TD de Statistiques - Séance N°3
Corrélation et régression linéaires

1 Corrélation linéaire

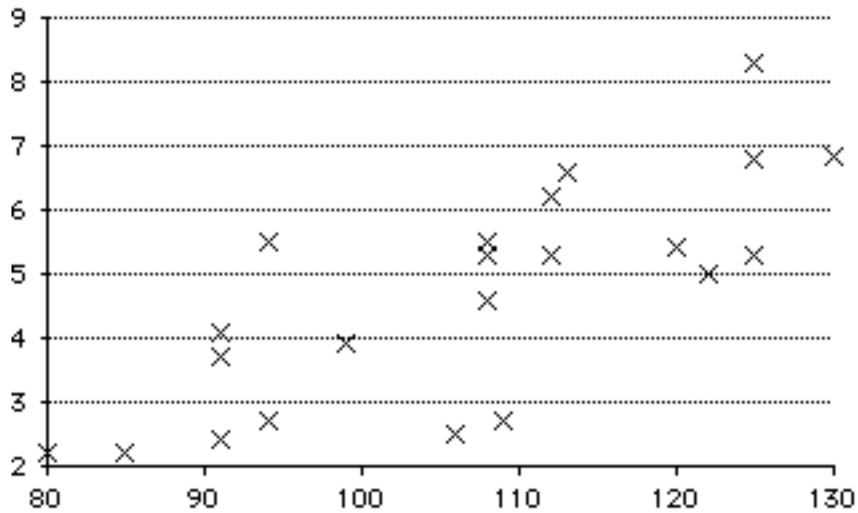
1.1 Statistiques bivariées : nuage de points

On se place dans la situation suivante : on a observé deux variables numériques sur une population ou un échantillon de sujets.

Données :

	X	Y
s_1	x_1	y_1
s_2	x_2	y_2
...	...	...

La situation peut être représentée graphiquement par un nuage de points : on place dans un repère les points de coordonnées (x_i,y_i).



1.2 Covariance et coefficient de corrélation de Bravais-Pearson

Le lien entre les variables peut être évalué à l'aide de la covariance des variables X et Y et de leur coefficient de corrélation.

Covariance des variables X et Y

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

ou

$$\mathrm{Cov}(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\,\overline{y}$$

La covariance dépend des domaines de variation et des unités des variables X et Y. Sa valeur est donc difficile à apprécier. C'est pourquoi on préfère généralement utiliser un paramètre sans unités : le coefficient de corrélation.

Coefficient de corrélation de Bravais Pearson

C'est le quotient de la covariance par le produit des écarts-types.

On désigne par $s(X)$ et $s(Y)$ les écarts types de X et Y. Le coefficient de corrélation est donné par :

$$r = \frac{\mathrm{Cov}(X,Y)}{s(X)s(Y)}.$$

Remarques

- r est un coefficient sans unités, compris entre -1 et 1.

- Les valeurs $r = -1$ et $r = 1$ correspondent à une relation fonctionnelle, déterministe (non statistique) entre X et Y.

- La valeur $r = 0$ correspond à l'*indépendance* des variables X et Y : la connaissance de l'une des variables ne renseigne absolument pas sur les valeurs possibles de l'autre.

- Une valeur positive de r indique que les variables X et Y varient dans le même sens : les grandes valeurs de X sont plutôt associées à de grandes valeurs de Y. Au contraire, une valeur négative de r indique que les variables X et Y varient en sens contraires : les grandes valeurs de X sont plutôt associées à de petites valeurs de Y.

- r mesure l'intensité de la relation entre X et Y, *lorsque cette relation est linéaire*. Mais, il existe des relations non linéaires.

- Corrélation n'est pas causalité.

1.2.1 Exemple

	X	Y	X^2	Y^2	XY
s_1	3	7	9	49	21
s_2	4	6	16	36	24
s_3	7	13	49	169	91
s_4	6	12	36	144	72
s_5	5	8	25	64	40
Σ	25	46	135	462	248

Moyenne de X : $\overline{X} = \frac{25}{5} = 5$

Moyenne de Y : $\overline{Y} = \frac{46}{5} = 9{,}2$

Variance de X : $s^2(X) = \frac{135}{5} - 5^2 = 2$

Ecart type de X : $s(X) = \sqrt{2} = 1{,}41$

Variance de Y : $s^2(Y) = \frac{462}{5} - 9{,}2^2 = 7{,}76$

Ecart type de Y : $s(Y) = \sqrt{7{,}76} = 2{,}66$

Covariance de X et Y : $\text{Cov}(X,Y) = \dfrac{248}{5} - 5 \times 9{,}2 = 3{,}6$

Coefficient de corrélation de X et Y : $r = \dfrac{3{,}6}{1{,}41 \times 2{,}66} = 0{,}91$

1.3 Une application de la corrélation linéaire : alpha de Cronbach

1.3.1 Définition du coefficient

Dans un questionnaire, un groupe d'items $X_1, X_2, ..., X_k$ (par exemple des scores sur des échelles de Likert) mesure un même aspect du comportement.

Problème : comment mesurer la cohérence de cet ensemble d'items ?

Dans le cas de 2 items : la cohérence est d'autant meilleure que la covariance entre ces items est plus élevée. Le rapport :

$$\alpha = 4\frac{\text{Cov}(X_1,X_2)}{\text{Var}(X_1 + X_2)}$$

- vaut 1 si $X_1 = X_2$
- vaut 0 si X_1 et X_2 sont indépendantes
- est négatif si X_1 et X_2 sont anti-corrélées.

Généralisation :

On introduit S=$X_1 + X_2 + ... + X_k$ et on considère le rapport :

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\text{Var}(X_1) + \text{Var}(X_2) + ... + \text{Var}(X_k)}{\text{Var}(S)}\right]$$

Ce rapport est le coefficient α de Cronbach.

Signification de α : on considère généralement que l'on doit avoir α ≥ 0,7 pour qu'il soit pertinent de considérer la somme ou la moyenne des scores sur les différents items. Mais une valeur trop proche de 1 révèle une pauvreté dans le choix des items.

1.3.2 Exemple :

On a mesuré 3 items (échelle en 5 points) sur 6 sujets.

	X_1	X_2	X_3	S
s_1	1	2	2	5
s_2	2	1	2	5
s_3	2	3	3	8
s_4	3	3	5	11
S_5	4	5	4	13
s_6	5	4	4	13
Var.	2,167	2,000	1,467	13,77

On obtient : $\alpha = \dfrac{3}{2}\left[\dfrac{2{,}167 + 2 + 1{,}467}{13{,}77}\right] = 0{,}886$.

2 Régression linéaire

La situation envisagée est différente de celle du paragraphe précédent. On s'intéresse au rôle "explicatif" de l'une des variables par rapport à l'autre : les variations de Y peuvent-elles (au moins en partie) être expliquées par celles de X ? Peuvent-elles être prédites par celles de X ?

Autrement dit, on recherche un modèle permettant d'estimer Y connaissant X.

2.1 Droite de régression de Y par rapport à X

Le meilleur modèle linéaire (c'est-à-dire sous la forme d'une équation $y = ax + b$) au sens "des moindres carrés est la droite de régression de Y par rapport à X.

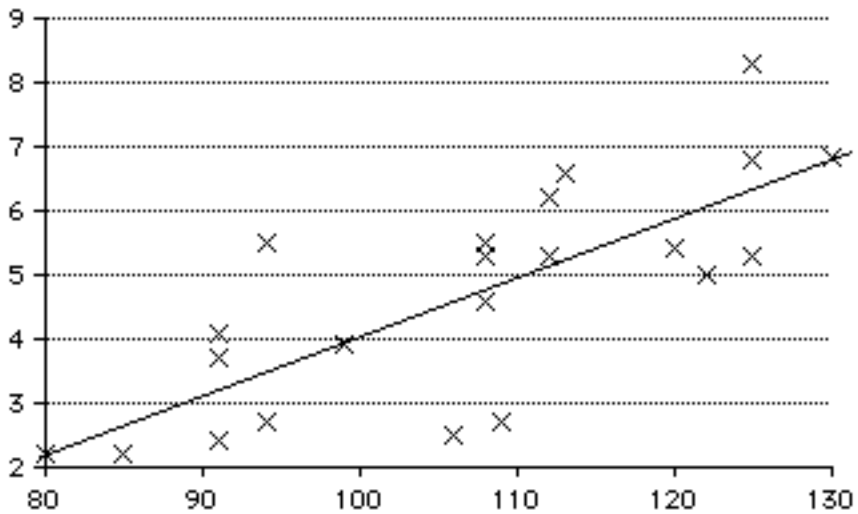
Cette droite a pour équation :

$$y = b_0 + b_1 x$$

avec :

$$b_1 = \frac{\mathrm{Cov}(X,Y)}{s^2(X)} \quad ; \quad b_0 = \overline{Y} - b_1 \overline{X}$$

N.B. dans la formule précédente, $\overline{X}$ et $\overline{Y}$ désignent les moyennes des variables X et Y, $s^2(X)$ désigne la variance (carré de l'écart type) de X.



Remarques :

Si les variables X et Y sont centrées et réduites, l'équation de la droite de régression est : $Y = r\,X$

On définit le *coefficient de régression standardisé* par :

$$\beta_1 = b_1 \frac{s(X)}{s(Y)}$$

Dans le cas de la régression linéaire à deux variables (régression linéaire simple) : $\beta_1 = r$.

2.2 Comparaison des valeurs observées et des valeurs estimées

Le modèle fourni par l'équation de régression linéaire permet de définir une nouvelle variable statistique $\hat{Y}$ dont les valeurs sur les individus statistiques sont données par : $\hat{y}_i = b_0 + b_1 x_i$

La variable E "erreur" ou "résidu" est définie quant à elle par : $e_i = y_i - \hat{y}_i$

Les variables $\hat{Y}$ et E sont indépendantes et on montre que :

$$s^2(Y) = s^2\left(\hat{Y}\right) + s^2(E)$$

avec :

$$\frac{s^2\left(\hat{Y}\right)}{s^2(Y)} = r^2 \quad ; \quad \frac{s^2(E)}{s^2(Y)} = 1 - r^2$$

$s^2\left(\hat{Y}\right)$ est la variance expliquée (par le modèle, par les variations de X), $s^2(E)$ est la variance perdue ou résiduelle.

r^2 est la part de la variance de Y qui est expliquée par la variance de X. r^2 est appelé coefficient de détermination.

2.2.1 Exemple

Sur notre mini-exemple, les coefficients de la droite de régression sont :

$$b_1 = \frac{3,6}{2} = 1,8 \quad \text{et} \quad b_0 = 9,2 - 1,8 \times 5 = 0,2$$

L'équation de la droite de régression est : $y = 0,2 + 1,8\,x$.

Valeurs observées, valeurs prévues et résidus:

	X	Y	$\hat{Y}$	E
s_1	3	7	5,6	1,4
s_2	4	6	7,4	-1,4
s_3	7	13	12,8	0,2
s_4	6	12	11	1
s_5	5	8	9,2	-1,2
Σ	25	46	46	0

Coefficient de détermination : $r^2 = 0,835$.