Préparation à l'agrégation externe de Sciences Sociales

Statistique inférentielle - Modèle linéaire

2023-2024

Exercice 1

On souhaite mettre en évidence une corrélation entre le temps passé chaque jour devant la télévision (`time_tv`, en heures) et le taux de cholestérol (`cholesterol`, en mmol par litre de sang).

1. Rappeler les hypothèses du modèle linéaire gaussien dans le cas d'une variable explicative et d'une variable expliquée.
2. Énoncer les propriétés des estimateurs des coefficients du modèle linéaire gaussien par la méthode des moindres carrés.
3. Commenter en détail les deux lignes inférieures du tableau de résultats suivant :

```
. regress cholesterol time_tv

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   17.47
       Model |  5.04902329     1  5.04902329            Prob > F      =  0.0001
    Residual |  28.3220135    98  .289000137           R-squared     =  0.1513
-------------+------------------------------           Adj R-squared =  0.1426
       Total |  33.3710367    99  .337081179           Root MSE      =  .53759

------------------------------------------------------------------------------
 cholesterol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     time_tv |   .0440691   .0105434     4.18   0.000     .0231461    .0649921
       _cons |  -2.134777   1.813099    -1.18   0.242    -5.732812    1.463259
------------------------------------------------------------------------------
```

4. Proposez quelques pistes pour améliorer le R^2 de la régression.

Exercice 2 (2009)

Le tableau ci-dessous fournit pour treize films, le nombre y d'entrées la première semaine d'exploitation ainsi que le nombre x de salles dans lesquelles le film est projeté (avec i le numéro du film, y_i le nombre d'entrées en milliers, et x_i le nombre de salles) :

i	1	2	3	4	5	6	7	8	9	10	11	12	13
y_i	129,0	95,7	89,9	890,5	138,7	60,9	340,5	137,8	44,4	30,7	260,2	86,7	79,5
x_i	226	249	296	695	485	181	494	171	168	159	332	175	98

Les résultats statistiques pourront être déterminés à la calculatrice.

1. Étude des séries statistiques.
   (a) Calculer le nombre moyen d'entrées par film ainsi que la variance du nombre d'entrées par film.
   (b) Calculer le nombre moyen de salles ainsi que la variance du nombre de salles.
2. Le nombre de salles dans lequel le film est projeté est-il corrélé linéairement avec le nombre d'entrées réalisées par le film en première semaine d'exploitation ? Justifier votre réponse.

3. Modélisation du lien entre les séries statistiques.
   (a) Représenter le nuage de points $M_i(x_i; y_i)$. Déterminer une équation de la droite de régression de y en x, obtenue par la méthode des moindres carrés sous la forme $y = \hat{a}x + \hat{b}$ et représenter cette droite sur le graphique précédent.

   Dans la suite on notera $\hat{y} = \hat{a}x + \hat{b}$ l'estimation ponctuelle de y obtenue par la méthode des moindres carrés.

   (b) En supposant ce modèle utilisé par les producteurs/distributeurs de films pour anticiper le nombre d'entrées en fonction du nombre de salles qu'ils réservent, le score du film 5 est-il conforme aux attentes de son producteur ?
4. Analyse de la variance.
   (a) Calculer $\text{Var}(\hat{y})$, variance expliquée par le modèle.
   (b) Écrire l'équation de l'analyse de la variance et en déduire la variance résiduelle.
   (c) Déduire des résultats précédents le coefficient de détermination R^2. Interpréter votre résultat.