2 Régression Linéaire Simple I

Exercice 1
Supposez qu'un chercheur utilise des données sur la taille des classes (TC) et la note moyenne d'examen pour 100 classes de lycée et qu'il obtienne les résultats de régression suivants:

Note d'examen = 520,4 +5,82 x TC, $R^2 = 0,08$, SER = 1,5

(a) Une classe a 22 étudiants. Quelle est la prédiction de la note moyenne d'examen de cette classe d'après la régression?

(b) L'année dernière, une classe avait 19 étudiants, et cette année elle en a 23. Quel est la changement de la note moyenne prédit par la régression pour cette classe?

(c) La taille moyenne des classes dans l'échantillon de 100 classes est de 21,4. Quelle est la note moyenne dans l'échantillon des 100 classes? (Indice: Revoir les formules des estimateurs des MCO)

(d) Quel est l'écart-type des notes d'examen dans les 100 classes? (Indice: Revoir les formules du R^2 et de l'écart-type de la régression SER)

Exercice 2
Une régression des salaires hebdomadaires moyens (ISS, en dollars) sur l'âge moyen (en années) utilise un échantillon aléatoire de travailleurs, avec un diplôme universitaire, travaillant à temps complet et qui ont entre 25 et 65 ans. Les résultats sont les suivants:

$$\hat{ISS} = 696,7 + 9,6\text{x}Age,\ R^2 = 0,023,\ SER = 624,1$$

(a) Expliquez ce que signifient les valeurs des coefficients, 696,7 et 9,6.

(b) Quelles sont les unités de mesure de l'écart-type de la régression (SER)? des dollars? des années? ou est-ce qu'il n'y a pas d'unités?

(c) Quelles sont les unités de mesure du R^2 ?des dollars? des années? ou est-ce qu'il n'y a pas d'unités?

(d) Quels sont les écarts de salaire prédits par la régression pour un travailleur de 25 ans? Et pour un travailleur de 45 ans?

(e) Est-ce cette régression donnerait des prédictions fiables pour un travailleur de 99 ans? Pourquoi?

(f) Compte tenu de ce que vous savez sur la répartition des revenus, pensez-vous qu'il est possible que la distribution des erreurs de cette régression soit normale? (Indication: Réfléchissez à la symétrie de la distribution et à la valeur minimale des revenus, et ensuite voyez si ces caractéristiques sont compatibles avec une distributon normale)

(g) L'âge moyen dans cet échantillon est de 41,6 ans. Quelle est la moyenne d'ISS dans cet échantillon?

Exercice 3
Un enseignant décide de mener une expérience destinée à mesurer l'effet du temps disponible sur les résultats aux examens. Il donne le même examen à chacun des 400 élèves de sa classe, mais les uns ont 90 minutes tandis que les autres ont 120 minutes pour terminer l'examen. Chaque élève se voit attribuer aléatoirement un des deux temps disponibles pour l'examen par tirage au sort. Soit Y_i la note du i-ème élève ($0 \leq Y_i \leq 100$). Soit X_i le temps disponible pour terminer l'examen ($X_i = 90$ ou $X_i = 120$). Considérez le modèle de régression suivant:

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

(a) Expliquez ce que représente le terme U_i: Pourquoi des élèves différents auront-ils des valeurs différentes de U_i?

(b) Expliquez pourquoi $E(U_i|X_i) = 0$ dans ce modèle de régression.

(c) La régression estimée est $\hat{Y}_i = 49 + 0,24X_i$.

(i) Calculez la prédiction estimée de la note moyenne des étudiants qui ont eu 90 minutes pour terminer leur examen.

(ii) Calculez le gain estimé dans la note d'un élève qui a eu 10 minutes de plus pour terminer son examen.

Exercice 4 Sur la page web du livre de Stock et Watson (2012)
`http://wps.aw.com/aw_stock_ie_3/178/45691/11696965.cw/index.html`,
vous trouverez un fichier de données "CollegeDistance" qui contient des données provenant d'un échantillon aléatoire d'élèves de lycée interrogés en 1980 et réinterrogées en 1986. Dans cet exercice, vous allez utiliser ces données pour étudier la relation entre le nombre d'années de scolarité de jeunes adultes et la distance de l'école de chaque élève à l'université la plus proche. (La proximité de l'université réduit le coût de l'éducation, et donc les étudiants qui vivent près de l'université devraient, en moyenne faire plus d'années d'études supérieures). Le fichier "CollegeDistance_description", disponible sur la même page contient une description détaillée des variables.

(a) Effectuez une régression des moindres carrés ordinaires des années d'éducation (ED) en fonction de la distance à l'université la plus proche (Dist), où Dist est mesurée en dizaines de miles. (Par exemple Dist = 2 correspond à une distance de 20 miles). Quelle est l'estimation de la constante? Quelle est la pente estimée? Utilisez la régression estimée pour répondre à la question suivante: De combien le nombre moyen d'années de scolarité varie-t-il quand les universités sont construites près du lieu où les élèves vont à l'école secondaire?

(b) L'école de Bob est située à 20 miles de l'université la plus proche. Prédisez le nombre d'années d'études effectuées par Bob à l'aide de la régression estimée. Comment cette valeur changerait-elle si Bob vivait à 10 miles de l'université la plus proche?

(c) La distance à l'université explique-t-elle une grande proportion de la variance du niveau de scolarité entre les individus? Expliquez.

(d) Quelle est la valeur de l'erreur-type de l'estimation? Quelles sont les unités de l'écart-type de la régression SER (mètres, grammes, années, dollars, etc)?

Exercice 5

Sur la page web du livre de Stock et Watson (2012):
`http://wps.aw.com/aw_stock_ie_3/178/45691/11696965.cw/index.html`,
vous trouverez un fichier de données TeachingRatings qui contient des données d'évaluation de cours, des caractéristiques des cours, des caractéristiques des professeurs pour 463 cours de l'université du Texas à Austin. Les données sont décrites dans le fichier TeachingRating_Description, disponible sur la même page. Une des caractéristiques est un indice de beauté des enseignants (Beauty) tel qu'attribué aux professeurs par un panel de 10 juges. Dans cet exercice, nous allons étudier le lien entre les évaluations de cours et la beauté de l'enseignant.

(a) Faites un diagramme de dispersion des évaluations moyennes de cours (Cours_Eval) en fonction de la beauté de l'enseignant (Beauty). Est-ce qu'à première vue il semblerait qu'il existe une relation entre ces variables?

(b) Régressez les évaluations moyennes de cours (Cours_Eval) sur la beauté de l'enseignant (Beauty). Quelle est la constante estimée? Quelle est la pente estimée? Expliquez pourquoi l'estimation de la constante est égale à la moyenne de l'échantillon de la variable Course_Eval? (Indication: Quelle est la moyenne de l'échantillon de la variable Beauty?)

(c) L'enseignant A a une valeur moyenne pour sa variable Beauty, alors que l'enseignant B a une valeur de Beauty d'un écart-type au-dessus de la moyenne. Prédire les évaluations de cours des enseignants A et B.

(d) Que pensez-vous de la taille de la pente estimées? L'effet de Beauty sur Course_Eval est-il grand ou petit? Expliquez ce que vous entendez par "grand" et "petit".

(e) La beauté de l'enseignant explique-t-elle une grande proportion de la variance des évaluations de cours des enseignants? Expliquez.

SOLUTIONS:
1. a) 392,36,b) 23,28; c) 395,85; d) 11,5.
2. b) Dollars par semaine; c) Pas d'unités de mesure; d) i) $936,7, ii)$1.128,7; e) Non, l'âge maximal de l'échantillon est de 65 ans; f) Non, la distribution est asymétrique avec une kurtosis supérieure à celle de la normale; g) $1.096,06.
3. b) Etant donné que l'attribution du temps d'examen est aléatoire U_i et X_i sont indépendantes, donc $E(U_i|X_i) = E(U_i) = 0$; c) i) 70,06, ii) 2,4.
4. a) Ed = 13,96+0,073 x Dist. La régression prédit que si les universités sont construites 10 milles plus près des lycées, les années d'éducation augmentent de 0,073 années; b) 13,81; c) R2=0.007, donc la distance explique seulement un pourcentage très faible des années d'éducation; d) 1,8 années.
5. a) Il semble qu'il n'y ait pas de relation; b) Course_Eval = 4,00+0,133 x Beauty. La moyenne de l'échantillon de la variable Beauty est 0; le terme d'erreur est la moyenne

de la variable dépendante (Course_Eval) moins la pente estimée (0,133) multipliée par la moyenne du régresseur (Beauty). Donc l'estimation de la constante est égale à la moyenne de Course_Eval; c) A 4,00 et B 4,150; d) L'écart-type de Course_Eval est 0,55 et l'écart-type de Beauty est 0,789. Une augmentation d'un écart-type de Beauty devrait augmenter Course_Eval de 0,133 x 0,789 = 0,105, ou 1/5 de l'l'écart-type de Course_Eval. L'effet est faible; e) R2 =0.036, donc Beauty n'explique que 3.6% de la variance de Course_Eval.