Cours 7 : Régression linéaire simple

=Prédire / expliquer les valeurs d'une variable quantitative Y à partir d'une autre variable X= expliquer Y par une fonction affine de X

La régression linéaire simple est une méthode statistique qui permet de résumer et d'étudier les relations entre deux variables continues (quantitatives).
Ce chapitre présente le concept et les procédures de base de la régression linéaire simple.
Une variable, notée **x**, est considérée comme la variable prédictive, explicative ou indépendante.
L'autre variable, notée **y**, est considérée comme la réponse, le résultat ou la variable dépendante. Comme les autres termes sont utilisés moins fréquemment aujourd'hui, nous utiliserons les termes "prédicteur" et "réponse" pour faire référence aux variables rencontrées dans ce manuscrit.
Les autres termes mentionnés, vous les rencontrez dans d'autres domaines. La régression linéaire simple prend son adjectif "simple", car elle ne concerne que l'étude d'une seule variable prédictive.

Définition d'un modèle de régression linéaire simple :

Le but de la régression linéaire est de modéliser une variable continue y en tant que fonction mathématique d'une ou de plusieurs variables x , de sorte que nous puissions utiliser ce modèle de régression pour prédire le y lorsque seul x est connu.
Cette équation mathématique peut être généralisée comme suit :

$$y_j = a \times x_j + b + \varepsilon_j$$

i=1,…,n
avec
$\mathbf{y_i}$: la variable endogène (dépendante, à expliquer) à l'indice i.
$\mathbf{x_i}$: la variable exogène (indépendante, explicative) à l'indice i.

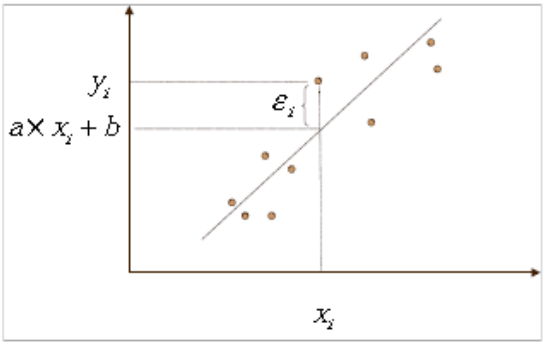
a*; *b : les paramètres inconnus du modèles où
b : représente le point x de la droite de régression avec l'ordonnée à l'origine.
a : est la pente de la droite de régression.
ε_i : l'erreur aléatoire du modèle.
n : nombre d'observations.

Hypothèses du modèle :

Le modèle repose sur les hypothèses suivantes :

1. $E(\varepsilon_i) = 0$, l'erreur est centrée.
2. $E(\varepsilon_i^2) = \sigma_\varepsilon^2$, la variance de l'erreur est constante.
3. $Cov(\varepsilon_i, \varepsilon_{i'}) = 0, i \neq i'$ les erreurs ne sont pas autocorrélées.
4. $Cov(x_i, \varepsilon_{i'}) = 0$, l'erreur n'est pas corrélée avec la variable exogène.
5. La variable exogène X_i n'est pas aléatoire.
6. Le modèle est linéaire en X par rapport aux paramètres.

Principe de l'ajustement des moindres carrées :



Critère des moindres carrés : trouver les valeurs de **a** et **b** qui minimise la somme des carrés des écarts entre les vraies valeurs de Y et les valeurs prédites avec le modèle de prédiction.

$$S = \sum_{i=1}^{n} \varepsilon_i^2$$
$$S = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$
$$S = \sum_{i=1}^{n} [y_i - ax_i - b]^2$$

SOLUTION

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

$$\begin{cases} \sum_i x_i y_i - a\sum_i x_i^2 - b\bar{x} = 0 \\ \bar{y} - a\bar{x} - b = 0 \end{cases}$$

Equations normales

$$\begin{cases} \hat{a} = \dfrac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\ \hat{b} = \bar{y} - \hat{a}\bar{x} \end{cases}$$

Estimateurs des moindres carrés

Equation d'analyse de variance -Décomposition de la variance- :

Objectif de la régression : minimiser S.
Mais $0 \leq S \leq +\infty$; à partir de quand peut-on dire que la régression est de « bonne qualité » ?

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Somme des écarts à la moyenne

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$
$$= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\underbrace{\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{=0}$$

S'il y a une association linéaire entre X et Y, quand la valeur X dévie de sa moyenne, la valeur Y correspondante tend à dévier aussi de sa moyenne.
Donc, nous pouvons utiliser la droite du meilleur ajustement pour expliquer en partie la variabilité totale dans les réponses. Nous séparons la déviation totale en deux composantes : une déviation expliquée et une déviation résiduelle.

L'équation de l'analyse de la variance s'écrit comme suit :

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2$$

⇨ Décomposition de la variance

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$
$$SCT = SCR + SCE$$

SCT : somme des carrés totaux
SCE : somme des carrés expliqués par le modèle
SCR : somme des carrés résiduels, non expliqués par le modèle

Le test de **Fisher** s'intéresse à **la significativité globale d'un modèle.**
Dans le cas de la régression simple, seul le paramètre **a** est concerné.
On a déjà défini SCT, SCE et SCR. Ces sommes peuvent être utilisées pour tester les hypothèses suivantes :

$$\boldsymbol{H_0 : a = 0 \; vs \; H_1 : a \neq 0}$$

Sous l'hypothèse $\boldsymbol{H_0 : a = 0}$, on a

$$\boldsymbol{E(SCT) = (n-1)\sigma^2};$$
$$\boldsymbol{E(SCR) = (n-2)\sigma^2}$$
$$\boldsymbol{et}$$
$$\boldsymbol{E(SCE) = (1)\sigma^2};$$

avec $\mathbf{(n-1)}$; $\mathbf{(1)}$ **et** $\mathbf{(n-2)}$ les degrés de liberté de SCT, SCE et SCR respectivement.
D'autre part, lorsque $\boldsymbol{H_0}$ est vérifiée on a :

$$\frac{SCT}{n-1} \rightsquigarrow \chi^2_{n-1}$$
$$\frac{SCR}{n-2} \rightsquigarrow \chi^2_{n-2} \; et \; \frac{SCE}{1} \rightsquigarrow \chi^2_1.$$

Du moment que SCR et SCE sont indépendantes, on peut déduire alors **la statistique F (Fisher)** qui est le rapport entre deux Khi deux indépendants et leurs degrés de libertés. Il est défini comme suit :

$$F_c = \frac{SCE/1}{SCR/n-2} = \frac{SCE \times (n-2)}{SCR \times 1} \rightsquigarrow F_{(\alpha,1,n-2)},$$

$$\boldsymbol{F_c = (n-2)\frac{R^2}{1-R^2}}$$

avec
$\boldsymbol{F_C}$: Désigne la valeur calculée de Fisher (**la statistique**).
$\mathbf{F_{Tab}}$: Désigne la valeur de Fisher lue à partir de la table statistique de Fisher aux degrés de liberté $(1; n-2)$.
$\boldsymbol{\alpha}$: Le seuil donné %.
Si $\boldsymbol{|F_C| < F_{Tab}(\alpha; 1; n-2)}$, on **accepte** l'hypothèse $\boldsymbol{H_0}$, cela signifie que $\boldsymbol{H_1}$est **rejetée**. C'est-à-dire le modèle n'est pas globalement significatif(i.e. **il n' y a pas une régression**).

$Si\ |F_C| > F_{Tab}(\alpha; 1; n-2)$, on **rejette** l'hypothèse H_0 et cela signifie que H_1 est acceptée. C'est-à-dire le modèle est globalement significatif.

Source de variation	Sommes des carrées	d.d.l	carrées Moyennes	F calculée
Variabilité à expliquer	SCE	K=1	$S_E^2 = \frac{SCE}{1}$	$F_C = \frac{S_E^2}{S_R^2}$
Variabilité résiduelle	SCR	n-2	$S_R^2 = \frac{SCR}{n-2}$	
Variabilité totale	SCT=SCE+SCR	n-1		

Coefficient de détermination R^2 :

Pour mesurer la qualité d'ajustement on utilise le coefficient de détermination R^2 qui est un indicateur de qualité du modèle, il exprime la proportion de variabilité de Y.

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

$$R^2 = 1 - \frac{SCR}{SCT}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Si on a :

$0 \leq R^2 \leq 1$, plus la valeur de R^2 est proche de 1, plus le modèle est significatif.

$R^2 = 1$, correspond au cas ou y_i et coïncide avec $\hat{y}_i$, c.à.d que les données sont parfaitement alignées sur la droite de régression.

$R^2 = 0$, le modèle n'est pas adéquat.

Intervalles de prédiction :

Un des buts de la régression est de proposer des prédictions pour la variable à expliquer Y. Soit $\boldsymbol{x_{n+1}}$ une nouvelle valeur de la variable **X**, nous voulons prédire y_{n+1}.
Le modèle indique que

$$y_{n+1} = \mathrm{a}.x_{n+1} + b + \varepsilon_{n+1}$$

Nous pouvons prédire la valeur correspondante grâce au modèle estimé

$$\hat{y}_{n+1} = \hat{\mathrm{a}}.x_{n+1} + \hat{b}$$

En utilisant la notation $\hat{y}_{n+1}$:
Deux types d'erreurs vont entacher notre prévision, la première due à la non connaissance de ε_{n+1} et l'autre due à l'estimation des paramètres.
La variance augmente lorsque x_{n+1} s'éloigne du **centre de gravité** du nuage. Faire de la prévision lorsque $\boldsymbol{x_{n+1}}$ est loin de $\bar{x}$ est donc périlleux, la variance de l'erreur de prévision peut alors être très grande.
Un intervalle de confiance **IC de $\boldsymbol{y_{n+1}}$ au niveau $1-\alpha$ est donné par :**

$$\left[\hat{y}_{n+1} \pm (t_{\alpha,n-2}) S_R \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}\,;\right]$$

avec

$\boldsymbol{t_{\alpha,n-2}}$: Désigne la valeur t de Student lue à partir de la table statistique au seuil ($\boldsymbol{\alpha}$) et de degré de liberté ($\boldsymbol{n-2}$).

$\hat{\sigma} = \sqrt{S_R^2} = \sqrt{\frac{SCR}{n-2}}$: Désigne l'écart-type de l'erreur en valeur connue (l'écart type du carré moyens de la variation résiduelle).

$x_j = x_{n+1}$: Désigne la valeur de la variable exogène (explicative) à l'horizon (n+1) :
$\hat{y}_{n+1}$: Désigne la valeur de la variable endogène estimée à l'horizon (n+1).

Exercice :

On a mesuré le poids (en kg) et le périmètre thoracique (en cm) de 8 taureaux à l'engrais:

	Périmètre thoracique (X)	Poids corporel (Y)
	158	540
	164	544
	167	553
	170	549
	171	560
	176	557
	179	556
	183	565
Somme:	1368	4424
Moyenne:	171	553

a) On demande de mesurer la relation qui existe entre ces deux variables **et** d'en tester la signification.
b) Deux animaux ont un périmètre thoracique de **172 cm** et **186 cm**, respectivement. Pouvez-vous prédire leur poids?