Objectifs et données

STT-7330 Analyse de données

Véronique Tremblay

Département de mathématiques et de statistique

Hiver 2020

Objectifs

Inventaire des données

La constitution de la base de données

Exploration et traitement préliminaire

Les étapes d'un projet d'analyse de données

1. Définition des objectifs
2. Données
   - Inventaire des données
   - Constitution de la base de données
   - Exploration et traitement préliminaire
3. Élaboration et validation des modèles
4. Documentation et présentation
5. Mise en oeuvre
6. Suivi de la performance et amélioration

Exemple

Temps requis pour chaque étape (selon Pyle, 1999)

Étape	Temps requis
Identifier le problème	10%
Explorer la solution	9%
Spécification de l'implantation	1%
Préparation des données	60%
Analyse descriptive des données	15%
Modélisation	5%

Objectifs

Définir les objectifs

Est-ce qu'on veut...

- Comprendre
- Prédire
  - `Prédire quoi?`
- Comprendre et prédire

Exemple 1

Une compagnie qui analyse le comportement des usagers sur le web vous demande de faire un modèle pour prédire si l'utilisateur va acheter ou non un produit.

Exemple 2

On vous demande de construire un modèle permettant de prédire le type de forêt à partir d'une image satellite.

Exemple 3

On vous demande de prédire le volume de vente pour une période donnée de façon à mieux planifier les ressources humaines pendant cette période.

Exemple 4

Un groupe de recherche vous demande de faire un modèle permettant de prédire si un patient va bien répondre à un certain traitement.

Exemple 5

On veut prédire le volume d'une crue millénaire avec des données observées depuis une centaine d'années.

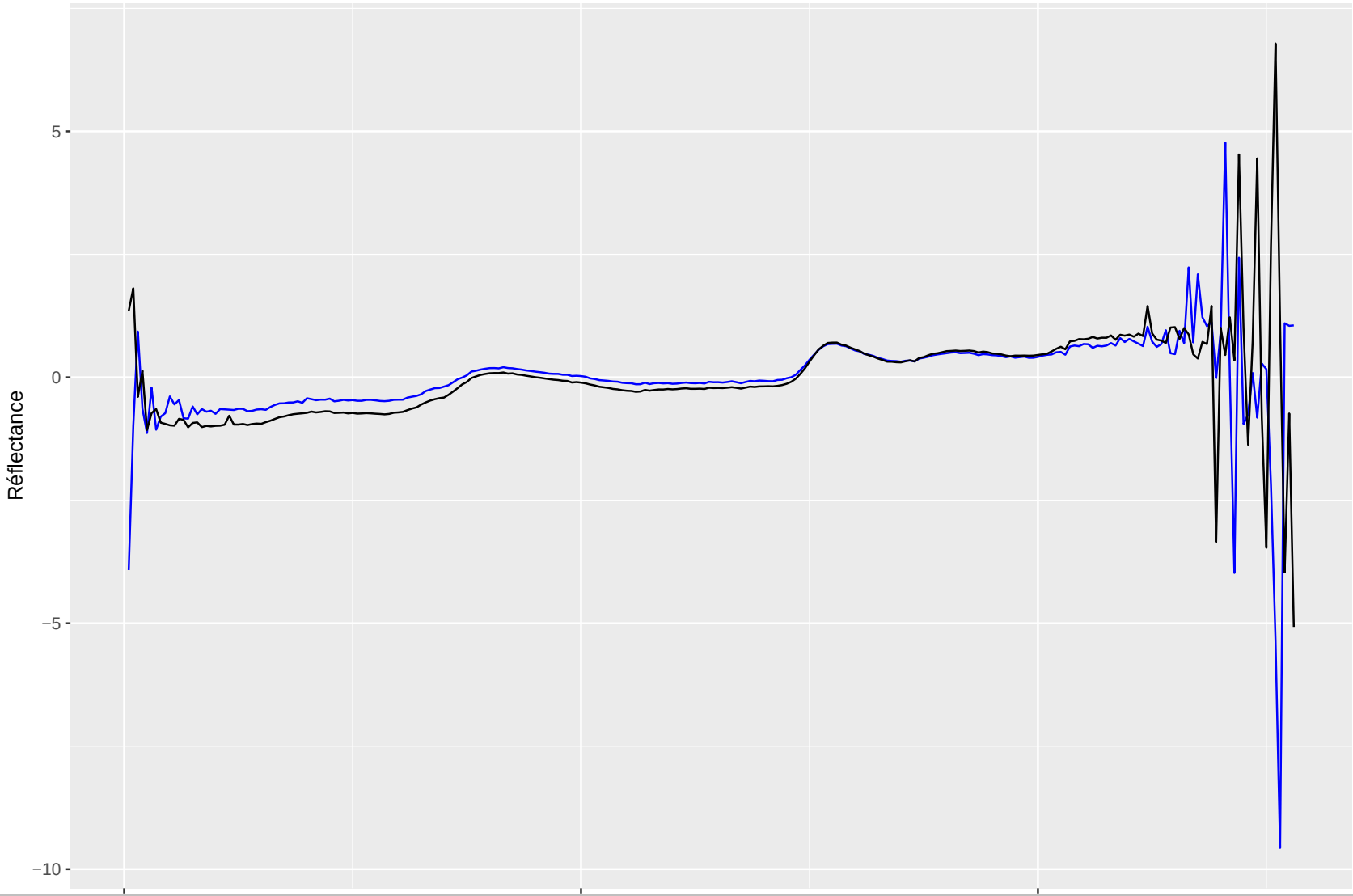
Exemple 6

Un assureur souhaite estimer les chances d'avoir un accident pour fixer la prime d'un individu.

Le choix du modèle (et du logiciel) dépend aussi des données et l'utilisation qu'on en fera.

Exemple 7

Exemple 8



Inventaire des données

Les types de données

- Structurées
- Non-structurées

Qualité des données

Qu'est-ce qu'on veut dire par qualité des données?

Qualité des données

- Échantillon aléatoires
- Expérience randomisées
- Représentative / variées
- Fiables

Principales sources de données

- Données expérimentales
- Données internes
- Données officielles
- Données de sondage
- Autres données sur le web
- Données payantes

Principales sources de données

Données internes

- Base de données des clients
- Listes de transactions
- Base de données des employés
- Information sur les visites web (Google analytics)
- Listes de clients potentiels

Principales sources de données

Principales sources officielles

- Statistique Canada
  https://www150.statcan.gc.ca/n1/fr/type/donnees
  - Recensement https://www12.statcan.gc.ca/datasets/Index-fra.cfm?Temporal=2016&Theme=-1&VNAMEF=&GA=-1&S=0
  - `CANSIM2R`
- Institut de la statistique du Québec
  http://www.bdso.gouv.qc.ca
- Ressources naturelles Canada https://geogratis.gc.ca/

Principales sources de données

Autres sources officielles

- Ministères et organismes

- Gouvernement ouvert
  - https://ouvert.canada.ca/fr
  - https://www.donneesquebec.ca/fr/

Principales sources de données

Quelques trucs:

- Appelez les ministères et organismes

- https://toolbox.google.com/datasetsearch

La constitution de la base de données

Extraire les données

Les données peuvent être emmagasinées de différentes façons:

- Fichiers plats
- BD Relationnelle
- Entrepôts pour données massives (type Hadoop)
- ...

Extraire les données (suite)

Librairies pour l'importation de données en R

Format	Extension	Librairie
Texte	.txt, .csv	readr
Excel	.xlsx	readxl
SAS	.sas7bdat	haven
SPSS	.sav, .zsav	haven
JSON	.json	jsonlite
Shapefile (spatial vectoriel)	.shp	shapefile
Raster (spatial matriciel)	raster	.grd (et autres)
Page web	.html	xml, httr, RCurl, rvest

Formater, nettoyer et fusionner les données

Jeu de données **structuré**: [1]

- Une et une seule ligne par observation

- Une colonne par variable (caractéristique)

  exercice

[1] Les données de survie et les données transactionnelles ont un format plus spécifique

Formater, nettoyer et fusionner les données

Jeu de données **propre**

- Pas de doublons
- Une même modalité est exprimée de la même façon
- Chaque variable est enregistrée dans le bon format:
  - Les dates sont sous forme de date (ex. `POSIXlt` ou `POSIXct`);
  - Les variables quantitatives sont sous forme numérique (`numeric` ou `double`);
  - Les variables catégorielles sont dans le format approprié pour le logiciel (`factor`);
  - Les identifiants sont dans un format approprié (`character`)
- Les valeurs spéciales comme les valeurs manquantes sont enregistrées correctement.

Formater, nettoyer et fusionner les données

Les métadonnées

- Une ligne par variable
- Le type de variable (continu, catégoriel...)
- La source
- Les modalités
- Les unités
- Une note explicative
- Des éléments descriptifs

Formater, nettoyer et fusionner les données

- Format large et format long
- Les jointures

```
# Joindre à A les variables de B.
leftJoin(A, B, by = "ID")
# Joindre à B les variables de A.
rightJoin(A, B, by = "ID")
# Garde toutes les observations
fulljoin(A, B, by = "ID")
# Ne conserve que les lignes communes
innerjoin(A,B, by = "ID")
```

Exploration et traitement préliminaire

- Valeurs extrêmes, aberrantes et influentes
- Valeurs manquantes, valeurs absentes et imputation
- Transformation des variables

Valeurs extrêmes, aberrantes et influentes

Valeur aberrante:

- Valeur erronée causée par une erreur de saisie, une erreur de calcul, une mauvaise mesure ou une fausse déclaration.

Exemples :

- Nombre total de pièces dans une maison est plus petit que le nombre de chambres à coucher
- La date du premier achat est antérieure à la date de naissance du client
- La personne a déclaré avoir 17 ans et être retraitée

Valeurs extrêmes, aberrantes et influentes

Valeur extrême

- Valeur éloignée des autres valeurs dans la population.

- Pas nécessairement aberrante
  - `Peut correspondre à un profil particulier et plus rar`

Valeurs extrêmes, aberrantes et influentes

Détecter les valeurs extrêmes et aberrantes.

- Faire des statistiques descriptives unidimensionnelles
  - Minimum et maximum
  - Histogramme, boite à moustaches (box-plot)
  - Tableau de fréquences
- Faire des statistiques descriptives multidimensionnelles
  - Tableaux croisés
  - Diagrammes de dispersion
  - Distance de Mahalanobis

Valeurs extrêmes, aberrantes et influentes

Est-ce que ça influence le modèle?

Valeurs extrêmes, aberrantes et influentes

Quoi faire avec les valeurs aberrantes?

- Les remplacer si on peut connaitre la bonne valeur
- Les classer dans une catégorie «inconnus» pour les variables catégorielles
- Retirer l'individu, surtout si plusieurs éléments sont incohérents
- Imputer

Valeurs extrêmes, aberrantes et influentes

Quoi faire avec les valeurs extrêmes?

- Retirer les observations
- Utiliser des méthodes robustes
- Imputer par une valeur plus «réaliste»

Valeurs manquantes et absence de valeur

Une valeur manquante implique qu'il y a une valeur pour cette observation, mais qu'elle n'a pas été capturée.

À ne pas confondre avec l'absence de valeur.

Mécanisme de non-réponse

- Données manquantes complètement au hasard (MCAR)
- Données manquantes au hasard (MAR)
- Données manquantes pas au hasard (NMAR)

Notation

Soit Y une matrice de données de dimension n par p.

L'entrée Y_{ij} donne la valeur de la variable j pour l'observation i.

On divise cette matrice en deux parties

$$Y = (Y_{obs}, Y_{mis}).$$

Y_{obs} contient les données observées,

Y_{mis} contient les données manquantes.

On définit également une matrice indicatrice R de taille n par p nommée **matrice de réponse** où

$$R_{ij} = \begin{cases} 1 \text{ si } Y_{ij} \text{ est observé} \\ 0 \text{ si } Y_{ij} \text{ est manquant} \end{cases}$$

On s'intéresse à la distribution de R, soit le mécanisme de réponse. On peut écrire

$$f(R|Y, \theta)$$

où θ est un vecteur de paramètres.

MCAR (Missing completely at random)

Lorsque la probabilité de réponse ne dépend pas de Y, i.e.

$$f(R|Y,\theta) = f(R|\theta).$$

Exemple: On perd au hasard 20% des valeurs mesurées. Alors

$$f(R_{ij}|\theta) \sim \text{Bernoulli}(0,8)$$

pour tout $i = 1,\ldots,n$, $j = 1,\ldots,p$.

MAR (Missing At Random)

Lorsque la probabilité de réponse dépend uniquement de la valeur des variables qui ont été observées, i.e.

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

Exemple: On fait remplir un questionnaire à plusieurs individus, en leur demandant leur sexe et leur salaire. Si les femmes ont une plus grande probabilité que les hommes d'accepter de fournir leur salaire, alors on aurait un mécanisme MAR.

NMAR (Not Missing At Random)

Lorsque la probabilité de réponse dépend de Y_{mis} également. Dans ce cas, on ne peut pas simplifier

$$f(R|Y, \theta)$$

NMAR (Not Missing At Random)

- La probabilité de réponse dépend de la variable elle-même.
  - Les individus avec des salaires plus élevés ont une plus grande probabilité de refuser de déclarer leur salaire.

- La probabilité de réponse dépend d'une variable non-observée.
  - La probabilité de déclarer son salaire dépend de l'âge de l'individu, mais on n'a pas mesuré cette variable.

Quoi faire?

- Documenter la non-réponse.
  - Faire des statistique descriptives
- Identifier quel mécanisme de non-réponse serait défendable.
  - Tests pour MCAR
  - Connaissance du domaine
- Choisir une méthode de traitement de la non-réponse.

Identifier le mécanisme de non-réponse

Test de Welch (test de t)

Adaptation du test de t de Student dans le cas où les deux populations ont des variances inégales.

Technique : Séparer les cas observés et manquants pour une variable et tester les différences de moyennes pour les autres variables.

Identifier le mécanisme de non-réponse

Test de Welch (test de t)

- Implique beaucoup de tests; problème de comparaisons multiples.
- Ne tient pas compte des corrélations entre les variables.
- Ne garantie pas l'hypothèse MCAR
- Peut nous aider à identifier les variables pour lesquelles ajuster dans les procédures de traitement de données manquantes.

Identifier le mécanisme de non-réponse

Test de Little

Compare en un seul test les moyennes et les variances à l'intérieur de chaque patron de non-réponse.

Statistique de test :

$$d^2 = \sum_{j=1}^{J} n_j (\hat{\mu}_j - \hat{\mu}_j^{ML}) \hat{\Sigma}_j^{-1} (\hat{\mu}_j - \hat{\mu}_j^{ML})$$

où $j = 1, \ldots, J$ dénote le patron de non-réponse, μ_j est la moyenne pour les cas du patron j, $\hat{\mu}_j^{ML}$ est l'estimateur du maximum de vraisemblance global de la moyenne et Σ_j l'estimateur du maximum de vraisemblance global de la matrice de covariance. Le nombre de variables change d'un patron à l'autre, d'où l'index j.

Identifier le mécanisme de non-réponse

Test de Little (suite)

Sous l'hypothèse nulle que les données sont MCAR, on a

$$d^2 \overset{approx}{\sim} \chi^2 \text{ avec} \sum_{j=1}^{J} k_j - k \text{ degrés de liberté}$$

où k_j est le nombre de variables complètement observées dans le patron j et k est le nombre de variables du jeu de données.

Identifier le mécanisme de non-réponse

Test de Little

- Le test ne nous indique pas quelles variables sont problématiques.
- Le test suppose la même matrice de covariance pour tous les patrons de non-réponse.
- Des études de simulations ont mis en évidence un manque de puissance de ce test.
- Ne permet pas de garantir l'hypothèse MCAR.

Identifier le mécanisme de non-réponse

Traiter la non-réponse

- Retirer les variables qui contiennent trop de données manquantes
  - Perte d'information

- Retirer les observations avec des données manquantes
  - Risque de biais si MAR ou NMAR

Traiter la non-réponse

- Si la variable est catégorielle : créer une catégorie «inconnu»
- Imputer

Traiter la non-réponse - Imputer

- Par la moyenne, la médiane ou le mode
  - Modifie la distribution
  - Altère la structure de corrélation

Traiter la non-réponse - Imputer

- Par la distribution
  - On remplace par une autre valeur pigée au hasard dans les données
  - Préserve la distribution, mais on perd la structure de corrélation

Traiter la non-réponse - Imputer

- Avec un modèle prédictif
  - Préserve la distribution et la structure de dépendance
  - Réduit la variabilité (on surestimera la précicion)

Quelques solutions

- Ajouter une erreur au modèle de régression

- Faire de l'imputation multiple

Transformer les variables

- Transformations simples ($\log(X)$)
- Standardisation

Transformer les variables

- Regroupement
  - Regroupement des marques de voiture en sous-compacte, compacte, intermédiaire, grande et très grande
  - Regroupement des entreprises :« SIC » standard industry classification
  - Regroupement des codes postaux en régions en utilisant les 3 premiers caractères seulement

Transformer les variables

- Combinaison de variables
  - Date du premier achat - date de naissance
  - Date d'aujourd'hui - date du dernier achat
  - Somme de tous les achats dans les 12 derniers
  - Prix payé l'an passé - prix payé cette année

Transformer les variables

- Réduction du nombre de dimensions
  - ACP, ACM, sélection de variables

Quelques conseils

- Ne modifiez pas les bases de données originales
- Automatisez
- Documentez