UQÀM | Département de mathématiques
FACULTÉ DES SCIENCES
Université du Québec à Montréal

PLAN DE COURS
MAT 7381 - RÉGRESSION - 3 CRÉDITS (HIVER 2017)

Professeur: Jean-François Coeurjolly
coeurjolly.jean-francois@uqam.ca
Département de Mathématiques, Pavillon Kennedy,
Faculté des Sciences, UQAM
Bureau 5525.
Disponibilités: Lundi 14h00-15h30.

1 Description succincte du cours

Le cours vise à présenter quelques modèles standard de régression paramétrique: modèles de reégression linéaire, modèles de régression linéaire généralisés, modèles linéaires mixtes.

2 Horaire et format du cours

Les cours débutent la semaine du 9 janvier et se terminent la semaine du 28 avril. Ils auront lieu les mardi de 10h30 à 12h00 et de 13h30 à 15h00 dans la salle SH-3540. Veuillez noter qu'il n'y aura pas cours la semaine du 27 février au 03 mars (semaine de relâche).

Chaque séance de cours prendra la forme d'un exposé magistral assuré par le professeur. D'une semaine sur l'autre, l'étudiant se verra proposer des séries d'exercices en lien avec le cours. Il pourra s'agir d'exercices théoriques ou pratiques et dans ce cas ils seront à réaliser avec le logiciel R. L'étudiant aura à charge d'imprimer les notes de cours (copie partielle des acétates présentées en cours) qu'il pourra compléter pendant la présentation du professeur.

3 Evaluation des apprentissages

- *Examens écrits (40% de la note finale)*: deux examens en classe seront envisagés pour ce cours comptant pour 20% chacun. Ces deux examens testeront davantage les aptitudes théoriques. Il s'agira d'un examen écrit vraisemblablement de 2h.
- *Projet (40% de la note finale)*: Les projets seront réalisés en équipe de deux, cet effectif pouvant toutefois être modulé par le professeur en fonction du nombre d'étudiants dans le groupe. Une recherche sur un sujet relié à la régression sera entreprise et les

résultats seront présentés sous la forme d'un article rédigé à l'aide de l'environnement Rmarkdown de Rstudio et d'une présentation orale effectuée par exemple de type Beamer (utilisant l'environnement Rmarkdown). La lecture d'articles de revues scientifiques pourrait s'avérer nécessaire. Les détails de ce projet seront annoncés en cours de session et disponibles sur le site web associé à ce cours.

- *Devoirs (20% de la note finale)*: Au cours de la session, deux devoirs maison seront à préparer, comptant pour 10% chacun de la note finale. Ces deux devoirs (et cela est conseillé) peuvent être préparés en binôme. Ils devront être rendus sous format .pdf et utiliser l'environnement Rmarkdown. Le sujet sera déposé environ dix jours avant sa date de remise sur le moodle associé au cours. La non remise d'un devoir dans le temps imparti impliquera nécessairement une note 0 au devoir correspondant. Un devoir maison permet de s'assurer que les éléments théoriques et pratiques vus en cours ont été correctement assimilés.

Remarques importantes: La présence en classe est fortement recommandée pour bien réussir le cours. Tout acte de tricherie pendant les examens ou tout plagiat entraînera de lourdes sanctions.

Calendrier prévisionnel: (le calendrier est susceptible d'être modifié légèrement)

- Devoir 1 : à rendre pour le mardi 7 février.
- Examen 1 : le mardi 7 mars de 10h30 à 12h30.
- Mardi 7 mars de 13h30 à 15h00 : discussion autour du projet de fin de session.
- Devoir 2 : à rendre pour le mardi 28 mars.
- Examen 2 : le mardi 18 avril de 10h30 à 12h30.
- Mardi 18 avril de 13h30 à 15h00 : soutien aux projets de fin de session.
- Mardi 25 avril : soutenance des projets.
- Mardi 2 mai : date limite de remise des projets de fin de session.

4 Objectifs

La régression permet de modéliser la relation qui peut exister entre une (ou plusieurs) variable(s) que l'on tente d'expliquer et une ou plusieurs autres variables explicatives. Par exemple, quelle est la relation entre la consommation d'essence et différentes autres variables telles que la taxe sur l'essence, la proportion de gens avec un permis de conduire, etc. ou encore, quelle est la relation entre le fait de guérir ou non d'une maladie et l'âge du patient, son sexe et le fait qu'il fume ou non. Dans un premier cours, on s'intéresse

généralement à la régression linéaire, c'est-à-dire que la relation entre les variables est linéaire. Dans ce deuxième cours, nous allons voir certains des aspects les plus poussés de la régression linéaire multiple de même que certaines généralisations où, par exemple, la variable à expliquer n'est plus continue; dans ce cas elle pourrait être binaire (et l'on parle en général de régression logistique) ou de comptage (et l'on parle alors de régression de Poisson). Par ailleurs, de nombreux jeux de données sont constituées de mesures répétées pour la variable réponse ou plus simplement, les données peuvent être catégorisés en plusieurs groupes qui ne suivent pas tous parfaitement le même modèle. Pour ce type de données, il peut être intéressant d'introduire une couche d'aléatoire supplémentaire en utilisant des modèles linéaires mixtes.

La régression offre plusieurs défis intéressants. Par exemple, une seule observation peut-elle grandement influencer les résultats obtenus? Comment doit-on choisir les variables qui feront partie du modèle? Doit-on vérifier l'influence d'une observation avant ou après avoir choisi les variables dans le modèle? A-t-on le droit d'interpréter chaque p-valeur associée à un coefficient de régression de manière individuelle et de façon classique? Les réponses à ces questions sont rarement uniques ! Ce cours s'adresse à tout étudiant de maîtrise ou de doctorat en statistique qui veut se familiariser davantage avec certains des outils fondamentaux de la modélisation statistique de même qu'avec certains des développements les plus récents.

5 Plan du cours

1. Introduction et rappels: introduction générale au travers de plusieurs exemples; rappels d'algèbre linéaire de probabilité et de statistique; présentation succincte des logiciels R/RStudio et de l'environnement RMarkdown.

2. Modèles de régression linéaire multiple:

   (a) sans hypothèse gaussienne: estimation par moindres carrés; valeurs ajustées; résidus; coefficient de détermination et prédiction; propriétés mathématiques et théorèmes limite; estimation par intervalles de confiance et tests d'hypothèses asymptotiques; bootstrap paramétrique.

   (b) avec hypothèse gaussienne: estimateurs du maximum de vraisemblance; loi des estimateurs; intervalles et régions de confiance; intervalles de prédiction; tests d'hypothèses paramétriques; test du rapport de vraisemblance.

3. Détection et correction des écarts au modèle: Analyse des résidus; Analyse de la matrice de projection; mesures d'influence; détection et (premier) traitement de la colinéarité; détection et traitement de l'hétéroscédasticité par la méthode des moindres carrés généralisés.

4. Sélection de variables et autres questions liées à la grande dimension: critères de qualité d'un modèle; critère de sélection de variables; Liens entre les différents

critères; Méthodes algorithmiques de sélection; Courte introduction au problème des tests multiples; Régression ridge et régression lasso.

5. Introduction aux modèles linéaires généralisés:
   (a) définition générale du modèle et quelques exemples; fonction de lien canonique; estimateur du maximum de vraisemblance; déviance et propriétés asymptotiques; tests et intervalles de confiance; analyse des résidus;
   (b) Régression pour une variable réponse binaire: modèles logit et probit; outils et tests spécifiques pour ce modèle.
   (c) Régression pour une variable quantitative discrète: ex. de la régression poissonienne; outils et tests spécifiques pour ce modèle.
6. Introduction aux modèles linéaires mixtes.

6 Bibliographie

Le cours ne suit pas un ouvrage en particulier. Aussi, il n'est pas obligatoire de se procurer et lire les ouvrages suivants pour réussir avec succés le cours. Les notes de cours préparés par le professeur devraient être suffisantes. Néanmoins, elles pourraient parfois pour des lectures complémentaires renvoyer à certains chapitres de ces ouvrages.

- Lafaye de Micheaux, P., Drouilhet, R. et Liquet, B. (2010). Le logiciel R - Maîtriser le langage - Effectuer des analyses statistiques, Springer.
- Cornillon, P.-A. et Matzner-Løber, E. (2007). Régression. Théorie et applications, Springer, France.
- Cornillon, P.-A. et Matzner-Løber, E. (2011). Régression avec R, Springer, France.
- McCullagh, P. et Nelder, J.A. (1989). Generalized Linear Models. Second Edition. Monographs on Statis- tics and Applied Probability 37. Chapman and Hall.
- Madsen H. and Thyregord, P (2011). Introduction to General and Generalized Linear Models, CRC Press, Taylor and Francis Group, Boca Raton, Florida, USA.
- Weisberg, S (2005). Applied linear regression. Vol. 528. John Wiley & Sons.