Université de Picardie Jules Verne **2015-2016**
UFR des Sciences

Licence Sciences, Technologies, Santé toutes mentions - Semestre 2
Probabilités et Statistique

Régression - Droite des moindres carrés

Le chapitre précédent traitait de la **statistique descriptive univariée**, c'est-à-dire de la description d'une série statistique selon un seul caractère (la taille par exemple). On veut maintenant étudier, visualiser et mesurer (le cas échéant) les liens existant entre deux variables : c'est l'objet de la **statistique descriptive bivariée**.

On considère une population sur laquelle on étudie deux variables (ou caractères) quantitatives X et Y. On étudiera donc des **séries statistiques à deux variables** ; autrement dit un couple de variables (X,Y). On veut savoir si les deux variables sont liés par une liaison fonctionnelle du type $Y = f(X)$ (c'est-à-dire que l'on peut prévoir les valeurs de Y à partir des valeurs de X), ou $X = f(Y)$ (c'est-à-dire que l'on peut prévoir les valeurs de X à partir des valeurs de Y).

Précisons dès maintenant que l'existence d'une telle liaison entre les deux variables X et Y ne signifie pas obligatoirement un lien de cause à effet entre elles.

Exemple fondamental : $Y = aX + b$ (liaison fonctionnelle affine).

Représentation graphique : **nuage de points**.

Sur un échantillon de n individus extrait de la population, on observe n couples (x_1,y_1), $(x_2,y_2),\dots,(x_n,y_n)$ de valeurs de X et Y.

Ces observations peuvent être représentées dans le plan. A chaque couple (x_i,y_i), $i = 1,\dots,n$, on fait correspondre un point M_i. On obtient un nuage de point.

La forme du nuage obtenu peut indiquer le type de dépendance possible entre X et Y. Si les points sont "plutôt" alignés, on peut envisager une relation de type $Y = aX + b$ (équation de droite). Si le nuage "forme" une parabole, on peut envisager une relation de type $Y = aX^2 + bX + c$.

On dit que l'on cherche à ajuster une courbe au nuage de points.

1. Droite des moindres carrés (ou de régression) de y en x

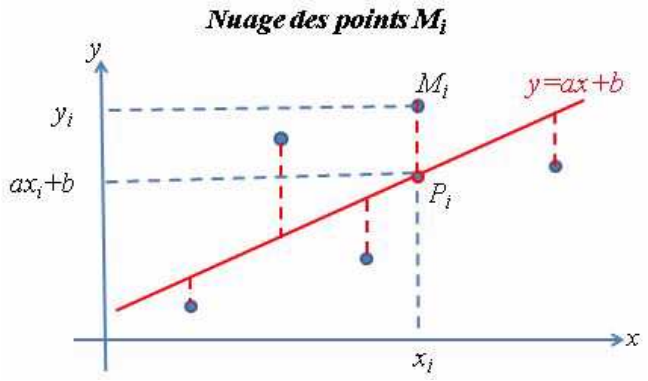
On cherche à ajuster une droite d'équation $y = ax + b$ au nuage de points.

Le critère d'ajustement est la distance totale entre les points du nuage $M_i(x_i,y_i)$ et les points $P_i(x_i, ax_i + b)$ correspondant sur la droite d'ajustement.

On cherche donc le couple $\left(\hat{a},\hat{b}\right)$ qui minimise $f(a,b) = \sum_{i=1}^{n}(y_i - (ax_i + b))^2$.

On peut démontrer (on l'admettra ici) qu'il existe un unique couple $\left(\hat{a},\hat{b}\right)$ rendant $f(a,b)$ minimum, et donc une seule droite répondant au problème.

C'est la droite des moindres carrés de y en x ; on dit aussi droite de régression de y en x.



Equation de la droite des moindres carrés de y en x :

$D_{y/x}$: $y = \hat{a}x + \hat{b}$, avec $\hat{a} = \dfrac{cov(x,y)}{s_x^2}$ et $\hat{b} = \bar{y} - \hat{a}\bar{x}$.

Notations :

Moyennes : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i$. *Covariance :* $cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}x_iy_i - \bar{x}\bar{y}$.

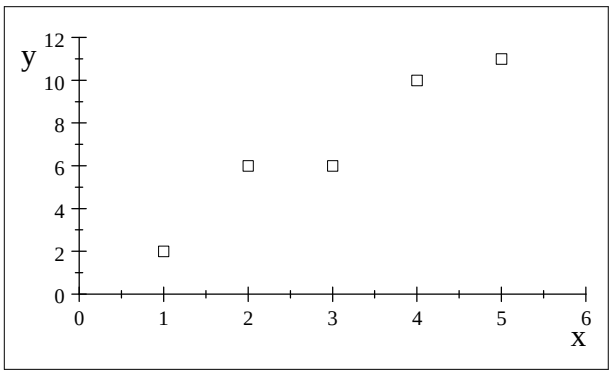
Variances : $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2, s_y^2 = \frac{1}{n}\sum_{i=1}^{n}y_i^2 - (\bar{y})^2$.

Exemple.

On considère la série double statistique suivante :

x_i	2	3	5	1	4
y_i	6	6	11	2	10

Le nuage de points correspondant est représenté sur le graphe ci-dessous.



Nuage de points

La droite de regression de y en x a pour équation : $y = \widehat{a}x + \widehat{b}$, avec $\widehat{a} = \dfrac{cov(x,y)}{s_x^2}$ et $\widehat{b} = \bar{y} - \widehat{a}\bar{x}$.

x_i	y_i	x_iy_i	x_i^2
2	6	12	4
3	6	18	9
5	11	55	25
1	2	2	1
4	10	40	16
Total	**15**	**35**	**127**	**55**

On a $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i = \dfrac{1}{5} \times 15 = 3$, $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \dfrac{1}{5} \times 35 = 7$,

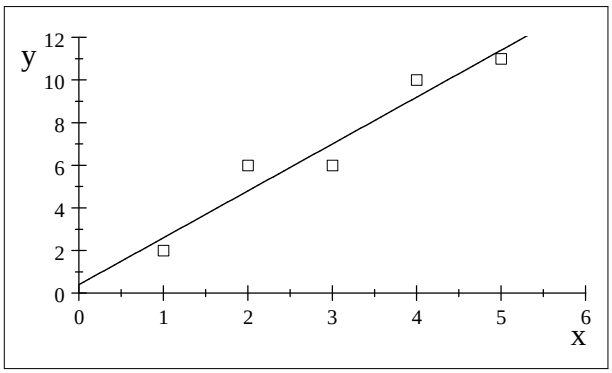
$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = \frac{1}{5} \times 127 - 3 \times 7 = 4,4\,,$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{1}{5} \times 55 - (3)^2 = 2.$$

On en déduit : $\widehat{a} = \dfrac{cov(x,y)}{s_x^2} = \dfrac{4,4}{2} = 2,2$

et $\widehat{b} = \bar{y} - \widehat{a}\bar{x} = 7 - 2,2 \times 3 = 0,4.$

La droite de regression de y en x a donc pour équation : $y = 2,2x + 0,4$.



Nuage de points et droite de régression de y en x

2. Droite des moindres carrés de x en y.

On suit une démarche analogue à celle qui a donné la droite des moindres carrés de y en x :

$$D_{y/x} : \ y = \widehat{a}x + \widehat{b}, \text{ avec } \widehat{a} = \frac{cov(x,y)}{s_x^2} \text{ et } \widehat{b} = \bar{y} - \widehat{a}\bar{x}.$$

On cherche à ajuster une droite $D_{x/y}$ d'équation $x = a'y + b'$ au nuage de points.
On obtient la droite des moindres carrés de x en y :

$$D_{x/y} : \ x = \widehat{a'}y + \widehat{b'}, \text{ avec } \widehat{a'} = \frac{cov(x,y)}{s_y^2} \text{ et } \widehat{b'} = \bar{x} - \widehat{a'}\bar{y}.$$

Remarque. Ces équations peuvent aussi s'écrire :

$$D_{y/x} : \ y - \bar{y} = \hat{a}(x - \bar{x})$$

$$D_{x/y} : \ x - \bar{x} = \widehat{a'}(y - \bar{y})$$

Les droites $D_{y/x}$ et $D_{x/y}$ se coupent donc au point $G(\bar{x},\bar{y})$.

Exemple.

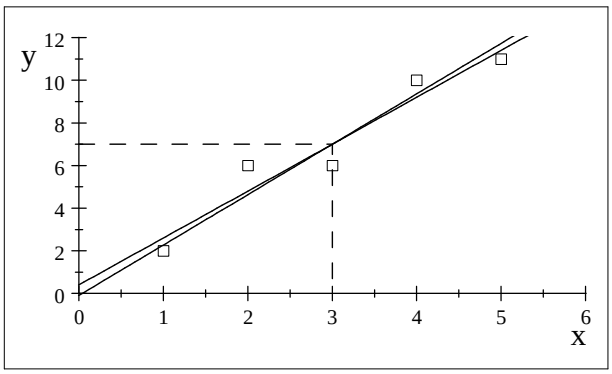
Reprenons l'exemple précédent. On a toujours $\bar{x} = 3$, $\bar{y} = 7$, $cov(x,y) = 4,4$, $s_x^2 = 2$ et $\hat{a} = 2,2$.

On calcule $s_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\bar{y})^2 = \frac{1}{5} \times 297 - (7)^2 = 10,4$, d'où $\widehat{a'} = \frac{cov(x,y)}{s_y^2} = \frac{4,4}{10,4} = \frac{1,1}{2,6}$.

La droite de regression de x en y a donc pour équation $x - \bar{x} = \widehat{a'}(y - \bar{y})$, soit $x - 3 = \frac{1,1}{2,6}(y - 7)$, c'est-à-dire $y = 2,3637x - 0,0909$.

On retrouve également une équation de la droite de régression de y en x : $y - \bar{y} = \hat{a}(x - \bar{x})$, soit $y - 7 = 2,2(x - 3)$, c'est-à-dire $y = 2,2x + 0,4$.

Les droites $D_{y/x}$ et $D_{x/y}$ se coupent au point $G(\bar{x},\bar{y}) = G(3,7)$.



Droites de régression de y en x et de x en y

3. Coefficient de corrélation linéaire entre x et y

Le coefficient de corrélation linéaire est donné par : $r_{x,y} = \frac{cov(x,y)}{s_x s_y}$.

Qualité de l'ajustement.

On peut démontrer que $r_{x,y}^2 = 1 - \frac{1}{s_y^2}\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{a}x_i - \hat{b}\right)^2 \leq 1$. On en déduit que $r_{x,y}^2 = 1$ si et seulement si $\sum_{i=1}^{n}\left(y_i - \hat{a}x_i - \hat{b}\right)^2 = 0$, c'est-à-dire $y_i - \hat{a}x_i - \hat{b} = 0$, pour tout $i = 1,\ldots,n$, soit $M_i(x_i,y_i) \in D_{y/x}$. Ainsi, $r_{x,y}^2 = 1$ si et seulement si les points M_i sont alignés sur $D_{y/x}$.

De façon générale, plus $r_{x,y}^2$ est proche de 1, meilleur est l'ajustement de la droite des moindres carrés au nuage de points. Dans la pratique, on dit qu'il y a une bonne corrélation linéaire entre X et Y si $\frac{\sqrt{3}}{2} \leq |r_{x,y}| \leq 1$, c'est-à-dire si $r_{x,y}^2 \geq \frac{3}{4}$.

Le signe de $r_{x,y}$ (même signe que celui de $\hat{a}$) indique le sens de la liaison (croissante si $r_{x,y} > 0$, décroissante si $r_{x,y} < 0$) entre X et Y.

Signification de $r_{x,y}$.

La question se pose de savoir si une forte valeur de $r_{x,y}$ (en valeur absolue) ou de $r_{x,y}^2$ prouve qu'il y a une forte corrélation entre les deux caractères X et Y (par exemple lorsque l'ajustement est bon) ou si elle est due au hasard de l'échantillonage (par exemple lorsque n est petit). Pour obtenir une réponse, on peut utiliser des tests statistiques (question non abordée ici).

Formule de décomposition.

La notion de liaison entre X et Y signifie qu'une variation de X entraîne une variation de Y. La formule de décomposition permet de préciser la part de variation de Y expliquée par la variation de X :

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\left(\widehat{y_i}-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2, \text{ avec } \widehat{y_i}=\hat{a}x_i+\hat{b}.$$

La démonstration repose sur le fait que le double produit s'annule : $\sum_{i=1}^{n}\left(\widehat{y_i}-\bar{y}\right)\left(y_i-\widehat{y_i}\right)=\hat{a}\sum_{i=1}^{n}(x_i-\bar{x})e_i=0$, avec $e_i=y_i-\widehat{y_i}$ (erreur observée), et grâce aux équations définissant $\hat{a}$ et $\hat{b}$.

La **somme des carrés totale** : $\sum_{i=1}^{n}(y_i-\bar{y})^2$ mesure la variation globale des y_i autour de leur moyenne $\bar{y}$.

La **somme des carrés expliquée** par la variable X : $\sum_{i=1}^{n}\left(\widehat{y_i}-\bar{y}\right)^2=\hat{a}^2\sum_{i=1}^{n}(x_i-\bar{x})^2$ mesure la variation de Y expliquée par la variable X. Ce terme n'est d'ailleurs fonction que de la pente de la droite des moindres carrés et des valeurs de X.

La **somme des carrés résiduelle** : $\sum_{i=1}^{n}\left(y_i-\widehat{y_i}\right)^2=\sum_{i=1}^{n}e_i^2$ mesure la variation de Y non expliquée par la variable X.

Coefficient de détermination.

Il est naturel de mesurer la force de la liaison entre les variables X et Y à l'aide du coefficient de détermination :

$$R^2=\frac{\sum_{i=1}^{n}\left(\widehat{y_i}-\bar{y}\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\text{somme des carrés expliquée}}{\text{somme des carrés totale}}$$

On peut vérifier que $R^2=r_{x,y}^2$. Ce qui explique que $r_{x,y}$ mesure la force de la liaison entre X et Y.

Position relative de $D_{y/x}$ et $D_{x/y}$.

Si $r_{x,y}^2=0$, alors $cov(x,y)=0$, et donc $\hat{a}=\widehat{a'}=0$. Ainsi, $D_{y/x}$: $y=\bar{y}$ et $D_{x/y}$: $x=\bar{x}$.

Si $r_{x,y}^2\neq 0$, alors $\hat{a}\neq 0$ et $\widehat{a'}\neq 0$. On a alors : $D_{y/x}$: $y-\bar{y}=\hat{a}(x-\bar{x})$ et $D_{x/y}$: $y-\bar{y}=\frac{1}{\widehat{a'}}(x-\bar{x})$.

On a : $\hat{a}\widehat{a'}=\frac{cov(x,y)}{s_x^2}\frac{cov(x,y)}{s_y^2}=\left(\frac{cov(x,y)}{s_xs_y}\right)^2=r_{x,y}^2$.

Si $r_{x,y}^2=1$, alors $\hat{a}=\frac{1}{\widehat{a'}}$ et donc $D_{x/y}$ a pour équation : $(y-\bar{y})=\frac{1}{\widehat{a'}}(x-\bar{x})=\hat{a}(x-\bar{x})$, soit l'équation de $D_{y/x}$. Ainsi, $D_{x/y}=D_{y/x}$.

Si $0<r_{x,y}^2<1$, alors $0<\hat{a}\widehat{a'}<1$ et deux cas sont possibles :

- soit $0<r_{x,y}<1$ et alors $\hat{a}>0$, $\widehat{a'}>0$ et $\hat{a}<\frac{1}{\widehat{a'}}$;
- soit $-1<r_{x,y}<0$ et alors $\hat{a}<0$, $\widehat{a'}<0$ et $\hat{a}>\frac{1}{\widehat{a'}}$.

4. Transformation de variable

Lorsque la corrélation linéaire entre x et y est mauvaise, l'ajustement d'une droite d'équation $y=ax+b$ au nuage de points n'est pas bon. En observant le nuage de points, on peut alors penser à d'autres types de relation entre x et y ; par exemple $y=\beta e^{\alpha x}$, $y=a\ln x+b$, ... Par transformation d'une des variables x ou y, ou des deux variables, on peut se ramener à une relation affine entre les variables transformées et utiliser les résultats précédents.

Sur ce sujet, voir les exemples traités en cours et les exercices 2 à 4.

5. Exercices

Exercice 1.

Dans la série statistique suivante, x représente le nombre de jours d'exposition au soleil d'une feuille et y le nombre de stomates aérifères au millimètre carré :

x	2	4	8	10	24	40	52
y	6	11	15	20	39	62	85

1) Représenter graphiquement le nuage de points correspondant.
2) Déterminer une équation de la droite de regression de y en x.
3) Calculer le coefficient de corrélation linéaire entre x et y. Commenter le résultat.
4) Quel nombre de stomates peut-on prévoir après 30 jours d'exposition au soleil ? après 60 jours ?

Exercice 2. (*D'après partiel de novembre 2007)*

Le tableau ci-dessous donne une estimation du montant des achats en ligne des ménages français :

Année	1998	1999	2000	2001	2002	2003	2004
Rang de l'année : x_i	0	1	2	3	4	5	6
Montant d'achats en millions d'euros : y_i	75	260	820	1650	2300	4000	5300

1) a) Préciser la population, la(es) variable(s) étudiée(s) et la taille de l'échantillon.
b) Donner une équation de la droite de régression de y en x.
c) Donner le coefficient de corrélation linéaire entre x et y. Interpréter le résultat obtenu.
d) Quelle prévision du montant d'achats peut-on faire pour l'année 2005 ? Est-elle fiable ?
2) On considère la nouvelle variable $z = \sqrt{y}$.

a) Déterminer une équation de la droite de régression de z en x, ainsi que le coefficient de corrélation linéaire entre x et z. Interpréter le résultat obtenu.

b) En déduire une expression de y en fonction de x, puis une prévision du montant d'achats pour l'année 2005.

3) A partir du tableau de données, le logiciel Excel propose un ajustement polynomial par l'équation $y = 130x^2 + 100x + 68$.
a) S'agit-il du même ajustement que celui obtenu dans le 2) ? Expliquer cette situation.
b) Déduire de cet ajustement une prévision du montant d'achats pour l'année 2005.

4) Le montant des achats en ligne en 2005 a été de 7700 millions d'euros. Lequel des trois ajustements précédents vous parait-il le plus conforme à la réalité ? Jutifier votre réponse.

Exercice 3. (*D'après examen de mai 2013)*

Les résultats numériques et les coefficients demandés seront donnés avec trois décimales.

Le tableau ci-dessous donne la fréquentation des lignes aériennes, en millions de passagers, entre la France métropolitaine et les pays étrangers depuis 1980 (source INSEE).

Année	1980	1985	1990	1995	2000	2005	2008
Rang de l'année : x_i	0	5	10	15	20	25	28
Nombre de passager (en millions) : y_i	21,9	26,4	36,9	44,7	67,0	82,0	97,9

On cherche à étudier l'évolution du nombre de passagers y entre la France métropolitaine et les pays étrangers en fonction du rang x de l'année.

1) a) Représenter graphiquement la série statistique (x_i, y_i).
b) Quelle(s) courbe(s) d'ajustement suggère cette représentation graphique ? Justifier la réponse.
2) Un premier ajustement.

a) Donner une équation de la droite de régression de y en x (obtenue par la méthode des moindres carrées).

b) Donner le coefficient de corrélation linéaire entre x et y. Interpréter le résultat obtenu.

3) Un deuxième ajustement. On considère la nouvelle variable $z = \ln y$.

a) Donner une équation de la droite de régression de z en x, et le coefficient de corrélation linéaire entre x et z. Interpréter le résultat obtenu.

b) En déduire une nouvelle expression de y en fonction de x.

c) En utilisant ce nouveau modèle, déterminer une estimation du nombre de passagers pour l'année 2011.

d) Les compagnies aériennes prévoient que le pourcentage d'augmentation entre 2008 et 2011 sera de 30%. Cela est-il cohérent avec ce deuxième ajustement ?

Exercice 4. (*D'après partiel de novembre 2009)*

Le tableau ci-dessous donne les valeurs expérimentales du volume V et de la pression P d'un gaz.

Volume (en cm^3) : v_i	620	890	1013	1186	1454	1944	2313	3179
Pression (en Kg par cm^3) : p_i	6.7	4.3	3.48	2.644	1.997	1.35	1.1	0.7

D'après les lois de la thermodynamique de Laplace pour un gaz parfait, on a la relation $PV^\gamma = C$, où γ et C sont des constantes.

1) Préciser la population, la(es) variable(s) étudiée(s) et la taille de l'échantillon.

2) On considère les variables $X = \ln V$ et $Y = \ln P$. Démontrer que $Y = -\gamma X + \ln C$.

Le tableau ci-dessous donne les valeurs expérimentales transformées :

$x_i = \ln v_i$	6,430	6.791	6.921	7.078	7,282	7,573	7,746	8,064
$y_i = \ln p_i$	1,902	1,459	1,253	0,956	0,693	0,336	0,095	–0,357

3) Donner une équation de la droite de régression de y en x. Donner le coefficient de corrélation linéaire entre x et y. Interpréter le résultat obtenu.

4) En déduire, en justifiant, la valeur de γ et de C, puis une équation de P en fonction de V.

5) Déterminer une estimation de la pression du gaz pour un volume de 2000 cm^3, puis pour 4000 cm^3. Ces deux estimations sont-elles fiables ?

Exercice 5.

On sélectionne 12 personnes inscrites à un stage de formation. Avant le début de la formation, ces stagiaires subissent une épreuve A notée de 0 à 20. A l'issue du stage, une épreuve B identique à la première est aussi notée de 0 à 20. Considérant les deux variables X = note de A et Y = note de B, on a obtenu les résultats suivants :

stagiaire	1	2	3	4	5	6	7	8	9	10	11	12
x_i	3	4	6	7	9	10	9	11	12	13	15	4
y_i	8	9	10	13	15	14	13	16	13	19	6	19

1) a) Représenter ces résultats par un nuage de points.

b) Quelle courbe d'ajustement ce nuage vous suggère-t-il ?

2) A partir des résulats obtenus, on a déterminé la droite de régression de y en x, ainsi le coefficient de corrélation linéaire entre x et y. On a obtenu l'équation $y = 0,108\,x + 11,990$ et le coefficient $r = 0,101$.

A partir de ces résultats, expliquer pourquoi l'ajustement n'est pas bon.

3) On décide d'éliminer les stagiaires 11 et 12, et donc de ne tenir compte que des stagiaires 1 à 10.

a) Déterminer une équation de la droite de régression de y en x par la méthode des moindres carrés.

b) Calculer le coefficient de corrélation linéaire entre x et y. Interpréter le résultat obtenu.

Exercice 6.

On a procédé à l'ajustement affine d'un nuage de points. Les équations obtenues sont les suivantes :

- droite d'ajustement de y en x : $D : y = x + 30$
- droite d'ajustement de x en y : $D' : x = 1/4y + 60$

1) Calculer le coefficient de corrélation linéaire.

2) Calculer les moyennes arithmétiques de x et de y.

3) Calculer la covariance entre x et y et la variance de x, sachant que la variance de y est égale à 40.