Problem Set #2

ECO 2273

Automne 2025

Date limite : 5-déc-2025. Vous pouvez remettre votre travail en personne en classe ou par courriel. Si vous remettez par courriel, envoyez-le à `sam.gyetvay@gmail.com` et `djamaldiev.oumar@courrier.uqam.ca` avec pour objet ECO2273 + TP1 + votre nom de famille + numéro étudiant.

Rappel : La collaboration est autorisée (et encouragée), mais chaque étudiant doit rédiger et remettre sa propre solution.

Instructions : écrivez lisiblement et gardez vos réponses aussi courtes que possible. Si vous scannez ou prenez une photo de votre solution, assurez-vous qu'elle soit lisible. Pour les questions impliquant l'analyse d'un jeu de données, vous pouvez utiliser le logiciel de votre choix, mais je recommande STATA, car certaines questions d'examen vous demanderont d'interpréter du code et des résultats STATA.

Vous pouvez trouver ce document en ligne à

`www.samgyetvay.com/teaching/eco2273/tp1.pdf`

Question 1: Cigarettes et poids à la naissance (20 points)

Une économiste de la santé a recueilli une base de données contenant des informations sur les naissances au Québec. Deux variables d'intérêt sont la variable dépendante, le poids du nouveau-né en onces ($bwght$), et une variable explicative, le nombre moyen de cigarettes que la mère a fumées par jour pendant la grossesse ($cigs$). La régression linéaire simple suivante a été estimée à partir de données sur $n = 1,388$ naissances:

$$\hat{bwght} = 119.77 - 0.514cigs$$

Question 1 (a) Quel est le poids prévu lorsque $cigs = 0$? Et lorsque $cigs = 20$? Commentez la différence.

Question 1 (b) Cette régression simple capture-t-elle nécessairement une relation causale entre le poids de naissance de l'enfant et les habitudes tabagiques de la mère? Expliquez pourquoi ou pourquoi pas.

Question 1 (c) Pour prédire un poids de 125 onces, quelle devrait être la valeur de $cigs$? Commentez.

Question 1 (d) La proportion de femmes dans l'échantillon qui ne fument pas pendant la grossesse est d'environ 0.85. Cela aide-t-il à concilier votre résultat de la partie (c)?

Question 2: Salaires des PDG (30 points)

Pour cette question, utilisez la base de données sur les salaires des PDG située à

`www.samgyetvay.com/teaching/eco2273/ceo.dta`
`www.samgyetvay.com/teaching/eco2273/ceo.csv`

la variable $salary$ correspond à la rémunération annuelle, en milliers de dollars, et $coeten$ est le nombre d'années précédentes en tant que PDG de l'entreprise.

Question 2 (a) Trouvez le salaire moyen et l'ancienneté moyenne dans l'échantillon.

Question 2 (b) Combien de PDG en sont à leur première année en tant que PDG (c'est-à-dire $ceoten = 0$)? Quelle est l'ancienneté maximale comme PDG?

Question 2 (c) Estimez le modèle de régression simple

$$\log(salary) = \beta_0 + \beta_1 ceoten + u$$

et rapportez vos résultats. Quelle est l'augmentation en pourcentage approximative du salaire associée à une année supplémentaire comme PDG? Quelle fraction de la variation du salaire est expliquée par l'ancienneté? Pouvez-vous rejeter l'hypothèse nulle selon laquelle il n'y a aucune relation entre l'ancienneté et le salaire?

Question 2 (d) Créez un nuage de points montrant la relation entre $\log(salary)$ et $ceoten$, contenant la droite d'ajustement estimée à la partie (c).

Question 2 (e) Parmi les autres variables dans la base de données (age, $college$, $grad$, $sales$, $profits$, $profmarg$), laquelle explique la plus grande fraction de la variation du salaire des PDG?

Question 3: Simulation (50 points)

Dans cette question, vous analyserez des données que vous simulez vous-même. Dans STATA, vous pouvez simuler des données en utilisant les fonctions $runiform()$ et $rnormal()$.

Question 3 (a) Commencez par générer 500 observations de x_i provenant d'une distribution uniforme sur l'intervalle [0,10]. Quelle est la moyenne échantillonnale et l'écart-type échantillonnal des x_i? Quel est l'intervalle de confiance à 95% pour la moyenne de x_i? Contient-il la vraie moyenne?

Question 3 (b) Générez aléatoirement 500 erreurs u_i provenant de la distribution $N(0, 36)$. La moyenne échantillonnale des u_i est-elle exactement égale à zéro? Pourquoi ou pourquoi pas? Quel est l'écart-type échantillonnal des u_i?

Question 3 (c) Maintenant, générez la variable y_i selon

$$y_i = 1 + 2x_i + u_i = \beta_0 + \beta_1 x_i + u_i$$

c'est-à-dire que l'ordonnée à l'origine dans la population est égale à un et la pente dans la population est égale à deux. Utilisez les données pour estimer la régression de y_i sur x_i. Quelles sont vos estimations de l'ordonnée à l'origine et de la pente? Sont-elles égales aux valeurs de population dans l'équation ci-dessus? Expliquez.

Question 3 (d) Obtenez les résidus MCO $\hat{u_i}$ et vérifiez que

$$\sum_{i=1}^{n} \hat{u}_i = 0, \quad \text{et} \quad \sum_{i=1}^{n} x_i \hat{u_i} = 0$$

Question 3 (e) Calculez les mêmes quantités qu'à la partie (d) mais utilisez les erreurs u_i à la place des résidus. Que concluez-vous maintenant?

Question 3 (f) Répétez les parties (a), (b) et (c) avec un nouvel échantillon de données, en recommençant par générer les x_i. Que trouvez-vous maintenant pour β_0 et β_1? Sont-ils différents de ce que vous avez obtenu à la partie (c)? Si oui, pourquoi? Sinon, pourquoi pas?