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

Plan de cours
STT 5100 - Modèles linéaires appliqués
3 crédits (Hiver 2020)

Professeur: Arthur Charpentier
charpentier.arthur@uqam.ca
Département de Mathématiques, Pavillon Kennedy,
Faculté des Sciences, UQAM
Bureau 5330.
https://github.com/freakonometrics/STT5100

1 Description succincte du cours

Méthode d'estimation par moindre carrés, Régression linéaire simple et multiple, Tests d'hypothèses et intervalles de confiance, tests d'ajustement, introduction aux modèles linéaires généralisés (GLM): régression logistique, régression Poisson, régression Gamma. Applications numériques à l'aide de R
(via https://etudier.uqam.ca/cours?sigle=STT5100)

2 Horaire et format du cours

Les cours débutent la semaine du 6 janvier et se terminent le 24 avril. Ils auront lieu les vendredi de 9h00 à 12h00. Chaque séance de cours prendra la forme d'un exposé magistral assuré par le professeur. Il n'y a pas de notes de cours, le cours sera fait au tableau, et tiendra compte du ryhtme de la classe, et des questions posées. Des références complémentaires sont mentionnées sur la page https://github.com/freakonometrics/STT5100, avec une ébauche de document reprenant les éléments essentiels du cours, et le code R.

3 Evaluation des apprentissages

Petit rappel : comme le disait Charles Goodhart, membre du Comité de politique monétaire de la Banque d'Angleterre, "when a measure becomes a target, it ceases to be a good measure". Autrement dit, les examens ne sont pas l'objectif du cours. Les examens servent juste a vérifier que la matière présentée a été travaillée, et comprise.

A des fins pédagogiques, des exercices seront donnés d'une semaine sur l'autre. Il s'agira de simple quizz, non-notés, mais dont la participation sera intégrée dans les travaux de session.

- Examen #1 le 21 février 2020 (3 heures : 35% de la note finale) : modèle linéaire
- Examen #2 le 24 avril 2020 (3 heures : 35% de la note finale) : GLM
- Travail de session #1, à remettre le 1er mars à midi (15% de la note finale)
- Travail de session #2, à remettre le 1er mai à midi (15% de la note finale)

Les instructions au sujet des deux travaux seront en ligne sur la page
`https://github.com/freakonometrics/STT5100`
À titre indicatif, les examens et devoirs des années passées sont en ligne[1]

- Examen #1 automne 2019, `sujet` et `correction`
- Examen #1 hiver 2019, `sujet` et `correction`
- Examen #1 automne 2018, `sujet` et `correction`
- Examen #2 automne 2019, `sujet` et `correction`
- Examen #2 hiver 2019, `sujet` et `correction`
- Examen #2 automne 2018, `sujet` et `correction`
- Devoir #1 automne 2019, `sujet`
- Devoir #1 hiver 2019, `sujet`
- Devoir #1 automne 2018, `sujet`
- Devoir #2 automne 2019, `sujet`
- Devoir #2 hiver 2019, `sujet`
- Devoir #2 automne 2018, `sujet`

[1]Le terme "correction" utilisé ic désigne des éléments partiels de correction

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 (voir régelement no 18).

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 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 Prérequis

Les prérequis suivants sont indispensables pour suivre le cours:

- Avoir suivi un ou deux cours d'algèbre linéaire;
- Avoir suivi des cours de statistique inférentielle au baccalauréat;
- Avoir connaissance du logiciel R.

Avoir connaissance du l'environnement de RStudio et du langage Rmarkdown (permettant la création de documents reproductibles) sont des plus. Ils seront présentés dans les séances de démonstration.

6 Plan du cours

1. Rappels de probabilités, de statistiques, d'optimisation (~5% du temps)

   Espérance, Variance, Espérance Conditionnelle, Variance Conditionnelle, Loi Gaussienne, loi de Poisson, loi Gamma, famille exponentielle, vecteur Gaussien, introduction aux méthodes de simulation (Monte Carlo), maximum de vraisemblance et théorie asymptotique, tests basés sur la vraisemblance, optimisation numérique, optimisation avec contraintes (linéaires), rappels d'algèbre linéaire

2. Modèle linéaire Gaussien (~45% du temps)

   Variables continues et variables catégorielles, Méthode des moindres carrés, Maximum de vraisemblance, Propriétés sous hypothèse de normalité, théorème de Gauss-Markov, Significativité, Prévision et Intervalle de confiance, Simulation dans le modèle linéaire et incertitude, Choix de modèle et choix de variables (stepwise), Moindres carrées pondérés, Introduction aux méthodes non-linéaire : approche locale et décomposition dans une base (splines), modèle additif vs. multiplicatif et transformation de Box-Cox, diagnostique et analyse des résidus, tests de normalité

   Bonus : méthode de pénalisation et lasso (sélection de variable)

3. Régression logistique et méthodes de classification (~20% du temps)

   Maximum de vraisemblance, méthodes numériques d'estimation, logistique vs. Probit, Incertitude dans le modèle logistique et prévision

   Bonus : introduction aux arbres de classification

4. Régression de Poisson (~15% du temps)

   Maximum de vraisemblance, méthodes numériques d'estimation, équi-dispersion, offset, méthode des marges et tableaux de contingences

5. Modèles GLM (~15% du temps)

   Régression dans la famille exponentielle, maximum de vraisemblance, déviance et analyse des résidus, choix de modèles

   Bonus : quasi-vraisemblance et sur/sous-dispersion

7 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. Néanmoins, certains chapitres permettent d'avoir les preuves complètes de certains résultats, et d'aller plus loin.

- Cameron, R. & Trivedi, P.K. (2013). Regression Analysis of Count Data, Cambridge University Press.
- 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.
- de Jong, P. & Heller, G. (2008). Generalized Linear Models for Insurance Data[2], Cambridge University Press.
- 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.
- Zeileis, A. (2008) Applied Econometrics with R, Springer Verlag.

Pour se familiariser avec RMarkdown

- `https://toltex.u-ga.fr/cours/RMarkdown/rmd`
- `https://daringfireball.net/projects/markdown/syntax`
- `https://rmarkdown.rstudio.com/lesson-1.html`

8 Compléments

8.1 Politique 16 sur le harcèlement sexuel

Le harcèlement sexuel se définit comme étant un comportement à connotation sexuelle unilatéral et non désiré ayant pour effet de compromettre le droit à des conditions de travail et d'études justes et raisonnables ou le droit à la dignité.

La Politique 16 identifie les comportements suivants comme du harcèlement sexuel :

[2] L'ouvrage contient néanmoins un nombre important d'imprécisions

1. Manifestations persistantes ou abusives d'un intérêt sexuel non désirées.
2. Remarques, commentaires, allusions, plaisanteries ou insultes persistants à caractère sexuel portant atteinte à un environnement propice au travail ou à l'étude.
3. Avances verbales ou propositions insistantes à caractère sexuel non désirées.
4. Avances physiques, attouchements, frôlements, pincements, baisers non désirés.
5. Promesses de récompense ou menaces de représailles, implicites ou explicites, représailles liées à l'acceptation ou au refus d'une demande d'ordre sexuel.
6. Actes de voyeurisme ou d'exhibitionnisme.
7. Manifestations de violence physique à caractère sexuel ou imposition d'une intimité sexuelle non voulue.
8. Toute autre manifestation à caractère sexuel offensante ou non désirée.

Pour plus d'information, Pour rencontrer une personne ou faire un signalement : Bureau d'intervention et de prévention en matière de harcèlement : 514-987-3000, poste 0886
`http://www.harcelement.uqam.ca`

8.2 Règlement no 18 sur les infractions de nature académique

Tout acte de plagiat, fraude, copiage, tricherie ou falsification de document commis par une étudiante, un étudiant, de même que toute participation à ces actes ou tentative de les commettre, à l'occasion d'un examen ou d'un travail faisant l'objet d'une évaluation ou dans toute autre circonstance, constituent une infraction au sens de ce règlement La liste non limitative des infractions est définie comme suit :

- la substitution de personnes ;
- l'utilisation totale ou partielle du texte d'autrui en le faisant passer pour sien ou sans indication de référence ;
- la transmission d'un travail pour fins d'évaluation alors qu'il constitue essentiellement un travail qui a déjà été transmis pour fins d'évaluation académique à l'Université ou dans une autre institution d'enseignement, sauf avec l'accord préalable de l'enseignante, l'enseignant ;
- l'obtention par vol, manœuvre ou corruption de questions ou de réponses d'examen ou de tout autre document ou matériel non autorisés, ou encore d'une évaluation non méritée ;
- la possession ou l'utilisation, avant ou pendant un examen, de tout document non autorisé ;

- l'utilisation pendant un examen de la copie d'examen d'une autre personne ;
- l'obtention de toute aide non autorisée, qu'elle soit collective ou individuelle ;
- la falsification d'un document, notamment d'un document transmis par l'Université ou d'un document de l'Université transmis ou non à une tierce personne, quelles que soient les circonstances ;
- la falsification de données de recherche dans un travail, notamment une thèse, un mémoire, un mémoire - création, un rapport de stage ou uns rapport de recherche.

Les sanctions reliées à ces infractions sont précisées à l'article 3 du Règlement no 18 Pour plus d'information sur les infractions académiques et comment les prévenir : `http://www.integrite.uqam.ca`

8.3 Respect de l'intégrité académique

Face à l'importance et à l'ampleur du phénomène de la tricherie et du plagiat dans les universités, ici et à l'étranger, l'UQAM a amorcé, en janvier 2007, une démarche visant à prom ouvoir le respect de l'intégrité académique. Dans ce contexte et inspirée d'une philosophie de « tolérance zéro », la Commission des études de l'UQAM a modifié son Règlement sur les infractions de nature académique (R. 18) à sa réunion du 2 décembre 2008. Endossant cette philosophie de « tolérance zéro » relativement aux actes de plagiat, de fraude et de tricherie, la Faculté des sciences de l'UQAM souhaite sensibiliser ses étudiants à l'importance du respect de l'intégrité académique. Puisqu'en sollicitan t son admission à l'UQAM, toute candidate, tout candidat s'engage à suivre les politiques et règlements de l'Université, la Faculté souhaite informer ses étudiants des différents articles de ce règlement, des actes répréhensibles et des sanctions applicabl es. Un extrait de ces articles se trouve ci - dessous. Le Règlement complet et son application à la Faculté des sciences sont disponibles à l'adresse Web suivante : `http://www.sciences.uqam.ca/decanat/reglements.php` Tous ces efforts visent à assurer la validité de la formation dispensée par la Faculté, ainsi qu'un traitement équitable de tous afin de maintenir la qualité de ses diplômes.

Article 2 - Infractions de nature académique

2.1 Infraction

Tout acte de plagiat, fraude, copiage, tricherie, falsifica tion de document ou création d'un faux document commis par une candidate, un candidat, une étudiante, un étudiant de même que toute participation à ces actes ou tentative de les commettre, à l'occasion d'un examen, d'un travail ou d'un stage faisant l'obje t d'une évaluation ou dans toute autre circonstance, constitue une infraction au sens de ce règlement.

2.2 Liste non limitative des infractions

Sans limiter la généralité de ce qui précède, constitue notamment une infraction le fait de poser ou tenter de poser l'un des actes suivants ou le fait d'y participer :

- la substitution de personnes ou l'usurpation d'identité;
- le plagiat : l'utilisation totale ou partielle du texte ou de la production d'autrui en le faisant passer pour sien ou sans indication de référence;
- l'autoplagiat : le dépôt d'un travail pour fins d'évaluation alors que ce travail constitue essentiellement un travail qui a déjà été soumis pour fins d'évaluation académique à l'Université ou dans une autre institution d'enseignement, sauf avec l'accord préalable de l'enseignante, l'enseignant;
- la possession ou l'obtention par vol, manoeuvre ou corruption de questions ou de réponses d'examen;
- la possession ou l'utilisation de tout document ou matériel non autorisé préalablement, pendant un examen ou lors de la réalisation de travaux, incluant le recours aux outils informatiques ou moyens technologiques;
- l'utilisation pendant un examen de la copie d'examen ou de tout autre matériel provenant d'une autre personne;
- l'obtention de toute aide non autorisée, qu'elle soit collective ou individuelle;
- l'obtention d'une évaluation non méritée notamment par corruption, chantage, intimidation ou toute forme de harcèlement ou la tentative d'obtenir une telle évaluation;
- la falsification d'un document ou la création d'un faux document, notamment d'un document transmis à l'Université ou d'un document de l'Université transmis ou non à une tierce personne, quelles que soient les circonstances;
- la falsification de données de recherche dans un travail, notamment une thèse, un mémoire, un mémoire - création, un rapport de stage ou un rapport de recherche.

Article 3 - Sanctions

3.1 L'attribution de la mention «P»

L'étudiante, l'étudiant qui commet une infraction est mis en probation et peut se voir imposer une ou plusieurs sanctions. La mise en probation génère l'attribution de la mention «P» au dossier informatisé de l'étudiant, de l'étudiante. La mention «P» n'apparaît pas au relevé de notes de l'étudiante, l'étudiant mais figure en tout temps à son dossier.

Lorsque la sanction est la suspension, une mention à cet effet apparaîtra au relevé de notes pour la durée de la suspension. Dans le cas d'une expulsion définitive de l'Université, une mention à cet effet apparaîtra de manière permanente au relevé de notes.

3.2 La mise en probation et autres sanctions

- la mise en probation; La mise en probation constitue la reconnaissance que l'étudiante, l'étudiant a commis une infraction au présent règlement. La mise en probation peut être imposée sans autre sanction, auquel cas, l'enseignant, l'enseignante se voit inviter à attribuer une notation à l'étudiante, l'étudiant pour le cours conformément au résultat obtenu pour les prestations complétées. La mise en probation sans autre sanction signifie que la mention «P» est inscrite au dossier de l'étudiante, l'étudiant et que celle, celui qui en est l'objet ne doit commettre aucune autre infraction au présent règlement, à défaut de quoi, l'une ou l'autre des sanctions suivantes lui sera imposée. Outre la mise en probation, l'étudiante, l'étudiant peut se voir imposer une ou plusieurs des sanctions suivantes :
- l'échec au cours ou à l'activité créditée;
- l'obligation de réussir de trois à six crédits additionnels, hors programme, afin d'obtenir son grade, diplôme, certificat ou attestation; les cours doivent être identifiés;
- la suspension de toute activité à l'Université, pour une période maximale de neuf trimestres consécutifs;
- son expulsion définitive de l'Université