Plan $\mathcal{S} < \mathcal{A}_a > * \mathcal{B}_b$

Plan à mesures partiellement répétées ou plan split-plot

$\mathcal{A}$ et $\mathcal{B}$: facteurs fixes.

Notations
a, b, n, x_{ijk}

Présentation des résultats

Source	S. carrés	ddl	C. moyen	F
	Entre les sujets			
$\mathcal{A}$	SC_A	$a-1$	CM_A	$\dfrac{CM_A}{CM_{S(A)}}$
$\mathcal{S}(\mathcal{A})$	$SC_{S(A)}$	$a(n-1)$	$CM_{S(A)}$	
	Dans les sujets			
$\mathcal{B}$	SC_B	$b-1$	CM_B	$\dfrac{CM_B}{CM_{BS(A)}}$
Int. $\mathcal{AB}$	SC_{AB}	$(a-1)(b-1)$	CM_{AB}	$\dfrac{CM_{AB}}{CM_{BS(A)}}$
Résid.	$SC_{BS(A)}$	$a(n-1)(b-1)$	$CM_{BS(A)}$	
Total	SC_T	$N-1$		

F_A : loi de Fisher à $a-1$ et $a(n-1)$ ddl
F_B : loi de Fisher à $b-1$ et $a(n-1)(b-1)$ ddl
F_{AB} : loi de Fisher à $(a-1)(b-1)$ et $a(n-1)(b-1)$ ddl

Exemple : Expérimentation de Bahrick (reconnaissance de portraits)

Facteurs : sexe du sujet, sexe du portrait
VD : nombre de portraits reconnus

Nom du sujet	Portrait masculin	Portrait féminin
Albert	6	6
Henri	6	6
Jules	5	5
Paul	5	5
Octave	5	6
Albertine	6	8
Henriette	7	8
Julie	6	6
Paule	7	7
Octavie	6	6

Source	S. carrés	ddl	C. moyen	F
	Entre les sujets			
$\mathcal{X}$	7.2	1	7.2	10.28*
$\mathcal{S}(\mathcal{X})$	5.6	8	0.7	
	Dans les sujets			
$\mathcal{P}$	0.8	1	0.8	3.2 NS
Int. $\mathcal{XP}$	0.2	1	0.2	0.8 NS
Résid.	2	8	0.25	
Total	15.8	19		

Remarques et conclusion

Modèle basé sur l'hypothèse d'additivité des effets

Conditions théoriques d'application de la méthode :

– Normalité de la VD dans les populations parentes

– Egalité des variances dans les populations parentes

Tests permettant de vérifier que les conditions sont remplies :

– Test de normalité de Lilliefors ou test d'Anderson Darling

– Tests de O'Brien ou de Bartlett sur les variances

La méthode est robuste : elle fournit des résultats corrects, même si les conditions ne sont qu'approximativement vérifiées.

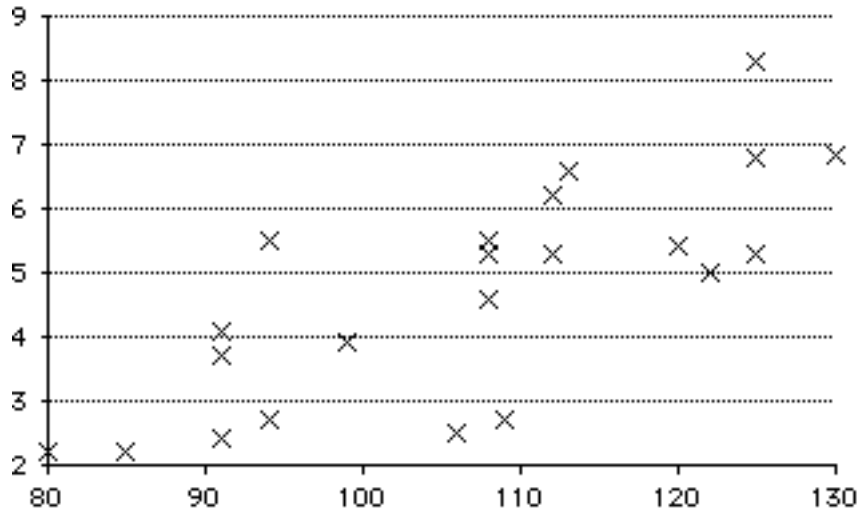
Il existe également des méthodes non paramétriques : travail sur des rangs (test de Kruskal-Wallis)

Corrélation linéaire

Données :

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

Nuage de points : points (x_i, y_i)



Covariance des variables X et Y

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

ou

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

Coefficient de corrélation de Bravais Pearson

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

Remarques

– Formules analogues avec covariance et écarts types corrigés. La valeur de r est la même dans les deux cas.

– Il existe des relations non linéaires

– Corrélation n'est pas causalité

Significativité du coefficient de corrélation

- Les données (x_i, y_i) constituent un échantillon
- r est une statistique
- ρ : coefficient de corrélation sur la population

H_0 : Indépendance sur la population ; $\rho = 0$
H_1 : $\rho \neq 0$ (bilatéral) ou $\rho > 0$ ou $\rho < 0$ (unilatéral)

Statistique de test

– Petits échantillons : tables spécifiques. $ddl = n - 2$
– Grands échantillons :

$$T = \sqrt{n-2}\ \frac{r}{\sqrt{1-r^2}}$$

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

Régression linéaire

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 ?

Modèle permettant d'estimer Y connaissant X

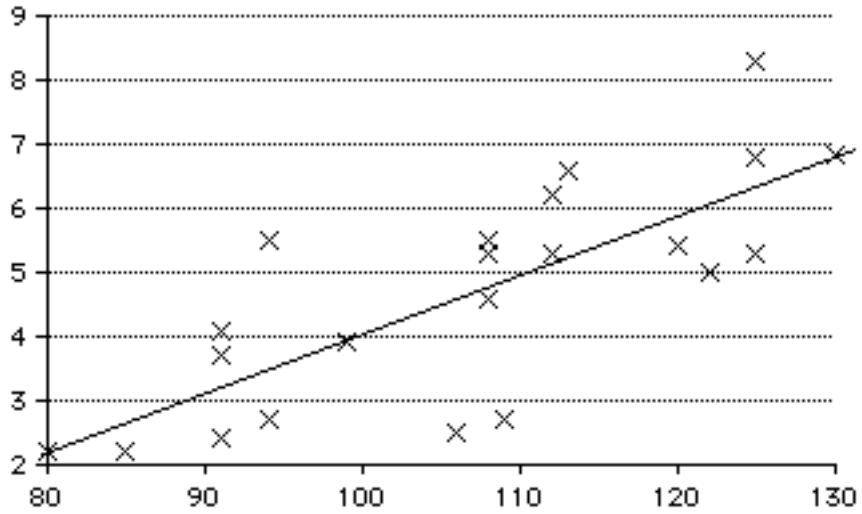
Droite de régression de Y par rapport à X :

La droite de régression de Y par rapport à X a pour équation :

$$y = ax + b$$

avec :

$$a = \frac{Cov(X,Y)}{s^2(X)} \quad ; \quad b = \overline{Y} - a\overline{X}$$



Comparaison des valeurs observées et des valeurs estimées

Valeurs estimées : $\widehat{y}_i = ax_i + b$: variable $\widehat{Y}$
Erreur (ou résidu) : $e_i = y_i - \widehat{y}_i$: variable E

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

$$s^2(Y) = s^2(\widehat{Y}) + s^2(E)$$

avec :

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

$s^2(\widehat{Y})$: variance *expliquée* (par la variation de X, par le modèle)

$s^2(E)$: variance *perdue* ou *résiduelle*

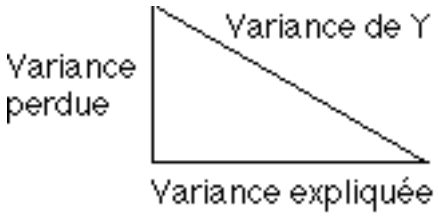
r^2 : part de la variance de Y qui est expliquée par la variance de X. r^2 est appelé *coefficient de détermination*.

Exemple : $r = 0.86$

$r^2 = 0.75$; $1 - r^2 = 0.25$; $\sqrt{1 - r^2} = 0.5$.

• La part de la variance de Y expliquée par la variation de X est de 75%.

• L'écart type des résidus est la moitié de l'écart type de Y.



Remarque : test du coefficient de corrélation

Rappel

Valeurs estimées : $\widehat{y}_i = ax_i + b$
Erreur (ou résidu) : $e_i = y_i - \widehat{y}_i$

$$s^2(Y) = s^2(\widehat{Y}) + s^2(E)$$

avec :

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

La plupart des logiciels de statistiques utilisent une analyse de variance pour tester la significativité du coefficient de corrélation.

Test du coefficient de corrélation à l'aide d'une analyse de variance

Source	SC	ddl	CM	F
Modèle	$ns^2(\widehat{Y})$	1	CM_1	F_{obs}
Résiduelle	$ns^2(E)$	$n-2$	CM_2	
Total	$ns^2(Y)$	$n-1$		

$F_{obs} = \dfrac{CM_1}{CM_2} = (n-2)\dfrac{r^2}{1-r^2}$ suit une loi de Fisher à 1 et $n-2$ ddl.
On retrouve : $F_{obs} = T_{obs}^2$

Régression linéaire multiple

Position du problème

Une population (ou un échantillon) sur laquelle on a observé un ensemble de variables numériques.

	X_1	X_2	...	X_p
s_1	x_{11}	x_{12}	...	x_{1p}
...	...	...	...	...

Exemple avec trois variables

x_i : âge de la mère
y_i : rang de l'enfant dans la fratrie
z_i : poids de l'enfant à la naissance
n : nombre d'observations (ici : $n = 200$)

	X	Y	Z
s_1	26.5	1	2100
...	...	...	...
s_{200}	34.5	2	4500

Nuage de points

Pour trois variables : représentation dans l'espace.

Pour plus de trois variables, détermination des directions de "plus grande dispersion du nuage" : analyse en composantes principales.

Paramètres associés aux données

Matrice des covariances, matrice des corrélations.

Exemple : *Coefficients de corrélation des variables prises 2 à 2 :*

$$\begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{bmatrix} X & Y & Z \\ 1 & 0.60 & 0.24 \\ 0.60 & 1 & 0.28 \\ 0.24 & 0.28 & 1 \end{bmatrix}$$

$r_{xz} = 0.24$ $**$: âge et poids sont corrélés
$r_{yz} = 0.28$ $**$: rang et poids sont corrélés
$r_{xy} = 0.60$ $**$: rang et âge sont fortement corrélés

"Hyperplan" de régression

L'une des variables (Z) est la variable "à prévoir". Les autres ($X_1, X_2, \ldots, X_p$) sont les variables "prédicatives".

$$Z = a_0 + a_1 X_1 + \ldots + a_p X_p$$

Avec trois variables :

$$Z = c + aX + bY$$

Passe par le point moyen, c'est-à-dire : $c = \overline{Z} - a\overline{X} - b\overline{Y}$

Coefficient de corrélation multiple

$\hat{Z}$: valeurs estimées à l'aide de l'équation précédente.

$$R = r_{Z\hat{Z}} = \frac{Cov(Z, \hat{Z})}{s(Z)s(\hat{Z})}$$

Dans l'exemple proposé : $R = 0.29$

Comme précédemment, R^2 est la part de la variance "expliquée par le modèle".

Coefficients de corrélation partielle

Corrélations obtenues en contrôlant la troisième variable. Pour calculer $r_{yz.x}$, par exemple :
– On calcule les résidus de la régression de Z par rapport à X
– On calcule les résidus de la régression de Y par rapport à X
– On calcule le coefficient de corrélation entre les deux séries obtenues.

$r_{yz.x} = 0.18$ $**$: A âge constant, rangs et poids sont corrélés
$r_{xz,y} = 0.09$ NS : A rang constant, pas de corrélation entre âge et poids.

Seul le rang de naissance intervient. L'âge de la mère n'est lié au poids de l'enfant que par le rang de naissance.