ECOLE DES MINES DE NANCY

Décision et Prévision Statistiques

Thierry Verdel (2007)

Chapitre 7

La régression linéaire

Les sciences exactes sont fondées sur la notion de ***relations répétables,*** *qui peut s'énoncer ainsi : dans les mêmes conditions, les mêmes causes produisent les mêmes effets. Notant alors x la mesure des causes, et y celle des effets, la liaison entre y et x s'écrit suivant la relation fonctionnelle* $y = f_c(x)$ *: à une valeur donnée de x correspond une valeur bien déterminée de y.*

Or, pour de nombreux phénomènes (notamment industriels), une étude exhaustive de tous les facteurs est impossible, à cause de leur grand nombre ou de leur complexité. Il en résulte que la reproductibilité des conditions, d'une expérience à une autre, ne peut être garantie. Partant de cette constatation, la statistique va permettre d'étendre la notion de relation fonctionnelle répétable, à celle de corrélation où la relation entre x et y est entachée d'une certaine dispersion due à la variabilité des conditions d'expérience : on écrira $y = f(x) + \varepsilon$*, où* ε *est une variable aléatoire.*

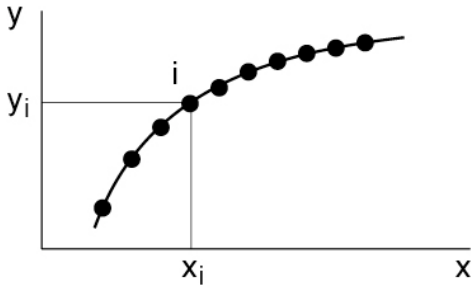
---

1. La droite des moindres carrés

1.1. Nuage des individus

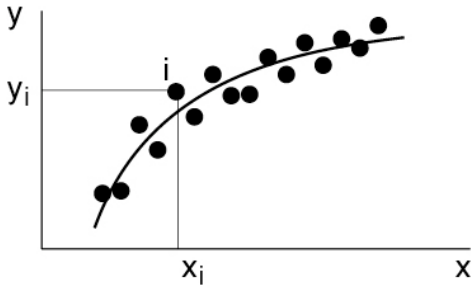
Le problème est d'étudier l'influence d'une variable quantitative X sur une autre variable quantitative Y. La première est souvent appelée variable *explicative* (ou encore exogène) et la seconde est appelée variable *expliquée* (ou encore endogène). Pour résoudre ce problème, on a réalisé une expérimentation qui consiste à prélever un échantillon de n individus, et à mesurer sur chacun d'eux les valeurs prises par chacune des deux variables. En vue, par exemple, d'étudier l'influence de la teneur en carbone d'un acier sur sa résistance à la traction, on a procédé à la mesure de ces deux variables sur 100 éprouvettes. On dispose donc d'un échantillon de n couples d'observations (x_i, y_i) que l'on peut représenter sur un graphique, dans le plan $\mathbb{R}^2$, où chaque point i, d'abscisse x_i et d'ordonnée y_i, correspond à un couple d'observations. Plusieurs cas peuvent se présenter.

Les points s'alignent sur une courbe qui, dans l'hypothèse la plus simple est une droite. On dit que la relation entre Y et X est *fonctionnelle* : lorsque la valeur de X est donnée, celle de Y est déterminée sans ambiguïté. C'est le cas idéal qui, expérimentalement, n'est jamais réalisé de façon parfaite.



(7.1)

Les mesures sont en effet toujours entachées de quelque imprécision. Les points forment alors un nuage. Mais celui-ci présente une orientation qui suggère, par exemple, que lorsque X augmente, la valeur moyenne de Y augmente également.

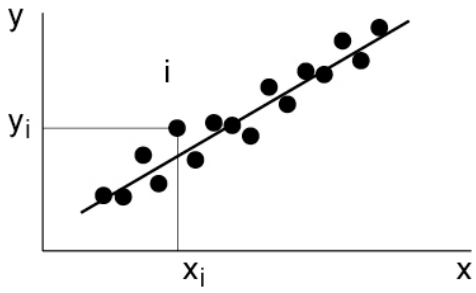


(7.2)

Lorsque X est donné, Y n'est pas complètement déterminé : ses valeurs se dispersent autour d'une certaine valeur moyenne. Mais les valeurs moyennes décrivent, lorsque X varie, une courbe qui est appelée la *ligne de régression* de Y par rapport à X :

$E(Y / X = x) = f(x)$.

La liaison entre Y et X est alors appelée *stochastique* (ou statistique). Un cas particulièrement important est celui où le nuage se dispose suivant une forme allongée et exhibe une tendance sensiblement linéaire. C'est à ce cas de *régression linéaire* que nous allons nous attacher dans ce chapitre.



(7.3)

Cette condition de linéarité n'est pas aussi restrictive qu'il pourrait paraître : une transformation mathématique appropriée permettra toujours de passer d'une ligne de régression d'équation quelconque à une droite de régression. Si la tendance est, par exemple, de la forme $y = b\,x^a$, il suffira d'effectuer les changements de variable $y' = \log(y)$ et $x' = \log(x)$ pour retrouver une relation linéaire : $\log(y) = a \log(x) + \log(b)$.

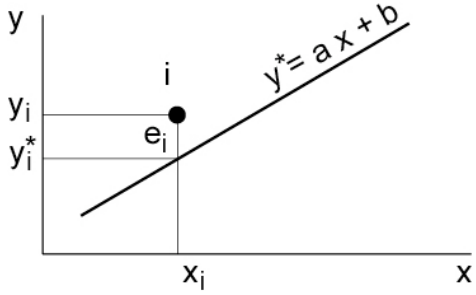
1.2. Caractérisation de la droite de régression

Nous cherchons une droite $y^* = a\,x + b$ qui décrive au mieux la tendance du nuage observé. La démarche la plus couramment utilisée consiste à :

- faire l'hypothèse que, pour chaque individu i, on a : $y_i = a\,x_i + b + e_i$, où e_i est une certaine « erreur », appelée *résidu*, qui s'ajoute à la valeur $y_i^* = a\,x_i + b$ qui résulterait d'une relation fonctionnelle linéaire entre Y et X,

- à rechercher la droite $y^* = a\,x + b$, qui est dite *droite des moindres carrés*, telle que la somme quadratique des résidus e_i soit minimale, c'est-à-dire que :

$S = \sum_{i=1}^{n} e_i^2$ soit minimale.



(7.4)

Cette quantité S s'écrit en fonction de a et b : $S = \sum_{i=1}^{n} (y_i - a\,x_i - b)^2$. Elle est minimale si les dérivées partielles par rapport à a et b sont nulles :

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^{n} x_i(y_i - a\,x_i - b) = -2 \sum_{i=1}^{n} x_i\, e_i = 0 \qquad (1)$$

$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a\,x_i - b) = -2 \sum_{i=1}^{n} e_i = 0 \qquad (2)$$

En appelant $\overline{y}$ et $\overline{x}$ les moyennes des valeurs x_i et y_i :

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \text{ et } \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

l'équation (2) permet d'obtenir la valeur de l'ordonnée à l'origine de la droite :

$$b = \overline{y} - a\,\overline{x},$$

ce qui signifie que la droite des moindres carrés passe par le point moyen du nuage, de coordonnées $(\overline{x},\ \overline{y})$, puisque son équation devient :

$$y^* - \overline{y} = a(x - \overline{x})$$

Le système des deux équations (1) et (2) devient alors, en remplaçant b par sa valeur :

$$\sum_{i=1}^{n} x_i[(y_i - \overline{y}) - a(x_i - \overline{x})] = 0 \qquad (3)$$

$$\sum_{i=1}^{n} [(y_i - \overline{y}) - a(x_i - \overline{x})] = 0 \qquad (4)$$

Multipliant (4) par $\overline{x}$ et retranchant de (3), il vient :

$$\sum_{i=1}^{n} (x_i - \overline{x})\,[(y_i - \overline{y}) - a(x_i - \overline{x})] = 0 \qquad (5)$$

d'où l'on déduit la valeur de a :

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

On peut noter que : $\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})$ est la covariance empirique de X et Y, et que : $\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$ est la variance empirique de X. Par conséquent, l'expression de a peut s'écrire :

$$a = \frac{s(X,Y)}{s^2(X)}$$

1.3. Analyse de la variance

Rapportant l'observation y_i à la moyenne $\overline{y}$ des observations, on peut écrire :

$$(y_i - \overline{y}) = a\,(x_i - \overline{x}) + e_i$$

Dans cette expression, la quantité $a\,(x_i - \overline{x})$ représente ce qui est « expliqué » par X, et la quantité e_i est une erreur qu'on a appellée un *résidu*.

En élevant au carré et en sommant pour toutes les observations, il vient :

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = a^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} e_i{}^2 \qquad (6)$$

En effet, le double produit est nul puisqu'il peut s'écrire :

$$2\,a \sum_{i=1}^{n} (x_i - \overline{x})[(y_i - \overline{y}) - a(x_i - \overline{x})],$$

forme sous laquelle on reconnait le premier membre de l'équation (5) précédente.

La relation (6) est appelée *équation d'analyse de la variance*. En fait, il s'agit de sommes de carrés : il faudrait diviser par n pour obtenir des variances. La relation (6) s'écrit souvent :

$$\text{SCT} = \text{SCE} + \text{SCR}$$

Elle décompose la somme des carrés *totale* SCT en une somme *expliquée* SCE et une somme *résiduelle* SCR.

1.4. Coefficient de corrélation

On définit alors le carré du coefficient de corrélation noté r^2 comme le ratio :

$$r^2 = \frac{\text{SCE}}{\text{SCT}}$$

Il représente donc la part relative de la variabilité totale de Y qui est expliquée par X :

$$\text{SCE} = r^2 \, \text{SCT}$$

Et, symétriquement, $(1 - r^2)$ représente la part résiduelle :

$$\text{SCR} = (1 - r^2) \, \text{SCT} \qquad (7)$$

En explicitant SCE et SCT puis a, on peut écrire :

$$r^2 = \frac{a^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{s^2(X,Y)}{s^2(X)\, s^2(Y)}.$$

Le coefficient de corrélation a le signe de la covariance :

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{s(X,Y)}{s(X)\, s(Y)}$$

de telle façon que, si X et Y varient dans le même sens, r est positif ; sinon, il est négatif.

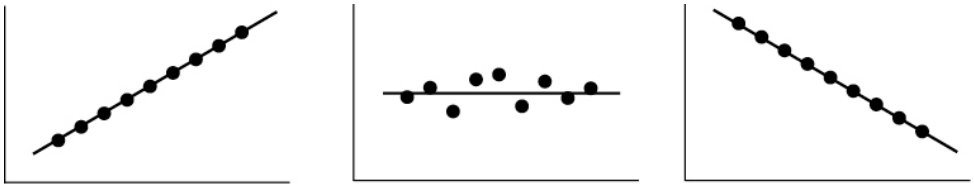
Il résulte de la relation (7) que le coefficient de corrélation est toujours compris entre -1 et 1, puisqu'une somme de carrés est nécessairement positive.

Le coefficient de corrélation présente les valeurs remarquables suivantes :

- si $|r| = 1$, il y a une relation fonctionnelle linéaire entre X et Y;

- si r = 0, Y est indépendante de X : la covariance est nulle et la droite de régression est horizontale.

- la liaison entre X et Y est d'autant plus intime que $|r|$ est voisin de 1, et d'autant plus faible que $|r|$ est voisin de 0 .



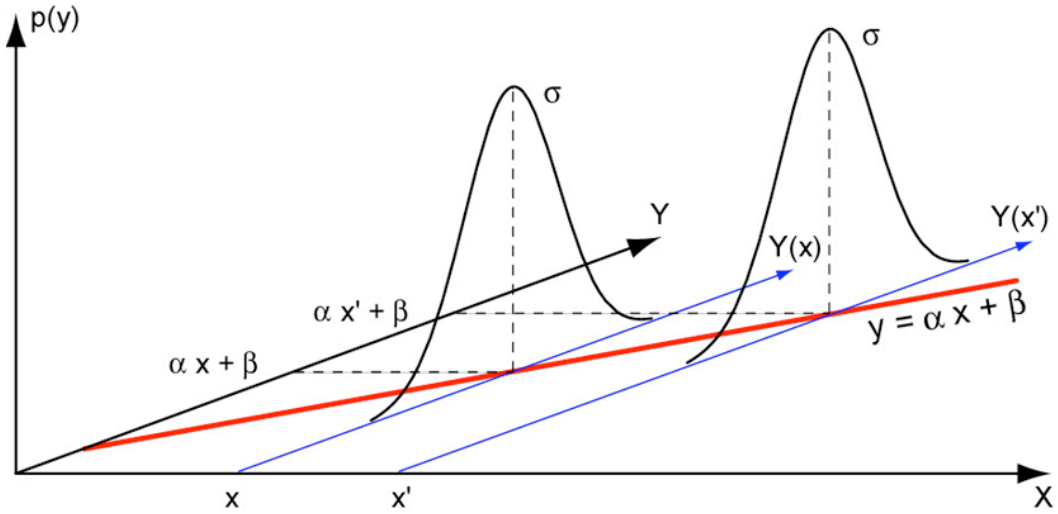
(7.5)

2. Propriétés statistiques de la droite des moindres carrés

2.1. Le modèle de la régression linéaire

Nous nous sommes jusqu'ici limités à décrire l'échantillon des valeurs observées (x_i, y_i) sans faire d'hypothèses sur la structure de la population dans laquelle il a été prélevé. Pour pouvoir pratiquer l'inférence, c'est-à-dire émettre des conclusions qui soient valables pour cette population, nous sommes obligés d'adopter un modèle de population.

Nous admettrons que, pour un individu i prélevé au hasard dans la population, x_i est *connu sans erreur*, et que y_i est une réalisation d'une variable aléatoire : $Y_i = \alpha x_i + \beta + \varepsilon_i$. Les paramètres α et β sont des quantités certaines, mais inconnues qu'il faudra estimer.



(7.6)

Les quantités ε_i sont des variables aléatoires avec les propriétés suivantes :

- elles sont centrées : $E(\varepsilon_i) = 0$,

- elles ont même variance : $E(\varepsilon_i^2) = \sigma^2$,

- elles sont indépendantes : $E(\varepsilon_i \varepsilon_j) = 0$ si $i \neq j$.

Pour une valeur donnée x_i, on a :

$$E[Y(x_i)] = \alpha x_i + \beta.$$

La ligne de régression est donc la droite d'équation $y = \alpha x + \beta$. La dispersion autour de cette droite correspond à un écart-type σ : elle est indépendante de X.

Rappelons que nous avions écrit, à partir de la droite des moindres carrés que :

$$y_i = a x_i + b + e_i$$

Sous les hypothèses ci-dessus, nous allons montrer que a et b sont des estimations sans biais de α et β et qu'il est possible d'estimer σ^2 à partir de SCR = $\sum_{i=1}^{n} e_i^2$.

2.2. Propriétés de *a* et *b*

Conformément au modèle adopté, a est à considérer comme une réalisation de la variable aléatoire :

$$A = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,(Y_i-\overline{Y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

et b comme une réalisation de la variable aléatoire :

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

A et B sont des *estimateurs sans biais et convergents* de α et β. Les calculs qui permettent de le montrer, peuvent être omis mais ils constituent toutefois un bon entrainement à la pratique des opérateurs " espérance mathématique " et " variance ".

Tenant compte de ce que $Y_i = \alpha\, x_i + \beta + \varepsilon_i$ *, on peut mettre A et B sous la forme :*

$$A = \alpha + \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,(\varepsilon_i-\overline{\varepsilon})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \alpha + \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,\varepsilon_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

$$B = \beta - \overline{x}(A-\alpha) + \overline{\varepsilon}$$

On en déduit tout d'abord les espérances mathématiques de A et B :

$$E(A) = \alpha + \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,E(\varepsilon_i)}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \alpha$$

$$E(B) = \beta - \overline{x}\,E(A-\alpha) + E(\overline{\varepsilon}) = \beta$$

On peut calculer ensuite les variances de A et B :

$$\sigma^2(A) = \frac{\sum_{i=1}^{n}(x_i-\overline{x}^2)\,\sigma^2(\varepsilon_i)}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

$$\sigma^2(B) = \overline{x}^2\,\sigma^2(A) + \sigma^2(\overline{\varepsilon}) - 2\,\sigma(\overline{\varepsilon},\,A) = \frac{\overline{x}^2\,\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2} + \frac{\sigma^2}{n}$$

puisque la covariance de $\overline{\varepsilon}$ *et A est nulle* :

$$\sigma(\overline{\varepsilon},\,A) = E\left[\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\,\frac{\sum_{i=1}^{n}(x_i-\overline{x})\,\varepsilon_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right] = \frac{1}{n}\sum_{j=1}^{n}\frac{\sum_{i=1}^{n}(x_i-\overline{x})\,E(\varepsilon_i\,\varepsilon_j)}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = 0$$

On peut enfin calculer la covariance de A et B :

$$\sigma(A,\,B) = E[(A-\alpha)\,(B-\beta)] = E[(A-\alpha)\,(\overline{\varepsilon}-\overline{x}(A-\alpha))] = \sigma(\overline{\varepsilon},\,A) - \overline{x}\,\sigma^2(A) = -\frac{\overline{x}\,\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

On constate ainsi que A et B sont des estimateurs de α et β : *sans biais* $(E(A) = \alpha,\ E(B) = \beta)$ et *convergents* $\left(\sigma^2(A) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \underset{n\to\infty}{\longrightarrow} 0 \text{ et } \sigma^2(B) = \left(\frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2} + \frac{1}{n}\right)\sigma^2 \underset{n\to\infty}{\longrightarrow} 0\right)$ mais qu'*ils ne sont pas indépendants* $(\sigma\,(A,\ B) \neq 0)$.

Par contre, A et $\overline{Y}$ sont *indépendants* puisque $\sigma(\overline{\varepsilon},\,A) = 0$. Ce résultat sera exploité un peu plus loin.

2.3. Estimation de σ^2

Montrons maintenant que :

$$\sigma^{*2} = \frac{\text{SCR}}{n-2} = \frac{\sum_{i=1}^{n} e_i^{\,2}}{n-2} = \frac{(1-r^2)\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-2}$$

est une *estimation sans biais* de σ^2.

Utilisant conjointement :

$$(y_i - \overline{y}) = a(x_i - \overline{x}) + e_i$$

$$(y_i - \overline{y}) = \alpha(x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon})$$

on peut écrire que :

$$\sum_{i=1}^{n} (\varepsilon_i - \overline{\varepsilon})^2 = (a-\alpha)^2 \sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} e_i^{\,2} + 2\,(a-\alpha) \sum_{i=1}^{n} (x_i - \overline{x})\, e_i$$

On en déduit que :

$$E(\sum_{i=1}^{n} e_i^{\,2}) = (n-1)\,\sigma^2 - \sigma^2 - 0 = (n-2)\,\sigma^2,$$

et finalement :

$$E(\sigma^{*2}) = \sigma^2$$

3. La prévision statistique

3.1. Objectifs

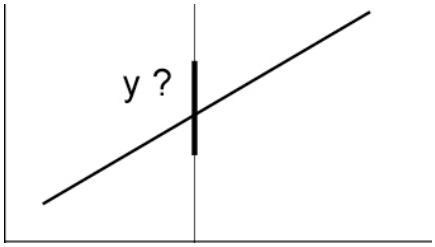
Dans une entreprise, on peut distinguer trois fonctions essentielles que nous allons brièvement illustrer par des exemples.

Décision : les performances d'un matériel dépendent de son âge. Au-dessous d'un certain seuil de performance, il convient de le réformer. Etant donné l'âge d'un matériel, il faudra décider de sa réforme ou de son maintien en activité.

Prévision : la consommation en matière première (ou en énergie) dépend de la quantité produite. Visant, pour une période future, une certaine production, quel stock de matière première faut-il prévoir ?

Contrôle : dans le même contexte, une certaine production ayant été assurée pour une certaine consommation, cette dernière est-elle « normale », faible, élevée ?

Ces trois problèmes se formulent finalement de la même façon. Pour une valeur donnée de X, quelle valeur attribuer à Y, et *avec quelle précision* ? D'un point de vue pratique, c'est l'objectif principal de ce qui suit.



(7.7)

Nous chercherons à apporter des réponses aux questions suivantes :

- la liaison entre les deux variables Y et X est-elle significative ? Autrement dit, peut-on ou non admettre que $\alpha = 0$?

- quels intervalles de confiance retenir pour les paramètres du modèle α et β ?

- pour une valeur donnée de X, comment estimer la valeur correspondante de Y ?

3.2. Hypothèse de normalité

La résolution de ces problèmes nécessite de compléter le modèle, en admettant pour les ε_i, outre les hypothèses précédentes (variance constante, indépendance), l'hypothèse de *normalité*.

Cette dernière hypothèse va permettre d'établir les lois de probabilité des estimateurs A et B, et celle d'un point quelconque $A\,x + B$. En effet, les x_i étant fixés, ces trois quantités sont des combinaisons linéaires des ε_i, donc suivent elles aussi des lois normales.

On peut montrer d'autre part que, sous l'hypothèse de normalité des ε_i, la quantité : $\frac{\sum_{i=1}^{n} {e_i}^2}{\sigma^2}$ suit une loi du χ^2 à $(n - 2)$ degrés de liberté, et qu'elle est indépendante des quantités A et B.

3.3. Test d'indépendance des variables

La variable aléatoire A suit une loi normale dont nous avons montré en 2.2 que l'espérance était égale à α et la variance à :

$$\sigma^2(A) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Faisons l'hypothèse qu'il n'y a pas de liaison entre les variables, c'est-à-dire que $\alpha = 0$. Il en découle que A suit une loi de moyenne nulle, donc que la quantité $\frac{A}{\sigma(A)}$ suit une loi normale centrée réduite.

Par suite, si on estime σ^2 par :

$$\sigma^{*2} = \frac{(1-r^2)\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2}$$

la quantité :

$$T = \frac{A}{\sigma^* \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

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

Il suffit de calculer sa valeur expérimentale :

$$t = \frac{a}{\sigma^* \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

et de la comparer au seuil lu dans la table correspondante pour le risque choisi.

Les mêmes propriétés permettent de calculer un intervalle de confiance pour α.

3.4. Test de nullité de l'ordonnée à l'origine

La variable aléatoire B suit une loi normale d'espérance égale à β, ordonnée à l'origine de la droite de régression $y(x) = \alpha x + \beta$, et nous avons montré en 2.2 que sa variance était égale à :

$$\sigma^2(B) = \left(\frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} + \frac{1}{n} \right) \sigma^2.$$

En procédant comme précédemment, on peut tester si la droite de régression passe par l'origine ($\beta = 0$), ou calculer un intervalle de confiance pour l'estimation de β.

3.5. Intervalles de confiance pour une valeur donnée *x* de *X*

Nous allons envisager successivement les intervalles de confiance :

- d'un point de la droite de régression $y(x) = \alpha x + \beta$, c'est-à-dire de la *valeur moyenne des observations* pour une valeur x donnée,

- puis d'un point du nuage $Y(x)$, c'est-à-dire de la *valeur d'une observation* pour une valeur x donnée.

Il est absolument nécessaire de bien prendre conscience de la différence fondamentale entre ces deux problèmes, dont les applications sont nombreuses et importantes : il s'agit dans le second cas de l'intervalle de confiance de $Y(x)$, alors que dans le premier cas il s'agit de l'intervalle de confiance de $E[Y(x)]$.

3.5.1. Intervalle de confiance d'un point de la droite $y = \alpha x + \beta$

L'équation de la droite de régression s'écrit :

$$y(x) = \alpha x + \beta.$$

Celle de la droite des moindres carrés s'écrit :

$$y^*(x) = a x + b$$

que l'on considère comme une réalisation de la variable aléatoire

$$Y^*(x) = A x + B.$$

D'après les résultats établis en 2.2, $E[Y^*(x)] = y(x)$. Donc le point $y^*(x)$ de la droite des moindres carrés est une estimation du point correspondant de la droite de régression $y(x)$.

Pour calculer maintenant la variance de $Y^*(x)$, on l'écrit sous la forme :

$$Y^*(x) = A(x - \overline{x}) + \overline{Y}$$

qui sera commode puisqu'on a vu que A et $\overline{Y}$ sont indépendantes. D'où la variance :

$$\sigma^2[Y^*(x)] = (x - \overline{x})^2 \sigma^2(A) + \sigma^2(\overline{Y}) = \left(\frac{(x-\overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} + \frac{1}{n} \right) \sigma^2$$

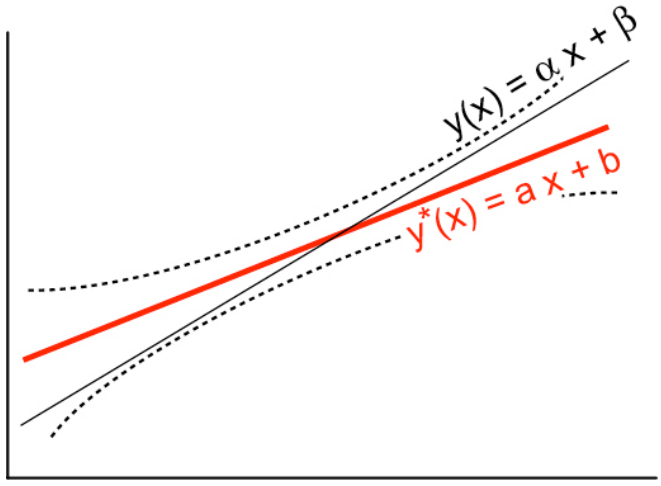
La quantité $\frac{Y^*(x)-y(x)}{\sigma[Y^*(x)]}$ suit une loi normale réduite. Et en estimant σ par σ^*, le quotient :

$$T = \frac{Y^*(x)-y(x)}{\sigma^* \sqrt{\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+\frac{1}{n}}}$$

suit une loi de Student-Fisher à $(n-2)$ degrés de liberté. Cette propriété permet de trouver un intervalle de confiance pour $y(x)$. Pour un risque α donné, on a :

$$y_\alpha(x) \in a\,x + b \pm t_{\alpha/2}\,\sigma^* \sqrt{\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+\frac{1}{n}}$$

Lorsque x varie, les limites $y_\alpha(x)$ décrivent une hyperbole. La droite de régression inconnue $y(x) = \alpha\,x + \beta$ se situe dans la zone comprise entre les deux branches de cette hyperbole.



(7.8)

3.5.2. Intervalle de confiance d'une observation

Un échantillon de n points a permis de déterminer les estimations a, b et σ^*. Nous cherchons à faire des prévisions sur l'ordonnée y_{n+1} d'un $(n+1)$ ème point d'abcisse x_{n+1} donnée. Cela revient à estimer y_{n+1} et à déterminer son intervalle de confiance.

Pour estimer y_{n+1} on prendra :

$$y^*(x_{n+1}) = a\,x_{n+1} + b$$

qui est une estimation sans biais. En effet :

- y_{n+1} est une réalisation de $Y_{n+1} = y(x_{n+1}) + \varepsilon_{n+1}$ avec $y(x_{n+1}) = \alpha\,x_{n+1} + \beta$,
- $y^*(x_{n+1})$ est une réalisation de $Y^*(x_{n+1}) = A\,x_{n+1} + B$

et, d'après 2.2, on a bien $E[Y_{n+1} - Y^*(x_{n+1})] = 0$.

Pour déterminer maintenant la précision de cette estimation, il faut caractériser l'erreur de prévision $(Y_{n+1} - Y^*(x_{n+1}))$ en calculant sa variance. Ecrivons pour cela que :

$$\begin{aligned} Y_{n+1} - Y^*(x_{n+1}) &= (Y_{n+1} - y(x_{n+1})) - (Y^*(x_{n+1}) - y(x_{n+1})) \\ &= \varepsilon_{n+1} - (Y^*(x_{n+1}) - y(x_{n+1})). \end{aligned}$$

Les deux quantités ε_{n+1} et $Y^*(x_{n+1})$ sont indépendantes puisque la seconde ne fait intervenir que les n premières observations, alors que la première concerne la $(n + 1)$ ème observation. Et, par conséquent, les variances s'ajoutent :

$$\sigma^2[Y_{n+1} - Y^*(x_{n+1})] = \sigma^2(\varepsilon_{n+1}) + \sigma^2[Y^*(x_{n+1})] = \sigma^2 + \left(\frac{(x_{n+1}-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2} + \frac{1}{n}\right)\sigma^2.$$

Il en résulte que la variable centrée, réduite :

$$T = \frac{Y_{n+1} - Y^*(x_{n+1})}{\sigma^* \sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}}$$

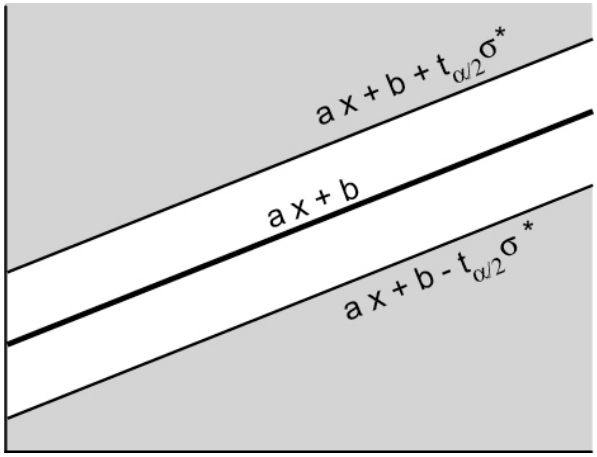
suit une loi de Student à $(n - 2)$ degrés de liberté, ce qui permet de trouver l'intervalle de confiance cherché.

On remarque, et c'est normal, que plus x_{n+1} est éloigné de $\overline{x}$, plus cet intervalle est grand. Il serait, de toute façon, illusoire et dangereux de prétendre faire des prévisions de $Y(x)$ pour des valeurs de x qui se trouveraient en dehors de l'intervalle de variation des données expérimentales ayant permis de calculer les relations sur lesquelles reposent ces prévisions.

En fait, on simplifie le plus souvent l'expression de l'intervalle de confiance d'une observation en notant que, si x_{n+1} n'est pas trop éloigné de $\overline{x}$, la quantité $(x_{n+1} - \overline{x})^2$ est généralement négligeable devant la quantité $\sum_{i=1}^{n}(x_i - \overline{x})^2$, et en admettant que n est suffisamment grand pour que l'on puisse négliger $\frac{1}{n}$ devant 1.

Dans ces conditions, la plage de confiance des observations, au risque α, est comprise entre les deux droites parallèles :

$$y = a\,x + b \pm t_{\alpha/2}\,\sigma^*$$



(7.9)

4. Comparaison de deux régressions

Soit deux groupes d'individus, sur lesquels ont été mesurées les valeurs de deux variables Y et X : n_1 individus pour le premier groupe, et n_2 pour le second.

Groupe 1		Groupe 2	
Y	X	Y	X
y_{11}	x_{11}	y_{12}	x_{12}
⋮	⋮	⋮	⋮
y_{i1}	x_{i1}	$y_{i'2}$	$x_{i'2}$
⋮	⋮	⋮	⋮
$y_{n_1 1}$	$x_{n_1 1}$	$y_{n_2 2}$	$x_{n_2 2}$

Désignons la droite des moindres carrés correspondant au premier groupe par :

$$y_1^* = a_1 x + b_1$$

et sa variance résiduelle estimée par ${\sigma_1}^{*2}$.

Désignons la droite des moindres carrés correspondant au second groupe par :

$$y_2^* = a_2 x + b_2$$

et sa variance résiduelle estimée par ${\sigma_2}^{*2}$.

La comparaison va porter successivement sur les *variances*, puis sur les *pentes* et, enfin, sur les *ordonnées à l'origine*. Les tests correspondants étant calqués sur ceux qui ont été mis en oeuvre pour la comparaison de deux populations, nous nous limiterons à leur principe.

4.1. Comparaison des variances

Le test à appliquer est celui de Snedecor, au quotient :

$$f = \frac{{\sigma_1}^{*2}}{{\sigma_2}^{*2}} ,$$

avec $(n_1 - 2)$ degrés de liberté au numérateur, et $(n_2 - 2)$ degrés de liberté au dénominateur.

Si l'hypothèse d'égalité des variances est acceptable (${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$), on peut adopter comme estimation de la variance commune la quantité :

$$\sigma^{*2} = \frac{(n_1-2)\,{\sigma_1}^{*2}+(n_2-2)\,{\sigma_2}^{*2}}{(n_1-2)+(n_2-2)}$$

4.2. Comparaison des pentes

a_1 est une réalisation de la variable aléatoire A_1 de moyenne α_1 et de variance $\frac{\sigma^2}{\sum_{i=1}^{n_1}(x_{1\,i}-\overline{x_1})^2}$.

a_2 est une réalisation de la variable aléatoire A_2 de moyenne α_2 et de variance $\frac{\sigma^2}{\sum_{i'=1}^{n_2}(x_{2\,i'}-\overline{x_2})^2}$.

Sous l'hypothèse $\alpha_1 = \alpha_2 = \alpha$ la variable aléatoire $(A_1 - A_2)$ suit une loi normale de moyenne nulle et de variance :

$$\sigma^2(A_1 - A_2) = \sigma^2\left(\frac{1}{\sum_{i=1}^{n_1}(x_{1\,i}-\overline{x_1})^2} + \frac{1}{\sum_{i'=1}^{n_2}(x_{2\,i'}-\overline{x_2})^2}\right)$$

Donc la variable aléatoire :

$$T = \frac{A_1 - A_2}{\sigma^* \sqrt{\frac{1}{\sum_{i=1}^{n_1} (x_{1\,i} - \overline{x_1})^2} + \frac{1}{\sum_{i'=1}^{n_2} (x_{2\,i'} - \overline{x_2})^2}}}$$

suit une loi de Student à $(n_1 + n_2 - 4)$ degrés de liberté, ce qui permet de tester l'égalité des pentes.

4.3. Comparaison des ordonnées à l'origine

La même démarche que ci-dessus permet d'établir que, sous l'hypothèse $\beta_1 = \beta_2 = \beta$, la variable aléatoire :

$$T = \frac{B_1 - B_2}{\sigma^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{\overline{x_1}^2}{\sum_{i=1}^{n_1} (x_{1\,i} - \overline{x_1})^2} + \frac{\overline{x_2}^2}{\sum_{i'=1}^{n_2} (x_{2\,i'} - \overline{x_2})^2}}}$$

suit une loi de Student à $(n_1 + n_2 - 4)$ degrés de liberté, ce qui permet de tester l'égalité des ordonnées à l'origine.

Exercices du chapitre 7

Exercice 1

On a relevé pour chacune des années t de 1920 à 1938, numérotées de 1 à 19, la température moyenne x des mois d'été (en degrés centigrades) et la mortalité infantile y (nombre de décès d'enfants de moins d'un an pour 1000 naissances vivantes).

t	1	2	3	4	5	6	7	8	9	10
x	15.9	18.8	15.4	18	14.6	16.2	17.9	16.5	18.1	19.8
y	98	116	87	96	85	89	97	83	91	95

... / ...

t	11	12	13	14	15	16	17	18	19
x	17.2	15.9	19	19.2	19	18.31	17.1	18.2	17.9
y	77	76	77	75	69	69	67	65	65

Après examen de ces chiffres et des graphiques auxquels ils peuvent donner lieu, indiquer les problèmes statistiques qu'ils vous paraissent poser et les calculs à faire pour les traiter.

Exercice 2

Le tableau ci-après donne les résultats d'un certain nombre de déterminations de la distance nécessaire (y en mètres) à l'arrêt par freinage d'une automobile lancée à différentes vitesses (x en km/h). Une étude graphique montre que la courbe représentant y en fonction de x est manifestement concave vers les y positifs, mais que si l'on utilise x^2 au lieu de x, la liaison apparaît sensiblement linéaire. Peut-on justifier ce fait par une loi physique ?

Admettant la validité de ce type de liaison entre y et x^2, on suppose de plus que la vitesse x peut être déterminée avec une grande précision et que les écarts constatés sont dûs à des fluctuations aléatoires de y autour d'une vraie valeur correspondant à une liaison linéaire représentée par l' équation $y = \alpha x^2 + \beta$.

Vitesse x	33	49	65	33	79	49	93
Distance y	5.3	14.45	20.26	6.5	38.45	11.23	50.42
x^2	1089	2401	4225	1089	6241	2401	8649

$\Sigma y = 146.61$ $\quad \Sigma x^2 = 26, 095$

$\Sigma y^2 = 4836.3019$ $\quad \Sigma x^4 = 145507351$

$\Sigma x^2 y = 836155.41$

a) Quelle est la meilleure estimation de α et β ? Quelle hypothèse supplémentaire suppose cette estimation ?

b) Déterminer les limites de confiance à 95% pour les estimations de α et β.

c) Considérant le cas d'une voiture dont la vitesse est de 85 km/h, estimer la valeur moyenne correspondante de y. En donner une limite supérieure au seuil de confiance 99%.

d) On suppose que pour une voiture se déplaçant à 85 km/h, on observe une distance de freinage $y = 55$ mètres. Cette valeur peut-elle être considérée comme étant, à des fluctuations aléatoires admissibles près, d'accord avec l'équation d'estimation trouvée ?

Exercice 3

On a déterminé sur une série de 18 coulées Thomas la température y du bain d'acier liquide à la fin de l'opération (à l'aide d'un pyromètre à immersion) et la température x du centre de la flamme (à l'aide d'un pyromètre à flamme) juste avant le rabattement du convertisseur. Le tableau ci-dessous donne les résultats obtenus. Les températures sont exprimées en degrés centigrades.

Bain y	1610	1590	1600	1600	1593	1570	1608	1580
Flamme x	1504	1490	1505	1495	1490	1475	1508	1480

... / ...

Bain y	1580	1592	1608	1612	1606	1595	1590	1597	1618
Flamme x	1480	1482	1510	1520	1510	1492	1485	1495	1515

a) Vérifier graphiquement que la régression de y en x peut être considérée comme linéaire.

b) Estimer l'équation de la droite de régression de y en x et l'écart-type de y lié par x. Avec quelle précision la température de la flamme permet-elle de connaître la température du bain d'acier dans les conditions des essais ?

c) Peut-on considérer que la différence entre y et x ne dépend pas de x ?

$$\Sigma x = 26931 \qquad \Sigma y = 28744$$

$$\Sigma x^2 = 40296243 \qquad \Sigma y^2 = 45903604$$

$$\Sigma x y = 43008448$$

Exercice 4

Les données ci-dessous sont relatives à des mesures de la limite élastique y et de la résistance à la traction x en MPa d'alliages d'or destinés à des prothèses dentaires.

x	1148	1638	1678	1292	1422	1285	1152	1357	867	1158	1082	907
y	724	1293	1296	925	1078	948	893	1077	550	870	669	517

... / ...

x	752	1115	1307	1528	1357	1405	1127	1073	1308	812	1260	1008	875
y	495	692	1014	1282	1007	978	849	670	953	497	798	657	580

On donne les résultats de calculs suivants :

$$m_x = 1196.52 \qquad m_y = 852.48$$

$$\Sigma (x - m_x)^2 = 1450472.24 \qquad \Sigma (y - m_y)^2 = 1451542.24$$

$$\Sigma (x - m_x)(y - m_y) = 1406707.76$$

a) En admettant que $E(Y/x) = \alpha x + \beta$, estimer α et β par la méthode des moindres carrés.

b) Calculer les intervalles de confiance à 95% de α et β.

c) Estimer la valeur moyenne de la limite élastique pour une résistance égale à 1290 MPa et calculer son intervalle de confiance à 90%.

d) Calculer l'intervalle de confiance à 90% pour la limite élastique correspondant à une résistance égale à 1290 MPa.

Exercice 5

Les données ci-dessous sont relatives à l'étalonnage d'une méthode gravimétrique pour le dosage de la chaux en présence de magnésium. La variable en x est la teneur vraie et la variable en y est la teneur mesurée (en mg).

Vraie x	20.0	22.5	25.0	28.5	31.0	35.5	33.5	37.0	38.0	40.0
Mesurée y	19.8	22.8	24.5	27.3	31.0	35.0	35.1	37.1	38.5	39.0

a) En admettant que $E(Y/x) = \alpha x + \beta$, estimer α et β par la méthode des moindres carrés.

b) Caractériser la précision de la méthode gravimétrique.

c) Tester l'hypothèse $\alpha = 1$ de telle façon que la probabilité d'accepter l'hypothèse si elle est vraie soit égale à 90%.

d) Tester l'hypothèse $\beta = 0$ de telle façon que la probabilité d'accepter l'hypothèse si elle est vraie soit égale à 90%.

e) Bâtir et mettre en oeuvre un test permettant de tester simultanément que $\alpha = 1$ et que $\beta = 0$, la probabilité d'accepter l'hypothèse si elle est vraie étant encore égale à 90%.

$\Sigma x = 311.0$ $\Sigma y = 310.1$

$\Sigma x^2 = 10100.00$ $\Sigma y^2 = 10055.09$

$\Sigma x y = 10074.80$

Exercice 6

L'étude d'une méthode de dosage d'un élément dans des aciers a montré que, pour des échantillons de poids différents d'un même produit, les poids y de l'élément à doser variaient bien linéairement avec le poids x de l'échantillon, mais que la droite obtenue ne passait pas par l'origine.

Afin de vérifier l'hypothèse selon laquelle la solution acide utilisée en quantité fixe pour attaquer les échantillons contiendrait elle-même une certaine quantité de l'élément à doser, on s'est proposé de comparer les ordonnées à l'origine des droite de régression de y sur x obtenues pour deux aciers différents.

On a utilisé pour chacun des deux aciers 5 échantillons de poids équidistants, les mêmes pour les deux aciers soit 1, 2, 3, 4, 5. Les résultats obtenus sont les suivants :

$x_1 = x_2 = 3$ $\quad n_1 = n_2 = 5$

$a_1 = 24.21$ $\quad b_1 = 12.11$ $\quad {\sigma_1}^{*2} = 9.637$

$a_2 = 8.82$ $\quad b_2 = 9.58$ $\quad {\sigma_2}^{*2} = 4.209$

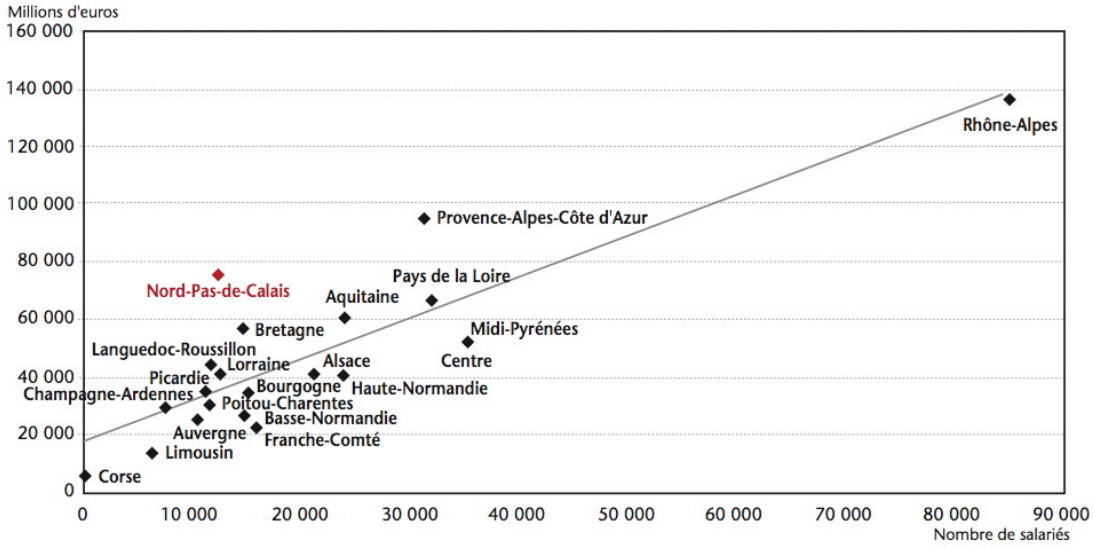
a) Comparer les variances résiduelles

b) Peut-on considérer que les pentes sont statistiquement égales ?

c) Peut-on considérer que les ordonnées à l'origine sont statistiquement égales ?

Exercice 7

La figure suivante indique, pour les 21 régions françaises de province et de métropole, le PIB (y) par région en fonction du nombre d'emplois (x) dans la haute technologie, pour l'année 2000 (source : INSEE Nord-Pas-de-Calais). Le nuage de points, de forme allongée, suggère l'existence d'une relation linéaire (figurée par la droite des moindres carrées) entre ces deux variables.



On donne par ailleurs les résultats intermédiaires suivants :

$\sum x$	$\sum y$	$\sum x^2$	$\sum y^2$	$\sum x\,y$
431200	992600	15078020000	64038160000	29144300000

a) Calculer les coefficients a et b, estimations des paramètres α et β de la relation linéaire ($\alpha\, x + \beta$) qu'on cherche à mettre en évidence.

b) La relation obtenue est elle significative au risque 5% ?

c) Pour 12000 emplois de haute technologie, quelle est l'espérance mathématique du PIB et son intervalle de confiance à 95 % ?

d) Dans cette étude, la région Nord-Pas-de-Calais affiche un PIB de 76 Milliards d'euros pour environ 12000 emplois de haute technologie. Que pensez de cette région par rapport aux autres ?

e) La région Nord-Pas-de-Calais ainsi que la région Provence-Alpes-Côte d'Azur sont en effet assez éloignées du modèle obtenu. Selon vous, quelles raisons structurelles propres à ces régions pourraient expliquer cet écart ?