STT 6415–Régression

Hiver 2021
Professeur : Florian Maire
florian.maire@umontreal.ca

Horaires et déroulement du cours

Les cours ont lieu les **mercredi de 9h00 à 10h30** et **vendredi de 8h30 à 10h00** sur zoom, du 15 janvier jusqu'au 30 avril (sauf pour la semaine du 1 au 5 mars). Il y aura 29 cours de 1h30 chaque. Pour y assister, vous devez vous connecter sur votre compte zoom de l'UdeM et vous rendre à la réunion suivante :

- numéro de réunion : 964 1762 8899
- mot de passe : 690174

Les cours se font en direct au tableau en s'appuyant sur des notes de cours. Tout ce qui est dans les notes de cours ne sera pas nécessairement discuté dans le cours et, inversement, des éléments (exemples, remarques, simulations, etc.) peuvent être discutés dans le cours sans être présents dans les notes. Les cours seront enregistrés et accessibles depuis Studium et les notes de cours seront également disponibles sur Studium. Chaque semaine, il y a une période de questions les **lundi de 10h30 à 12h30** sur zoom :

- numéro de réunion : 993 7041 2381
- mot de passe : 223412

Présentation

En statistique, une régression désigne toute technique permettant d'extraire de l'information d'une variable (la/les variable/s explicative/s ou prédicteur), facilement mesurable, afin de renseigner sur une autre variable (la variable réponse), difficilement mesurable, et sur laquelle se porte un intérêt particulier.

Ce cours offre un éclairage différent sur les questions de régression vues au bac (STT2400, régression linéaire). Le cours STT2400 a montré comment effectuer une analyse de données en présence d'une variable réponse et de variables explicatives de manière systématique en suivant les étapes (i) modélisation par modèle linéaire, (ii) estimation, (iii) validation, (iv) inférence et (v) prévision. Cette démarche, bien que simpliste et limitée, va nous servir de point de départ. Les techniques de régression étant tellement variées, il serait illusoire et peut-être même ennuyeux de vouloir en faire le tour en un cours. Nous procéderons donc par questions thématiques et montrerons comment mobiliser des concepts vus ailleurs en statistique (théorie du risque, analyse bayésienne, statistique robuste, inférence non-paramétrique, analyse asymptotique, régularisation, etc.) dans un contexte de régression. Les thèmes que nous aborderons et le type de questions que nous discuterons sont :

(1) **Généralisations de la régression linéaire.** Les hypothèses d'un bruit gaussien homoscédastique ainsi que d'une variable réponse réelle sont fortement réductrices. Comment généraliser l'analyse de données systématique (i),(ii),(iii),(iv),(v), au delà du modèle linéaire avec bruit Gaussien ?

(2) **Régime asymptotique en régression.** Dans le modèle linéaire usuel, les variables réponses sont aléatoires mais les variables explicatives sont déterministes. Que se passe-t-il si les variables explicatives sont elles aussi aléatoires ? Pourquoi toute analyse asymptotique de la régression (régime $n \to \infty$) se place naturellement dans ce cadre et quel sens donner aux estimateurs limites ?

(3) **Régularisation et régression bayésienne.** Le contexte *high-dimensional* où le nombre de variables explicatives p se rapproche de n mène à des modèles statistiques mal-posés et à des problèmes de multicolinéarité mais qui peuvent être régularisés. Pourquoi est-ce que la régularisation d'un problème de régression est intrinséquement liée à une philosophie Bayésienne de l'inférence statistique ?

(4) **Régression robuste.** L'adage "*Garbage in, garbage out*" bien connu en apprentissage statistique a une importance toute particulière en régression où les estimateurs usuels sont connus pour être sensibles aux données aberrantes. Comment les outils de statistiques robustes peuvent-ils être utilisés aussi en régression ?

(5) **Régression non-paramètrique.** Si l'on ne souhaite pas imposer de structure sur la relation entre les variables explicatives et la variable réponse et plutôt laisser les données construire elles-mêmes cette relation, on se place

dans un contexte de statistique non-paramètrique. Dans quelle mesure la régression non-paramétrique est-elle une alternative robuste et efficace à la régression paramètrique ? Quel est le prix à payer pour la flexibilité de ces modèles ?

La compréhension de l'essence de ces problèmes sera autant valorisée que celle des techniques existantes à mettre en oeuvre dans un contexte donné. À un niveau générale, l'objectif du cours est donc d'offrir une initiation à la résolution d'un problème de régression d'une manière indépendante, en insistant sur les aspects analytiques et computationnels.

Évaluations

Il y aura trois devoirs à faire en équipe ainsi que deux examens intra et final à faire individuellement. Les devoirs seront donnés une dizaine de jour à l'avance et à faire en équipe de trois ou quatre étudiants. Les examens, d'une durée de deux heures chaque, se dérouleront en ligne.

type	date	pondération
Devoir 1	mi février	10%
Intra	26 février	35%
Devoir 2	mi-mars	10%
Devoir 3	mi-avril	10%
Final	à confirmer	35%

Les étudiants inscrits au Bureau de Soutien aux Étudiants en Situation de Handicap (BSESH) désirant bénéficier de mesures d'accommodement aux examens (intra et final) sont priés de contacter le SAFIRE.

Plagiat

L'Université de Montréal a une politique très claire et ferme sur le plagiat, voir `https://integrite.umontreal.ca`. Elle ne concerne pas que les examens, mais également les devoirs. Ce rappel est d'autant plus valable car, par nature, l'environnement dans lequel les examens en ligne se déroulent est plus difficilement controllable. Plutôt que d'opter pour une méthode de surveillance disproportionnée, l'utilisation de toutes les ressources (livres, notes de cours, internet, logiciels) est permise lors des examens. En revanche, la communication entre étudiants est strictement interdite. À ce niveau, il sera demandé à ce que chaque étudiant écrive une déclaration sur l'honneur en introduction de leur copie d'examen, guarantissant le caractère personnel de leur travail. Il en va de la valeur de vos diplômes !

Bibliographie

Le programme de ce cours n'est pas calqué sur le développement d'un livre en particulier. Une liste non-exhaustive de références est donnée à titre indicative.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*, Springer. [Parties II et IV]

Efron, B. et Hastie, T. (2016). *Computer Age Statistical Inference*, Cambridge University Press. [Chap. 6,7,8 et 16]

Wasserman, L (2006). *All of Nonparametric Statistics*, Springer. [Chapitre 5]

Rousseeuw, P. J. et Leroy, A. M (2003). *Robust Regression and Outlier Detection*, Wiley Series in Probability and Statistics. [Chap. 1 à 3]

Tous ces livres sont disponibles en ligne sur `https://bib.umontreal.ca/` au format numérique.