y} :

$$f(R|\mathbf{Y}, \theta) = f(R|\theta)$$

Exemple avec un θ de 10%

On perd 10% des valeurs mesurées alors, $\forall i \in \{1, \dots, n\}, j \in \{1, \dots, p\}$, la distribution du mécanisme de non-réponse :

$$R_{ij} \sim \text{Bernoulli}(p = 1 - \theta = 90\%)$$

- › Tester la différence de moyennes :

$$\mathcal{H}_0 : \{p_{\text{Cat}, \text{mis}} - p_{\text{Cat}, \text{obs}} = 0\} \text{ et }$$

$$\{p_{\text{Maj}, \text{mis}} - p_{\text{Maj}, \text{obs}} = 0\}$$

Est équivalent à tester :

\mathcal{H}_0 : les données sont MCAR

avec un test du khi-carré de Pearson;

2. MAR : Missing at Random;

- La probabilité de réponse $f(R|Y, \theta)$ dépend seulement des variables qui ont été observées dans le jeu de données Y_{obs} :

$$f(R|Y, \theta) = f(R|Y_{\text{obs}}, \theta)$$

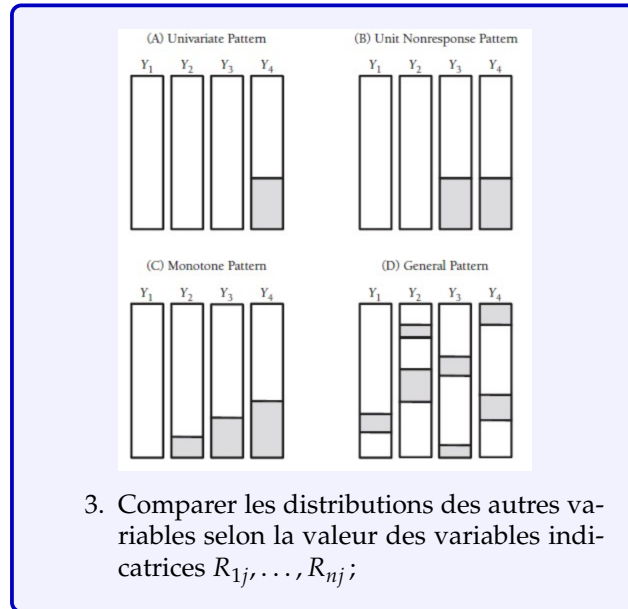
- Exemple de patients d'un hôpital : les données sont MAR lorsque la probabilité de non-réponse ne dépend pas de la qualité de vie sachant l'âge;

3. NMAR : Not Missing at Random;

- Le patron de non-réponse pour Y est relié à sa valeur et les variables observées;

Ce même si on conditionne sur les valeurs observées;

- La probabilité de réponse $f(R|Y, \theta)$ dépend également de Y_{mis} et ne peut pas être simplifiée;
- Pour exemple, les patients malades ne répondent pas aux sondages en plus des patients plus jeunes et donc la probabilité de réponse dépend de la qualité de vie;
- Pour exemple, la probabilité de réponse dépend d'une autre variable non observée;



Identification des types de non-réponse

- Pour les variables continues, on fait un **test t sur les différences de moyenne** au lieu du khi-carré de Pearson comme pour MCAR;
- Problème de comparaisons multiples;
- Le test MCAR de Little est peu utile, mais peut adresser le problème de comparaisons multiple avec un hypothèse testant toutes les variables;

Traitement des données manquantes

En continuant la mise en contexte, on suppose qu'on veut estimer le vecteur β des coefficients de la régression logistique pour prédire la probabilité de paiement; Une question valide est si les options pour le traitement des données manquantes dépendent du **type** de non-réponse

Options de traitement :

- Utiliser seulement les **cas complets** (*complete-case analysis*);

- L'option par défaut pour les fonctions :

`lm, glm, na.rm, na.omit`

- Impact :**

↓ taille de l'échantillon

↑ variance des estimateurs

↓ puissance des tests

- Uniquement valide sous **MCAR**;

2. Utiliser seulement les **cas disponibles** (*available-case analysis*);

- Utilise uniquement les données observées pour l'analyse;
- Rarement applicable;
- ↓ la taille de l'échantillon **moins** qu'en utilisant d'uniquement les cas complets;
- Sans biais** uniquement sous **MCAR**;

3. Imputation simple par la **moyenne** ou la **médiane**

- Substitue les NA par la moyenne ou médiane de la variable;

- Impact :**

↓ variabilité de la variable

↓ corrélation de la variable avec les autres

- Même sous MCAR, les données sont **sévérement** « distorted »;

4. Imputation simple par une régression;

- Substitue les NA par la prévision d'une régression de la variable sur les autres avec les cas complets;
- Si plusieurs variables ont des données manquantes, leurs patrons doivent être traités séparément;

Visualisation et détection

Traiter les données manquantes

- Détectez, visualisez et documentez les données manquantes;
- Identifier le patron de non-réponse;
 Pour exemple, voici quelques patrons de non-réponse pour quelques variables :

- › L'inter **corrélation** des variables est **conservée**, mais est **surestimée** (même si MCAR);
- › La variance est **sous-estimée**, mais **moins** qu'avec l'imputation par la moyenne;

5. Imputation stochastique par une régression;

- › Ajoute un terme d'erreur ε (normalement distribué) à la prévision de la régression;
- › *Si plusieurs variables sont manquantes dans un patron, les erreurs sont corrélées*
- › Corrige les biais pour la méthode d'imputation par la régression (sous-estimation de la variance et surestimation de l'inter corrélation des variables);
- › La variance des paramètres est **sous-estimée**, *sauf si on en tient compte dans les calculs*;
- › Fonctions R utiles du paquetage mice :

```
mice.impute.norm.nob(),
mice.impute.norm()
```

6. Imputation simple *hot-deck*;

- › Substitue les valeurs NA d'une observation par les valeurs observées d'une autre observation choisie aléatoirement;;
- › Habituellement, cette observation fait parmi d'un sous-ensemble d'observations *proches* (pensez au K-NN, clustering, etc.);
- › Souvent utilisée pour les sondages;
- › N'altère pas les distributions univariées
- › ↓ L'inter corrélation des variables;
- › Biais des estimations des coefficients **fi** de régression;

7. Imputation **multiple**;

- › Répète l'imputation stochastique et agrège les résultats;
- › Ce faisant, la variabilité additionnelle dûe à l'imputation des valeurs manquantes est adressée et la variance des estimateurs est *non biaisée*;

Autres méthodes :

- › MLE avec données manquantes;
- › Algorithme EM (expectation-maximisation)
- › Inférence bayésienne;

Conseils

- › Conserver un script pour le traitement de données manquantes et ne **pas hard-coder**;
- › *Utiliser une méthode d'imputation qui respecte le format de la variable*;
- › Plus la proportion de non-réponses est élevée, plus l'impact sur l'analyse sera important;
- › S'il y a plusieurs patrons de non-réponse différents, l'ordre dans lequel les données sont imputées est important;