AJUSTEMENT ANALYTIQUE RÉGRESSION - CORRÉLATION

1. INTRODUCTION

Il est fréquent de s'interroger sur la relation qui peut exister entre deux grandeurs en particulier dans les problèmes de prévision et d'estimation.

Trois types de problèmes peuvent apparaître:

1. On dispose d'un certain nombre de points expérimentaux (x_i ,y_i) $1 \le i \le n$,où x_i et y_i sont les valeurs prises par les grandeurs x et y et on essaye de déterminer une relation fonctionnelle entre ces deux grandeurs x et y. Cette relation, pour des raisons théoriques ou pratiques s'écrira y = f(x, a,b,c...) et le problème sera d'ajuster au mieux les paramètres a,b,c.... pour que la courbe représentative de f passe au plus près des points (x_i ,y_i). Il s'agit d'un **problème d'ajustement analytique**.

Exemple : Le nombre de particules émises par un élément radioactif varie en fonction du temps. On sait que la loi est de la forme $n = n_0 e^{-\lambda t}$. Les mesures expérimentales permettront d'estimer au mieux n_0 et λ .

2. On essaye de déterminer la relation statistique qui existe entre les deux grandeurs X et Y. Ce type d'analyse s'appelle **analyse de régression**. On considère que la variation de l'une des deux variables (par exemple X) explique celle de l'autre (par exemple Y). Chaque domaine d'application a baptisé de noms différents ces deux variables : On trouve ainsi :

X	Y
Variable explicative	Variable expliquée
Variable contrôlée	Réponse
Variable indépendante	Variable dépendante
Régresseur	....

Dans ce type d'analyse, on fixe *a priori* les valeurs de X. X n'est donc pas une variable aléatoire. Mais la deuxième grandeur Y, elle, est une variable aléatoire et sa distribution est influencée par la valeur de X. On a alors du point de vue statistique une relation de cause à effet. Le problème sera d'identifier cette relation.

Exemple : On veut déterminer si un produit est toxique. Pour cela on le fait absorber en quantités variables par des souris et on mesure leur temps de survie. On partage la population des souris en quatre lots qui absorberont respectivement 0, 1, 2, 3 doses de produit.
X = nombre de doses de produit est une variable contrôlée prenant les valeurs (0, 1, 2, 3).
Y = temps de survie d'une souris est une variable aléatoire (réponse à la variable contrôlée X).

Si Y est fonction de X, le produit est toxique. Connaître la relation entre X et Y nous permettra d'évaluer ses effets toxiques.

3. Les deux grandeurs X et Y sont aléatoires et on cherche à savoir si leurs variations sont liées. Il n'y a pas ici de variable explicative ni de variable expliquée. Les variables peuvent avoir des causes communes de variation, parmi d'autres, qui expliquent leur relation d'un point de vue statistique : on est en présence d'un **problème de corrélation**. On cherche alors à mesurer le degré d'association entre les variables.

Exemple : poids et taille d'un individu, résultats obtenus à deux examens par des étudiants...

2. AJUSTEMENT ANALYTIQUE

2.1. PRINCIPE DE L'AJUSTEMENT

On dispose d'un certain nombre de points (x_i ,y_i) $1 \le i \le n$, formant un nuage statistique, et on cherche à traduire la dépendance entre x et y par une relation de la forme y = f(x) ou x =g(y) selon ce qui a un sens, ou selon ce qui nous intéresse.

- Si une relation théorique s'impose à nous comme dans l'exemple des particules radioactives, on ajuste au mieux les paramètres de la loi théorique.

- S'il nous faut déterminer empiriquement f ou g, on privilégiera les **modèles linéaires** .Précisons que la linéarité du modèle ne doit pas prêter à confusion : le terme linéaire ne se réfère qu'aux paramètres du modèle: y = a + bx ou y = a + bx+ cx^2 sont des modèles linéaires alors que dans le deuxième exemple la relation entre x et y est quadratique. En revanche $y = \beta_0 \beta_1^{\,x}$ n'est pas un modèle linéaire. Pour déterminer cette relation empirique la forme du nuage statistique peut guider notre choix.

Mais quelle méthode utiliser pour déterminer au mieux les paramètres du modèle ?

2.2. MÉTHODE DES MOINDRES CARRES

- Soit y = f(x,a,b,c, ...) l'équation de la courbe que l'on cherche à ajuster au nuage statistique. Nous voudrions que les erreurs entre la valeur observée y_i et la valeur ajustée $f(x_i)$ soit minimale. Appelons e_i la différence : $e_i = y_i - f(x_i)$.

  e_i est le **résidu** de la $i^{ème}$ observation et sa valeur absolue représente la distance entre les points $M_i(x_i, y_i)$ et $P_i(x_i, f(x_i))$.

- Les résidus étant positifs ou négatifs, leur somme peut être de faible valeur pour une courbe mal ajustée. On évite cette difficulté en considérant la somme des carrés des résidus (la somme de valeurs absolues n'étant pas pratique pour des développements mathématiques).

- Cette somme $S(a,b,c,....) = \sum_{i=1}^{n} e_i^{\,2}$ dépend des paramètres a,b,c,... à ajuster . On choisira ces paramètres de manière qu'elle soit minimale. $\sum_{i=1}^{n} e_i^{\,2}$ est appelé **variation résiduelle** et nous donne une mesure de l'ampleur de l'éparpillement des observations y_i autour de la courbe d'ajustement.

2.3. CAS DU MODÈLE LINÉAIRE D'ORDRE UN

Dans ce cas la courbe d'ajustement sera une droite d'équation y = a + bx. Il nous faut déterminer les deux paramètres a et b.

$\Rightarrow$ La variation résiduelle s'écrit : $S(a,b) = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (y_i - a - bx_i)^2$

$\Rightarrow$ S(a,b) sera minimum lorsque : $\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = 0$

$\Rightarrow$ On obtient : $\begin{cases} \frac{\partial S}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \frac{\partial S}{\partial b} = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0 \end{cases}$

En distribuant l'opérateur $\sum$, il vient :

$$\begin{cases} \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^{\,2} = 0 \end{cases}$$

ce qui conduit ainsi à deux **équations** dites "**normales**":

$$\begin{cases} na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Nota: Ce système se généralise facilement aux modèles linéaires d'ordre n (courbes d'ajustement polynomiales à n paramètres). On obtient un système linéaire de n équations à n inconnues. L'utilisation des techniques matricielles facilite considérablement sa résolution.

En résolvant ce système, on obtient :

la pente de la droite $b = \dfrac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^{\,2} - (\sum_{i=1}^{n} x_i)^2}$

l'ordonnée à l'origine $a = \dfrac{(\sum_{i=1}^{n} x_i^{\,2})(\sum_{i=1}^{n} y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n\sum_{i=1}^{n} x_i^{\,2} - (\sum_{i=1}^{n} x_i)^2}$

$\Rightarrow$ Autres expressions pour a et b :

On a : $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ et $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$

Si on utilise le fait que : $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \dfrac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$

et que : $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^{\,2} - \dfrac{1}{n}(\sum_{i=1}^{n} x_i)^2$, l'écriture de b simplifie :

$$\boxed{b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (1)$$

et la première équation normale permet de déterminer a :

$$\boxed{a = \bar{y} - b\bar{x}} \qquad (2)$$

Remarques :

1. $\dfrac{\partial S}{\partial a} = \dfrac{\partial S}{\partial b} = 0$ ne sont que des conditions nécessaires de minimalité pour s. L'étude des dérivées secondes montre effectivement que les valeurs trouvées minimisent S(a,b)

2. L'équation (2) signifie que la droite d'ajustement passe par le point $(\bar{x}, \bar{y})$ appelé **point moyen du nuage.**

3. Cette droite des moindres carrés est appelée **droite de régression de y en x.** Elle est unique.

4. Si on avait cherché à exprimer la relation entre x et y par $x = a' + b'y$, on aurait obtenu la **droite de régression de x en y** qui minimise la somme des carrés des distances entre les points $M_i(x_i, y_i)$ et $Q_i(a' + b'y_i, y_i)$.

Historiquement, c'est sir Francis GALTON (1822-1911), cousin de Charles DARWIN , qui a introduit la notion de régression : Il a comparé la taille des enfants adultes à la taille moyenne de leurs parents (en multipliant celle des femmes par 1.08). En regroupant les données en classes et en représentant chaque classe par son centre, il a obtenu une relation presque linéaire dont la pente est une estimation de l'héritabilité et est d'environ 2/3. Ceci signifie que des parents très grands ont des enfants plus petits qu'eux, et que des parents très petits ont des enfants plus grands qu'eux. D'où une r**égression** vers la moyenne.

2.4. TRANSFORMATIONS SIMPLES PERMETTANT D'ÉTENDRE L'USAGE DE L'AJUSTEMENT LINÉAIRE

2.4.1. Schéma exponentiel

y et x sont liés par une relation du type : $y = y_0 \alpha^x$ (1).

On en déduit : $\ln y = \ln y_0 + x \ln \alpha$

En posant Y = lny, a = $\ln y_0$, b = lnα, on est ramené à la recherche des paramètres de la droite Y = a + bx qui représente (1) sur un graphique semi-logarithmique.

2.4.2. Schéma à élasticité constante :

Q et P sont liés par une relation du type $Q = AP^E$ (2)

avec : Q = quantité offerte d'un produit P = prix demandé

E = élasticité A = constante de normalisation

Ce schéma est courant en économie :

On tire de (2) la relation : $\ln Q = \ln A + E \ln P$

En posant Y = lnQ, X = lnP, a = lnA, on est ramené à la recherche des paramètres de la droite Y = a + EX qui représente la relation (2) sur un graphique à doubles coordonnées logarithmiques.

2.4.3. Schéma gaussien

Nous avons vu qu'il existe entre la valeur x d'une variable statistique distribuée normalement et sa fréquence cumulée y la relation : $y = \Pi(\frac{x-m}{\sigma})$ (3)

où Π est la fonction de répartition de la loi normale centrée réduite.

Cette relation peut s'écrire : $\Pi^{-1}(y) = \frac{x-m}{\sigma}$ et si l'on pose $t = \Pi^{-1}(y)$, elle devient $t = \frac{x-m}{\sigma}$.

La relation (3) est donc représentée par une droite, la droite de Henry (cf chapitre I), sur un papier gausso-arithmétique. m et σ , caractéristiques de la distribution normale peuvent alors être estimés par la méthodes des moindres carrés.

Conclusion : Cette méthode d'ajustement analytique est une méthode d'analyse numérique. Nous allons à présent la traiter sous l'angle statistique, en considérant d'abord, pour tout i entre 1 et n, y_i comme la réalisation d'une variable aléatoire Y_i, les x_i n'étant pas

aléatoires (analyse de régression), puis en considérant, pour tout i entre 1 et n, x_i et y_i comme les réalisations de deux variables aléatoires X_i et Y_i (problème de corrélation) .

3. LE MODÈLE DE RÉGRESSION LINÉAIRE SIMPLE

3.1. INTRODUCTION

Considérons un exemple.

Le directeur du personnel d'une compagnie de jouets, a découvert qu'il existe une relation logique et étroite entre le rendement fourni par les employés et le résultat obtenu à un test d'aptitudes qu'il a élaboré. Sur huit employés, il a obtenu les résultats suivants

Employés	A	B	C	D	E	F	G	H
Production (Y) (en douzaine d'unités)	30	49	18	42	39	25	41	52
Résultats au test d'aptitude (X)	6	9	3	8	7	5	8	10

Supposons de plus que ce directeur ait calculé, par la technique du paragraphe précédent, une équation d'estimation (l'équation de la droite de régression) pour prédire le rendement futur du candidat (la variable dépendante) en se fondant sur les résultats du test (la variable indépendante).

$$Y = 1.0277 + 5.1389X$$

On représente sur la figure ci-dessous le nuage de points et la droite de régression ainsi obtenus.

L'analyse de régression peut nous permettre de déterminer *le degré de fiabilité* des prédictions obtenues à l'aide de cette équation.

Au vu des résultats observés, il semble qu'il y ait une relation assez étroite entre les résultats au test et la productivité des employés. Mais les apparences sont parfois trompeuses. Qu'en serait-il si la population totale de l'ensemble des employés était répartie comme l'indique la figure (a) ci-contre ? Serait-il possible que, malgré l'échantillon obtenu, un tel diagramme représente l'ensemble des employés de la compagnie ? Si tel était le cas, il nous faudrait conclure qu'il n'existe pas de relation entre X et Y car dans une telle situation, la pente de la droite de régression pour la population (paramètre que nous représenterons par le symbole β) serait égale à 0.
En somme, il est possible que le directeur ait eu un coup de malchance et que l'échantillon qu'il a prélevé l'ait fortement induit en erreur.

Un tel cas n'est pas impossible. En effet, on pourrait représenter l'échantillon, perdu dans la population, sur la figure (b) ci-contre. La pente positive de la droite de régression échantillonnale nous indique une relation. On constate qu'alors le résultat obtenu grâce à l'échantillon est en contradiction avec la réalité de la population.

Comment conclure ? Il nous faut effectuer un calcul statistique pour constater qu'un tel cas n'est certes pas impossible mais fortement improbable. On pourra juger correctement grâce au calcul d'un intervalle de confiance, ou en effectuant un test d'hypothèses. On constatera que la vraie droite de régression (calculée à partir de la population toute entière) peut être, selon les données et la taille de l'échantillon, assez différente de la droite de régression obtenue à partir de l'échantillon mais que, statistiquement, elle se situe dans une région voisine, comme le montre la figure suivante.

Or pour pouvoir utiliser les techniques de l'analyse de régression linéaire simple en inférence statistique (intervalles de confiance, tests), il faut que certaines hypothèses soient vérifiées. Nous allons les préciser dans les deux paragraphes suivants.

3.2. DÉFINITION DU MODÈLE

Étant donné n couples d'observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, si l'on suppose que la relation plausible entre les deux grandeurs X et Y est linéaire et d'ordre un, alors le modèle de régression linéaire simple s'écrit :

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1,2 \ldots,n \quad \text{où}$$

- Y_i est la variable dépendante (ou expliquée) ayant un caractère aléatoire et dont les valeurs sont conditionnées par la valeur x_i de la variable explicative (ou contrôlée) X et par celles de la composante aléatoire ε_i. y_i représente la réalisation de la $i^{ème}$ variable aléatoire Y_i.
- x_i est la valeur de la variable explicative (ou régresseur) X mesurée sans erreur ou dont les valeurs sont fixées avant expérience à des valeurs arbitraires. Elle n'est pas susceptible de variations aléatoires et on la considère comme une grandeur certaine.
- ε_i dénote la fluctuation aléatoire non observable, attribuable à un ensemble de facteurs ou de variables non pris en considération dans le modèle. Cette fluctuation aléatoire n'est pas expliquée par le modèle et se reflète sur la variable dépendante Y_i

Conclusion : Pour chaque valeur particulière x_i prise par le régresseur X, la variable dépendante Y_i est une variable aléatoire caractérisée par :

- Une certaine distribution de probabilité.
- Des paramètres descriptifs : sa moyenne et sa variance.

3.3. CONDITIONS D'APPLICATION DU MODÈLE

Pour pouvoir étudier le modèle de régression linéaire simple, il faut préciser certaines hypothèses fondamentales qui assurent le fondement théorique des méthodes d'analyse que nous allons employer.

Hypothèses fondamentales du modèle linéaire simple : $Y_i = \alpha + \beta x_i + \varepsilon_i$

- La courbe joignant les moyennes des distributions des Y_i pour les différentes valeurs x_i est une droite. Dans ce cas l'équation de régression est de la forme : $E(Y_i) = \alpha + \beta x_i$. Comme nous savons que : $E(Y_i) = \alpha + \beta x_i + E(\varepsilon_i)$,on en tire que $E(\varepsilon_i) = 0$. Les erreurs aléatoires sont de moyenne nulle.

- La variance σ^2 de chaque distribution des Y_i est la même quelle que soit la valeur x_i prise par la variable explicative X. Pour tout i entre 0 et n, $\underline{Var(Y_i) = \sigma^2}$.Ceci est équivalent à dire que la variance des ε_i (c'est-à-dire des erreurs aléatoires) demeure constante et égale à σ^2 pour toutes les valeurs de X.
  $Var(Y_i) = Var(\alpha + \beta x_i) + Var(\varepsilon_i) = 0 + \sigma^2$.
  On suppose donc que l'ampleur de la dispersion de chaque distribution des Y_i est identique, quelle que soient les valeurs prises par la variable explicative X.

- Les Y_i sont des variables aléatoires indépendantes, c'est-à-dire que les observations de la variable expliquée ne dépendent pas des observations précédentes et n'influent pas sur les suivantes.

- La distribution des Y_i est une distribution normale . Ceci revient également à dire que les erreurs ε_i sont distribuées normalement.

Ces hypothèses fondamentales peuvent se résumer ainsi :

$Y_i \rightsquigarrow N(\alpha + \beta x_i, \sigma)$, $\varepsilon_i \rightsquigarrow N(0,\sigma)$, les variables Y_i étant indépendantes.

On peut schématiser ces hypothèses par la figure ci-dessous :

3.4. INFÉRENCE SUR LES PARAMÈTRES DU MODÈLE : ESTIMATIONS ET TESTS D'HYPOTHÈSE

- Les paramètres α et β sont appelés **paramètres de régression**. Ce sont des constantes, ordonnée à l'origine et pente de la droite de régression, que l'on pourrait calculer si le nuage de points englobait la population toute entière.
- Les valeurs a et b déterminées par l'ajustement analytique au paragraphe 2 permettent d'en donner une estimation. a et b sont même les meilleurs estimateurs possibles de α et β (en anglais, on utilise le terme BLUE (pour Best Linear Unbiaised Estimators).
- On considère que a et b sont les valeurs observées dans notre n-échantillon des variables aléatoires d'échantillon ,notées A et B, et définies par:

$$B = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \text{ où } \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ et } \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$A = \overline{Y} - B\overline{x}$$

Pour pouvoir effectuer des calculs statistiques, établir un intervalle de confiance ou exécuter un test statistique sur l'un ou l'autre des paramètres de régression α et β ,il nous faut connaître la distribution d'échantillonnage des deux variables A et B. Il faut donc en connaître la forme, la moyenne et la variance. C'est l'objet des deux paragraphes suivants.

3.4.1. Distribution d'échantillonnage de la variable B

- B est une combinaison linéaire de variables normales indépendantes. Elle est donc distribuée suivant une loi normale.

- $$E(B) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})[E(Y_i) - E(\overline{Y})]}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})[\alpha + \beta x_i - \alpha - \beta\overline{x}]}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\beta\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \beta$$

$\boxed{E(B) = \beta}$. Donc B est un estimateur sans biais de β.

- $$\mathrm{Var}(B) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 \mathrm{Var}(Y_i)}{[\sum_{i=1}^{n}(x_i - \overline{x})^2]^2} = \frac{\sigma^2\sum_{i=1}^{n}(x_i - \overline{x})^2}{[\sum_{i=1}^{n}(x_i - \overline{x})^2]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\boxed{Var(B) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

∗ **Cas des grands échantillons (n≥30)**

$$T = \frac{B - \beta}{\frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$ suit la loi normale N(0,1).

Si σ^2 n'est pas connu, on l'estime par $s^2 = \frac{\sum_{i=1}^{n}(y_i - a - bx_i)^2}{n-2}$ qui est la valeur prise dans l'échantillon de $S^2 = \frac{\sum_{i=1}^{n}(Y_i - A - Bx_i)^2}{n-2}$ estimateur non biaisé de σ^2 .

∗ **Cas des petits échantillons (n<30)**

$$T = \frac{B - \beta}{\frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$ suit une loi de Student à (n-2) degrés de liberté.

Remarque : On n'a que (n-2) degrés de liberté car dans S on a estimé les deux paramètres α et β par a et b. On peut trouver le même résultat en considérant qu'il y a deux relations entre A, B et Y_i .

3.4.2. Distribution d'échantillonnage de la variable A

Il est moins fréquent d'effectuer de l'inférence statistique concernant le paramètre α. Plusieurs situations ne comportent aucune valeur dans le voisinage de X = 0. De plus, dans certains cas, l'interprétation du paramètre α est dénuée de sens, dans d'autres, elle présente un certain intérêt. Donnons rapidement les résultats concernant la distribution d'échantillonnage de la variable A.

- De même que B, A a une distribution normale.
- On montre que E(A) = α , ce qui prouve que A est un estimateur sans biais de α.
- On montre également que $Var(A) = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]$.

Ici aussi, si σ est inconnu on l'estime par s.

Comme dans le cas précédent, lorsque nous aurons affaire à de petits échantillons et que σ sera inconnu on utilisera une loi de Student à (n-2) degrés de liberté.

3.4.3. Utilisation de ces lois

Grâce à la connaissance de ces lois, on pourra selon les cas :

⇒ obtenir des intervalles de confiance pour α et β ce qui nous permet de préciser la marge d'erreur de leurs estimations a et b.

⇒ effectuer un test d'hypothèse sur β, pente de la droite de régression.

Si par exemple, on veut déterminer si la dépendance linéaire entre X et Y est significative (en admettant que le modèle linéaire d'ordre 1 est plausible), il nous faudra savoir si la pente de la droite de régression est significativement différente de 0. Dans ce cas on dira que la composante linéaire permet d'expliquer d'une façon significative les fluctuations dans les observations de Y_i.

On testera alors l'hypothèse H_0 : $\beta = 0$
contre H_1 : $\beta \neq 0$

3.5. QUELQUES CONSIDÉRATIONS PRATIQUES DANS L'APPLICATION DES MÉTHODES DE RÉGRESSION

3.5.1. Extrapolation avec une équation de régression

Il faut être très prudent dans l'utilisation d'une équation de régression en dehors des limites du domaine étudié de la variable explicative. La droite de régression empirique est basée sur un ensemble particulier d'observations. Lorsque nous effectuons une prévision au delà des valeurs de X utilisées dans l'analyse de régression, nous effectuons une **extrapolation**.

Du point de vue purement statistique, toute extrapolation avec une équation de régression n'est pas justifiée puisqu'il n'est absolument pas évident que le phénomène étudié se comporte de la même façon en dehors du domaine observé. En effet la vraie fonction de régression peut être linéaire pour un certain intervalle de la variable explicative et présenter un tout autre comportement (du type curviligne par exemple) en dehors du champ observé.

Même si des considérations théoriques ou pratiques permettent de penser que l'équation de régression peut s'appliquer dans tout domaine, un autre inconvénient apparaît : la précision de nos estimations et de nos prévisions diminue à mesure que l'on s'éloigne de la valeur moyenne de la variable explicative , c'est-à-dire que la marge d'erreur augmente comme le montre la figure ci-après.

3.5.2. Relation de cause à effet

Le fait qu'une liaison statistique existe entre deux variables n'implique pas nécessairement une relation de cause à effet. Il faut s'interroger sur la pertinence de la variable explicative utilisée comme élément prédicteur de la variable dépendante et examiner s'il n'existe pas certains facteurs ou variables non incluses dans l'analyse et dont les variations provoquent sur les variables initiales de l'étude un comportement de régression illusoire.

3.5.3. Étude de la pertinence du modèle

Une manière simple de détecter les défaillances du modèle consiste à calculer les résidus donnés par la formule : $e_i = y_i - (a + bx_i)$ et surtout les résidus réduits $er_i = \frac{e_i}{s}$ où s est l'estimateur non biaisé de l'écart-type σ de la distribution des erreurs. Ces résidus réduits estiment les erreurs réduites $\frac{\varepsilon_i}{\sigma}$ qui sont distribuées normalement suivant une loi gaussienne centrée réduite.

Une étude systématique des résidus est un élément essentiel de toute analyse de régression. Un graphique de ces résidus révèle les gros écarts au modèle. On doit représenter le graphique des résidus en fonction de toute caractéristique qui peut avoir une action sur eux. Voici trois graphiques possibles qui paraissent naturels et qui permettent d'éviter bien des erreurs.

- Graphique des résidus er_i en fonction des valeurs ajustées $\hat{y}_i = a + bx_i$.
- Graphique des résidus er_i en fonction des valeurs x_i du régresseur.
- Graphique des résidus er_i dans leur ordre d'acquisition ; en effet le temps peut être la caractéristique essentielle dans de nombreuses études.

Si le modèle est correct, les résidus er_i doivent se trouver approximativement entre -2 et +2, étant entendu que leur moyenne est nulle. Ils ne doivent présenter aucune structure

particulière. Si jamais ils en présentent une, c'est que le modèle n'est pas complètement pertinent.

3.6. MESURE DE L'AJUSTEMENT : L'ANALYSE DE VARIANCE

Un autre objectif d'une étude de régression est de déterminer dans quelle mesure la droite de régression est utile à expliquer la variation existante dans les observations des Y_i . On veut donc évaluer la qualité de l'ajustement du modèle linéaire simple.

3.6.1. Analyse de la variance

- On rappelle que l'on note $\hat{y}_i$ la valeur estimée de y_i à l'aide de la droite de régression: $\hat{y}_i = a + bx_i$ .Pour chaque valeur y_i , l'écart total $(y_i - \bar{y})$ peut être décomposé en somme de deux écarts :

$(y_i - \bar{y})$	=	$(\hat{y}_i - \bar{y})$	+	$(y_i - \hat{y}_i)$
écart total	=	**écart expliqué** par la droite de régression lorsque $X = x_i$	+	**écart inexpliqué** par la droite de régression lorsque $X = x_i$: résidu

Cette décomposition peut être visualisée sur la figure ci-après.

- On peut donc exprimer la variation totale dans les observations de Y_i comme la somme d'une variation expliquée (attribuable à la droite de régression) et d'une variation inexpliquée (attribuable aux résidus). On démontre le résultat suivant :

$\sum_{i=1}^{n}(y_i-\bar{y})^2$	=	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	+	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$
Variation totale	=	**Variation expliquée** par la régression	+	**Variation résiduelle**

3.6.2. Le coefficient de détermination

Pour mieux apprécier la contribution de la variable explicative pour expliquer les fluctuations dans la variable dépendante on définit le **coefficient de détermination**, appelé

aussi **coefficient d'explication.** Ce nombre, noté r^2 , est la proportion de la variation totale qui est expliquée par la droite de régression.

$$r^2 = \frac{\text{variation expliquée}}{\text{variation totale}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

⇒ C'est un indice de la qualité de l'ajustement de la droite aux points expérimentaux.

⇒ Ce coefficient varie toujours entre 0 et 1. Cette propriété découle immédiatement de la définition.

⇒ $100r^2$ a une interprétation concrète : c'est le pourcentage de la variation de Y qui est expliquée par la variation de X.

⇒ On peut déduire de r^2 le **coefficient de corrélation linéaire simple** par :
$r = \pm\sqrt{r^2}$, le signe de r étant le même que celui de b, pente de la droite de régression.

On démontre qu'algébriquement : $$r = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(\hat{y}_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2} \sqrt{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}}$$

Nous définirons un coefficient analogue dans le paragraphe suivant qui concerne justement la corrélation pour mesurer l'intensité de la liaison linéaire entre deux variables .

4. CORRÉLATION LINÉAIRE

4.1. INTRODUCTION ET VOCABULAIRE

Nous prélevons d'une population un échantillon aléatoire de taille n et nous observons, sur chaque unité de l'échantillon les valeurs de deux variables statistiques que nous notons conventionnellement X et Y. On dispose donc de n couples d'observations (x_i ,y_i). On veut déterminer par la suite si les variations des deux variables sont liées entre elles, c'est à dire s'il y a **corrélation** entre ces deux variables.

L'existence de cette corrélation peut être déterminée graphiquement en traçant le nuage des points M_i (x_i ,y_i)($1 \leq i \leq n$) appelé **diagramme de dispersion**. La forme de ce nuage nous permettra de déceler le cas échéant, la nature de la liaison entre X et Y. Il nous renseignera sur la forme de la liaison statistique entre les deux variables observées ainsi que sur l'intensité de cette liaison.

- Dans ce cours, nous ne traiterons que de la forme linéaire. les points auront alors tendance à s'aligner selon une droite. On dit qu'il y a **corrélation linéaire**.

- Si Y croît en même temps que X, la **corrélation** est dite **directe** ou **positive**. Si Y décroît lorsque X croît, la **corrélation** est dite **inverse** ou **négative.**

Essayons d'associer aux différents nuages de points les conclusions qui s'y rattachent.

Conclusions : **(a)** Forte corrélation négative
(b) Absence de corrélation linéaire mais présence d'une liaison de forme parabolique.
(c) Faible corrélation linéaire mais présence d'une liaison de forme parabolique.
(d) Absence de corrélation et aucune liaison apparente. Indépendance entre ces deux variables.
(e) Corrélation positive marquée.

4.2. DÉFINITION DU COEFFICIENT DE CORRÉLATION LINÉAIRE

Définition : Le **coefficient de corrélation linéaire** , noté r est un nombre sans dimension qui mesure l'intensité de la liaison linéaire entre deux variables observées dans un échantillon de taille n.

On pose :

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

De même que le coefficient de détermination, r^2 représente le pourcentage de la variation de y expliquée par la variation de x.

Remarque : En raison de sa symétrie, r mesure aussi bien l'intensité de la liaison linéaire entre x et y qu'entre y et x.

4.3. PROPRIÉTÉS DU COEFFICIENT DE CORRÉLATION LINÉAIRE

Propriété 1 : On a toujours : $-1 \leq r \leq 1$

⇒ La corrélation parfaite , correspondant au cas $|r| = 1$, se rencontre très peu en pratique, mais sert de point de comparaison. Plus $|r|$ est proche de 1, plus les variables x et y seront étroitement liées.

⇒ Si x et y sont indépendantes, on a bien sûr r = 0. Mais la réciproque n'est pas nécessairement vraie. Si r = 0, on peut affirmer qu'il n'existe pas de liaison linéaire entre x et y. Mais il peut exister une liaison d'un autre type.

Exemple :

Propriété 2 : La droite d'ajustement linéaire de y en fonction de x dans notre échantillon a pour équation y = a + bx

avec $b = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$ donc $\boxed{b = r\dfrac{s_Y}{s_X}}$, s_X^2 et s_Y^2 représentant respectivement les variances de X et de Y dans l'échantillon des n points.

Propriété 3 : Le signe du coefficient de corrélation permet de savoir si la corrélation est positive ou négative puisque r et b sont de même signe.

Propriété 4 : Si la droite d'ajustement linéaire de y en fonction de x a pour équation y = a + bx et celle de x en y a pour équation x = a' + b'y ,

on a $b = r\dfrac{s_Y}{s_X}$ et aussi par symétrie $b' = r\dfrac{s_X}{s_Y}$. D'où : $\boxed{bb' = r^2}$

On retrouve ainsi que si la liaison entre x et y est forte, les pentes b et b' sont telles que $b' \approx \dfrac{1}{b}$ et les droites sont presque confondues.

4.4. COMMENT TESTER L'INDÉPENDANCE LINÉAIRE DE X ET Y

La question est de savoir si la valeur trouvée pour r est significativement différente de 0 ou pas. Le coefficient de corrélation r calculé à partir d'un échantillon de taille n donne une estimation ponctuelle du coefficient de corrélation de la population noté ρ. ρ est défini par :

$$\rho = E[(\frac{X - E(X)}{\sigma(X)})(\frac{Y - E(Y)}{\sigma(Y)})] = \frac{E(XY) - E(X)E(Y)}{\sigma(X)\sigma(Y)} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

- Appelons R la variable aléatoire : coefficient de corrélation de tous les échantillons de même taille n prélevés dans une même population : $R = (\frac{X - E(X)}{\sigma(X)})(\frac{Y - E(Y)}{\sigma(Y)})$.

  Bien sûr on a E(R) = ρ.

- On suppose que X et Y sont distribuées suivant une loi normale conjointe (loi binormale)

- Il s'avère que la distribution de R, coefficient de corrélation d'échantillon, ne dépend que de n et de ρ. On représente la fonction la fonction de densité de R sur la figure suivante pour n= 9, ρ = 0 et ρ = 0.8. On constate que la densité de R est symétrique pour ρ = 0 et c'est le seul cas où cette propriété est vérifiée.

- On pourra tester l'indépendance de X et Y en exécutant le test statistique suivant :

  ⇒ On teste <u>l'hypothèse H_0 contre l'hypothèse H_1</u> :
  Ho : X et Y ne sont pas corrélés linéairement c'est-à-dire $\rho = 0$.
  H_1 : X et Y sont corrélés linéairement c'est-à-dire $\rho \neq 0$.

  ⇒ On se place sous l'hypothèse H_0 , c'est-à-dire que l'on suppose que $\rho = 0$.
  La <u>fonction discriminante du test</u> est : $T = \dfrac{R - E(R)}{\sigma(R)} = \dfrac{R - \rho}{\sigma(R)} = \dfrac{R}{\sigma(R)}$
  On montre qu'elle peut s'écrire : $T = \dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$.
  T est distribuée suivant une loi de Student à (n-2) degrés de liberté.

  ⇒ On calcule la <u>valeur réduite</u> $t_0 = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, r étant la valeur du coefficient de corrélation de notre échantillon. Puis on cherche dans la table de Student $P_{t_0} = P(|T| \geq t_0)$.
  si $P_{t_0} < \alpha$: on rejette H_o : X et Y sont linéairement dépendants.
  si $P_{t_0} > \alpha$: on accepte H_o : X et Y sont linéairement indépendants.
  où α est le seuil de signification du test.

Remarques : **1)** Bien sûr, on constate que plus la taille de l'échantillon est petite, plus le coefficient de corrélation de l'échantillon doit être élevé pour être significativement différent de 0.

Ordre de grandeur au seuil de 5% : On accepte Ho pour $r < 0.378$ pour $n = 20$
On accepte Ho pour $r < 0.549$ pour $n = 10$

2) Si on veut tester d'autres valeurs de ρ , il faut employer une autre méthode et passer par ce qu'on appelle la transformation de FISCHER.

4.5. MISE EN GARDE A PROPOS DE LA CORRÉLATION

Il existe plusieurs explications du fait que deux séries varient en même temps, et dans certaines situations, l'analyste est en droit d'interpréter les mesures de corrélation dans le sens d'une relation de cause à effet. Cependant le seul fait que deux variables soient reliées entre elles, au sens statistique du terme, n'est pas suffisant pour conclure à l'existence d'une relation de cause à effet. En d'autres termes, la seule existence d'une relation statistique n'est pas une preuve de causalité.

Il existe une relation de **cause à effet** si la variation d'une variable est attribuable à la variation de l'autre variable. Par exemple la variation de température dans une réaction chimique peut causer une variation du résultat de cette réaction.

Par ailleurs, deux variables peuvent être conjointement influencées par un facteur de **cause commune**. Il est probable qu'il existe une relation étroite entre les ventes de bijoux et les ventes d'appareils stéréophoniques; toutefois, il serait farfelu d'imputer à l'une de ces variables les variations observées de l'autre variable. Ici, les variations des deux variables sont probablement le résultat d'une variation des revenus des consommateurs.

Certaines relations sont accidentelles. Même s'il existe une relation entre le nombre de voitures vendues au Canada et le nombre d'éléphants en Afrique, il serait absurde d'analyser le marché de l'automobile au Canada en se basant sur cette relation. Les relations de ce genre sont appelées **fausses corrélations**.