Tp de probabilité et statistiques

L3 informatique université Clermont Auvergne

2021/2022

1 Première partie : Régression linéaire

1.1 Régression Linéaire simple

On considère le modèle de régression linéaire suivant :

$$Y_i = \beta_0 + \beta_1 x_i.$$

Soit l'échantillon suivant qui représente le pourcentage de rendement, y_i, en fonction de la température x_i en degré celsius d'un procédé chimique.

x_i : 45 50 55 60 65 70 75 80 85 90
y_i : 43 45 48 51 55 57 59 63 66 68

Il est prévu que le pourcentage du rendement d'un procédé chimique est liée linéairement à la température. L'objectif est donc d'estimer un modèle de régression linéaire simple à partir de cet échantillon. Pour cela:

- Créez une fonction qui calcule et renvoie les coefficients de régression à partir des données, en utilisant la version non vectoriel. (calcul des β_i à l'aide de $\overline{x}$ et $\overline{y}$).
- Représentez graphiquement les données et la droite de régression obtenue (sur le même graphique).
- Comparez vos résultats avec ceux que donne la fonction polyfit python .

Modèle vectoriel
Maintenant considérons le même modèle sous sa forme vectoriel

$$y = A\beta.$$

où $\beta = (\beta_0, \beta_1)$ est le vecteur des coefficients de régressions inconnus à estimer, A la matrice de régression. y le vecteur des données des sorties. Afin d'estimer le vecteur paramètre β du modèle, Nous allons créer une fonction qui calcule et renvoie ce vecteur des coefficients de régression à partir des données en utilisant la version vectorielle:

- Implémentez la formule matricielle des $(A^T A)^{-1} A^T y$ pour estimer β.
- Représentez graphiquement les données et la droite de régression obtenue (sur le même graphique).
- Testez graphiquement la normalité des erreurs en utilisant La droite du QQ-Plot qui indique la position que devraient avoir les points s'ils obéissaient exactement à la distribution normale.
- Comparez vos résultats avec ceux que donne la fonction polyfit python .

1.2 Régression linéaire et descente de gradient

Nous allons utiliser l'algorithme de gradient pour estimer les paramètres de régression (bien que l'on ait une solution exacte, l'idée ici est de voir la notion d'optimisation globale et méthode itérative ...).

L'algorithme de gradient est comme suit. On se donne un paramètre initial β^0 et un seuil de tolérance $\epsilon > 0$ (pour le teste de convergence).

Initialisation : $\beta = \beta^0$.

Répéter

$$\beta^t = \beta^{t-1} - \lambda \frac{\partial f(\beta)}{\partial \beta}.$$

Tant que $\|\beta^t - \beta^{t-1}\| > \epsilon$.
λ étant le pas de descente $\in [0, 1]$.
avec,

$$\frac{\partial}{\partial \beta_0} f(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=0}^{n} (h(x_i) - y_i)$$

$$\frac{\partial}{\partial \beta_1} f(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=0}^{n} (h(x_i) - y_i) x_i$$

h est la fonction de la droite $h(x) = \beta_0 + \beta_1 x$.

- Créez une fonction implémentant l'algorithme de gradient pour la régression polynomiale.
- Considérez un des jeux de données précédents et comparez les résultats obtenus par la méthode de gradient avec le cas de la minimisation directe des moindres carrés.
- Faites une représentation graphique du résultat obtenu.
- Représentez graphiquement (sur le même graphique) le résultat obtenu après estimation par moindres carrés de β_0 et β_1. Que peut on déduire ?

2 Deuxième partie: Étude et manipulation de lois de probabilités

2.1 Loi Binomiale

Soit la loi Binomiale de paramètres (n, p).

- pour x=0:100, représenter graphiquement la loi probabilité pour les trois cas suivants (sur le même graphique) : $(n = 30, p = 0.5), (n = 30, p = 0.7) et (n = 50; p = 0.4)$
- utiliser la fonction binopdf pour le calcul de la loi de probabilité $B(n, p)$

2.2 Loi Normale univariée

- pour $x = -10 : 0.1 : 10$ (ou de votre choix), représenter graphiquement la fonction de densité probabilité pour les trois cas suivants pour (sur le même graphique) : $(\mu = 0, \sigma = 1)$ (loi centrée réduite), $(\mu = 2, \sigma = 1.5)$ et $(\mu = 2, \sigma = 0.6)$
- utiliser la fonction normpdf pour le calcul de la fonction de densité probabilité $\mathrm{N}(x; \mu, \sigma^2)$

2.3 Simulation de données à partir d'une loi

2.3.1 Cas de la loi Normale

- générer un échantillon i.i.d de taille n selon la loi normale centrée réduite (ou en choisissant par vous même les espérance et variance) ,utiliser pour cela soit la fonction randn soit la fonction normrnd
- afficher l'histogramme des données générées. Pour cela utiliser la commande hist (vous remarquerez que ça a une forme en cloche et donc Gaussienne)
- représenter graphiquement la vrai densité (qui est Gaussienne),prenez un support $x = -6 : 0.1 : 6$.

2.4 Estimation de densité

2.4.1 Cas de la loi Normale

- Soit la loi normale univariée centrée réduite N $(0, 1)$ ($\mu = 0$ et $\sigma = 1$) pour chacun des trois cas suivants : $n = 20; n = 80; n = 150$:
  - générer un échantillon i.i.d de taille n selon N (0, 1)
  - calculer les estimations par MV de μ et σ à partir de l'échantillon généré

- pour $x = -5 : 0.1 : 5$ calculer la la fonction de densité probabilité théorique et la la fonction de densité probabilité estimée pour x.
- représenter graphiquement la vraie densité (théorique) et la densité empirique (sur le même graphique).

2.4.2 Cas de la loi exponentielle

Soit la loi exponentielle de paramètre λ. Prenons $\lambda = 1.5$. Pour chacun des trois cas suivants $n = 20; n = 80; n = 150$:

- générer un échantillon i.i.d de taille n selon la loi exponentielle de paramètre λ. Utiliser pour cela la fonction exprnd.
- calculer l'estimation par MV de λ à partir de l'échantillon généré.
- pour $x = 0 : 0.1 : 8$, calculer la fonction de densité probabilité théorique et la fonction de densité probabilité estimée pour x.
- représenter graphiquement la vrai densité (théorique) et la densité empirique (sur le même graphique)
- faites le même travail avec la fonction de répartition (cdf).
- faites varier λ.

3 Troisième partie: Intervalles de confiance

Intervalle de confiance pour l'espérance d'une loi normale: Soit X une variable aléatoire suivant un loi normale $N(\mu, \sigma^2)$, et $X_1, X_2..., X_n$ n variables i.i.d selon la loi X.

- Si la variance de la loi de X est connue alors:

$$IC(\mu) = [\bar{X}_n - u\frac{\sigma}{\sqrt{n}}, \bar{X}_n + u\frac{\sigma}{\sqrt{n}}]$$

Où, u est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi N(0,1).
$\bar{X}_n$ est la moyenne empirique.

Exemple:

Si $\alpha = 5\%$, le fractile de la loi normale centrée réduite correspond à 1,96. (Réf la table des fractiles de la loi normale centrée réduite).

- Si la variance est inconnue:

Après centrage et réduction de la moyenne empirique on a, $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n}$ est équivalent à une loi de student à $n-1$ degrés de libertés : $St(n-1)$.

Où S_n est l'écart type empirique. Et on a :

$$IC(\mu) = [\bar{X}_n - t\frac{S_n}{\sqrt{n}}, \bar{X}_n + t\frac{S_n}{\sqrt{n}}]$$

.

t est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi de student $St(n-1)$.

Applications:

Problème 1: On considère un échantillon de 16 pots de confiture d'une certaine marque, on mesure le poids en Kg de chaque pot dans le tableau suivant:

Poids en kg : 0.499 0.509 0.501 0.494 0.498 0.497 0.504 0.506 0.502 0.496 0.495 0.493 0.507 0.505 0.503 0.491

Le poids en kg d'un pot de confiture peut être décrit par une variable aléatoire suivant une loi normale $N(\mu, \sigma^2)$.

- Calculer la moyenne empirique.
- Tracer l'histogramme des fréquences.
- Déterminer un intervalle de confiance pour μ au niveau 95% et 99 %.

  On se donne un deuxième jeu de données décrivant la masse de l'avocat provenant d'une ferme en Mexique

  Masse en g: 85.06 91.44 87.93 89.02 87.28 82.34 86.23, 84.16 88.56 90.45 84.91 89.90 85.52 86.75 88.54 87.90

- Donner un intervalle de confiance de la masse moyenne d'un avocat au niveau de 95%.

Problème 2:

Une compagnie aérienne souhaite étudier le pourcentage de voyageurs satisfaits par ses services, on en a interrogé 500 choisis au hasard. Parmi eux, 95 se disent satisfaits. Déterminer un intervalle de confiance pour la proportion inconnue de voyageurs satisfaits au niveau 99%.

Problème 3:

Simuler un échantillon de taille n de $\mathcal{B}(p = \frac{1}{2})$ indépendantes. Calculer numériquement un intervalle de confiance du paramètre, de niveau de confiance de 95%.