Méthodes quantitatives d'analyse (POL 2809)

Séance 5, 2 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.

Devoir 2 en ligne.

Aujourd'hui

Retour sur le Devoir 1.

Introduction à la régression multiple.

Effets d'interaction.

Retour sur les séances 3 et 4

Régression linéaire bivariée.

$$Y = \hat{\alpha} + \hat{\beta} * X + \varepsilon$$

Interprétation du coefficient $\hat{\beta}$.

Incertitude du coefficient $\hat{\beta}$.

Plus d'infos aux pages 63 à 80 du manuel.

La régression linéaire (OLS) multiple

La régression linéaire (OLS) multiple

Terme ε de la régression: toutes les variables qui ne sont pas prises en compte par le modèle.

La régression linéaire (OLS) multiple

Terme ε de la régression: toutes les variables qui ne sont pas prises en compte par le modèle.

Condition importante pour que $\hat{\beta}$ ne soit pas biaisé: X doit être indépendant de toutes les variables ignorées par le modèle.

La régression linéaire multiple permet...

La régression linéaire multiple permet...

... d'ajouter des variables importantes!

La régression linéaire multiple permet...

... d'ajouter des variables importantes!

Pour mieux expliquer la variation dans Y.

RL multiple

RL multiple

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

RL multiple

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

Ou, pour "k" variables indépendantes:

RL multiple

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

Ou, pour "k" variables indépendantes:

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

RL multiple

$$\boldsymbol{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_{\boldsymbol{k}} X_{\boldsymbol{k}} + \varepsilon$$

Une variable dépendante (pour l'instant, on s'en tient aux VD continues).

RL multiple

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

Une variable dépendante.

Un intercept (ou constante).

RL multiple

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

Une variable dépendante.

Un intercept (ou constante).

Des coefficients de régression (ou pentes).

RL multiple

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

Une variable dépendante.

Un intercept (ou constante).

Des coefficients de régression (ou pentes).

Des variables indépendantes.

RL multiple

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

Une variable dépendante.

Un intercept (ou constante).

Des coefficients de régression (ou pentes).

Des variables indépendantes.

Un terme résiduel (ou terme d'erreur).

Pourquoi?

Pourquoi?

En choisissant les bonnes variables à ajouter au modèle, on peut améliorer nos prédictions de Y ...

Pourquoi?

En choisissant les bonnes variables à ajouter au modèle, on peut améliorer nos prédictions de Y ...

... parce que les variables autrefois ignorées par le modèle ne le sont plus!

"Contrôler"

"Contrôler"

Il y a maintenant plus d'une variable qui permet d'expliquer (de prédire) Y.

"Contrôler"

Il y a maintenant plus d'une variable qui permet d'expliquer (de prédire) Y.

Mais X_1 explique aussi un peu X_2.

"Contrôler"

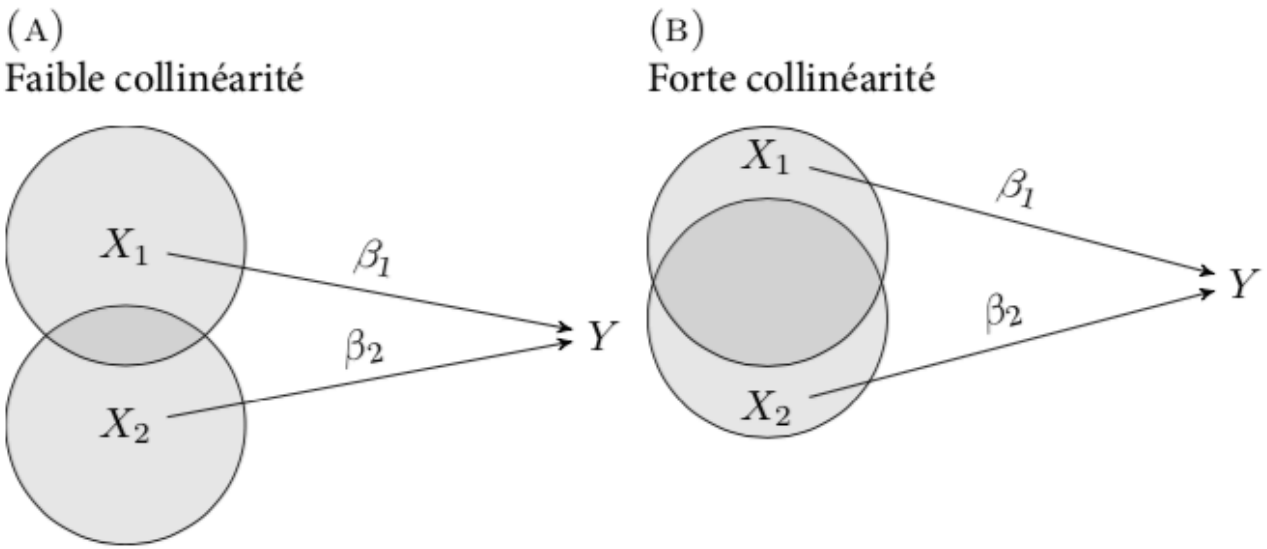
Il y a maintenant plus d'une variable qui permet d'expliquer (de prédire) Y.

Mais X_1 explique aussi un peu X_2.

Explications au tableau.

Mes explications au tableau, en résumé:

FIGURE 5.8. –
La régression multiple calcule les coefficients de régression en ignorant la variation commune des variables explicatives (l'intersection du diagramme de Venne). Lorsque les variables explicatives sont très corrélées (forte collinéarité), l'information 'indépendante' qui est disponible pour estimer les coefficients est pauvre, et l'incertitude est élevée. Les erreurs types reflèteront cette incertitude.



"Contrôler"

Le modèle utilise la partie de X_1 qui n'est pas expliquée par X_2 pour prédire Y.

"Contrôler"

Le modèle utilise la partie de X_1 qui n'est pas expliquée par X_2 pour prédire Y.

L'interprétation de $\hat{\beta}_1$ devient: "L'effet de X_1 sur Y après avoir contrôlé pour X_2".

"Contrôler"

Le modèle utilise la partie de X_1 qui n'est pas expliquée par X_2 pour prédire Y.

L'interprétation de $\hat{\beta}_1$ devient: "L'effet de X_1 sur Y après avoir contrôlé pour X_2".

Ou: "L'effet de X_1 sur Y en tenant X_2 constant".

Exercices

$$Y = 1500,2 + 2,05X_1 + 300X_2 + \varepsilon$$

Exercices

$$Y = 1500,2 + 2,05X_1 + 300X_2 + \varepsilon$$

Si $X_1 = 20$ et $X_2 = 2$, qu'elle est votre prédiction pour Y?

Si X_1 diminue de 3 et X_2 augmente de 10, quel changement cela entraîne-t-il dans Y?

Exercices

$$Y = -75 - 0,05X_1 + 20X_2 + 0,04X_3 + \varepsilon$$

Exercices

$$Y = -75 - 0,05X_1 + 20X_2 + 0,04X_3 + \varepsilon$$

Si $X_1 = 3, X_2 = 2$ et $X_3 = 1$, qu'elle est votre prédiction pour Y?

Si X_1 augmente de 5, X_2 diminue de 5 et X_3 augmente de 2, quel changement cela entraîne-t-il dans Y?

Questions?

Questions?

C'est la pause!

Les interactions!

Les interactions!

Mise en contexte: Je m'intéresse à l'effet du genre et du statut d'immigration sur le revenu.

Les interactions!

Mise en contexte: Je m'intéresse à l'effet du genre et du statut d'immigration sur le revenu.

Je développe le modèle suivant:

Les interactions!

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

Les interactions!

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

Y = revenu, en dollars

X_1 = le fait d'être un homme (codé 1, femmes sont codées 0)

$\hat{\beta}_1$ = effet du fait d'être un homme sur le revenu

X_2 = ne pas être immigrant (codé 1, les immigrants sont codés 0)

$\hat{\beta}_2$ = effet de ne pas être immigrant sur le revenu

ε = terme d'erreur

Les interactions!

$\hat{\beta}_1$ et $\hat{\beta}_2$ sont tous deux positifs.

Les interactions!

$\hat{\beta}_1$ et $\hat{\beta}_2$ sont tous deux positifs.

Les hommes ont un meilleur salaire, après avoir contrôlé pour le statut d'immigration.

Les interactions!

$\hat{\beta}_1$ et $\hat{\beta}_2$ sont tous deux positifs.

Les hommes ont un meilleur salaire, après avoir contrôlé pour le statut d'immigration.

Les non-immigrants ont un meilleur salaire, après avoir contrôlé pour le genre.

Les interactions!

Et si les hommes non-immigrants avaient un avantage supplémentaire?

Les interactions!

Et si les hommes non-immigrants avaient un avantage supplémentaire?

Et si avoir ces 2 caractéristiques en même temps procurait un avantage supplémentaire?

Les interactions!

Les interactions permettent de prendre en compte ce genre de "bonus".

Les interactions!

Les interactions permettent de prendre en compte ce genre de "bonus".

Comment?

Les interactions!

Les interactions permettent de prendre en compte ce genre de "bonus".

Comment?

En ajoutant un terme qui multiplie X_1 et X_2!

Les interactions!

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

$\hat{\beta}_3$ est le "bonus" pour ceux qui ont la valeur "1" aux variables X_1 et X_2, soit les hommes non-immigrants!

Exemple chiffré

Disons:

$$\text{revenu} = 30000 + 1000\text{homme} + 1500\text{non-imm} + 200\text{homme} * \text{non-imm} + \varepsilon$$

Explications au tableau.

Exemple chiffré

Pour les hommes non-immigrants, il y a un "bonus" de 200 dollars.

Exemple chiffré

Pour les hommes non-immigrants, il y a un "bonus" de 200 dollars.

Pour tous les autres, le terme interactif s'annule.

En bref, pour 2 variables X dichotomiques:

Le coefficient de l'interaction est un bonus pour les cas où les deux variables sont codées “1” (ici, les hommes non-immigrants).

Le terme interactif n'est pas toujours positif

Dans l'exemple, $\hat{\beta}_3$ est positif: il y a un avantage à être à la fois homme et non-immigrant.

Le terme interactif n'est pas toujours positif

Dans l'exemple, $\hat{\beta}_3$ est positif: il y a un avantage à être à la fois homme et non-immigrant.

Si $\hat{\beta}_3$ avait été négatif: il y aurait eu un désavantage à être à la fois homme et non-immigrant.

Le terme interactif n'est pas toujours positif

Dans l'exemple, $\hat{\beta}_3$ est positif: il y a un avantage à être à la fois homme et non-immigrant.

Si $\hat{\beta}_3$ avait été négatif: il y aurait eu un désavantage à être à la fois homme et non-immigrant.

Si $\hat{\beta}_3$ avait été $= 0$: il n'y aurait eu aucun avantage ou désavantage à être à la fois homme et non-immigrant.

Exercices

$$\%\ \text{parti sortant} = 40 + 2\text{croissance} + 3\text{guerre} + 0,4\text{croissance} * \text{guerre} + \varepsilon$$

% parti sortant: Pourcentage de votes pour le parti sortant

croissance: le fait d'être en période de croissance économique (=1, sinon 0)

guerre: le fait d'être en guerre (=1, sinon 0)

Exercices

Quel est le pourcentage prédit de votes pour le parti sortant dans un pays en croissance ET en guerre?

Quel est le pourcentage prédit de votes pour le parti sortant dans un pays en croissance mais en période de paix?

Exercices

$$\text{taux de fecon} = 2,2 - 0,03\text{indus} + 0,4\text{politique} + 0,01\text{indus} * \text{politique} + \varepsilon$$

taux de fecon: Taux de fécondité (nb d'enfants par femme)

indus: le fait d'être un pays industrialisé (=1, sinon 0)

politique: le fait d'avoir une politique d'incitation aux naissances (=1, sinon 0)

Exercices

Comment interprète-t-on le terme interactif?

Quelle est la valeur prédite du taux de fécondité pour les pays qui ne sont pas industrialisés et qui n'ont pas de politique des naissances?

Quelle est la valeur prédite du taux de fécondité pour les pays qui ne sont pas industrialisés mais qui ont une politique des naissances?

Questions?

Questions?

C'est la pause!

Et une interaction de variables continues?

Et une interaction de variables continues?

Mise en contexte: Je m'intéresse à l'effet de l'âge (en années) et du niveau d'éducation (en années) sur le revenu.

Et une interaction de variables continues?

Mise en contexte: Je m'intéresse à l'effet de l'âge (en années) et du niveau d'éducation (en années) sur le revenu.

Je développe le modèle suivant:

Interaction avec 2 variables X continues

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

Interaction avec 2 variables X continues

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

Y = revenu, en dollars

X_1 = l'âge, en années

$\hat{\beta}_1$ = effet de l'âge sur le revenu

X_2 = le niveau d'éducation, en années

$\hat{\beta}_2$ = effet du niveau d'éducation sur le revenu

ε = terme d'erreur

Interaction avec 2 variables X continues

$\hat{\beta}_1$ et $\hat{\beta}_2$ sont tous deux positifs.

Les personnes plus âgées ont un meilleur salaire, après avoir contrôlé pour le niveau d'éducation.

Interaction avec 2 variables X continues

$\hat{\beta}_1$ et $\hat{\beta}_2$ sont tous deux positifs.

Les personnes plus âgées ont un meilleur salaire, après avoir contrôlé pour le niveau d'éducation.

Les personnes qui ont passé plus de temps à l'école ont un meilleur salaire, après avoir contrôlé pour l'âge.

Interaction avec 2 variables X continues

Et si les personnes qui sont plus âgées et qui sont restées plus longtemps à l'école avaient un avantage supplémentaire?

Interaction avec 2 variables X continues

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

$\hat{\beta}_3$ est le "bonus" pour les valeurs élevées de l'âge et/ou du niveau d'éducation.

Exemple chiffré

Si:

$$\text{revenu} = 30000 + 750\text{âge} + 1000\text{educ} + 20\text{âge} * \text{educ} + \varepsilon$$

Explications au tableau.

Quand on interprète l'effet de l'âge, on ne peut plus ignorer l'éducation, et vice-versa

Quand on interprète l'effet de l'âge, on ne peut plus ignorer l'éducation, et vice-versa

L’effet de l’âge sur le revenu est plus grand pour les niveaux d'éducation plus élevés.

Quand on interprète l'effet de l'âge, on ne peut plus ignorer l'éducation, et vice-versa

L’effet de l’âge sur le revenu est plus grand pour les niveaux d'éducation plus élevés.

L’effet du niveau d'éducation sur le revenu est plus grand pour les personnes plus âgées.

Si vous aimez l'algèbre

Si vous aimez l'algèbre

L'effet de l'âge est maintenant égal à $\beta_1 + \beta_3$educ

L'effet de l'âge sur le revenu est plus grand pour les niveaux d'éducation plus élevés.

Si vous aimez l'algèbre

L'effet de l'âge est maintenant égal à $\beta_1 + \beta_3$educ

L'effet de l'âge sur le revenu est plus grand pour les niveaux d'éducation plus élevés.

L'effet de l'éducation est maintenant égal à $\beta_2 + \beta_3$âge

L'effet du niveau d'éducation sur le revenu est plus grand pour les personnes plus âgées.

En bref, pour 2 variables X continues:

Le coefficient de l'interaction est un bonus pour les valeurs élevées de X_1 et de X_2.

Comme tantôt: le terme interactif n'est pas toujours positif

Dans l'exemple, $\hat{\beta}_3$ est positif: il y a un avantage à être plus âgé et plus éduqué.

Si $\hat{\beta}_3$ avait été négatif: il y aurait eu un désavantage à être plus âgé et plus éduqué.

Si $\hat{\beta}_3$ avait été $= 0$: il n'y aurait eu aucun avantage ou désavantage à être plus âgé et plus éduqué.

Exercices

$$\%\ \text{parti sortant} = 40 + 0,02\text{pourc. croissance} + 2\text{pop PM} + 0,3\text{pourc. croissance} * \text{pop PM} + \varepsilon$$

% parti sortant: Pourcentage de votes pour le parti sortant

pourc. croissance: Croissance du PIB (en pourcentage)

pop PM: Popularité du Premier ministre, de 0 à 100

Exercices

Quel est le pourcentage prédit de votes pour le parti sortant dans un pays qui a eu une croissance de 2% et dont le PM a un niveau de popularité de 5?

Quel est le pourcentage prédit de votes pour le parti sortant dans un pays qui a eu une croissance de -1% et dont le PM a un niveau de popularité de 3?

Interprétez l'effet de la popularité sur le vote.

Exercices

$$\text{taux de fecon} = 2,2 - 0,002\text{PIB par hab} + 0,0003\text{incitatif} + 0,001\text{PIB par hab} * \text{incitatif} + \varepsilon$$

taux de fecon: Taux de fécondité (nb d'enfants par femme)

PIB par hab: PIB par habitant, en milliers de dollars

incitatif: valeur, en dollars, de l'incitatif financier à avoir des enfants

Exercices

Comment interpréter le terme interactif?

Quelle est la valeur prédite du taux de fécondité pour un pays où le PIB par habitant est de 35 mille dollars et où l'incitatif financier à avoir un enfant est de 100 dollars?

Quelle est la valeur prédite du taux de fécondité pour un pays où le PIB par habitant est de 40 mille dollars et où l'incitatif financier à avoir un enfant est de 100 dollars?

Exercices

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

Quel est est l’effet de la variable "incitatif" sur le taux de fécondité?

Questions?

Attention!

Régression multiple $\neq$ Causalité

Prochain cours

Lire le document explicatif pour le travail final (sur Studium)

Suite (et fin) de la régression multiple


À la semaine
prochaine!