5 Régression Linéaire Multiple: Estimation II

Exercice 1
(Y_i, X_{1i}, X_{2i}) satisfont les hypothèses du modèle de régression multiple RLM.1-RLM.4. On s'intéresse à β_1, l'effet causal de X_1 sur Y. Supposons que X_1 et X_2 ne sont pas corrélés. On estime β_1 dans la régression de Y sur X_1 (et donc X_2 n'est pas inclus dans la régression). Est-ce que cet estimateur souffre d'un biais de variables omise? Expliquez votre réponse.

Exercice 2
(Y_i, X_{1i}, X_{2i}) satisfont les hypothèses du modèle de régression multiple RLM.1-RLM.4. De plus, $Var(U_i|X_{1i}, X_{2i}) = 4$ et $Var(X_{1i}) = 6$. On dispose d'un échantillon aléatoire de la population de taille $n = 400$.

(a) Supposez que X_1 et X_2 ne sont pas corrélées. Calculez la variance de $\hat{\beta}_1$.

(b) Supposez que $Corr(X_1, X_2) = 0,5$. Calculez la variance de $\hat{\beta}_1$.

(c) Commentez les affirmations suivantes: "Si X_1 et X_2 sont corrélées, la variance de $\hat{\beta}_1$ est supérieure à ce qu'elle serait si X_1 et X_2 ne l'étaient pas. Par conséquent, si l'on s'intéresse à β_1, il vaut mieux enlever la variable X_2 de la régression si elle est corrélée avec X_1.

Exercice 3
Considérez le modèle de régression suivant:

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + U_i,$$

pour $i = 1, \ldots, n$ (Prenez bien en compte le fait qu'il n'y a PAS DE CONSTANTE dans ce modèle).

(a) Spécifiez la fonction de moindres carrés que l'on minimise par MCO.

(b) Calculez les dérivées partielles de la fonction objectif par rapport à b_1 et b_2.

(c) Supposez que $\sum_{i=1}^n X_{1i}X_{2i} = 0$. Montrez que $\hat{\beta}_1 = \frac{\sum_{i=1}^n X_{1i}Y_i}{\sum_{i=1}^n X_{1i}^2}$

(d) Supposez que $\sum_{i=1}^n X_{1i}X_{2i} \neq 0$. Obtenez une expression pour $\hat{\beta}_1$ en fonction des données (Y_i, X_{1i}, X_{2i}), $i = 1, \ldots, n$.

(e) Supposez que le modèle inclut une constante: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_i$. Montrez que les estimateurs MCO satisfont $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$.

(f) Supposez que le modèle inclut une constante comme pour le point (e). Supposez en outre que:

$$\sum_{i=1}^n (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2) = 0.$$

Montrez que

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$

Comparez cet estimateur avec l'estimateur MCO de β_1 de la régression dans laquelle on a omis X_2.

Exercice 4
Avec la base de données TeachingRatings, utilisée dans la série 2, répondez aux question suivantes:

(a) Effectuez une régression de la variable Course_Eval (rappelez-vous, il s'agit de notes d'examen) sur la variable Beauty (la variable qui mesure la beauté de l'enseignant). Quelle est la pente estimée?

(b) Effectuez une régression de la variable Course_Eval sur la variable Beauty, en incluant des variables de contrôle supplémentaires sur le type de cours et des caractéristiques des enseignants. En particulier, utilisez comme variables explicatives supplémentaires: OneCredit, Female, Minority et English. Quel est l'effet estimé de la variable Beauty sur la variable Course_Eval? Pensez-vous que la régression (a) comportait des biais de variables omises importants?

(c) Estimez le coefficient de la variable Beauty dans le modèle de régression multiple du point (b) en suivant les trois étapes du théorème Frisch-Waugh:

(1) Régressez la variable dépendante Course_Eval sur les variables de contrôles supplémentaires et obtenez les résidus $\tilde{Y}$.

(2) Régressez la variable explicative Beauty sur les variables de contrôles supplémentaires et obtenez les résidus $\tilde{X}$.

(3) Régressez les résidus $\tilde{Y}$ sur les résidus $\tilde{X}$, et vérifiez que vous obtenez le même coefficient pour Beauty, que celui obtenu au point en (b).

(d) Le professeur Smith est un homme noir avec une valeur moyenne de la variable Beauty et il enseigne, dans sa langue natale, un cours de trois crédits. Faites un prédiction de l'évaluation de cours du professeur Smith.

Exercice 5
Avec la base de données CollegeDistance, utilisée dans la série 2, répondez aux question suivantes:

(a) Effectuez la régression de la variable années d'étude (ED) sur la variable distance à l'université la plus proche (Dist). Quel est le coefficient estimé?

(b) Effectuez la régression de la variable ED sur la variable Dist, mais en incluant des variables explicatives de contrôle supplémentaires sur les caractéristiques de l'élève, la famille de l'étudiant et le marché du travail local. Plus précisément, incluez les variables Bytest, Female, Black, Hispanic, Incomehi, Ownhome, DadColl, Cue80, et Stwmfg80. Quel est l'effet estimé de la variable Dist sur la variable ED?

(c) L'effet estimé de la variable Dist sur variable ED est-il très différent dans les régressions sues points (b) et (a)? En utilisant cette information, est-ce qu'il semble que la régression du point (a) souffrait de biais de variable omise importants?

(d) Comparez la qualité de l'ajustement (le fit) des régressions des points (a) et (b) en utilisant les écarts-types de l'estimation, les R^2 et les $\bar{R}^2$. Pourquoi le R^2 et le $\bar{R}^2$ sont-ils aussi semblables dans la régression du point (b)?

(e) Le coefficient de la variable DadColl est positif. Que mesure ce coefficient?

(f) Expliquez pourquoi les variables Cue80 et Swmfg80 apparaissent dans la régression. Que pensez-vous que devraient être les signes de leurs coefficients estimés. Interprétez la taille de ces coefficients.

(g) Bob est un homme noir. Son école était à 20 miles de l'université la plus proche. Son résultat de test (Bytest) était de 58. Son revenu familial en 1980 était de \$26,000 et sa famille possédait une maison. Sa mère est allée à l'université, mais pas son père. Le taux de chômage dans sa région était de 7,5% et le salaire horaire moyen industriel dans l'État était 9,74\$. Estimez le nombre d'années d'études de Bob en utilisant la régression (b).

(h) Jim a les mêmes caractéristiques que Bob, sauf que son école secondaire était à 40 miles de l'université la plus proche. Estimez le nombre d'années d'études de Bob en utilisant la régression (b).

SOLUTIONS:
2. a) 0,00167; b) 0,0022.
4. a) 0,133; b) 0,166 la coefficient ne change pas beaucoup et l'effet ne parait pas très grand; d) 3,901.