
Utilisation en modélisation

Régression linéaire

La régression est l'une des notions basiques de la statistique et de l'analyse des données. Généralement, le problème consiste à décrire la dépendance entre deux variables aléatoires X et Y . Les mécanismes probabilistes qui se cachent derrière cette dépendance pourront être compliqués, et c'est pourquoi on se contente d'une approximation de Y en fonction de X

$$Y \approx f(X).$$

La fonction f est appelée *régression* et le problème consiste à reconstruire (estimer) f à partir de X et Y . Très souvent, cette estimation a pour but une prévision. Si un nouveau X' arrive, en sachant f on peut prédire la nouvelle valeur de Y' inconnue. La deuxième idée principale de l'estimation de régression consiste à chercher f dans une classe de fonctions assez étroite, par exemple, dans la classe de fonctions (opérateurs) linéaires.

Dans cet exercice on tentera d'analyser le marché immobilier des appartements T3 du point de vue statistique: on cherchera à estimer la régression entre le prix et la surface d'un appartement dans les 8^{ème} et 9^{ème} arrondissements de Marseille. Ce marché est d'une part assez homogène, et d'autre part il est grand et permet d'avoir des échantillons assez représentatifs.

1 Prix d'un appartement T3 dans les 8^{ème} et 9^{ème} arr. de Marseille

La valeur vénale d'un logement résulte de plusieurs composantes:

- l'emplacement
- l'orientation
- la qualité de la construction.
- les prestations
- l'étage de l'immeuble
- la surface.

Mais parmi ces six composantes, il y a une qui joue le rôle dominant: la surface. Sans aucun doute, la surface d'un appartement et son prix sont très fortement liés. Pour illustrer ce fait, on va construire un modèle de la régression linéaire en se basant sur les données suivantes issues du journal "Le 13" paru le 1 novembre 2004.

surface (m^2)	62	68	67	76	90	67	90	60	81	90
prix (mille euros)	190.8	245	200	350	274	242	420	195	220	265
surface (m^2)	70	63	80	84	73	70	80	78	63	66
prix (mille euros)	182	208	223	252	343	199	230	245	222	185

Tableau 1.

2 Régression linéaire simple

On notera par X_i la surface de l'appartement i et par Y_i son prix. Pour modéliser la dépendance entre la surface et le prix d'un appartement, nous choisissons le modèle de la régression linéaire simple

$$Y_i = a + bX_i + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

où a, b sont les paramètres inconnus et ε_i sont les v. a. gaussiennes indépendantes avec $\mathbf{E}\varepsilon_i = 0$, $\mathbf{E}\varepsilon_i^2 = \sigma^2$. Le sens statistique du bruit ε_i est le suivant: c'est la partie du prix qui résulte de l'emplacement, de l'orientation, des prestations, etc. Évidemment, la variance σ^2 n'est pas connue.

Notre but consiste à estimer, à partir des données

$$(\mathbf{X}, \mathbf{Y}) = (X_i, Y_i) \quad i = 1, \dots, n,$$

les paramètres inconnus a, b et σ^2 du modèle (1) et construire un ensemble de confiance de niveau α pour la valeur $a + bx$. (Cette valeur nous permet de prédire les prix des appartements de surface x .) Plus précisément, on cherchera un intervalle $[\underline{P}(x), \overline{P}(x)]$ t.q.

$$\mathbf{P}(a + bx \notin [\underline{P}(x), \overline{P}(x)]) \leq \alpha.$$

Soulignons que $\underline{P}(x)$ et $\overline{P}(x)$ sont des fonction des données $\mathbf{X}, \mathbf{Y}$. Pour construire un intervalle de confiance de taille raisonnable, nous choisissons l'estimateur du maximum de vraisemblance pour $a + bx$. Il est facile de voir que cet estimateur est donné par

$$\hat{P}(x, \mathbf{X}, \mathbf{Y}) = \hat{a}(\mathbf{X}, \mathbf{Y}) + \hat{b}(\mathbf{X}, \mathbf{Y})x,$$

où

$$\hat{b} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle - \bar{\mathbf{X}}\bar{\mathbf{Y}}}{\mathbf{var}(\mathbf{X})} \quad \hat{a} = \bar{\mathbf{Y}} - \hat{b}\bar{\mathbf{X}} \quad (2)$$

avec

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{\mathbf{Y}} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad \langle \mathbf{X}, \mathbf{Y} \rangle = \frac{1}{n} \sum_{i=1}^n X_i Y_i, \quad \mathbf{var}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{\mathbf{X}}^2.$$

Pour estimer la variance inconnue σ^2 nous utilisons l'estimateur sans biais

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{a} - \hat{b}X_i)^2.$$

Le résultat suivant constitue la base théorique pour la construction de l'ensemble de confiance $[\underline{P}(x), \overline{P}(x)]$.

Théorème 1 *La variable aléatoire*

$$\frac{(\hat{a} + \hat{b}x) - (a + bx)}{\Sigma(x)},$$

où

$$\Sigma(x) = \sqrt{\frac{\hat{\sigma}^2}{n} \left[1 + \frac{(x - \bar{\mathbf{X}})^2}{\mathbf{var}(\mathbf{X})} \right]}$$

suit une loi de Student à $n - 2$ degrés de liberté.

C'est pourquoi nous choisissons l'intervalle de confiance suivant

$$\underline{P}(x) = \hat{a} + \hat{b}x - t_\alpha \Sigma(x), \quad \overline{P}(x) = \hat{a} + \hat{b}x + t_\alpha \Sigma(x),$$

où t_α est le quantile d'ordre $\alpha/2$ de la loi de Student à $n - 2$ degrés de liberté, c'est-à-dire la racine de

$$\mathbf{P}\left(S_{n-2} \geq t_\alpha\right) = \alpha/2, \quad (3)$$

où S_{n-2} est la v.a. de loi de Student à $n - 2$ degrés de liberté.

Avant de démontrer le Théorème 1, rappelons les définitions des lois du chi-2 et de Student.

Définition. Si $\xi_1, \dots, \xi_n$ est une famille indépendante de n v.a. gaussiennes centrées réduites, alors la somme

$$C_n = \sum_{i=1}^n \xi_i^2$$

suit, par définition, une loi du chi-2 à n degrés de liberté. Cette loi est notée χ_n^2 . Le résultat suivant explique le sens géométrique de cette loi.

Lemme 1 Soit $\mathbf{Z} = (Z_1, \dots, Z_n)$ un vecteur dans $\mathbb{R}^n$ de composantes

$$Z_i = \mu_i + \sigma \xi_i, \quad i = 1, \dots, n,$$

où ξ_i sont i.i.d. de loi $\mathcal{N}(0, 1)$. Soit H_J un sous-espace dans $\mathbb{R}^n$ de dimension J et

$$D_J = \min_{\mathbf{y} \in H_J} \|\mathbf{Y} - \mathbf{y}\|.$$

Si le vecteur μ appartient à H_J alors D_J^2 suit une loi de χ_{n-J}^2 à $n - J$ degrés de liberté.

Définition. Si ξ est C_n sont deux v.a. indépendantes, respectivement de loi gaussienne centrée réduite et χ_n^2 , alors le rapport

$$S = \frac{\xi}{\sqrt{C_n/n}}$$

suit, par définition, une loi de Student à n degrés de liberté.

Démonstration du Théorème 1. Pour démontrer ce théorème il est suffit de vérifier les fait suivants:

1. la v.a. $\hat{a} + \hat{b}x - a - bx$ suit une loi gaussienne avec

$$\begin{aligned} \mathbf{E}(\hat{a} + \hat{b}x - a - bx) &= 0 \\ \mathbf{E}(\hat{a} + \hat{b}x - a - bx)^2 &= \frac{\sigma^2}{n} \left(1 + \frac{(\bar{\mathbf{X}} - x)^2}{\mathbf{Var}(\mathbf{X})} \right) \end{aligned} \quad (4)$$

2. les v.a. $\hat{a} + \hat{b}x - a - bx$ et $\hat{\sigma}^2$ sont indépendantes

3. la v.a. $(n - 2)\hat{\sigma}^2/\sigma^2$ suit une loi de χ_{n-2}^2 .

1. Il est facile de voir (voir (2)) que

$$\hat{a} - a + (\hat{b} - b)x = \frac{1}{n} \sum_{k=1}^n \varepsilon_k \left(1 + \frac{(\bar{\mathbf{X}} - x)(\bar{\mathbf{X}} - X_k)}{\mathbf{Var}(\mathbf{X})} \right). \quad (5)$$

D'où on obtient immédiatement (4).

2. Notons que (voir (5))

$$\begin{aligned} Y_i - \hat{a} - \hat{b}X_i &= \varepsilon_i + a - \hat{a} + (b - \hat{b})X_i \\ &= \varepsilon_i - \frac{1}{n} \sum_{k=1}^n \varepsilon_k \left(1 + \frac{(\bar{\mathbf{X}} - X_k)(\bar{\mathbf{X}} - X_i)}{\mathbf{Var}(\mathbf{X})} \right). \end{aligned} \quad (6)$$

Donc les v.a. $\hat{a} - a + (\hat{b} - b)x$ et $Y_i - \hat{a} - \hat{b}X_i$ sont gaussiennes centrées. De plus, elles sont indépendantes car elles ne sont pas corrélées :

$$\begin{aligned} \mathbf{E}[\hat{a} - a + (\hat{b} - b)x][Y_i - \hat{a} - \hat{b}X_i] &= \\ &= \mathbf{E}\varepsilon_i \frac{1}{n} \sum_{k=1}^n \varepsilon_k \left(1 + \frac{(\bar{\mathbf{X}} - x)(\bar{\mathbf{X}} - X_k)}{\mathbf{Var}(\mathbf{X})} \right) \\ &\quad - \mathbf{E} \frac{1}{n} \sum_{k=1}^n \varepsilon_k \left(1 + \frac{(\bar{\mathbf{X}} - x)(\bar{\mathbf{X}} - X_k)}{\mathbf{Var}(\mathbf{X})} \right) \frac{1}{n} \sum_{s=1}^n \varepsilon_s \left(1 + \frac{(\bar{\mathbf{X}} - X_s)(\bar{\mathbf{X}} - X_i)}{\mathbf{Var}(\mathbf{X})} \right) \\ &= \frac{\sigma^2}{n} \left(1 + \frac{(\bar{\mathbf{X}} - x)(\bar{\mathbf{X}} - X_k)}{\mathbf{Var}(\mathbf{X})} \right) \\ &\quad - \frac{\sigma^2}{n^2} \sum_{k=1}^n \left(1 + \frac{(\bar{\mathbf{X}} - x)(\bar{\mathbf{X}} - X_k)}{\mathbf{Var}(\mathbf{X})} \right) \left(1 + \frac{(\bar{\mathbf{X}} - X_k)(\bar{\mathbf{X}} - X_i)}{\mathbf{Var}(\mathbf{X})} \right) \\ &= \frac{\sigma^2}{n} \left(1 + \frac{(\bar{\mathbf{X}} - x)(\bar{\mathbf{X}} - X_k)}{\mathbf{Var}(\mathbf{X})} \right) - \frac{\sigma^2}{n} \left(1 + \frac{(\bar{\mathbf{X}} - x)(\bar{\mathbf{X}} - X_k)}{\mathbf{Var}(\mathbf{X})} \right) = 0. \end{aligned}$$

Ça veut dire que les v.a. $\hat{\sigma}^2$ et $\hat{a} - a + (\hat{b} - b)x$ sont indépendantes.

3. On déduit le fait que la v.a. $(n-2)\hat{\sigma}^2/\sigma^2$ suit une loi de χ_{n-2}^2 du Lemme 1 car

$$\hat{\sigma}^2 = \min_{a,b} \frac{1}{n-2} \sum_{i=1}^n (Y_i - a - bX_i)^2 = \min_{\mathbf{y} \in H_2} \|\mathbf{Y} - \mathbf{y}\|^2 / (n-2),$$

où H_2 est le sous-espace dans $\mathbb{R}^n$, engendré par deux vecteurs

$$(1, \dots, 1) \quad \text{et} \quad (X_1, \dots, X_n). \quad \square$$

3 Modélisation avec MATLAB

1. Programmer un fichier script qui estime les paramètres du modèle (1) et rend l'ensemble de confiance pour la valeur $a + bx$.

- Prendre les données (X_i, Y_i) , $i = 1, \dots, n$ du tableau 1. N'oublier pas de les ordonner!
- Calculer $\hat{a}, \hat{b}, \hat{\sigma}^2$.

- Trouver la valeur t_α pour $\alpha = 0.05$ définie par (3) (On peut utiliser la méthode de Monte-Carlo.)
2. Tracer sur le même graphique les données (X_i, Y_i) , la fonction $\hat{a} + \hat{b}x$ et l'intervalle de confiance $[\underline{P}(x), \overline{P}(x)]$ de niveau 0.95 en fonction de x (voir la figure 1).
 3. Enlever des données les appartements de grand luxe (prix plus que 300.000 euros). Faire le même exercice avec ces nouvelles données. Représenter les résultats comme sur la figure 2.
 4. Discuter la différence entre les deux graphiques. Par exemple, expliquer
 - pourquoi la valeur positive de a signifie que le prix du mètre carré des grands appartements est, en principe, moins cher que celui des petites surfaces
 - pourquoi, à votre avis, le modèle gaussien n'est pas adéquate aux données (comparer les valeurs de $\hat{a}$ et $\hat{b}$ avant et après l'enlèvement des trois appartement).

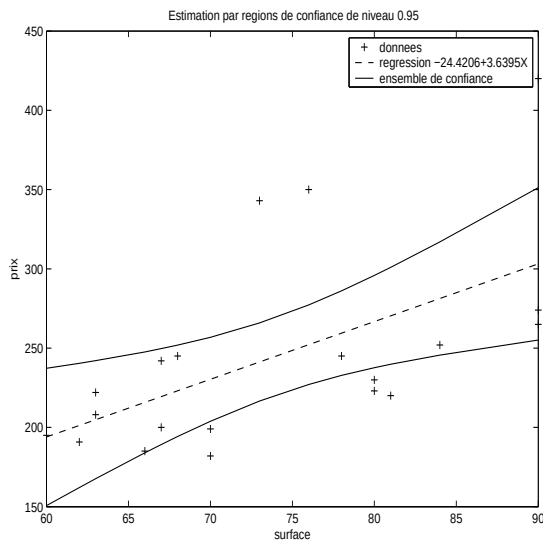


Figure 1: Régression linéaire pour tout les T3 dans le tableau 1.

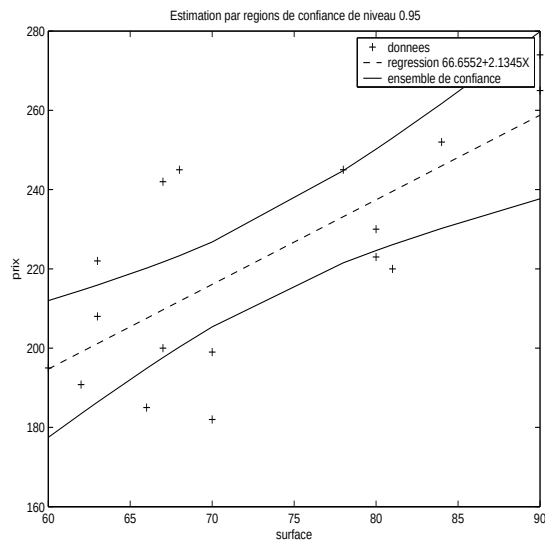


Figure 2: Régression linéaire sans les trois T3 de grand luxe.