Université de Montréal

Département de démographie

Démographie

PLAN DE COURS

DMO 6371 AUTOMNE 2017

MÉTHODES STATISTIQUES EN DÉMOGRAPHIE 3 CR.

COURS:	**lundi et mercredi, de 09h30 à 11h30**
EXAMEN INTRA:	**lundi, 16 octobre 2017, de 09h30 à 11h15**
EXAMEN FINAL:	**lundi, 18 décembre 2017, de 09h30 à 11h15**

- Modification du choix de cours: date limite le **20 septembre 2017** *(tout cours annulé pendant la période active de modification du choix de cours ne sera pas mentionné dans le relevé de notes et ne sera pas facturé à l'étudiant);*
- Abandon d'un cours: date limite le **10 novembre 2017** *(entre le 21 septembre et le 11 novembre, l'abandon de cours peut se faire en se présentant au Secrétariat de son département; tout cours abandonné fera l'objet d'une mention "ABA" sur le relevé de notes et la facturation des frais de scolarité sera maintenue).*

Professeur:	**LeGRAND, Thomas K.**
Courriel:	tk.legrand@umontreal.ca
Local:	C-5016
Disponibilité:	Lundi, de 14h00 à 15h30 ou sur demande

Une version électronique de ce plan de cours est disponible sur Internet. On peut y accéder par la page d'accueil du Département de démographie (**www.demo.umontreal.ca**). Cependant, noter que les informations qui suivent peuvent faire l'objet de modifications au cours du trimestre. Le cas échéant, le professeur vous avisera en classe ou, s'il y a lieu, au moyen du calendrier affiché sur le site StudiUM du cours (https://studium.umontreal.ca/).

OBJECTIF DU COURS

- Faire le survol des méthodes statistiques de base utilisées en démographie sociale (et en sociologie, santé publique…) pour l'analyse explicative, c'est-à-dire l'analyse des causes et conséquences des phénomènes démographiques.
- Sensibiliser les étudiants aux problèmes méthodologiques et conceptuels que l'on rencontre fréquemment dans les études empiriques et donner les outils pour les contourner ou les éviter.
- Donner une compréhension intuitive et, dans une certaine mesure, mathématique, de comment ces techniques fonctionnent et pourquoi elles sont appropriées (ou non) dans des situations différentes.
- Développer la capacité fonctionnelle et l'habitude d'utiliser ces techniques par les étudiants à travers des exercices sur micro-ordinateur.

Les étudiants apprendront les notions et méthodes statistiques de base les plus importantes utilisées en démographie pour l'analyse explicative: les causes et conséquences des phénomènes. Le cours ciblera les régressions linéaires et non-linéaires binomiales et polychotomiques, et mettra l'accent sur les enjeux conceptuels et problèmes méthodologiques que l'on rencontre fréquemment dans les études empiriques. Ces connaissances sont essentielles pour comprendre la plupart des études publiées dans les revues scientifiques et, le plus souvent, pour faire un mémoire ou une thèse en démographie. La maîtrise de ces méthodes est un prérequis pour accéder aux cours de statistiques plus avancés, tels que l'analyse multiniveaux ou les méthodes de risques et durées. Les étudiants auront à faire beaucoup d'exercices pratiques afin de perfectionner leurs compétences à utiliser ces méthodes et d'interpréter correctement les résultats. Les fondations mathématiques de ces méthodes ne seront pas présentées dans ce cours; pour ceux qui s'y intéressent, voir les cours offerts en statistique ou en économétrie.

Voici quelques-unes des questions auxquelles nous allons répondre au cours du trimestre:

- Comment utiliser les méthodes statistiques pour étudier les déterminants de la survie des enfants?
- Comment savoir si une variable « indépendante » a un impact significatif sur ce poids (par exemple, l'effet de la scolarité de la mère sur le poids à la naissance des enfants) et que veut dire significatif?
- Que sont les variables continues, dichotomiques (*dummies*), polychotomiques, *proxies* et les interactions? Comment les utiliser dans les régressions et ensuite interpréter les résultats?
- Quels sont les problèmes liés à l'utilisation d'une régression linéaire pour étudier le nombre d'enfants nés des femmes dans une population donnée?
- Pourquoi une régression logistique est-elle mieux qu'une régression linéaire pour l'analyse des déterminants de la participation des femmes au marché de travail (et ça veut dire quoi "mieux")?
- Pourquoi la fécondité des femmes ne peut-elle pas être considérée comme une simple cause de leur décision de travailler ou que les pratiques d'allaitement des enfants comme un déterminant « exogène » de leur santé ou survie? Quelle méthode d'estimation est adéquate pour évaluer les effets causaux dans ces situations?
- Pourquoi et de quelle manière les techniques statistiques habituelles donnent-elles les résultats biaisés lorsque vos données proviennent d'une enquête pondérée ou en grappe?
- Pourquoi et comment les régressions de type pas-à-pas ("*stepwise*") donnent-elles des résultats systématiquement biaisés dans les études de type « cause et effet »?
- Comment estimer les divers modèles de régression sur l'ordinateur et quels sont les problèmes que l'on rencontre souvent (valeurs manquantes, préparation de la base de données, etc.)?
- Et le plus important: comment procéder pour faire une étude empirique, soit conceptualiser les liens de causalité, choisir le modèle et ensuite estimer et interpréter les résultats?

PRÉREQUIS

Pour ceux qui ont besoin de revoir les notions statistiques et mathématiques de base, il est très important d'étudier durant la première semaine:

- Le livre de Wonnacott & Wonnacott, ch. 4, section sur les notions de variable « aléatoire », la moyenne et variance d'une variable, l'espérance mathématique et la loi normale.
- Les sections du livre de Duchêne et Vilquin (1992), *Mathématiques pour démographes: Rappels théoriques - exercices résolus* (en réserve à la bibliothèque) sur les notions de base en mathématique: les logarithmes, les fonctions exponentielles et le calcul différentiel.

LIVRES OBLIGATOIRES

Statistique: Économie-Gestion-Sciences-Médecine, par T.H. Wonnacott & R.J. Wonnacott (1998; Economica) [Livre élémentaire général de statistique. Ce livre n'est plus disponible; les copies des chapitres pertinents seront mises à la disposition des étudiants inscrits au cours.]

A Guide to Econometrics, par Peter Kennedy (2008; 6e éd: MIT Press).
[Ce livre fait un bon survol intuitif et largement non mathématique des sujets difficiles; ~50$]

AUTRES OUVRAGES UTILES POUR CE COURS

- Bressoux, Pascal (2010). *Modélisation statistique appliquée aux sciences sociales*, 2e ed., De Boeck Supérieur.
- Briscoe, Akin & Guilkey (1990). "People are not Passive Acceptors of Threats to Health: Endogeneity and its Consequences", *International Journal of Epidemiology*, 19(1): 146-153.
- Jan Kmenta (1997). *Elements of Econometrics* (2e édition: University of Michigan Press).
- Alfred Demaris (1992). *Logit Modeling: Practical Applications*, Beverly Hills, Sage Publications 86.
- V.T. Farewell (1979). "Some Results on the Estimation of Logistic Models Based on Retrospective Data", *Biometrika* 66(1): 27-32.
- J.H. Stock & M.W. Watson (2007). *Introduction to Econometrics* (2e édition: Addison-Westley).
- Fred Pampel (2000). *Logit Regression: A Primer*, Beverly Hills, Sage Publications 132.
- Lee, E.S & R.M. Forthofer (2006). Analyzing complex survey data, Beverly Hills, Sage Publications 07-071.
- Pétry, F. & Gélineau F (2009). *Guide pratique d'introduction à la régression en Sciences sociales,* Les presses de l'Université Laval.
- Riberdy et al. (2000). "Des concepts aux chiffres", ch. 4 (pp. 131-136) dans *Culture, santé et ethnicité*, (S. Gravel et A. Battaglini, Eds) Régie régionale de la santé et des services sociaux de Montréal-centre.

ÉVALUATION

Le plagiat à l'UdeM est sanctionné par le *Règlement disciplinaire sur la fraude et le plagiat concernant les étudiants*. Pour plus de renseignements, consultez le site www.integrite.umontreal.ca.

Selon le règlement pédagogique (article 9.9 reproduit ci-dessous), l'étudiant doit motiver toute absence à une évaluation; pour ce faire, il faut s'adresser au Secrétariat de son département et non au professeur. Seul un motif imprévu et hors du contrôle de l'étudiant peut être acceptable.

« L'étudiant doit motiver, par écrit, toute absence à une évaluation ou à un cours faisant l'objet d'une évaluation continue ***dès qu'il est en mesure de constater qu'il ne pourra être présent à une évaluation*** *et fournir les pièces justificatives. Dans les cas de force majeure, il doit le faire le plus rapidement possible par téléphone ou courriel* ***et fournir les pièces justificatives dans les cinq jours ouvrés suivant l'absence****.*
Le doyen ou l'autorité compétente détermine si le motif est acceptable en conformité des règles, politiques et normes applicables à l'Université.
Les pièces justificatives doivent être dûment datées et signées. De plus, le ***certificat médical doit préciser les activités auxquelles l'état de santé interdit de participer, la date et la durée de l'absence, il doit aussi permettre l'identification du médecin****. »*

Les notes seront déterminées selon les pondérations suivantes:

.....	30%	Examen intra
.....	40%	Examen final
.....	30%	TP (qui peuvent être faits conjointement avec au plus un autre étudiant; dans ce cas, veuillez indiquer les deux noms sur un seul TP)

6 septembre **Introduction**: les objectifs et le plan du cours

1) <u>Le modèle classique de régression linéaire</u>

A. <u>Notions de base</u>: relations exacte et stochastique (aléatoire)

11 septembre Fin: Notions de base. Les critères d'évaluation des méthodes statistiques différentes: biais, consistance, efficacité...

B. <u>La régression linéaire et les méthodes d'estimation</u>

❶ Conceptualisation de la régression linéaire simple

À lire sur les notions de base: Wonnacott and Wonnacott (W&W), pp. 260-279; Kennedy, chapitre 1 (les notes à la fin sont recommandées mais non obligatoires).

À lire sur la régression linéaire: W&W: pp. 408-414.

Lecture optionnelle sur les notions de base: Kennedy: §2.5-2.8 et *General Notes* associées (à la fin du chapitre); Kmenta: pp. 156-172 (très bon).

13 septembre ❷ Estimation par la méthode des moindres carrés (MCO): les hypothèses simplificatrices et la méthode

❸ L'estimation par maximum de vraisemblance (EMV): l'intuition et l'hypothèse simplificatrice additionnelle requise

À lire: W&W: pp. 423-431 et 452-460.

Lecture optionnelle: Kennedy: chapitre 3 et les *General Notes* associées.

15 septembre Atelier sur l'utilisation du logiciel *Stata* au laboratoire informatique, C-3115 (3e étage du pavillon Lionel Groulx) de 13h30 à 16h00.

18 septembre ❹ Comparaison des méthodes – Résumé des hypothèses classiques et les attributs souhaitables des estimateurs qu'elles impliquent

❺ Conceptualisation de la régression linéaire multiple, avec exemples pratiques

À lire: W&W, pp. 638-649; Kennedy § 2.9 (jeter un coup d'œil aux *General* et *Technical Notes* associés à la fin).

Lecture optionnelle: Kmenta, pp. 175-183.

À remettre en classe le TP #1: Initiation à Stata et exercice sur les régressions linéaires

20 septembre C. <u>Les tests d'hypothèses et les statistiques descriptives</u>: comment savoir si un des facteurs a un effet "significatif" sur la variable qui vous intéresse?

❶ Conceptualisation et définition de la significativité statistique – Variance des aléas et des coefficients estimés

❷ Le test "t" de Student bilatéral, les intervalles de confiance et le test unilatéral "t"

À lire: Kmenta: pp. 110-120 et 237-248; Kennedy, chapitre 4: §4.1 - §4.4, les *General Notes* associées et *Technical Note* 4.1.

Lecture optionnelle: Kmenta, pp. 120-135 (très bon); W&W, pp. 547-553 et 562-565 (de *a* à *c*); Kennedy §4.5 et 4.6.

25 septembre ❸ Le test F et le coefficient de détermination R^2

D. <u>Les types de variables indépendantes</u>: continues, dichotomiques (« dummies ») et les interactions. Comment interpréter leurs coefficients dans les régressions?

À lire: W&W: pp. 494-500, 510-527.

Lecture optionnelle: Kennedy, chapitre 15 sur les "Dummy Variables" (§15.1 - §15.4 et les *General Notes* associées à la fin).

27 septembre Fin: types de variables indépendantes

E. <u>L'utilisation du test F et les interactions pour examiner les changements *structurels*</u> (ex: le modèle explicatif des déterminants du divorce a-t-il changé de façon glo-

bale suite à une modification des lois au Canada?)
Quelques exemples de régression linéaire – Interprétation des résultats

À remettre en classe le TP #2*:* Intervalles de confiance et les tests t et F

2 octobre

F. Multicolinéarité parfaite et moins que parfaite

À lire: W&W, pp. 568-572 et Kennedy, chapitre 12; jeter un coup d'œil sur les *General Notes* à la fin, surtout §12.2.

4 octobre

2) Problèmes avec les hypothèses de base du modèle

A. Problèmes de spécification de l'équation

❶ Non linéarité

À lire: Kennedy, §6.1-§6.3: (1) Transformations et la section sur les diagrammes Venn-Ballentine, les *General Notes §3.3 & §6.2*.

Lecture optionnelle: Modèle Box-Cox: Kmenta, pp. 517-521 et Kennedy *General Notes* §6.3; item sur Box-Cox; variables pertinentes omises d'une régression: Kmenta: pp. 442-449; Comment faire une étude empirique: Kennedy, §5.1-§5.3 et ch. 22.

À remettre en classe le TP #3*:* Variables dichotomiques, R^2, etc.

9 octobre **Congé de l'Action de Grâce**

11 octobre

❷ Les variables indépendantes pertinentes omises de l'équation

❸ Les variables indépendantes superflues incluses dans l'équation

16 octobre **Examen intra (section 1 du cours)**

18 octobre **Cours annulé**

23 et 25 octobre: Semaine de relâche

30 octobre et 1er novembre **Cours annulés**

6 novembre

Corrigé de l'examen

❹ L'approche recommandée pour faire une étude empirique: comment procéder dans la pratique?
Conceptualisation du modèle causal, importance de la théorie, critères d'un bon modèle analytique et statistique, spécification du modèle, problèmes des prétests (spécification du modèle à partir des résultats préliminaires), etc. *Sujet important à ne pas manquer!*

TP #4 à remettre en classe: Multicolinéarité et spécification de l'équation…

8 novembre

B. Problèmes de mesure des variables

❶ Erreurs de mesure aléatoires et l'utilisation des variables "proxies"

À lire: Johnson *et al.* (1987), pp. 324-332.

Lecture optionnelle: Kmenta, pp. 346-352 & 579-581, Kennedy, §10.1 à 10.3 et Riberdy et al. (2000), chapitre 4 (pp. 131-136), "Un exemple sur comment passer de la conceptualisation à l'opérationnalisation des variables".

13 novembre

❷ Censure/troncature de la variable dépendante et le modèle Tobit classique (ex: le nombre d'émigrants d'une famille)

À lire: Kennedy §17.1 - §17.2 (jeter un coup d'œil sur §17.3) et les *General Notes* pour §17.1-17.2; Kmenta, pp. 560-563.

15 novembre C. L'endogénéité (causalité inverse, simultanéité, etc.) des variables indépendantes (exemple: comment mesurer l'effet de l'utilisation des soins de santé sur la mortalité, étant donné que cette utilisation est elle-même fonction de l'état de santé de l'individu? Conceptualisation de la causalité et introduction à l'estimation par la méthode des variables instrumentales (moindres carrés à 2 étapes). Aperçu de la méthodologie des vraies expériences.

À lire: W&W, chapitre 25 (pp. 819-831); Kennedy chapitre 9; Briscoe et al. (1990). "People are not passive acceptors of threats to health: Endogeneity and its Consequences".

Lecture optionnelle: Stock et Watson (2007), chapitre 12: "Instrumental Variables Regression" in *Introduction to Econometrics* (pp. 421-461); Kmenta, pp. 357-361.

À remettre en classe le TP #5: Le modèle Tobit

20 novembre Fin: endogénéité

D. Les problèmes liés à la variation aléatoire "ε"
Aperçu des problèmes de l'hétéroscédasticité et de l'autocorrélation. Introduction aux pondérations, aux effets des grappes et un bref aperçu de l'analyse multiniveaux. Utilisation de Stata pour analyser les données venant des enquêtes pondérés, en grappes ou stratifiées.

À lire: Kennedy, §8.1– §8.3 et Johnson *et al.* (1987), pp. 292-307. [Notez que dans leur présentation, Johnson *et al.* utilisent: x_i = (X_i - (moyenne de X)) et U_i = l'aléa ε_i.] Pour les pondérations avec *Stata*, voir les sections 23.16 (pp. 278-282), 30.2 (pp. 345-346) et 30.2.3 (pp. 348-350) du *User's Guide version 10*.

Lecture optionnelle: Lee et Forthofer (2006), pp. 1-14; Kmenta, pp. 366-373.

22 novembre **3) Les modèles non linéaires Logit et Probit**

A. Les modèles simples binomiaux: probabilité linéaire, logit et probit
Ces modèles concernent les variables dépendantes dichotomiques telles que la survie d'un enfant à un âge donné, le mariage ou le divorce au cours d'une année ou le statut d'activité (actif vs inactif/au chômage) au moment d'un recensement. Conceptualisation; estimation; tests d'hypothèses (rapport de vraisemblance, Wald); avantages et désavantages de présenter les résultats en forme de coefficients, de risques relatives prédites ou de « rapports de cote »; interprétation des résultats.

À lire: Kennedy, §16.1– §16.3.

À remettre en classe le TP #6*:* Endogénéité et échantillonnage

27 novembre Fin des modèles simples Logit/Probit

B. Introduction aux modèles polychotomiques non ordonnés et ordonnés

29 novembre Fin des présentations sur les régressions logit et probit polychotomiques

1er décembre À remettre au secrétariat au plus tard midi le TP #7a*:* Logit et probit simples

4 décembre Exemples d'analyse empiriques à partir des modèles Logit et Probit. Extension de ces modèles à l'analyse des transitions sur le temps (modèle de survie à temps discret).

5 décembre À remettre au secrétariat au plus tard à midi le TP #7b*:* Logit polychotomique

6 décembre *Disponibilité pour répondre aux questions des étudiants en préparation de l'examen final*.

18 décembre Examen final (sections 2 et 3 du cours)