Questions pratiques 5:
L’échantillon et ses problèmes

Jean-François Bickel

Statistique II – SPO8

- Deux caractéristiques de l’échantillon de données peuvent avoir de sérieuses conséquences sur la validité des résultats de la régression, leur interprétation et leur généralisation au delà de l’échantillon concerné

1. Sa taille
2. Son caractère biaisé ou non

1. La taille de l’échantillon

- La taille de l’échantillon influe fortement sur les tests de significativité statistique
- Avec un échantillon de 60 personnes, une corrélation doit être d’au moins .25 pour être significative au seuil de .05
- Dans un échantillon de 10000 personnes, n’importe quelle corrélation supérieure à .02 est significative au même seuil

A. Un échantillon de petite taille implique une double difficulté

1) la faiblesse de la puissance statistique
2) la détérioration de l'approximation réalisée par les tests statistiques

1) La puissance statistique d'un test peut être définie comme sa capacité à induire le rejet de l'hypothèse nulle quand celle-ci est (dans la réalité) effectivement fausse

- Autrement dit, un test a une puissance statistique forte si, sur la base du test, on conclut que x a un effet sur y et que cette conclusion est effectivement valide

- La puissance statistique dépend de n (effectif de l'échantillon)
- Il en résulte que dans un petit échantillon, la puissance statistique est faible

- D’où

i. un coefficient de régression statistiquement significatif permet une conclusion ferme de l’effet de x sur y

ii. l’absence de significativité statistique pour le coefficient de la variable x n’est pas forcément suffisant pour conclure à l’absence d’effet de cette variable x sur y

2) La grande majorité des tests que les chercheurs utilisent (test de t, test du chi-carré, etc.) sont seulement des approximations

- Ces approximations fonctionnent généralement bien quand l'échantillon est grand (théorème de la limite centrale)
- Mais le degré d'approximation peut se détériorer de manière marquée quand l'échantillon est petit

- Pour éviter une telle détérioration, il faut que la distribution du terme d’erreur soit normale
- Nous reprendrons ce point au moment d’examiner les assomptions de la régression

- Quand un échantillon est-il « petit »?
- Comme critère approximatif, on peut considérer comme « petit » un échantillon de moins de 100 observations
- Mais cela doit être pondéré par le nombre de paramètres dans l'équation de régression; quand ce nombre augmente, le seuil de « petitesse » croît également

- Au delà de 1000 observations, l’échantillon est grand et évite tout problème de petitesse

B. Lorsqu'un échantillon est très grand (plusieurs milliers d'observations), la significativité statistique peut, au contraire, ne pas être un critère suffisant pour conclure que la relation entre x et y est significative du point de vue substantiel

- Autrement dit, un coefficient différent de 0 peut impliquer un écart très petit dans la réalité

- Si les unités de mesure des variables sont familières (le revenu, le sexe, etc.), le raisonnement du chercheur (et le sens commun) peuvent permettre d’évaluer si une relation est substantiellement différente de 0
- Si les unités de mesures des variables ne sont pas familières, les coefficients standardisés peuvent aider pour cette évaluation

- Un coefficient standardisé très petit, disons égal ou inférieur à .05, indique qu'une part très petite de la variation de y est conditionnée par la variation de x

2. Les biais d’échantillon

- Le processus de sélection de l’échantillon peut conduire à deux types de biais d’échantillon
- Chaque type étant susceptible de porter atteinte à une caractéristique essentielle de l’échantillon

a) sa validité interne

b) sa validité externe

Nota Bene

- Ce sont là bien sûr des questions qui concernent toutes les méthodes statistiques, pas seulement la régression!

a) La validité interne réfère à la justesse et fiabilité des résultats des analyses portant sur l'échantillon lui-même

- Le processus de sélection de l'échantillon peut nuire à la validité interne et avoir pour conséquence des résultats erronés ou trompeurs

- Si l’échantillon est aléatoire et que le taux de non-réponses est faible, ce n’est généralement pas considéré comme un problème majeur
- Si l’échantillon a été sélectionné par convenance ou/et si une proportion importante de répondants potentiels refusent de répondre ou ne peuvent être localisés, l’enjeu devient beaucoup plus crucial

- Supposons une étude sur la relation entre éducation et revenu, et que les personnes qui se caractérisent par un niveau d'éducation élevé et un bas revenu ou à l'inverse un bas niveau d'éducation et un revenu élevé ont répondu beaucoup moins fréquemment au questionnaire
- Ceci peut conduire à faire apparaître une relation forte entre éducation et revenu alors même qu'en réalité cette relation est nulle ou faible

- Généralement, on ne s’attend pas à ce qu’une telle structure différentielle des réponses se produisent
- Mais, cela ne signifie pas qu’elle n’apparaisse jamais!

La validité externe de l’échantillon

- La validité externe se réfère au degré selon lequel les résultats des analyses menées sur l’échantillon peuvent être généralisés à d’autres groupes ou à la population de référence
- C’est donc la question de la représentativité de l’échantillon qui est ici en jeu

1) Un premier cas de figure est celui de l’échantillon aléatoire (probabiliste)

- Un échantillon aléatoire est un échantillon pour lequel la probabilité de sélectionner n’importe quel échantillon de taille n est connue ou peut être calculée

- Il existe trois types principaux d’échantillon aléatoire

i. simple (ou ordinaire)
ii. stratifié
iii. par cluster

- Généralement, il est assumé que dans le cas où l'échantillon est aléatoire (probabiliste), les résultats ont une validité externe et peuvent être généralisés à la population dont est issu l'échantillon
- Il existe un débat pour savoir si dans le cas d'échantillons stratifiés ou par cluster, des ajustements sont nécessaires pour que les résultats puissent être généralisés à la population

- Même pour un échantillon aléatoire, reste posée la question de la généralisation des résultats au-delà de la population dont est issu l'échantillon
- A ce titre, ce type d'échantillon est redevable du même examen critique que ceux du second cas de figure ci-après

2) Un second cas de figure est représenté par les échantillons constitués non aléatoirement

- échantillons de convenance, échantillons recrutés par voie de presse, etc.

- Il faut soigneusement analyser si les relations causales établies sur la base de l'échantillon ont de bonnes chances d'exister dans d'autres groupes auxquels on peut être intéressé
  - Par exemple, est-ce que les individus de l'échantillon se distinguent selon l'une ou l'autre caractéristiques importantes des individus d'autres groupes?