Statistiques paramétriques et non paramétriques Tests non paramétriques - Tests sur des groupes appariés - Corrélation et régression

1 Tests non paramétriques de comparaison de deux ou plusieurs groupes avec Statistica

On reprend l'exemple Conso-Crustaces.stw.
On veut comparer la consommation de crustacés dans les différents groupes définis par les variables catégorisées. Cependant, les distributions des données dans les différents groupes ainsi définis s'écartent notablement de celles de variables normales. Nous allons donc utiliser des tests non paramétriques pour réaliser ces comparaisons.

1.1 Comparer deux groupes indépendants : test de Mann Whitney, test de Kolmogorov Smirnov

Utilisez le menu Statistiques - Tests non paramétriques - Comparaison de deux échantillons (groupes).

Indiquez Conso comme variable dépendante et SD Sexe par exemple, comme variable de classement.

Le test le plus classiquement utilisé est celui de Mann Whitney. Cliquez sur le bouton correspondant.

On obtient :

	Test U de Mann-Whitney (Enquete Consommation dans Conso-crustaces.stw) Par var. SD sexe Tests significatifs marqués à p <,05000								
variable	SommeRgs Homme	SommeRgs Femme	U	Z	niv. p	Z ajusté	niv. p	N Actif Homme	N Actif Femme
Conso	45756,00	84549,00	26841,00	-2,35871	0,018339	-2,35871	0,018339	194	316

Il semble donc y avoir une différence significative entre les hommes et les femmes du point de vue de la consommation. Pour ce test, les hypothèses nulle et alternative portent sur les médianes. Il semble donc que la consommation médiane soit différente chez les hommes et chez les femmes.

Un autre test, moins couramment utilisé est le test de Kolmogorov-Smirnov. Pour ce dernier, les hypothèses portent sur les distributions elles-mêmes. On obtient :

variable	Test de Kolmogorov-Smirnov (Enquete Consommation dans Conso-crustaces.stw) Par var. SD sexe Tests significatifs marqués à p <,05000 Max Nég Différnc	Max Pos Différnc	niv. p	Moyenne Homme	Moyenne Femme	Ec-Type Homme	Ec-Type Femme	N Actif Homme	N Actif Femme
Conso	-0,156270	0,022772	p < .01	9,199364	12,63566	14,56083	19,14834	194	316

Là encore, il semblerait que les distributions de la consommation soient différentes selon le sexe.

1.2 Test non paramétrique pour plusieurs groupes indépendants : test de Kruskal et Wallis

On veut comparer la consommation selon le lieu de résidence. On peut utiliser le menu : Statistiques - Tests non paramétriques - Comparaison de plusieurs échantillons (groupes).

On indique Conso comme variable dépendante, SD Résidence comme variable de classement, et on obtient le résultat suivant :

Dépend. : Conso	ANOVA de Kruskal-Wallis par Rangs; Conso (Enquete Consommation dans Conso-crustaces.stw) Var. indépendante (classement) : SD residence Test de Kruskal-Wallis : H (3, N= 510) =28,01862 p =,000 Code	N Actifs	Somme Rangs
Nord	1	141	41694,50
Centre	2	156	40902,00
Sud	3	138	33715,50
Autre	4	75	13993,00

Là encore, les consommations médianes semblent différentes selon la zone de résidence.

2 Tests non paramétriques avec Modalisa

2.1 Exporter les données en format texte, les importer dans Modalisa

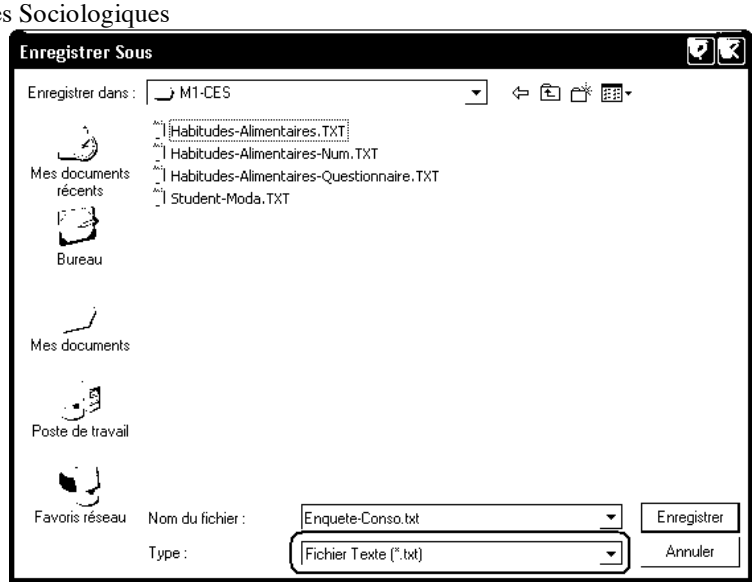
Statistica permet d'exporter une feuille de données au format Excel ou au format "texte seul". Pour cela :

- Modifiez la hiérarchie des objets du classeur de façon que la feuille à exporter soit un élément terminal de la hiérarchie.

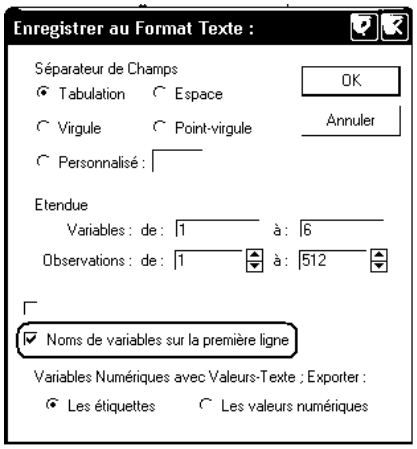
- A l'aide du bouton droit de la souris, extraire une copie de la feuille de données dans une fenêtre indépendante.

- L'export des valeurs numériques sera fait en utilisant les valeurs affichées des nombres et non leurs valeurs "internes". Veillez donc à ce que la variable "Conso" s'affiche avec un nombre de décimales suffisant.

- Affichez cette fenêtre en premier plan et utilisez le menu Fichier - Enregistrez sous.



Sélectionnez "Fichier Texte" comme type de fichier.



Pensez à indiquer à Statistica d'enregistrr les noms de variables en première ligne. Ils pourront ainsi être utilisés comme intitulés des questions dans Modalisa. On peut également indiquer à Statistica d'exporter les étiquettes et non des codes numériques pour les variables catégorielles. Modalisa pourra importer ces données comme des modalités de questions à réponse unique.

Chargez Modalisa et utilisez le menu Fichier - Import ASCII, puis le bouton "Importer des questionnaires dans une nouvelle enquête dont vous créez les questions et leurs types".

Veillez ensuite à ce que la case "La première ligne contient les libellés de colonnes" soit cochée, puis indiquez que la "question" Conso est de type numérique, et les autres questions sont de type "A réponse unique".

2.2 Tests non paramétriques sur deux groupes indépendants avec Modalisa

Utilisez ensuite le menu Analyse - Tests non paramétriques - Mann & Whitney U. Indiquez SD Sexe comme question fermée de départ et Conso comme question "à réponse de type numérique". Vous devriez obtenir le résultat suivant :

```
Mann & Whitney - U
X : SD sexe
Y : Conso
U       34463
U'      26841
Z       2,359
Groupe d'égaux   25
```

```
1) Homme    194
    Somme des rangs   45756
    Moyenne des rangs      235,856
2) Femme    316
    Somme des rangs   84549

    Moyenne des rangs      267,56
```

Modalisa nous donne la valeur de la statistique de test, mais n'indique pas de valeur critique de de p-value correspondante.

Faites également le test non paramétrique de Kolmogorov-Smirnov. Là encore, Modalisa n'indique pas la p-value de la statistique trouvée.

2.3 Tests non paramétriques sur plusieurs groupes indépendants avec Modalisa.

De même utilisez le menu Analyse - Tests non paramétriques - Kruskal et Wallis en indiquant SD Résidence comme question fermée et Conso comme question "à réponse de type numérique". Vous devriez obtenir :

```
Kruskal & Wallis
X : SD residence
Y : Conso
Degrés de liberté     3
Nombre de modalités   4
Nombre de valeurs     510
H      28,019
p      0,999
Sans réponses exclues      1
Groupe d'égaux   25
1) Centre   156
    Somme des rangs   40902
2) Nord     141
    Somme des rangs   41694,5
3) Sud      138
    Somme des rangs   33715,5
4) Autre    75
    Somme des rangs   13993
```

Comparer les résultats avec ceux indiqués par Statistica. Remarquez que, dans ce cas, Modalisa nous indique 1 - p-value comme probabilité, ce qui peut nous induire en erreur.

3 Comparaison de deux groupes appariés à l'aide d'un test paramétrique

3.1 Exemple

Source : A study on significant sources of the burnout syndrome in workers at occupational centres for mentally disabled, Pedro R. Gil-Monte and José Ma Peiró, Psychology in Spain, 1997, Vol. 2. No 1, 116-123.
Page Web : http://www.psychologyinspain.com/content/full/1997/6bis.htm

Subjects
Subjects were 95 employees in occupational institutions for mentally retarded people in the Valencia Autonomous Community (...).

Description des variables.

Self-confidence levels were measured by using five items of an adaptation of the Trait Sport-Confidence Inventory" (TSCI) (Vealey, 1986), in which the word "athlete" was replaced by "workmate". Cronbach's alpha coefficient for the present study was .84.
Social support at work was estimated using 6 items of the "Organisational Stress Questionnaire" (OSQ) (Caplan, Cobb, French, Van Harrison and Pinneau, 1975). These items reflect some aspects of social support coming from *workmates* (3 items) and *supervisors* (3 items). Reliability coefficient in this study was a=.86 for the supervisors' social support scale, and a=.76 for the workmates' social support scale.
Perceived *role conflict* and *role ambiguity* levels were measured by 3 items, for each of the variables, taken from their respective OSQ scales. Reliability values were a=.69 for role ambiguity and .68 for the role conflict scale.
The burnout syndrome was estimated by MBI (Maslach and Jackson, 1986). This instrument is comprised of 22 items measuring the three dimensions in the syndrome: *personal accomplishment* (8 items), *emotional exhaustion* (9 items), and *depersonalisation* (5 items). Reliability coefficients obtained in the study were: a=.76 for the personal accomplishment subscale, a=.87 for emotional exhaustion, and a=.52 for depersonalisation.

Ouvrez le classeur Valencia-Burnout.stw.

N.B. Les données figurant dans ce classeur ont été générées à partir des indications (moyennes, écarts-types, coefficients de corrélation) figurant dans l'article. Cela explique qu'il ne s'agisse pas de valeurs entières, comme on aurait pu le penser à la lecture de la description des variables.

Affichez les statistiques descriptives concernant ces variables. Vous devriez obtenir :

Variable	Statistiques Descriptives (Valencia-Burnout dans Valencia-Burnout.stw)				
	N Actifs	Moyenne	Minimum	Maximum	Ecart-type
Self-Confidence	95	6,4800	4,1632	9,4430	1,0656
Workmates Social Support	95	3,2500	1,5810	5,0078	0,6635
Supervisor Social Support	95	2,9000	0,5435	5,2818	0,8545
Role Conflict	95	2,7300	0,9488	4,7904	0,7942
Role Ambiguity	95	2,1100	0,1915	4,3977	0,7640
Personal Accomplishment	95	36,4300	23,2596	57,2632	6,9266
Emotional Exhaustion	95	17,5600	-5,4779	42,1888	10,1737
Depersonalisation	95	4,6900	-4,6758	15,3578	4,4636

3.2 Comparaison de moyennes sur deux groupes appariés

Les deux variables Role Conflict et Role Ambiguity ont le même intervalle de variation. On souhaite étudier si les moyennes des scores observés pour ces deux variables sont significativement différents.

Utilisez le menu Statistiques - Statistiques élémentaires - Test t pour des échantillons appariés.

Indiquez Role Conflict comme variable de la première liste, Role Ambiguity comme variable de la seconde liste. Vous devriez obtenir comme résultat :

Variable	Test t pour des Echantillons Appariés (Valencia-Burnout dans Valencia-Burnout.stw) Différences significatives marquées à p < ,05000							
	Moyenne	Ec-Type	N	Différ.	Ec-Type Différ.	t	dl	p
Role Conflict	2,730000	0,794191						
Role Ambiguity	2,110000	0,764032	95	0,620000	0,867943	6,962451	94	0,000000

Autrement dit : au vu de l'échantillon proposé, il semble exister une différence significative entre ces variables.

4 Corrélation et régression Linéaire

4.1 Corrélation entre les variables

Pour étudier s'il existe un lien entre les différentes variables numériques qui ont été recueillies, on peut également s'intéresser aux coefficients de corrélation entre ces variables.

Utilisez le menu Statistiques - Statistiques Elémentaires - Matrice de corrélations pour afficher les coefficients de corrélation entre les différentes variables prises deux à deux :

Variable	Corrélations (Valencia-Burnout dans Valencia-Burnout.stw) Corrélations significatives marquées à p < ,05000 N=95 (Observations à VM ignorées)							
	f-Confidenc	Workmates Social Support	Supervisor Social Support	Role Conflict	Role Ambiguity	Personal omplishme	Emotional Exhaustion	ersonalisati
Self-Confidence	1,00	0,08	0,16	-0,11	-0,33	0,35	-0,14	0,00
Workmates Soci	0,08	1,00	0,50	-0,41	-0,40	0,33	-0,45	-0,22
Supervisor Socia	0,16	0,50	1,00	-0,37	-0,43	0,22	-0,40	-0,12
Role Conflict	-0,11	-0,41	-0,37	1,00	0,38	-0,32	0,69	0,32
Role Ambiguity	-0,33	-0,40	-0,43	0,38	1,00	-0,48	0,40	0,28
Personal Accom	0,35	0,33	0,22	-0,32	-0,48	1,00	-0,40	-0,28
Emotional Exhau	-0,14	-0,45	-0,40	0,69	0,40	-0,40	1,00	0,40
Depersonalisatio	0,00	-0,22	-0,12	0,32	0,28	-0,28	0,40	1,00

Comparez avec les valeurs indiquées dans l'article :

Table 1
Descriptive Statistics, reliability values and variables correlations[a]

	M	SD	Range	1	2	3	4	5	6	7	8
1. Self-confidence	6.48	1.06	1-9	(.84)							
2. Workmates Social Support	3.25	.66	1-4	.08	(.76)						
3. Supervisor Social Support	2.90	.85	1-4	.16	.50	(.86)					
4. Role Conflict	2.73	.79	1-5	-.11	-.41	-.37	(.68)				
5. Role Ambiguity	2.11	.76	1-5	-.33	-.40	-.43	.38	(.69)			
6. Personal Accomplishment	36.43	6.89	0-48	.35	.33	.22	-.32	-.48	(.76)		
7. Emotional Exhaustion	17.56	10.12	0-54	-.14	-.45	-.40	.69	.40	-.40	(.87)	
8. Depersonalisation	4.69	4.44	0-30	-.00	-.22	-.12	.32	.28	-.28	.40	(.52)

4.2 La régression linéaire ordinaire

Effectuez ensuite une régression multiple ordinaire des 3 dernières variables sur les 5 premières :

Pour la variable Personal Accomplishment :

Le bouton "Synthèse de la régression" (onglet "Avancé") affiche les résultats suivants :

N=95	Synthèse de la Régression; Variable Dép. : Personal Accomplishment R= ,56173332 R?= ,31554432 R? Ajusté = ,27709175 F(5,89)=8,2061 p<,00000 Err-Type de l'Estim.: 5,8892					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(89)	niveau p
OrdOrig.			32,4726	7,3143	4,4396	0,0000
Self-Confidence	0,2293	0,0932	1,4906	0,6057	2,4610	0,0158
Workmates Social Support	0,1730	0,1074	1,8063	1,1213	1,6109	0,1108
Supervisor Social Support	-0,0923	0,1071	-0,7479	0,8678	-0,8618	0,3911
Role Conflict	-0,1351	0,1006	-1,1779	0,8773	-1,3426	0,1828
Role Ambiguity	-0,3235	0,1071	-2,9325	0,9710	-3,0201	0,0033

La colonne "B" donne les coefficients de l'équation de régression linéaire. Le modèle fourni par la régression linéaire est le suivant :
Personal Accomplishment = 32,47 + 1,49 *Self-Confidence +1,81 * Workmates Social Support - 0,75 * Supervisor Social Support - 1,18 * Role Conflict - 2,93 * Role Ambiguity

La valeur de R^2 est de 0,315 : 31,5% de la variance de la variable Personal Accomplishment est expliquée par le modèle.

Les coefficients de la colonne "Bêta" sont les coefficients standardisés, c'est-à-dire les coefficients que l'on observerait si on utilisait des variables centrées réduites au lieu des variables observées. On peut également les interpréter comme suit : lorsque "Self-Confidence" augmente d'un écart type, la variable "Personal Accomplishment" estimée augmente de 0,23 écart type, lorsque la variable "Role Conflict" augmente d'un écart type, "Personal Accomplishment" diminue de 0,135 écart type.
Par exemple, on pourra vérifier que

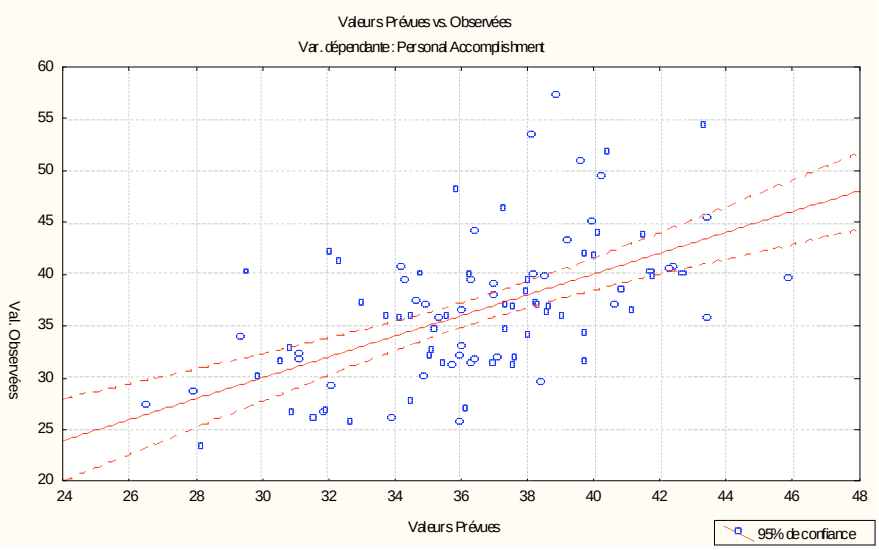
$$Beta(Self-Confidence) = \frac{Ecart\ type(Self-Confidence)}{Ecart\ type(Personal\ Accomplishment)} \times B(Self-Confidence) = \frac{1,0656}{6,9266} \times 1,4906 = 0,2293$$

Les valeurs de t sont obtenues en divisant la valeur correspondante de B par son erreur type. Autrement dit, on teste si le coefficient B est significativement différent de 0.

On peut afficher les résultats de l'ANOVA (bouton ANOVA) montrant qu'ici, le coefficient de régression multiple est significativement différent de 0, ou encore qu'il existe un lien linéaire significatif entre la variable dépendante et les autres variables :

	Analyse de Variance (Valencia-Burnout dans Valencia-Burnout.stw)				
Effet	Sommes Carrés	dl	Moyennes Carrés	F	niveau p
Régress.	1423,058	5	284,6115	8,2061	0,0000
1Résidus	3086,793	89	34,6831		
Total	4509,850				

Sous l'onglet "Nuage", on pourra obtenir différentes représentations graphiques dont, par exemple, le graphique illustrant l'adéquation entre les valeurs observées et les valeurs théoriques :



4.3 La régression linéaire pas à pas

Dans l'article, les auteurs indiquent qu'ils ont fait une régression linéaire pas à pas des dimensions du MBI sur les 5 premières variables.

4.3.1 Principe de la méthode

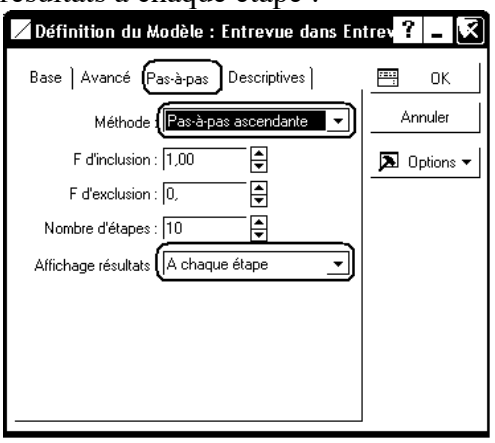
Les données sont formées par une VD Y et plusieurs variables explicatives X1, X2, ..., Xp.
On choisit, parmi les variables explicatives, celle qui est le mieux corrélée à Y. Pour simplifier les notations, nous supposerons qu'il s'agit de la variable X1.

On calcule l'équation de régression linéaire de Y sur X1 : Y = b1 X1 + b0.
On calcule alors les résidus : R1 = Y - b1 X1.- b0
On choisit, parmi les variables explicatives restantes, celle qui est le mieux corrélée à R1. Nous supposons ici qu'il s'agit de la variable X2.
On calcule l'équation de régression linéaire de Y sur X1 et X2 : Y = b'1 X1 + b2 X2 + b'0.
On calcule les nouveaux résidus : R2 = Y - (b'1 X1 + b2 X2 + b'0) et on poursuit la méthode jusqu'à ce que les variables explicatives restantes ne soient plus significativement corrélées aux résidus.

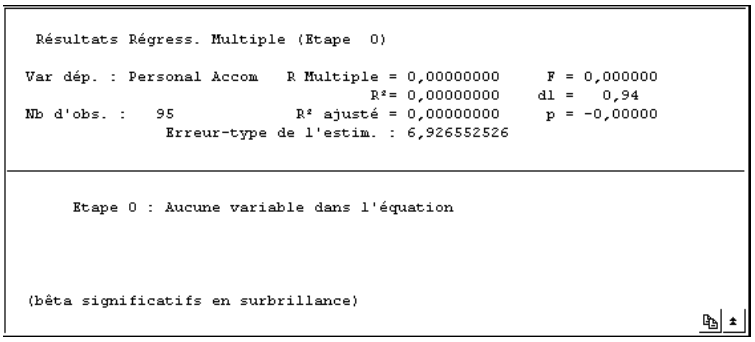
4.3.2 La régression linéaire pas à pas pour la variable Personal Accomplisment

Utilisez de nouveau le menu Statistiques - Régression Multiple
Sous l'onglet "Avancé", spécifiez "Personal Accomplishment" comme variable dépendante, les 5 premières variables comme variables indépendantes. Cochez l'option "régression ridge ou pas-à-pas".

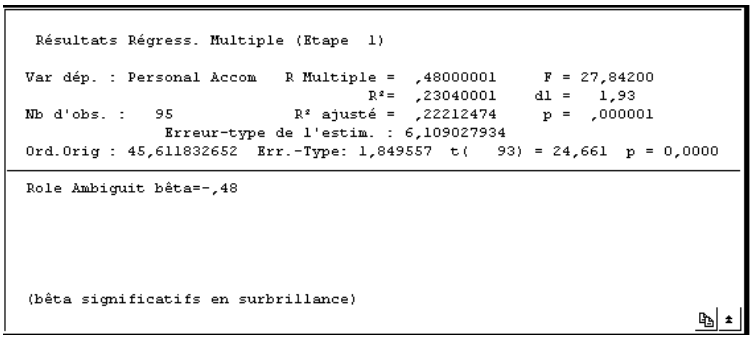
Dans le dialogue suivant, activez l'onglet "pas-à-pas" et sélectionnez la méthode "pas à pas ascendante", et l'affichage des résultats à chaque étape :



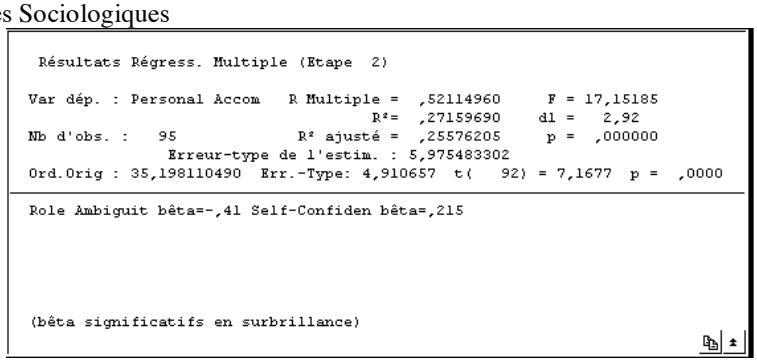
A la première étape, Statistica affiche les résultats suivants :



Cliquez sur "suivant". On obtient :



Puis :



```
 Résultats Régress. Multiple (Etape  2)

Var dép. : Personal Accom   R Multiple =  ,52114960     F = 17,15185
                                R²=  ,27159690    dl =   2,92
Nb d'obs. :   95             R² ajusté =  ,25576205     p =  ,000000
                Erreur-type de l'estim. : 5,975483302
Ord.Orig : 35,198110490  Err.-Type: 4,910657  t(   92) = 7,1677  p =  ,0000

Role Ambiguit bêta=-,41 Self-Confiden bêta=,215



(bêta significatifs en surbrillance)
```

Statistica accepte encore de faire rentrer deux autres variables dans la régression. Cependant, en affichant les résultats disponibles sous le bouton "Synthèse de la régression", on se rend compte que seules ces deux premières variables sont significativement corrélées aux résidus :

N=95	Synthèse de la Régression; Variable Dép. : Personal Accomplishment R= ,55662608 R?= ,30983259 R? Ajusté = ,27915848 F(4,90)=10,101 p<,00000 Err-Type de l'Estim.: 5,8808					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(90)	niveau p
OrdOrig.			30,8772	7,0660	4,3698	0,0000
Role Ambiguity	-0,3029	0,1043	-2,7456	0,9451	-2,9050	0,0046
Self-Confidence	0,2253	0,0929	1,4647	0,6041	2,4247	0,0173
Workmates Social Support	0,1406	0,1005	1,4680	1,0489	1,3996	0,1651
Role Conflict	-0,1225	0,0994	-1,0682	0,8668	-1,2323	0,2210

On peut alors reprendre la méthode en ne spécifiant que deux étapes et retrouver les résultats indiqués par les auteurs :

N=95	Synthèse de la Régression; Variable Dép. : Personal Accomplishment R= ,52114960 R?= ,27159690 R? Ajusté = ,25576205 F(2,92)=17,152 p<,00000 Err-Type de l'Estim.: 5,9755					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(92)	niveau p
OrdOrig.			35,1981	4,9107	7,1677	0,0000
Role Ambiguity	-0,4090	0,0943	-3,7083	0,8545	-4,3395	0,0000
Self-Confidence	0,2150	0,0943	1,3976	0,6127	2,2811	0,0249

Résultats indiqués dans l'article :

Table 2
Stepwise regression analysis for MBI dimensions

Variable Step	R2 increase	Beta	F for equation
	personal Accomplishment		
1 Role ambiguity	.23	-.41	
2 Self-confidence	.04	.22	17.27***
	Emotional Exhaustion		
1 Role conflict	.47	.60	
2 Workmates' social support	.03	-.20	47.27***
	Depersonalisation		
1 Role conflict	.10	.32	10.45***

*** p < 001