Régression
STT 6415

Hiver 2020 — 3 crédits
Professeur : Florian Maire — 4243, Pavillon André-Aisenstadt
Tél. : 343-7977 — maire@dms.umontreal.ca

Horaire : Mercredi de 9h00 à 10h30 au 4186 du Pavillon André Aisenstadt et Vendredi de 8h30 à 10h00 au 4186 du Pavillon André Aisenstadt, jusqu'au 30 avril (sauf pour la semaine du 2 au 6 mars).

Des notes de cours et d'autres informations seront disponibles via le site StudiUM du cours : voir `https://studium.umontreal.ca/`

Présentation : L'idée de ce cours est d'améliorer la compréhension des concepts de régression vus en STT2400 (régression linéaire) mais aussi de s'approprier d'autres outils de régression dont est supposé disposer un statisticien moderne. En statistique, une régression désigne toute méthode permettant de décrire l'aléatoire d'un phénomène d'intérêt à partir de l'aléatoire d'autres phénomènes explicatifs et observables. Les deux objectifs essentiels de la régression sont de comprendre les liens entre une collection de plusieurs variables (potentiellement) explicatives du phénonène d'intérêt et de pouvoir prédire de nouvelles observations du phénomène d'intérêt ayant observé les variables explicatives de nouveaux individus.

Le cours sera divisé en quatre parties de taille variable. L'objectif sera de présenter les concepts et les méthodes ainsi que leur philosophie sans systématiquement effectuer une étude théorique poussée de ces dernières. Une part importante des discussions sera laissée à l'illustration de la pratique de la régression.

1. Nous commencerons par revoir la régression linéaire multiple (estimation, tests d'hypothèses, régions de confiance) en insistant sur les hypothèses de la méthode des moindres carrés et l'importance de bien les vérifier après l'étape d'inférence. Ceci nous conduira naturellement à étudier les solutions possibles lorsque des problèmes d'hétéroskédasticité, de non-indépendence des perturbations et de colinéarité se présentent, comme par exemple les méthodes de régression ridge et LASSO. Nous traiterons également l'approche bayésienne pour effectuer une régression linéaire et discuterons des liens avec le point de vue fréquentiste.
2. Les modèles de régression linéaire ne sont pas satisfaisants pour tous les contextes. Il est généralement nécessaire des les rendre plus flexibles et nous étudierons le cas des modèles linéaires généralisés qui permettent, entre autre, d'intégrer des processus non Gaussien à l'analyse de la régression. Nous nous intéresserons au cas de modèles de régression dérivant des distributions exponentielles (incluant les lois normal, gamma, Poisson, binomial) et en particulier la régression logistique (utile quand la variable de réponse est binaire) et les modèles log-linéaires (quand la variable de réponse est catégorielle). En contraste avec la régression linéaire, la plupart des estimateurs ne sont pas connus explicitement et des méthodes numériques d'estimation seront donc présentées.
3. La méthode des moindres carrés peut être très sensible à des valeurs aberrantes. La notion de point de rupture d'un estimateur sera introduite et nous permettra de quantifier cette instabilité. Cela nous conduira naturellement à l'étude des techniques de régression robuste qui sont moins vulnérables aux données aberrantes. Plus particulièrement, nous introduirons les M-estimateurs ainsi que d'autres approches liées comme les S-estimateurs et les MM-estimateurs.
4. Nous introduirons les méthodes de régression non-paramètrique. Dans ce cadre, le prédicteur ne suit pas une forme imposée par le statisticien mais est directement déterminé par les données. L'intérêt est que le statisticien n'est pas ici obligé de connaître (ou de supposer) la nature de la relation entre la variable de réponse et les variables explicatives. Nous envisagerons dans ce cours les techniques de régression locale, régression sur base de splines et les estimateurs à

noyaux. Si le temps le permet, nous discuterons également l'intérêt des méthodes basées sur des arbres de régression.

Description détaillée :

- Régression linéaire multiple : hypothèses générales des moindres carrés, Théorème de Markov-Gauss, Théorème de Cochran, analyse dans le cas Gaussien (intervalle de confiance, tests d'hypothèses et prédiction), sélection de variable (tests sur les modèles emboités), moindres carrés pondérés, moindres carrés généralisés et approche bayésienne.
- Modèle linéaire généralisé : composantes du modèles, estimation, qualité d'ajustement, diagnostique (test de Wald), modèle log-linéaire, régression logistique, régression de Poisson, algorithmes Fisher scoring et Newton-Raphson.
- Régression robuste : point de rupture, estimation robuste dans le modèle de régression linéaire simple, moindre médiane des carrés, moindre résidus carrés tronqués, M-estimateurs, S-estimateurs et MM-estimateurs, diagnostiques.
- Régression non-paramètrique : régression linéaire locale, régression par polynome locaux, régression par noyaux (estimateur Nadaraya-Watson), régression par splines et méthodes par arbres de régression.

Évaluation :

type	date	pondération
Devoir 1	début février	10%
Intra	26 février	35%
Devoir 2	fin-mars	10%
Final	29 avril [1]	45%

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 consulter le lien suivant pour connaitre la procédure à suivre : `https://safire.umontreal.ca/reussite-et-ressources/mesures-daccommodement-aux-examens-pour-les-etudiants-en-situation-de-handicap/`.

Autres règles : La date limite pour abandonner le cours « sans frais » est le 21 janvier alors « qu'avec frais » (sans être remboursé), c'est le 13 mars. Par la suite, si vous abandonnez, vous aurez un échec. Vous avez également l'obligation de motiver une absence prévisible à l'intra ou au final dès que vous seriez en mesure de constater que vous ne pourriez être présent. Il appartiendra à l'autorité compétente de déterminer si le motif est acceptable.

Plagiat : L'Université de Montréal a une politique très claire sur le plagiat que vous êtes invités à consulter au `www.integrite.umontreal.ca`. Elle ne concerne pas que les examens, mais également les devoirs.

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.

Cornillon, P-A. et Matzner-Løber, E. (2011). *Régression avec R*, Springer

Sen, A. et Srivastava, M. (2012). *Regression Analysis Theory, Methods and Applications*, Springer Science & Business Media.

Chatterjee, S. et Simonoff, J. (2013). *Handbook of Regression Analysis*, John Wiley & Sons.

Rousseeuw, P. J. et Leroy, A. M (2003). *Robust Regression and Outlier Detection*, Wiley Series in Probability and Statistics.

Eubank, R. L (1999). *Nonparametric Regression and Spline Smoothing*, CRC Press.

Pour la disponibilité des livres en bibliothèque, contactez le comptoir de prêt `www.bib.umontreal.ca/nous-joindre/mi.htm`.