Régression linéaire simple

Qu'est-ce que la régression ?

Comme on l'a vu en cours, la régression simple est une méthode de machine learning qui permet d'étudier les relations entre deux variables quantitatives.

— Une des variables, classiquement notée x, est la *variable explicative* ou prédicteur.

— L'autre variable, classiquement notée y, est la *variable à prédire* ou réponse ou encore cible.

En mathématique et en physique, on connait des relations déterministes entre deux variables. Par exemple

— Circonférence $= \pi \times$ diamètre

— Loi de Ohm : $V = I/r$ avec V le voltage, I l'intensité et r la résistance.

En statistique et en machine learning, on s'intéresse à des données expérimentales et on a rarement une relation déterministe entre les variables. Les graphes de la figure 1 montrent que la relation entre les deux variables considérées n'est pas parfaitement linéaire.

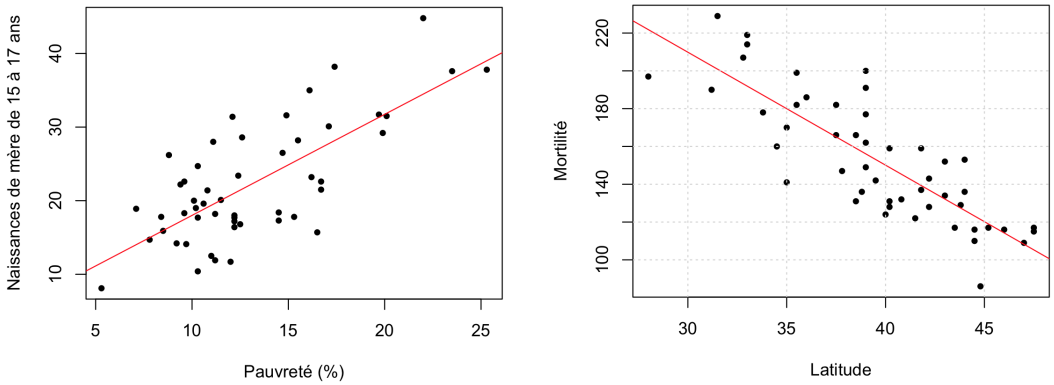

Figure 1 – Dans chacune des figures un point représente les données d'un état des États-Unis. A gauche : nombre de naissances de mères âgées de 15 à 17 ans (pour mille naissances) en fonction du pourcentage de ménages pauvres dans la population. A droite : nombre de décès par cancer de la peau en fonction de la latitude au centre de l'état.

Bien que la relation ne soit pas parfaitement linéaire entre les deux variables, on observe par exemple que le taux de naissance de mères jeunes augmente avec le pourcentage de ménages pauvres et que la croissance moyenne du taux de naissances est bien représentée par une droite. De même le nombre de décès par cancer de la peau décroît linéairement avec la latitude.

Dans les deux cas, on peut faire l'hypothèse qu'il existe une relation linéaire entre les deux variables, à une petite erreur près. On écrit alors le modèle suivant

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

Si on dispose d'un échantillon de n observations $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, on considère alors les équations

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \cdots, n \tag{2}$$

où x_i et y_i représentent respectivement la valeur de la variable explicative et la réponse pour la $i$ème observation. Le terme ϵ_i représente une erreur qui prend en compte différents effets aléatoires et/ou une erreur de modèle.

Droite de régression

Pour expliquer la relation entre les variables x et y ou faire de la prédiction, on doit estimer les paramètres de la droite de régression $y = \beta_0 + \beta_1 x$ associée au modèle (1), ie les paramètres inconnus β_0 et β_1.
On suppose que l'on dispose d'observations (x_i, y_i) pour $i = 1, \cdots, n$.

1. Dans un tel problème de machine learning, on commence par choisir tout d'abord la fonction de perte à minimiser. S'aider du cours et écrire la fonction de perte utilisée classiquement en régression linéaire. Donner son nom et son expression.

   Dans la suite on notera L la fonction de perte des moindres carrés[1]

   $$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

2. Montrer que

   $$L(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i^2 + n\beta_0^2 + \beta_1^2 \sum_{i=1}^{n} x_i^2 - 2\beta_0 \sum_{i=1}^{n} (y_i - \beta_1 x_i) - 2\beta_1 \sum_{i=1}^{n} y_i x_i \tag{3}$$

3. Estimateur de l'ordonnée à l'origine β_0
   (a) Déduire du développement (3) que $L_0 : \beta_0 \mapsto \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ est un polynôme de degré deux qui admet un minimum fini.

1. On note L pour *loss*, terme qu'on retrouvera quand on travaillera sur des problèmes de machine learning et en particulier avec les réseaux de neurones.

(b) Calculer la dérivée de

$$\begin{aligned} L_0 : \quad & \mathbb{R} \to \mathbb{R} \\ & \beta_0 \mapsto \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned}$$

(c) En déduire que le minimum de $L(\beta_0, \beta_1)$ s'écrit, pour une valeur quelconque de β_1 fixé

$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

4. Estimateur de la pente β_1

(a) Déduire du développement 3 que $L_1 : \beta_1 \mapsto \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ est un polynôme de degré deux qui admet un minimum fini.

(b) Calculer la dérivée de

$$\begin{aligned} L_1 : \quad & \mathbb{R} \to \mathbb{R} \\ & \beta_1 \mapsto \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \end{aligned}$$

(c) En déduire que l'optimmum de $L(\beta_0, \beta_1)$ vérifie

$$\beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\sum_{i=1}^{n} y_i x_i - \beta_0 \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

5. Solution du problème aux moindres carrés

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg \min_{(\beta_0,\beta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{6}$$

(a) Écrire le système linéaire qui donne la solution du problème (6). Donner aussi son expression sous forme matricielle.

(b) Et retrouver que

$$\widehat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} y_i \frac{1}{n}\sum_{i=1}^{n} x_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_i x_i)^2}$$

6. Données sur le nombre de naissances de mère jeunes

(a) On a $n = 51$ observations. En notant y_i le nombre de naissances de mères jeunes dans l'état i et x_i l'indice de pauvreté correspondant, on a

$$\sum_i y_i = 1136.4, \ \sum_i x_i = 669, \ \sum_i y_i x_i = 16163.14, \ \sum_i x_i^2 = 9690.44$$

Calculer $\widehat{\beta}_0$ et $\widehat{\beta}_1$.

Régression linéaire multiple

Il est assez rare qu'on se contente d'expliquer un phénomène ou une variable à partir d'une seule autre variable. La plupart du temps on prend en compte plus d'un facteur explicatif. Par exemple, on va chercher à expliquer le niveau de pauvreté en fonction de la répartition démographique, du nombre d'entreprises, du nombre de diplômés, etc.

Dans la suite, nous nous intéressons à un modèle classique en économie qui régit les relations entre production, travail et capital.

Modèle de Cobb-Douglas

On considère les variables suivantes, chacune concernant la totalité des États-Unis (i étant l'indice d'une année) :

— P : production

— K : capital (valeur des usines, etc.)

— T : travail fourni (basé sur un calcul du nombre total de travailleurs)

On cherche à expliquer P à l'aide des variables (K, T). Un modèle classique est alors le modèle de Cobb et Douglas :

$$P = \beta_0 K^{\beta_1} T^{\beta_2} \tag{7}$$

Le paramètre β_0 s'interprète comme un indice de niveau technologique. β_1 et β_2 s'interprètent comme des mesures de la part de production due au capital et au travail.

1. Vérifier qu'en transformant, par application d'un $\log$ aux membres de gauche et de droite de l'équation (7), on se ramène à un modèle linéaire de la forme

$$y = \beta_0' + \beta_1 x_1 + \beta_2 x_2. \tag{8}$$

   Donner les expressions respectives de y, x_1 et x_2 en fonction de P, K et T.

2. On suppose qu'on dispose de n observations du triplet (P, K, T). Pour $i = 1, \cdots, n$, on les note P_i, K_i et T_i. Pour ces observations, on déduit de (8) le système d'équations

$$y_i = \beta_0' + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \tag{9}$$

   où u_i représente l'erreur de modèle.
   En s'inspirant des questions 2. et 3.(a) de l'exercice précédent, résoudre le problème aux moindres carrés qui permet d'estimer les paramètres $\widehat{\beta_0'}$, $\widehat{\beta_1}$ et $\widehat{\beta_2}$.

Année	P	K	T	Année	P	K	T	Année	P	K	T
1899	100	100	100	1907	151	176	138	1915	189	266	154
1900	101	107	105	1908	126	185	121	1916	225	298	182
1901	112	114	110	1909	155	198	140	1917	227	335	196
1902	122	122	118	1910	159	208	144	1918	223	366	200
1903	124	131	123	1911	153	216	145	1919	218	387	193
1904	122	138	116	1912	177	226	152	1920	231	407	193
1905	143	149	125	1913	184	236	154	1921	179	417	147
1906	152	163	133	1914	244	244	149	1922	240	431	161

TABLE 1 – Données de l'article de Cobb et Douglas (1928)

(a) Écriture du problème d'optimisation des moindres carrées.

(b) Calcul des dérivées par rapport à chacun des paramètres inconnus

(c) Annulation des dérivées pour obtenir un système de 3 équations à 3 inconnues. En option : résoudre le système.

En utilisant les données de la Table 1, Cobb et Douglas ont obtenu $\beta_1 = 1/4$ et $\beta_2 = 3/4$.

Référence

D. Cobb. A theory of production. In : American Economic Review (1928), p. 139-165.