Cours - Informatique - Régression linéaire

Gustave Cortal

Exemple introductif

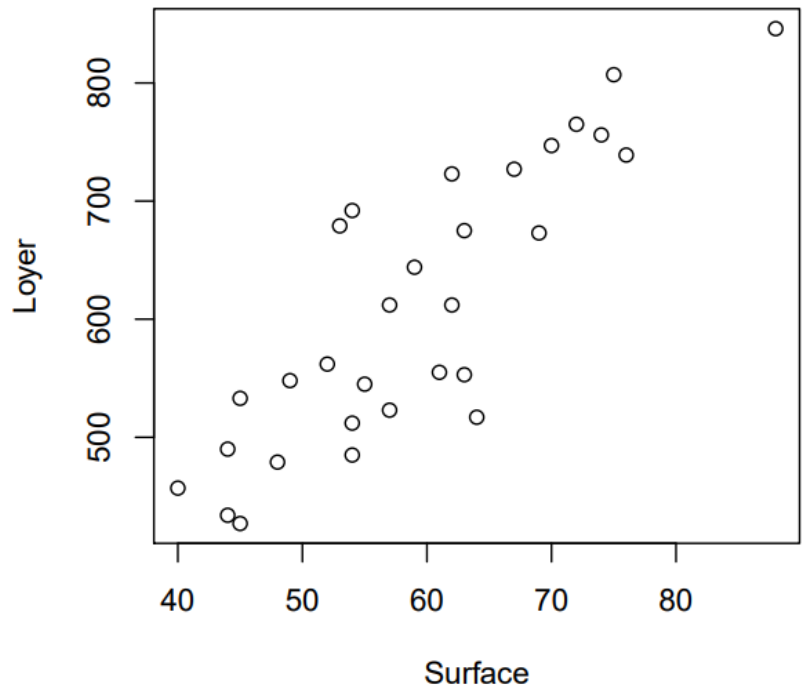

Figure: Prix du loyer en fonction de la surface de l'appartement.

Le prix d'un appartement est une variable aléatoire Y dont l'espérance dépend de la surface x. Y est la *variable à expliquer* et x est la *variable explicative*, supposée connue et non aléatoire.

Problèmes abordés

Les problèmes sont multiples :

- spécifier le modèle ;
- estimer les paramètres du modèle ;
- vérifier qu'il y a bien une relation entre les deux variables ;
- vérifier la validité du modèle retenu ;
- prédire le prix d'un nouvel appartement en fonction de sa surface.

Modèle de la régression linéaire simple

L'espérance de la variable aléatoire à expliquer Y est une fonction linéaire de la variable x :

$$E(Y_i) = a + bx_i \tag{1}$$

Nous avons :

$$Y_i = a + bx_i + \epsilon_i \qquad i = 1, ..., n. \tag{2}$$

avec $E(\epsilon_i) = 0$. On suppose que les ϵ_i sont **indépendants** et suivent **la même loi normale** $N(0, \sigma^2)$. Le modèle est alors :

$$Y_i \sim N(a + bx_i, \sigma^2), \qquad i = 1, ..., n. \tag{3}$$

Hypothèses fondamentales du modèle

- **Linéarité** : $E(Y_i)$ est une forme linéaire des paramètres : $E(Y_i) = a + bx_i$
- **Normalité des erreurs** : les erreurs suivent une loi normale
- **Homoscédasticité** : les erreurs partagent la même variance σ^2
- **Indépendance** : les erreurs sont indépendantes

Estimation des paramètres du modèle (1)

On estime les paramètres du modèle avec la méthode du maximum de vraisemblance. Soit la fonction de densité de Y_i :

$$f(Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{1}{2\sigma^2}(Y_i - a - bx_i)^2) \qquad i = 1, ..., n. \tag{4}$$

Les variables Y_i sont indépendantes. On en déduit la fonction de vraisemblance de l'échantillon :

$$L(a, b, \sigma^2; y_1, ..., y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(Y_i - a - bx_i)^2}{2\sigma^2}) \tag{5}$$

et la fonction de log-vraisemblance :

$$l(a, b, \sigma^2; y_1, ..., y_n) = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - a - bx_i)^2 \tag{6}$$

Estimation des paramètres du modèle (2)

Pour maximiser la vraisemblance par rapport à a et b, il faut minimiser :

$$Q(a,b) = \sum_{i=1}^{n} (Y_i - a - bx_i)^2 \tag{7}$$

$Q(a,b)$ est le *critère des moindres carrés*. Pour le minimiser, il faut résoudre le système à deux équations :

$$\frac{\partial Q}{\partial a} = 0 \qquad \frac{\partial Q}{\partial b} = 0 \tag{8}$$

On trouve les *estimateurs des moindres carrés* $\hat{a}$ et $\hat{b}$:

$$\hat{b} = \frac{S_{xY}}{s_x^2} \qquad \hat{a} = \overline{Y} - \frac{S_{xY}}{s_x^2}\overline{x} \tag{9}$$

S_{xY} est la covariance empirique des x_i et des Y_i, et s_x^2 la variance empirique des x_i. $\overline{Y}$ est la moyenne empirique des Y_i et $\overline{x}$ la moyenne empirique des x_i.

Estimation des paramètres du modèle (3)

Nous obtenons la droite des moindres carrés de Y en x :

$$\hat{Y}_i = \hat{a} + \hat{b}x_i \qquad et \qquad \hat{\epsilon} = Y_i - \hat{Y}_i \qquad i = 1, ..., n \tag{10}$$

$\hat{\epsilon_i}$ sont appelées *résidus*.

L'estimateur du maximum de vraisemblance de σ^2 est :

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon_i} \tag{11}$$

C'est la moyenne des carrés des résidus.

Analyse de la variance

Équation d'analyse de la variance :

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (12)$$

que l'on peut noter :

$$S_Y^2 = S_{reg} + S_{res} \qquad (13)$$

La variance totale est égale à la variance expliquée par la régression et la variance résiduelle. Les *estimateurs des moindres carrés* minimisent la variance résiduelle, donc maximise la variance expliquée par la régression.

Mesure de l'ajustement

On mesure la qualité de l'ajustement de la régression par le *coefficient de détermination* :

$$R^2 = \frac{S_{reg}}{S_Y^2} \tag{14}$$

C'est une quantité variant entre 0 et 1. Une valeur proche de 1 indique que la proportion de variance expliquée par la régression est grande, la régression est donc très explicative.

Prédiction sur de nouvelles valeurs

Soit une nouvelle valeur x_0 (e.g., une nouvelle surface), on peut prédire grâce au modèle de régression Y_0 (e.g., le prix de l'appartement correspondant) :

$$\hat{Y_0} = \hat{a} + \hat{b}x_0 \tag{15}$$

Aller plus loin

- Propriétés des estimateurs $\hat{a}$, $\hat{b}$ et $\hat{\sigma^2}$ (e.g., quelles lois suivent-ils?)
- Estimateur sans biais de $\hat{\sigma^2}$
- Intervalles de confiance sur $\hat{a}$ et $\hat{b}$
- Tester la non-nullité de la pente $\hat{b}$
- Vérification des hypothèses du modèle