Régression linéaire simple

Data Engineer

INPT

Veuillez combiner toutes vos réponses, le code `R` et les figures dans un seul fichier PDF.

Date d'échéance : 9 novembre 2021 à 23 h 59 (mardi soir).

Exercice 1

Afin d'étudier la faisabilité de lancer une édition dominicale pour un grand journal métropolitain, des informations ont été obtenues auprès d'un échantillon de 34 journaux concernant leurs tirages quotidiens et dominicaux (en milliers). Les données peuvent être lues sur la plateforme Moodle : [Data_Exercice1.txt].

1. Construisez un nuage de points de la diffusion dominicale par rapport à la diffusion quotidienne.
2. Le graphique suggère-t-il une relation linéaire entre la diffusion quotidienne et dominicale ?
3. Ajustez une droite de régression prédisant la diffusion du dimanche à partir de la diffusion quotidienne.
4. Existe-t-il une relation significative entre la diffusion dominicale et la diffusion quotidienne ? Justifiez votre réponse par un test statistique (Utilisez le test de Fisher).
5. Indiquez quelle hypothèse vous testez et votre conclusion pour le test de la partie (4).
6. À l'aide du tableau `anova`, calculez la proportion de la variabilité de la diffusion du dimanche qui est prise en compte par la diffusion quotidienne.

Exercice 2

Soient Y et X des variables dans une régression linéaire simple des prix médians des maisons par rapport au revenu médian des États (aux États-Unis). Supposons que le modèle

$$Y = \beta_0 + \beta_1 X + \epsilon$$

satisfait les hypothèses de régression habituelles.

Le tableau ci-dessous est un tableau similaire à la sortie de `anova` lorsqu'un modèle de régression linéaire simple est réalisé.

```
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1     NA    5291      NA        NA
Residuals 48 181289      NA
```

1. Calculez les valeurs manquantes dans le tableau ci-dessus.
2. Testez l'hypothèse nulle $H_0 : \beta_1 = 0$ au niveau $\alpha = 0,05$ en utilisant le tableau ci-dessus.
3. Est-il possible de tester l'hypothèse $H_0 : \beta_1 < 0$ en utilisant le tableau ci-dessus ?
4. Quelle proportion de la variabilité de Y est expliquée par X ?
5. Si Y et X étaient inversés dans la régression ci-dessus, à quoi vous attendriez-vous pour le R^2 ?

Exercice 3

Dans ce problème, nous étudierons la t-statistique pour l'hypothèse nulle $H_0 : \beta = 0$ en régression linéaire simple sans la constante. Pour commencer, nous générons un prédicteur x et une réponse y comme suit.

```r
set.seed(1)
x=rnorm(100)
y=7/4*x+rnorm(100)
```

1. Effectuez une régression linéaire simple de y sur x, sans la constante. Donnez l'estimation du coefficient $\hat{\beta}$, l'erreur standard de l'estimation de ce coefficient, et la t-statistique et la `p-value` associées à l'hypothèse nulle $H_0 : \beta = 0$. Commentez ces résultats. (Vous pouvez effectuer une régression sans la constante à l'aide de la commande `lm(y~x+0)`.)
2. Effectuez maintenant une régression linéaire simple de x sur y sans la constante, et donnez l'estimation du coefficient, son erreur standard, ainsi que la statistique t et la `p-value` correspondante associée à l'hypothèse nulle $H_0 : \beta = 0$.
3. Quelle est la relation entre les résultats obtenus en (1) et (2) ?
4. Pour la régression de Y sur X sans la constante, la statistique t pour $H_0 : \beta = 0$ prend la forme $\hat{\beta}/\operatorname{SE}(\hat{\beta})$, où $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ et où

$$\operatorname{SE}(\hat{\beta}) = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - x_i \hat{\beta}\right)^2}{(n-1)\sum_{i=1}^{n} x_i^2}}.$$

   Montrez algébriquement, et confirmez numériquement dans `R`, que la statistique t peut être écrite sous la forme

$$\frac{(\sqrt{n-1})\sum_{i=1}^{n} x_i y_i}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i^2) - (\sum_{i=1}^{n} x_i y_i)^2}}.$$

5. En utilisant les résultats de (4), argumentez que la statistique t pour la régression de y sur x est la même que la statistique t pour la régression de x sur y.
6. Dans `R`, montrez que lorsque la régression est effectuée avec la constante, la statistique t pour $H_0 : \beta_1 = 0$ est la même pour la régression de y sur x que pour la régression de x sur y.

Exercice 4

1. Vérifiez que l'estimation $\hat{\beta}$ pour la régression linéaire de Y sur X sans la constante est donnée par $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$. Dans quel cas l'estimation du coefficient pour la régression de X sur Y est-elle la même que l'estimation du coefficient pour la régression de Y sur X ?
2. Générez un exemple dans `R` avec $n = 100$ observations pour lesquelles l'estimation du coefficient pour la régression de X sur Y est différente de l'estimation du coefficient pour la régression de Y sur X.
3. Générez un exemple dans `R` avec $n = 100$ observations pour lesquelles l'estimation du coefficient pour la régression de X sur Y est la même que l'estimation du coefficient pour la régression de Y sur X.