Introduction à la modélisation par Equations structurelles

Master 2 recherche

R. Trouillet
MCF Psychologie clinique
Laboratoire Epsylon EA 4556

Sources bibliographiques

Byrne, B.M. (2010). *Structural Equation Modeling with AMOS*. Mahwah, NJ, Lawrence Erlbaum Associates.

Kline, R.B. (2010). *Principles and practice of structural equation modeling*. New York, NY, The Guilford Press.

Site internet : www.davidakenny.net

Objectifs du cours:

- Connaissances de base en Modélisation par équations structurales (*Structural Equation Modeling* – SEM)
- Initiation au logiciel AMOS
- Elaboration d'un plan de recherche appropriée à la SEM

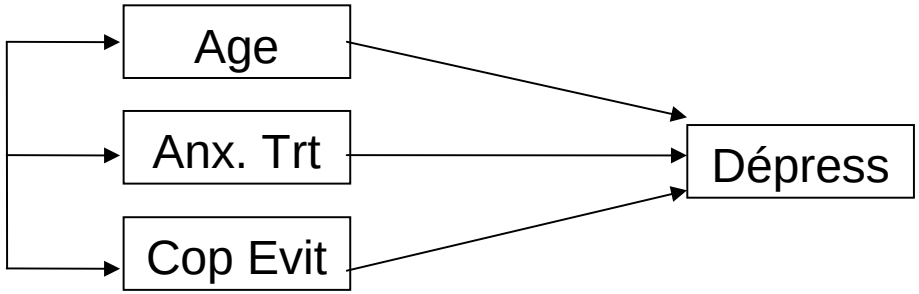
Pourquoi la SEM ?

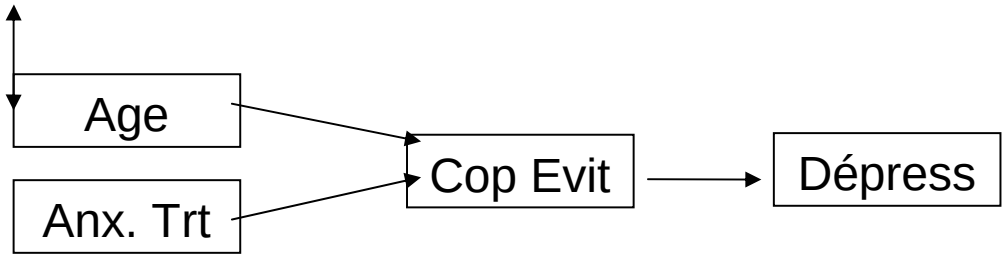
La SEM permet de représenter, estimer et tester un réseau de relations entre plusieurs variables.

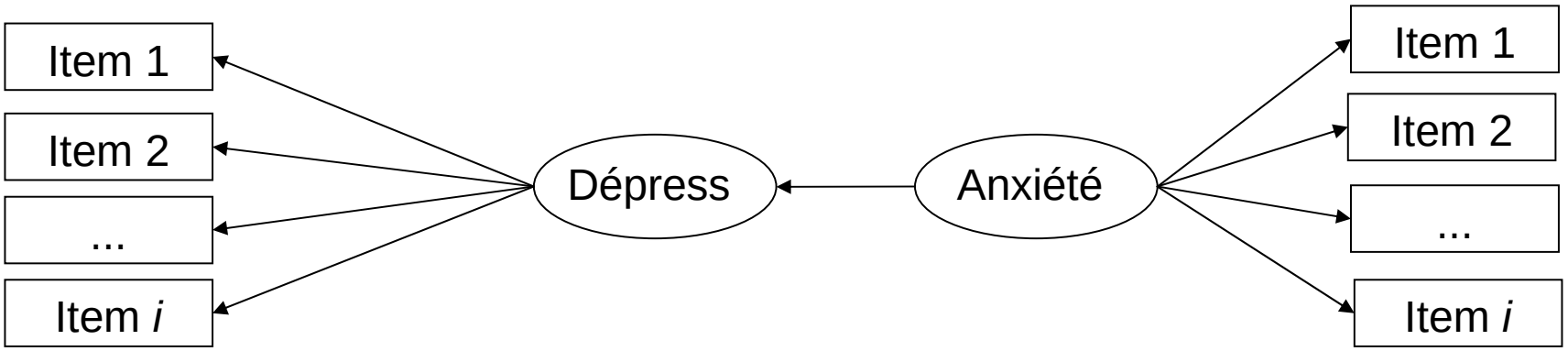
Objectifs SEM :

- Tester un ensemble hypothétique de relations directionnels et non-directionnels entre un ensemble de variables (un modèle doit être justifié théoriquement).
- Comprendre un ensemble de corrélations/covariances entre plusieurs variables
- Expliquer autant que possible la variance des ces variables par le modèle spécifié

Exemples de modèles (les erreurs ne sont pas représentées)







La SEM et les techniques traditionnelles *(analyses de corrélations, régressions).*

Des similarités

- Des modèles de statistiques linéaires
- Assomption d'une distribution normale des données (taux d'aplatissement et d'asymétrie)
- Ces approches ne permettent pas une explication causale.

Des différences

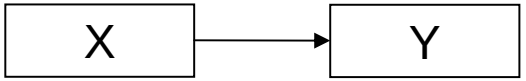
- La SEM est plus flexible
- La SEM permet une analyse plus complètes des données
- La SEM nécessite une spécification formelle du modèle à tester
- La SEM nécessite une limitation des liens pouvant être testés
- La SEM intègre des variables observées et latentes (technique adaptée à la CFA).
- La SEM prend en compte l'erreur associée aux variables endogènes.
- Construction d'un modèle (i.e., un diagramme) spécifiant les relations entre les variables.
- Le diagramme dessiné par le chercheur est automatiquement transformé en un ensemble d'équations qui sont résolues simultanément pour tester l'ajustement du modèle et l'estimation des paramètres (e.g., corrélations...).

Rappels statistiques

Equation de régression : $Y = a_1 + b_1X$

a = constante (ordonnée à l'origine)

b = coeff. régression



La SEM se réfère aux équations issues de la régression multiple où la valeur d'une variable dépendante est fonction de la valeur d'une ou plusieurs variables indépendantes et de leurs coefficients respectifs (la pente de la droite de régression).

Il s'agit de prédire théoriquement la valeur de la VD et d'obtenir l'équation permettant à cette valeur d'être la plus proche possible de la valeur de cette VD issue des données.

Covariance (x, y) = Σ (x_i – moy x) (y_i – moy y) / N - 1
= la moyenne des produits des écarts

Corrélation (x, y) = Cov (x, y) / $s_y s_x$

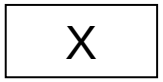
Trois stratégies pouvant s'inscrire dans un projet de modélisation structurale

1) Démarche confirmatoire stricte : rejet ou acceptation du modèle testé

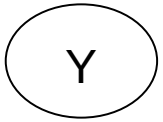
2) Stratégie des modèles compétitifs : proposition de plusieurs modèles alternatifs afin de ne retenir que celui qui s'ajuste le mieux aux données.

3) La génération de modèle : proposition d'un modèle d'essai dont on connaît à l'avance les limites + modification re-test autant de fois que nécessaire pour obtenir un modèle ajustant les données.

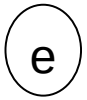
Conventions de base pour diagrammer



– Fenêtres rectangulaires = variables mesurées (manifestes ou indicateurs)

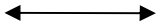


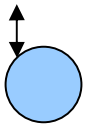
– Les cercles = variables latentes (ou facteurs)



– Petits cercles = variables résiduelles (variance d'erreur, erreur de mesure). Ces résidus = dérivés de l'équation structurale

– Flèches orientées = désignent sens de l'effet d'une variable sur une autres.

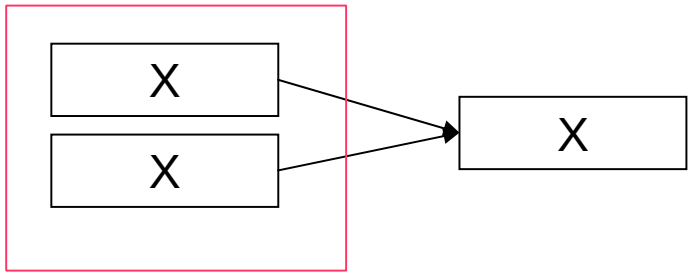




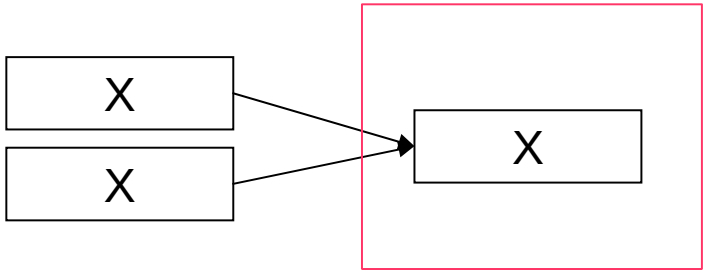
– Doubles flèches = liaisons non-dirigés pouvant désigner la variance ou la covariance (corrélation).

Positionnement des variables dans le modèle

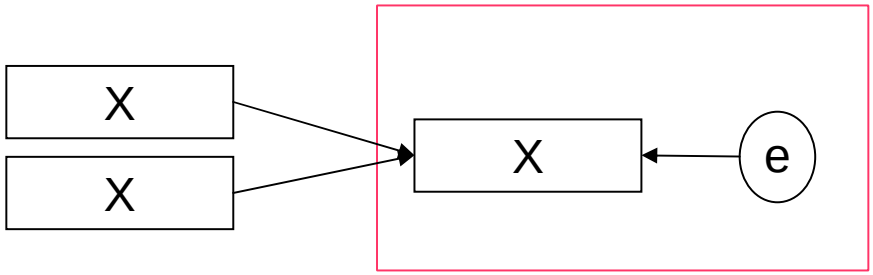
Les variables exogènes



Les variables endogènes



Une variance résiduelle doit être pointée sur chaque variable endogène



Principes de la SEM

Objectif 1 = reproduire la matrice de covariance telle que postulée par le modèle théorique à tester afin de la comparer à la matrice de covariance issue des données empiriques.

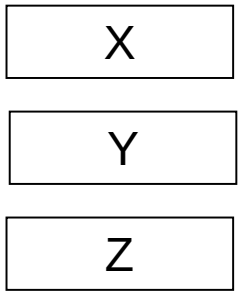
Objectif 2 = cette comparaison doit évaluer l'adéquation entre les variables observées et le modèle théorique.

Lorsque les deux matrices sont égales, le modèle reproduit de façon adéquate la matrice de covariance, et le modèle s'ajuste alors aux données.

Lorsque l'écart entre les deux matrices est important, le modèle n'est pas plausible car il ne reproduit pas de façon adéquate la matrice observée.

La différence entre ces deux matrices conduit à une matrice résiduelle.

Reproduction de la matrice de corrélation : *exemple d'un modèle en pistes causales*

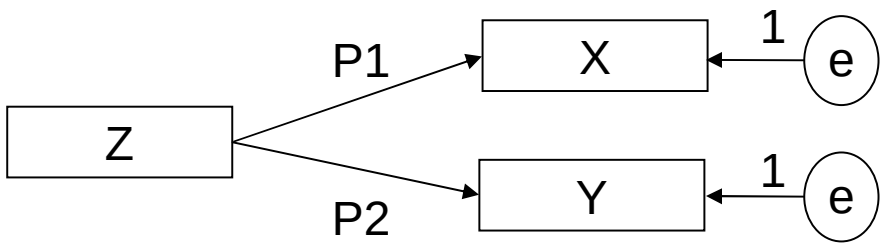


= Anxiété

= Neuroticism

= Age

	X	Y	Z
X	1.00	-	-
Y	.49	1.00	-
Z	.79	..59	1.00



Hypothèse testée : l'âge prédit les niveaux de neuroticism et d'anxiété

Ce modèle contient (notamment) deux paramètres (coeff. régress. partielles) à estimer + deux paramètres fixés.

Nous devons alors utiliser une méthode d'estimation pour trouver une valeur à chaque paramètre libre. Ces valeurs doivent permettre la reproduction d'une matrice de corrélations aussi proche que possible de la matrice de corrélations observée.

Nous disposons de plusieurs méthodes procédant par itérations qui sont effectuées par des algorithmes complexes. Parmi ces méthodes :

- méthode des moindres carrés

- méthode de distribution asymptotique (ADF) = distribution non-normale, mesures catégorielles MAIS + « gourmande »

- méthode du maximum de vraisemblance (ML).

	Valeurs provisoires		Corrélations observées			Fonction de divergence
	P1	P2	Rzx=.79	Rzy=.59	Rxy=.49	Moindres carrés
Cycles D'itérations			*Corrélations reproduites*			Σd^2
1	.50	.50	.50	.50	.250	.149
2	.49	.49	.49	.49	.249	.162
...	...	...	...	...	...	...
5	**.80**	**.61**	**.80**	**.61**	**.480**	**.0006**
6	.81	.61	.81	.61	.494	.0008

$\Sigma d^2 = (.79-.50)^2 + (.59-.50)^2 + (.49-.25)^2 = .149$

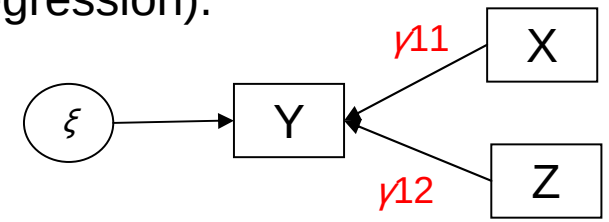
Variables		Corrélations reproduites		Corrélations résiduelles		
X	1.00	.48	.80	.00		
Y	.49	1.00	.61	.01	.00	
Z	.79	.59	1.00	-.01	.-.02	.00

Corrélations observées

Identification d'un modèle

= transposition de la matrice variance-covariance observée dans les paramètres structuraux du modèle (i.e., coefficients de régression).

Ex : $Y = \gamma_{10} + \gamma_{11}X + \gamma_{12}Z + \xi$



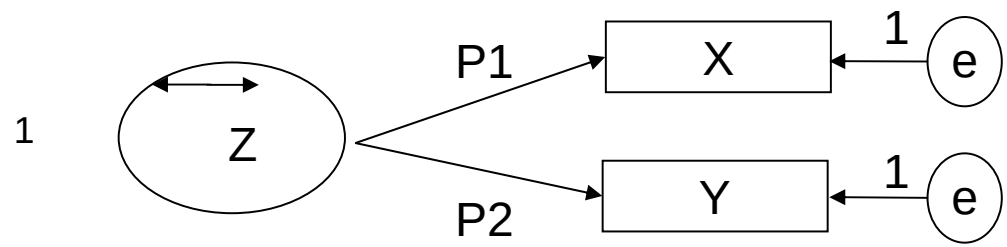
Renvoie au nombre de paramètres inconnus / nombre de données disponibles (i.e., nombre de variances et covariances des variables mesurées).

Etape 1 : k(k+1)/2 où « k » désigne le nombre de variables mesurées.
Etape 2 : compter le nombre de paramètres libres à estimer
Etape 3 : nbre paramètres observés – nbre paramètres libres = Df du modèle

Paramètres à estimer = variance erreurs, variance var. exogène, covariances, pistes.

Conclusion
1) Le nbre des données dispo = nbre des paramètres libres (modèle juste-identifié)
2) Le nbre de données dispo < nbre paramètres libres (modèle sous-identifié)
3) nbre de données dispo > nbre de paramètres libres (modèle sur-identifié) **BEST**

Exemple d'un modèle non-identifié



Pourquoi ce modèle n'est pas identifié ?

2 mesures / 4 paramètres à estimer (2 pistes + 2 variances résidus)
= 2(2+1)/2 = 6/2 = 3 or 3 < 4 paramètres à estimer.

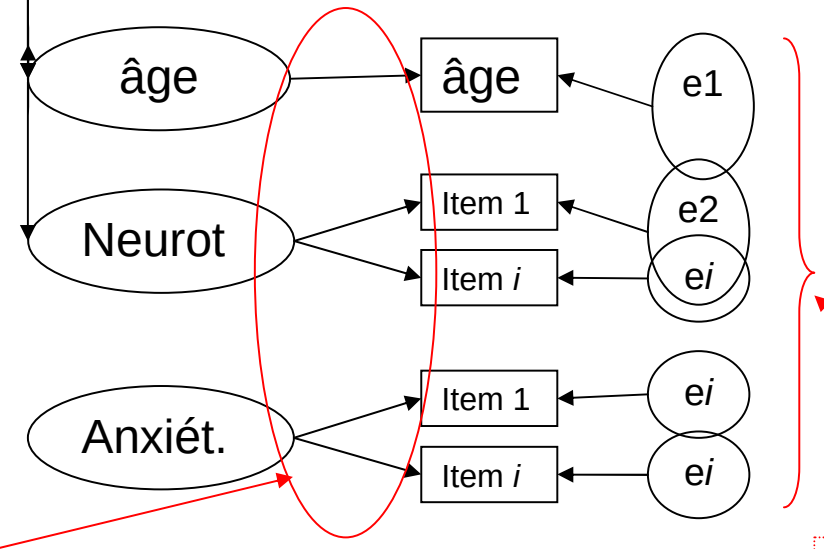
Contraindre 1 nouveau paramètre = juste-identifié (pas conseillé).

L'identification dépend du type de modèle

Modèles pour CFA = 2 (3 si un facteur unique) mesures / facteur sont requises
Si contrainte = 1 piste $\Rightarrow$ variable mesurée = marqueur (le facteur prend sa métrique)

Comment estimer l'ajustement aux données d'un modèle ?

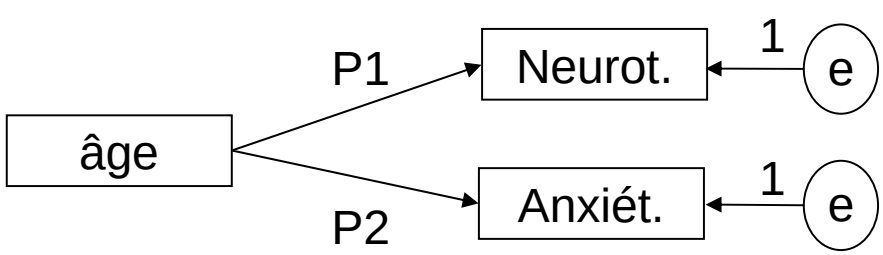
Etape 1 : estimer les modèles de mesure = CFA (toute mesure est entachée d'erreur)



bétas = interprétés comme des saturations (≥.70 ou ≥.40)

ATTENTION : ici e*i* = erreur de mesure

Etape 2 : si les CFAs sont satisfaisantes, nous estimons le modèle théorique postulé



ATTENTION : ici e*i* = le résidu = erreur de prédiction de la VI

Séance TD1 (13h-17h)

Objectifs :

1) Apprendre à effectuer un diagramme avec AMOS à partir d'hypothèses théoriques

1) Estimer le niveau d'identification du modèle

1) Vérifier la nécessité d'ajuster le ratio paramètres libres vs. fixés afin d'identifier le modèle.

Modèle à tester :

Nous postulons que l'âge prédit une augmentation du niveau de stress perçu et une diminution de la satisfaction ressentie à l'égard du soutien social. De plus l'âge corrèle positivement avec le nombre des évènements de vie, lesquels prédisent une augmentation du niveau de stress perçu.

La satisfaction / soutien social est évaluée avec 6 items (SSQ)
Le stress perçu est évalué avec14 items (PSS : l'item 12 est exclu de la structure)
Les évènements de vie sont évalués par un inventaire de 43 évènements

Comment estimer si le modèle s'ajuste aux données ?

L'objectif premier de la SEM est de déterminer le niveau d'ajustement entre le modèle hypothétique et les données issues de l'échantillon.

Résumé du processus d'ajustement du modèle :

Σ = matrice de covariance observée
Θ = vecteur comprenant les paramètres du modèle
Σ(Θ) =matrice de covariance restreinte par le modèle (i.e., la structure spécifiée par le modèle)

H_0 = le modèle postulé est supporté dans la population [i.e., Σ = Σ(Θ)].

En conséquence, en SEM, le chercheur **espère ne pas rejeter H_0** .
Le processus d'estimation cherche alors les valeurs des paramètres reduisant les résidus (e) entre les matrices de covariance observées (Σ) et les matrices de covariance induites par le modèle [Σ(Θ)].

Nous voulons une faible fonction de divergence (minimum fit function) F_{min}=Σ – Σ(Θ) = mini.

Nous utilisons alors des statistiques (i.e., indices) d'ajustement permettant d'estimer cette divergence. Nous parlons de « *goodness-of-fit statistics* ».

Comment estimer si le modèle s'ajuste aux données ? *(suite)*

Le soft teste trois modèles :

Le modèle d'indépendance (complète indépendance des variables : les corrélations = .00)
Le modèle postulé (i.e., votre modèle d'étude)
Le modèle saturé (paramètres estimés = variances et covariances observées).

Ces 3 modèles représentent un continuum (du plus restreint au moins retreint). Nous n'étudions que les résultats du **modèle postulé**.

Les indices d'ajustement absolu : ils n'utilisent pas un modèle alternatif comme base de comparaison

Le CMIN (ou Chi²) et le CMIN/DF

CMIN (minimum discrepancy) = estimation de la divergence entre la matrice de covariance Σ et la matrice Σ(Θ). Il s'apparente à test de Chi² où H_0 = les paramètres spécifiés dans le modèle testé sont valides [Σ – Σ(Θ) = résidus = 0].

En conséquence : ⇘ CMIN = ⇗ ajustement Σ(Θ) / Σ = OK !

La *p*. value associée = la probabilité que la valeur du Chi²,s i H_0 est vraie, soit plus élevée que la valeur du Chi² obtenue.

En conséquence :⇗*p*.value =ajustement modèle testé / modèle parfait=OK !

DF = degrés de liberté (cf. diapo. « identification »).

Problèmes :

CMIN est sensible à l'échantillon car $Chi^2 = (N-1) F_{min} \Rightarrow \nearrow N \Rightarrow \nearrow Chi^2$ (i.e., indique à tort le rejet du modèle testé). **PBL** si N > 200.

CMIN teste la divergence modèle testé et modèle parfait (i.e., modèle juste identifié) $\Rightarrow \nearrow$ nbre paramètres à estimer $\Rightarrow \nearrow$ saturation modèle $\Rightarrow \nearrow$ correspondance notre modèle / modèle juste identifié $\Rightarrow \searrow Chi^2$.

Une alternative = CMIN/DF (valeurs ≤ **3.00 = OK** !) (PBL = ce seuil manque de validation)

Standardized Root Mean Square Residual (SRMR) = différence standardisée entre la matrice de corrélation observée et la matrice prédite par le modèle. **SRMR < .08 = OK ! (attention le SRMR na pas de pénalité / complexité)**

Akaike's Information Criterion (AIC) & the Bayesian Information Criterion (BIC) = indices à utiliser pour comparer des modèles compétitifs : **le meilleur modèle obtient le + faible indice. Le BIC augmente la pénalité / complexité modèle.**

<u>Les indices de non-centralité</u>.
L'objectif n'est pas de tester si l'hypothèse nulle est vraie (cf. CMIN, SRMR…) mais de tester le rejet d'une hypothèse alternative.

<u>Root Mean Square Error of Approximation (RMSEA)</u> = indice de mauvais ajustement du modèle testé ⇒ ⇘ RMSEA = ⇗ ajustement du modèle.
Le RMSEA est assorti d'un IC à 90%.

Valeurs seuil de RMSEA :
< .06 = bon ajustement ; de .08 à .10 = ajustement médiocre ; > .10 = ajustement pauvre.

Les valeurs seuils de l'IC : **Limite inf. ≤ .05 ; limite sup. < .10**

<u>Comparative Fit Index (CFI)</u> = indice d'ajustement incrémentiel estimant dans quelle mesure le modèle testé est meilleur que le modèle d'indépendance.

Valeur seuil du **CFI ≥ .95**.

<u>Tucker Lewis Index (TLI)</u> = indice estimant, dans quelle mesure, notre modèle améliore l'ajustement au regard du modèle d'indépendance. Il est similaire au CFI mais le TLI est plus pénalisant / complexité modèle.

Valeur seuil **TLI ≥ .95**

Les résultats analytiques.

B = coefficient de corrélation non-standardisé (reflète corrélation entre 2 Variables)

S.E. = indique l'erreur type de l'estimation (⇘ S.E. = ⇗ valeur prédictive du prédicteur)
Ex : pour S.E. = .98 signifie que l'utilisation du prédicteur réduit l'erreur type de 2 %

Critical Ration (C.R.) = B/S.E. : estime si la cov_{xy} est significativement différente de 0. Si C.R. $\geq$ 1.96 = cov_{xy} significative à $p \leq .05$.

β = coefficient de régression standardisé