Méthodes quantitatives d'analyse (POL 2809)

Séance 6, 9 octobre 2019

Enseignante: Florence Vallée-Dubois

Bureau: C-3114

Dispos: mercredis, 10h-11h30

florence.vallee-dubois@umontreal.ca

Rappels

Prochains devoirs: format PDF s.v.p.; brouillons non acceptés.

Remise du Devoir 3: 25 octobre (repoussé d'une semaine).

Aujourd'hui

Présentation du travail final.

Fin des interactions.

Variable dépendante binaire.

Quelques précisions sur la régression OLS.

Réalisation du Devoir 2 en classe.

Travail final

Objectifs:

Travail final

Objectifs:

(1) formuler une question de recherche pouvant être testée avec des données quantitatives

Travail final

Objectifs:

(1) formuler une question de recherche pouvant être testée avec des données quantitatives

(2) réaliser les analyses,

Travail final

Objectifs:

(1) formuler une question de recherche pouvant être testée avec des données quantitatives

(2) réaliser les analyses,

(3) interpréter les résultats et

Travail final

Objectifs:

(1) formuler une question de recherche pouvant être testée avec des données quantitatives

(2) réaliser les analyses,

(3) interpréter les résultats et

(4) réfléchir aux biais potentiels et aux méthodes alternatives pour tester cette question de façon causale.

Travail final

7 parties:

Travail final

7 parties:

1. question de recherche

Travail final

7 parties:

1. question de recherche
2. hypothèse

Travail final

7 parties:

1. question de recherche
2. hypothèse
3. GOA et biais potentiels

Travail final

7 parties:

1. question de recherche
2. hypothèse
3. GOA et biais potentiels
4. description des variables

Travail final

7 parties:

1. question de recherche
2. hypothèse
3. GOA et biais potentiels
4. description des variables
5. exécution et présentation de la régression

Travail final

7 parties:

1. question de recherche
2. hypothèse
3. GOA et biais potentiels
4. description des variables
5. exécution et présentation de la régression
6. interprétation des résultats

Travail final

7 parties:

1. question de recherche
2. hypothèse
3. GOA et biais potentiels
4. description des variables
5. exécution et présentation de la régression
6. interprétation des résultats
7. discussion des limites et recherche de la causalité

Travail final

25 points au total

22,5 pour le contenu

2,5 pour le respect des consignes, la mise en forme et la qualité du français

Retour sur les séances précédentes

Régression linéaire multiple.

Plus d'une variable: on "contrôle" pour chacune.

Incertitude des coefficients $\hat{\beta}$.

RL multiple

RL multiple

$$Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \varepsilon$$

Ou, pour "k" variables indépendantes:

$$Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k + \varepsilon$$

Les interactions (suite et fin)

Les interactions (suite et fin)

Nous avons déjà vu les interactions de 2 variables dichotomiques et les interactions entre 2 variables continues.

Les interactions (suite et fin)

Nous avons déjà vu les interactions de 2 variables dichotomiques et les interactions entre 2 variables continues.

Qu'en est-il d'une interaction entre une variable continue et une variable dichotomique?

Interaction var. continue*dichotomique

Exemple:

$$Y = \hat{\alpha} + \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \hat{\beta}_3 X_1 * X_2 + \varepsilon$$

Interaction var. continue*dichotomique

Exemple:

$$Y = \hat{\alpha} + \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \hat{\beta}_3 X_1 * X_2 + \varepsilon$$

Y = taux de participation aux élections

X_1 = l'âge d'une démocratie, en années

X_2 = mode de scrutin majoritaire (= 1, les autres modes de scrutin sont codés 0)

Interaction var. continue*dichotomique

$\hat{\beta}_1$ est positif.

Les plus vieilles démocraties ont généralement un meilleur taux de participation électorale.

Interaction var. continue*dichotomique

$\hat{\beta}_1$ est positif.

Les plus vieilles démocraties ont généralement un meilleur taux de participation électorale.

$\hat{\beta}_2$ est négatif.

Les pays où le mode de scrutin est majoritaire ont généralement un moins bon taux de participation électorale.

Interaction var. continue*dichotomique

$\hat{\beta}_3$ est négatif.

L'effet positif de l'âge de la démocratie est MOINS fort pour les pays qui ont un mode de scrutin majoritaire.

Interaction var. continue*dichotomique

Explications au tableau.

Exemple chiffré

Disons:

$$\text{participation} = 60 + 0,5\text{age} - 0,4\text{maj} - 0,05\text{age} * \text{maj} + \varepsilon$$

Explications au tableau.

Interprétation

Pour les pays où le mode de scrutin est majoritaire ($X_2 = 1$):

Interprétation

Pour les pays où le mode de scrutin est majoritaire ($X_2 = 1$):

L'effet positif de l'âge de la démocratie (X_1) est moins fort.

Interprétation

Pour les pays où le mode de scrutin est majoritaire ($X_2 = 1$):

L'effet positif de l'âge de la démocratie (X_1) est moins fort.

L'effet du mode de scrutin (X_2) est plus faible pour les démocraties plus vieilles.

Interprétation

Pour les pays où le mode de scrutin est majoritaire ($X_2 = 1$):

L'effet positif de l'âge de la démocratie (X_1) est moins fort.

L'effet du mode de scrutin (X_2) est plus faible pour les démocraties plus vieilles.

Pour les pays où le mode de scrutin n'est pas majoritaire ($X_2 = 0$): il n'y a pas d'effet d'interaction (β_3 s'annule).

En chiffres

$$\text{participation} = 60 + 0,5\text{age} - 0,4\text{maj} - 0,05\text{age} * \text{maj} + \varepsilon$$

Effet de l'âge de la démocratie sur le taux de participation = 0,5-0,05 si le mode de scrutin est majoritaire

Effet du mode de scrutin majoritaire sur le taux de participation = 0,4-(0,05*âge de la démocratie)

Le terme interactif n'est pas toujours négatif

Dans l'exemple, $\hat{\beta}_3$ est négatif: il y a un désavantage à avoir un mode de scrutin majoritaire.

Le terme interactif n'est pas toujours négatif

Dans l'exemple, $\hat{\beta}_3$ est négatif: il y a un désavantage à avoir un mode de scrutin majoritaire.

Si $\hat{\beta}_3$ avait été positif: il y aurait eu un avantage à avoir un mode de scrutin majoritaire.

Le terme interactif n'est pas toujours négatif

Dans l'exemple, $\hat{\beta}_3$ est négatif: il y a un désavantage à avoir un mode de scrutin majoritaire.

Si $\hat{\beta}_3$ avait été positif: il y aurait eu un avantage à avoir un mode de scrutin majoritaire.

Si $\hat{\beta}_3$ avait été $= 0$: il n'y aurait aucun avantage ou désavantage à avoir un mode de scrutin majoritaire

Exercices

$$\text{taux fécondité} = 2,2 - 0,02\text{pib par hab} + 0,03\text{politique} + 0,01\text{pib par hab} * \text{politique} + \varepsilon$$

taux fécondité: taux de fécondité dans un pays

pib par hab: PIB par hab. en milliers de dollars

politique: avoir une politique des naissances (=1, sinon 0)

Exercices

$$\text{taux fécondité} = 2,2 - 0,02\text{pib par hab} + 0,03\text{politique} + 0,01\text{pib par hab} * \text{politique} + \varepsilon$$

Comment interpréter le terme interactif?

Quelle est la valeur prédite du taux de fécondité pour un pays qui a PIB par hab. de 35 mille dollars et une politique de naissances?

Exercices

$$\text{taux fécondité} = 2,2 - 0,02\text{pib par hab} + 0,03\text{politique} + 0,01\text{pib par hab} * \text{politique} + \varepsilon$$

Quelle est la valeur prédite du taux de fécondité pour un pays qui a PID par hab de 35 mille dollars mais aucun incitatif de naissance?

Comparer les 2 valeurs obtenues.

Exercices

$$\text{taux fécondité} = 2,2 - 0,02\text{pib par hab} + 0,03\text{politique} + 0,01\text{pib par hab} * \text{politique} + \varepsilon$$

Quel est l'effet de la variable "PIB par hab." sur le taux de fécondité?

Quel est est l'effet de la variable "politique" sur le taux de fécondité?

Questions?

Questions?

C'est la pause!

Variable dépendante (Y) binaire

Variable dépendante (Y) binaire

Jusqu'ici, on s'en est tenu à des VD continues.

Variable dépendante (Y) binaire

Jusqu'ici, on s'en est tenu à des VD continues.

Les phénomènes à expliquer peuvent parfois prendre 2 valeurs seulement.

Variable dépendante (Y) binaire

Jusqu'ici, on s'en est tenu à des VD continues.

Les phénomènes à expliquer peuvent parfois prendre 2 valeurs seulement.

Exemple: avoir voté ou pas; avoir subi une crise économique ou pas; avoir subi une guerre civile ou pas, etc.

Variable dépendante (Y) binaire

Variable dépendante (Y) binaire

Avec une VD binaire, on appelle le modèle linéaire un “modèle de probabilité linéaire”.

Variable dépendante (Y) binaire

Avec une VD binaire, on appelle le modèle linéaire un “modèle de probabilité linéaire”.

Y a une probabilité d'être égal à 1 et une probabilité d'être égal à 0.

Variable dépendante (Y) binaire

Interprétation des coefficients: une augmentation d'une unité dans X est associée à une augmentation de $100 * \beta$ points de pourcentage dans la probabilité que Y soit égale à 1.

Variable dépendante (Y) binaire

Variable dépendante (Y) binaire

Exemple: VD = avoir été en récession en 2008 (= 1), ou pas (= 0).

Variable dépendante (Y) binaire

Exemple: VD = avoir été en récession en 2008 (= 1), ou pas (= 0).

Si $Y = 0,5 + 0,04X_1 + \varepsilon$

Variable dépendante (Y) binaire

Exemple: VD = avoir été en récession en 2008 (= 1), ou pas (= 0).

Si $Y = 0,5 + 0,04X_1 + \varepsilon$

X_1 est le nombre de maisons vendues entre 2004 et 2008, en milliers.

Variable dépendante (Y) binaire

Exemple: VD = avoir été en récession en 2008 (= 1), ou pas (= 0).

Si $Y = 0,5 + 0,04X_1 + \varepsilon$

X_1 est le nombre de maisons vendues entre 2004 et 2008, en milliers.

Pour chaque millier de maisons vendues de plus entre 2004 et 2008, la probabilité que le pays ait été en récession augmente de 4 points de pourcentage.

Exercice

VD: Part. électorale	
Constante	0,7
Milléniaux	-0,08

VD = être allé.e voter aux dernières élections (= 1), ou pas (= 0).

Milléniaux = le fait d'appartenir à la génération des milléniaux (= 1, sinon 0).

Exercice

VD: Part. électorale	
Constante	0,7
Milléniaux	-0,08

VD = être allé.e voter aux dernières élections (= 1), ou pas (= 0).

Milléniaux = le fait d'appartenir à la génération des milléniaux (= 1, sinon 0).

Interprétez le coefficient de la variable "Milléniaux".

Donnée aberrante et/ou influente

Donnée aberrante et/ou influente

Aberrante: son résidu est élevé; le modèle prédit mal la VD pour cette donnée.

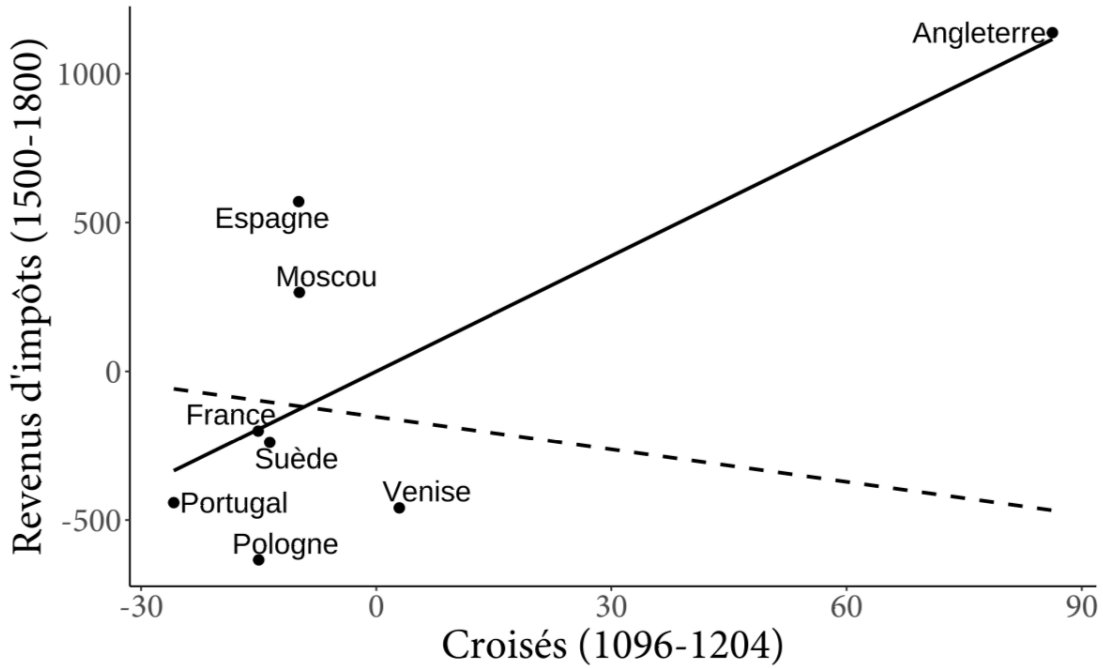
Donnée aberrante et/ou influente

Aberrante: son résidu est élevé; le modèle prédit mal la VD pour cette donnée.

Influente: exclure cette observation a un effet important sur le coefficient de régression.

Donnée aberrante et/ou influente

FIGURE 5.11. –
Relation entre le nombre de croisés issus d'une région et les revenus d'impôts du gouvernement de cette région. La ligne pleine reproduit la droite de régression de Blaydes et Paik (2016). La ligne pointillée représente un modèle de régression alternatif estimé en excluant l'Angleterre et le Portugal.



Comment les éviter?

Comment les éviter?

Visualiser vos données (je peux vous montrer comment la semaine prochaine).

Comment les éviter?

Visualiser vos données (je peux vous montrer comment la semaine prochaine).

Certains tests statistiques permettent de les identifier, pour décider ou non de les enlever.

Qualité de l'ajustement

Qualité de l'ajustement

Le modèle prédit-il bien les valeurs de Y?
Les erreurs sont-elles grandes?

Qualité de l'ajustement: R^2

Qualité de l'ajustement: R^2

Mesure la proportion de la variance de Y qui est “expliquée” par le modèle.

Qualité de l'ajustement: R^2

Mesure la proportion de la variance de Y qui est “expliquée” par le modèle.

R^2 élevé: le modèle est mieux ajusté aux données de l’échantillon.

Les limites du R^2

Les limites du R^2

Ça ne nous indique pas si le modèle est biaisé (rappelez-vous les crèmes glacées et les noyades).

Les limites du R^2

Ça ne nous indique pas si le modèle est biaisé (rappelez-vous les crèmes glacées et les noyades).

Il peut être augmenté simplement en ajoutant des variables dans le modèle.

R^2 (ajusté): exemple

Déterminants du vote par circonscription, 2007-2012

	PLQ	PQ	ADQ-CAQ
Anglophone (%)	0,45 (0,04)***	-0,44 (0,05)***	-0,08 (0,05)***
Minorité (%)	0,30 (0,05)***	-0,25 (0,06)***	-0,17 (0,07)***
Montréal-Laval	-1,98 (1,37)	-2,09 (1,54)	-2,05 (1,65)
Couronne	-2,48 (1,07)***	-2,15 (1,20)***	5,09 (1,29)***
Saguenay	3,03 (1,65)***	0,30 (1,86)	-2,34 (2,00)
Outaouais	2,63 (1,76)	-5,65 (1,98)***	-1,70 (0,43)
Sud-Est	3,91 (1,29)***	-16,94 (1,45)***	12,60 (1,56)***
Mauricie	4,67 (1,69)***	-6,95 (1,90)***	2,05 (2,04)
Estrie	-1,56 (1,96)	-4,59 (2,21)***	4,02 (2,37)*
Région de Québec	2,67 (1,25)***	-13,90 (1,41)***	11,05 (1,51)***
Sortant PLQ	12,61 (1,16)***	2,81 (1,30)***	-12,00 (1,39)***
Sortant PQ	1,32 (1,17)	10,49 (1,31)***	-11,99 (1,41)***
Candidats effectifs (t-1)	-10,06 (1,00)***	1,83 (1,12)*	2,79 (1,20)**
Nombre de candidats (t)	-0,51 (0,30)***	0,12 (0,34)	-0,30 (1,14)
2008	16,35 (0,95)***	6,83 (1,06)***	-18,82 (1,14)***
2012	-1,89 (0,94)***	4,79 (1,06)***	-2,99 (1,14)***
Constante	50,47 (2,97)***	26,43 (3,34)***	34,15 (3,58)***
R^2 ajusté	0,86	0,73	0,71
N	362	362	362

Figure: Godbout, Jean-François. "Les élections au Québec de 1973 à 2012," dans Les Québécois aux urnes (2013), p.37

La régression linéaire est flexible!

La régression linéaire est flexible!

Elle peut être courbée pour traduire des relations quadratiques.

La régression linéaire est flexible!

Elle peut être courbée pour traduire des relations quadratiques.

Par exemple: effet de l'âge sur la propension à aller voter; effet du nombre de secondes depuis la frappe sur la position d'une balle de baseball; effet de l'idéologie (gauche-droite) sur la force des convictions politique; effet de l'ambivalence sur la participation.

La régression linéaire est flexible!

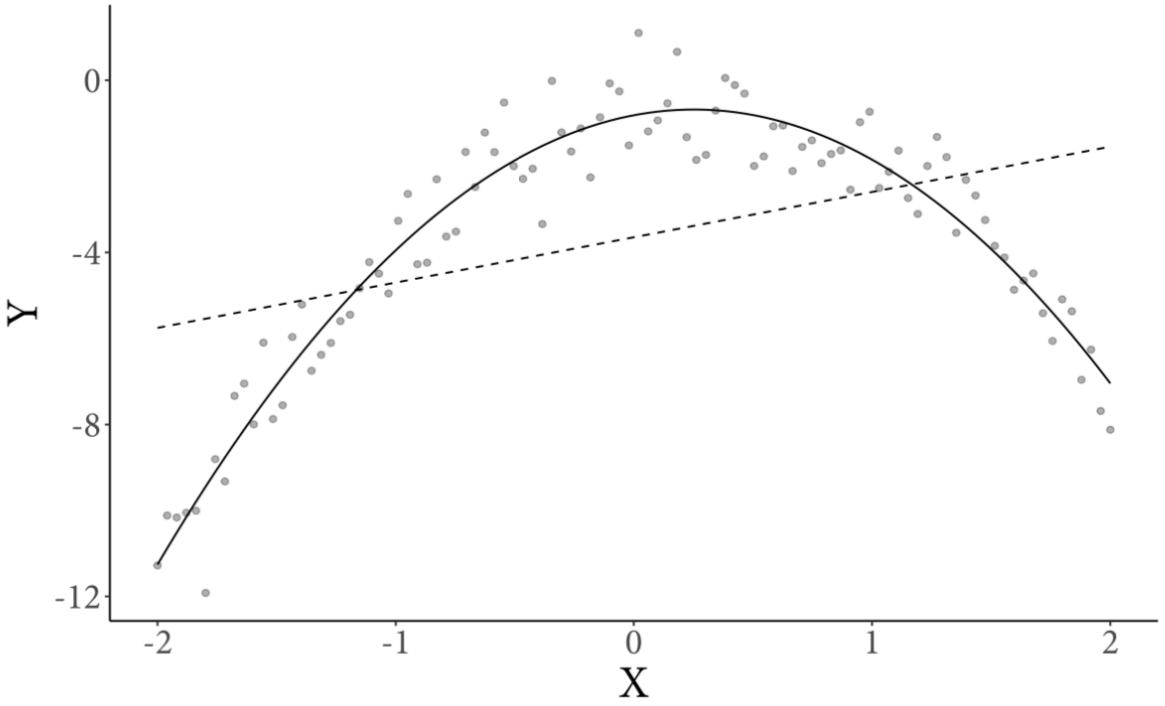
Elle peut être courbée pour traduire des relations quadratiques.

Par exemple: effet de l'âge sur la propension à aller voter; effet du nombre de secondes depuis la frappe sur la position d'une balle de baseball; effet de l'idéologie (gauche-droite) sur la force des convictions politique; effet de l'ambivalence sur la participation.

Dans ces circonstances, un terme X^2 est rajouté à la formule.

La régression linéaire est flexible!

FIGURE 5.12. –
Relation quadratique entre X et Y. La ligne pointillée correspond aux prédiction d'un modèle purement linéaire. La ligne pleine correspond aux prédictions du modèle 5.16



La régression linéaire est flexible!

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

La régression linéaire est flexible!

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Si β_2 est positif, la courbe descend, puis monte.

La régression linéaire est flexible!

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Si β_2 est positif, la courbe descend, puis monte.

Lorsque X est faible, une augmentation de X est associée à une diminution de Y.

La régression linéaire est flexible!

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Si β_2 est positif, la courbe descend, puis monte.

Lorsque X est faible, une augmentation de X est associée à une diminution de Y.

Lorsque X est élevé, une augmentation de X est associée à une augmentation de Y.

La régression linéaire est flexible!

$$Y = \alpha + \beta_1 X - \beta_2 X^2 + \varepsilon$$

La régression linéaire est flexible!

$$Y = \alpha + \beta_1 X - \beta_2 X^2 + \varepsilon$$

Si β_2 est négatif, la courbe monte, puis redescend.

La régression linéaire est flexible!

$$Y = \alpha + \beta_1 X - \beta_2 X^2 + \varepsilon$$

Si β_2 est négatif, la courbe monte, puis redescend.

Lorsque X est faible, une augmentation de X est associée à une augmentation de Y.

La régression linéaire est flexible!

$$Y = \alpha + \beta_1 X - \beta_2 X^2 + \varepsilon$$

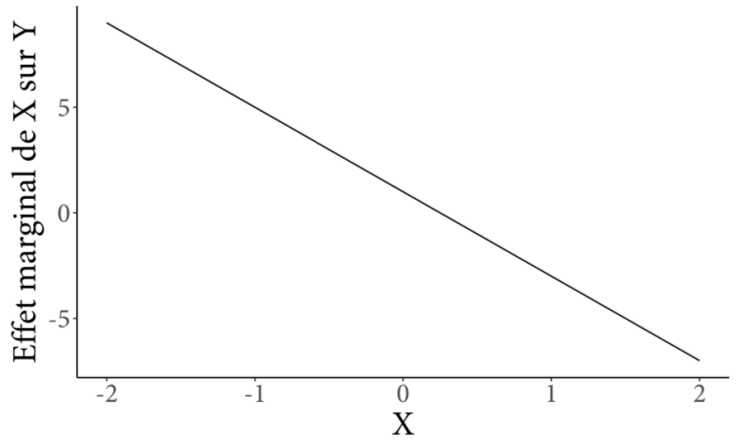
Si β_2 est négatif, la courbe monte, puis redescend.

Lorsque X est faible, une augmentation de X est associée à une augmentation de Y.

Lorsque X est élevé, une augmentation de X est associée à une diminution de Y.

La régression linéaire est flexible!

FIGURE 5.13. –
Effet marginal dans un modèle de régression avec variable explicative quadratique (équation 5.16).



L'effet de X sur Y est plus fort pour les faibles valeurs de X. L'effet de X sur Y diminue au fur et à mesure que X augmente (il peut éventuellement devenir négatif).

Questions?

Attention!

C'est la théorie qui nous indique quelles variables intégrer au modèle.

Régression multiple $\neq$ Modèle non-biaisé

Étoiles $\neq$ Modèle non-biaisé

R^2 élevé $\neq$ Modèle non-biaisé

Prochain cours

Avoir lu les notes de cours sur R.

Apporter votre ordinateur (si possible).

Pour le reste du cours

Réalisation du Devoir 2 en classe.


À la semaine
prochaine!