STT 2400–Régression linéaire

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

Horaires et déroulement du cours

Les cours ont lieu les **lundi de 8h30 à 10h30** et **mardi de 14h30 à 15h30** sur zoom, du 18 janvier jusqu'au 27 avril (sauf pour la semaine du 1 au 5 mars). Pour y assister, vous devez vous connecter sur votre compte zoom de l'UdeM et vous rendre à la réunion suivante :

- numéro de réunion : 978 3974 0490
- mot de passe : 192199

Les cours se font en direct au tableau en s'appuyant sur des notes de cours et sur le livre de référence. 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. Il y a une séance de TP chaque semaine les **mardi de 8h30 à 10h30**, à partir de la semaine du 25 janvier. Ces séances seront un mélange d'exercices théoriques et de pratique de la régression avec le logiciel R. Elles se dérouleront sur zoom :

- numéro de réunion : 931 2974 6664
- mot de passe : 004165

Enfin, il y a une période de questions chaque semaine sur zoom les **vendredi de 10h30 à 12h30**

- 953 0974 2523
- 868904

Présentation

Ce cours a pour objet d'étude le modèle de régression linéaire, un modèle statistique aussi élémentaire que puissant, utilisé dans les études statistiques d'un grand nombre de domaines (biologie, économie, santé, etc.). Ce modèle décrit une variable d'intérêt, appelée la variable réponse, comme une combinaison linéaire de variables explicatives.

La simplicité du modèle, qui permet entre autre une certaine flexibilité d'utilisation, fait de la régression linéaire un cas d'école qui permet de revoir et d'appliquer un certain nombre de concepts statistiques importants : variable aléatoire, estimateur, intervalle de confiance, test d'hypothèse mais aussi des outils d'analyse, d'algèbre linéaire et de géométrie. Au-delà de son aspect académique, ce modèle constitue, toujours aujourd'hui, généralement le point de départ de l'étude statistique d'un phénomène d'intérêt pour un grand nombre de statisticiens. Pour cette raison, la partie illustration et mise en pratique avec le logiciel R occupera une place importante dans la formation. Par conséquent, nous couvrirons principalement les chapitres 1 à 6 du livre *Régression avec R*, référencé plus bas.

Parmi les méthodes permettant d'effectuer une régression linéaire, nous nous concentrerons essentiellement sur la méthode des moindres carrés ordinaires. En effet, pourvu que certaines hypothèses sur les données soient vérifiées, l'analyse théorique de cette méthode est accessible au premier cycle et offre des résultats relativement explicites. L'analyse théorique de la régression occupera une place importante dans le cours. De plus, la philosophie des méthodes de régression plus sophistiquées se basent, dans une certaine mesure, sur la méthode des moindres carrés. Nous évoquerons également les limites de cette méthode et principalement son manque de robustesse.

Objectifs généraux : À l'issue de ce cours, l'étudiant sera familier avec les concepts de régression linéaire. Elle ou il saura en particulier appliquer la méthode des moindres carrés pour

- décrire des observations d'intérêt à partir de variables explicatives,
- identifier des liens d'association entre toutes ces variables,
- prédire de nouvelles observations.

Pour atteindre ces objectifs, elle ou il saura estimer les paramètres du modèle de régression linéaire en pratique à l'aide du language R et aussi en théorie (intervalle de confiance et tests d'hypothèses). En outre, elle ou il aura acquis les réflexes essentiels permettant de valider l'analyse : vérification des hypothèses et, le cas échéant, ajuster le modèle. L'étudiant aura une connaissance précise des limites de cette méthode, ce qui devrait éveilleur leur curiosité

pour les méthodes dérivées, plus récentes et plus robustes. À un autre niveau, l'étudiant devra être à l'aise avec les démonstrations et les techniques de calculs vues en cours.

Description détaillée

- Régression linéaire simple : hypothèses générales ; estimateur des moindres carrés ordinaire et leurs propriétés : optimalité parmi une certaine classe d'estimateurs (théorème de Gauss-Markov). Cas Gaussien : distributions des estimateurs ; tests d'hypothèses, intervalle de confiance et évaluation de la qualité de la régression.
- Régression linéaire multiple : hypothèses générales ; généralisation des résultats théoriques précédents ; interprétation géométrique de la régression (application des résultats d'algèbre linéaire) ; évaluation de la qualité de régression dans ce cas général. Cas Gaussien : estimateur du maximum de vraisemblance, Théorème de Cochran, intervalle de confiance, tests d'hypothèses, tests sur les modèles emboités.
- Validation des hypothèses du modèle de régression : analyse des résidus pour valider l'homoscédasticité/indépendence des erreurs et recherche de valeurs aberrantes ; analyse de l'influence des observations sur (i) leur estimation (point de levier), (ii) sur les paramètres de régression (distance de Cook) et (iii) sur tous les paramètres du modèle de régression (écart Welsh-Kuh). Ajustement au modèle : ajout de variable (régression partielle), transformation des variables explicatives et transformation des réponses (Box-Cox).
- Sélection de variables : comparaison de modèles (tests sur modèles emboités) ; compromis biais/variance des estimateurs ; critères de sélection (en fonction de l'objectif de la régression : estimer les paramètres ou décrire les données ou prédire des données) ; recherche non-exhaustive ; problèmes de multicolinéarité des variables explicatives.

Démonstratrice

Yuxi Wang, yuxi.wang@umontreal.ca

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 cinq étudiants. Les examens, d'une durée de deux heures chaque, se dérouleront en ligne.

type	date	pondération
Devoir 1	début février	10%
Intra	lundi 22 février 8h30–10h30	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

L'ouvrage de référence est

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

est disponible en ligne sur `https://bib.umontreal.ca/` au format numérique.

Autres ouvrages :

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

Weisberg, S. (2005). *Applied linear regression* Wiley Series in Probability and Statistics.