Master de Psychologie - 1ère année
PSY73B : Informatique : traitement des données - TD N°4
Corrélation et régression

18. Corrélation linéaire

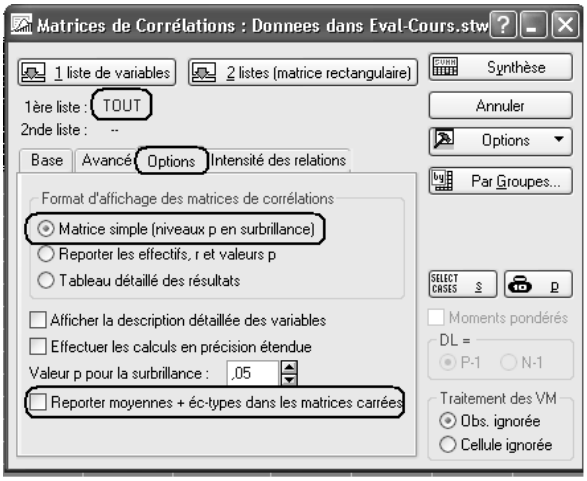
18.1. Coefficient de corrélation

L'association des étudiants d'une grande université (américaine) a publié une évaluation de plus de cent cours enseignés durant le semestre précédent. Les étudiants de chaque cours avaient rempli un questionnaire d'évaluation portant sur différents aspects du cours; l'évaluation se faisait sur une échelle en cinq points (1=très mauvais, 5=excellent).
Les données saisies dans le fichier Eval-Cours.stw sont les données réelles. Elles représentent les scores moyens enregistrés sur 6 variables pour un échantillon de 50 cours.
Ces variables étaient :

- la qualité globale des exposés (Qual-Glob)
- les aptitudes pédagogiques du professeur (Pédagogie)
- la qualité des tests et examens (Examen)
- la connaissance de la matière dont témoigne le professeur, telle qu'elle est perçue par les étudiants (Connaissance)
- les résultats auxquels s'attendent les étudiants pour ce cours (Résultat, de très bon à insuffisant)
- le nombre d'inscriptions à ce cours (Inscription)

On souhaite étudier les liens qui existent entre ces différentes variables.

Pour obtenir les coefficients de corrélation entre les différentes variables, on pourra utiliser le menu Statistiques - Statistiques Elémentaires - Matrices de corrélation. On peut utiliser l'onglet "Options" pour limiter l'affichage à la matrice des corrélations :



Variable	Corrélations (Donnees dans Eval-Cours.stw) Corrélations significatives marquées à p < ,05000 N=50 (Observations à VM ignorées)					
	Qual-Glob	Pédagogie	Examen	Connaissance	Résultat	Inscription
Qual-Glob	1,000	0,804	0,596	0,682	0,301	-0,240
Pédagogie	0,804	1,000	0,720	0,526	0,469	-0,451
Examen	0,596	0,720	1,000	0,451	0,610	-0,558
Connaissance	0,682	0,526	0,451	1,000	0,224	-0,128
Résultat	0,301	0,469	0,610	0,224	1,000	-0,337
Inscription	-0,240	-0,451	-0,558	-0,128	-0,337	1,000

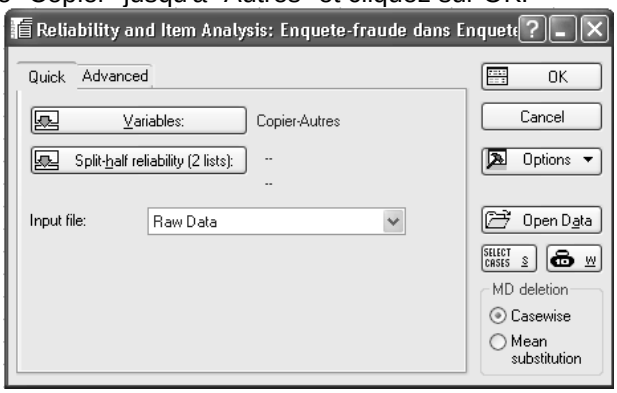
On voit que les coefficients de corrélation entre les 5 premières variables sont positifs, alors que la 6ème variable est corrélée négativement (anti-corrélée) avec les 5 autres.

18.2. Alpha de Cronbach

On reprend les données Enquete-Fraude.stw, décrites dans le polycopié précédent. On souhaite mesurer la cohérence des réponses des sujets quant aux techniques de fraude, afin d'estimer s'il est pertinent de construire une variable telle que ScoreTricheTotal.

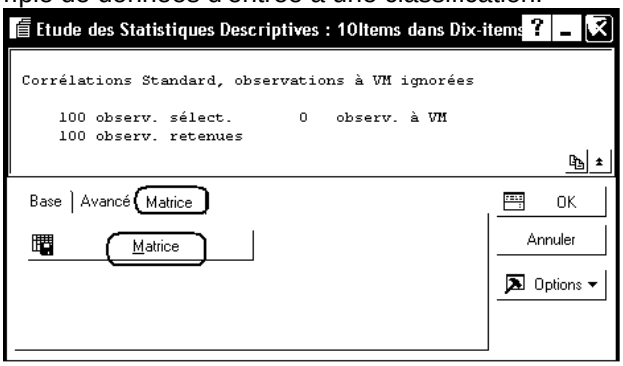
Utilisez le menu Statistiques - Techniques exploratoires multivariées - Fiabilité et analyse d'échelle.

Sélectionnez les variables de "Copier" jusqu'à "Autres" et cliquez sur OK.



On peut alors afficher les corrélations entre les variables à l'aide du bouton "corrélations". Toutefois, le menu Statistiques - Statistiques Elémentaires - Matrices de corrélation permet également de visualiser quels sont les coefficients de corrélation qui sont significatifs d'un lien entre les variables.

L'onglet "Matrice" permet d'afficher les données dans une feuille de données d'un type particulier, une *matrice*, pour servir par exemple de données d'entrée à une classification.

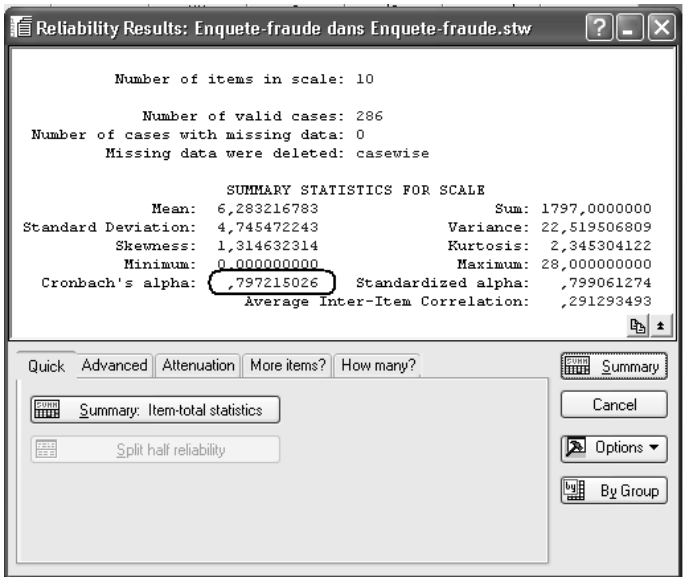


Une telle feuille est caractérisée par la présence d'observations supplémentaires dans le bas du tableau :

	Enquete-fraude dans Enquete-fraude.st	
	1 Copier	2 Communiquer
VolerSujet	0,170	0,158
Autres	0,209	0,156
Means	1,500	1,724
Std.Dev.	1,039	0,990
Nb Obs.	286,00000	
Matrice	1,00000	

et s'enregistre dans un format particulier (fichiers d'extension .smx).

Cliquez ensuite sur le bouton OK. On affiche ainsi la fenêtre de dialogue suivante :



La valeur du coefficient Alpha de Cronbach pour l'ensemble des items est 0,79. Le coefficient standardisé est celui que l'on obtiendrait en effectuant une transformation par centrage et réduction sur chaque variable avant de faire la somme.
Le bouton "Synthèse" permet d'avoir des résultats plus détaillés :

variable	Summary for scale: Mean=6,28322 Std.Dv.=4,74547 Valid N:28 Cronbach alpha: ,797215 Standardized alpha: ,799061 Average inter-item corr.: ,291294				
	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
Copier	4,78	16,20	4,02	0,62	0,76
Communiquer	4,56	16,60	4,07	0,60	0,76
EchangeBrouillon	5,41	16,83	4,10	0,60	0,76
Antiseche	5,24	16,18	4,02	0,55	0,77
SMS	6,04	19,01	4,36	0,50	0,78
CoursGenoux	5,96	19,17	4,38	0,49	0,78
GarderCopie	6,19	21,19	4,60	0,33	0,80
PreparerSalle	6,08	20,03	4,48	0,40	0,79
VolerSujet	6,14	20,23	4,50	0,39	0,79
Autres	6,15	20,44	4,52	0,27	0,80

On voit, par exemple, que l'on pourrait améliorer le coefficient Alpha en retirant la variable "GarderCopie" ou la variable "Autres".

18.3. Corrélation des rangs

La distribution des variables évaluant les différentes techniques de fraude s'écarte notablement d'une loi normale. Pour mesurer les liens existant entre ces variables il peut sembler préférable de calculer des coefficients de corrélation non paramétriques.

Utilisez le menu Statistiques - Tests non paramétriques - Corrélations (Spearman, tau de Kendall, Gamma).

Vous obtenez pour le R de Spearman :

Variable	Coeffs de Corrélations de Rangs de Spearman Cellules à VM ignorées Corrélations significatives marquées à p <,0500		
	Copier	Communiquer	EchangeBrouillon
Copier	1,000	0,606	0,487
Communiquer	0,606	1,000	0,565
EchangeBrouillon	0,487	0,565	1,000

et, pour le tau de Kendall :

Variable	Corrélations du Tau de Kendall (Enquete-fraude dans Enquete-fraude.st Cellules à VM ignorées Corrélations significatives marquées à p <,05000		
	Copier	Communiquer	EchangeBrouillon
Copier	1,000	0,541	0,431
Communiquer	0,541	1,000	0,508
EchangeBrouillon	0,431	0,508	1,000

Quant à la statistique Gamma, l'aide de Statistica 7 indique :

Gamma. La statistique Gamma (Siegel & Castellan, 1988) est préférable au R de Spearman ou au Tau de Kendall lorsque les données contiennent de nombreux ex-aequo. En termes d'hypothèses sous-jacentes, Gamma est équivalent au R de Spearman ou au Tau de Kendall ; en termes d'interprétation et de calculs, il est plus proche du Tau de Kendall que du R de Spearman. En résumé, Gamma est également une probabilité ; plus précisément, il se calcule comme la différence entre la probabilité que le rang de deux variables soit identique, moins la probabilité qu'il soit différent, divisé par 1 moins la probabilité d'ex-aequo. C'est pourquoi, Gamma est en fait équivalent au Tau de Kendall, à la différence que les ex-aequo sont ici, explicitement pris en compte.

19. Régression linéaire à deux ou plusieurs variables

19.1. Régression linéaire à deux variables

On reprend les données Eval-Cours.stw utilisées au paragraphe 18.1. On souhaite déterminer la droite de régression de Qual-Glob par rapport à Pédagogie.

19.1.1 Equation de la droite de régression

On peut, pour cela, utiliser le menu Statistiques - Régression linéaire multiple :

On indique Qual-Glob comme variable dépendante, Pédagogie comme variable indépendante et on clique sur OK.
Le bouton "Synthèse : résultats de la régression" du dialogue suivant permet d'obtenir l'équation de la droite de régression :

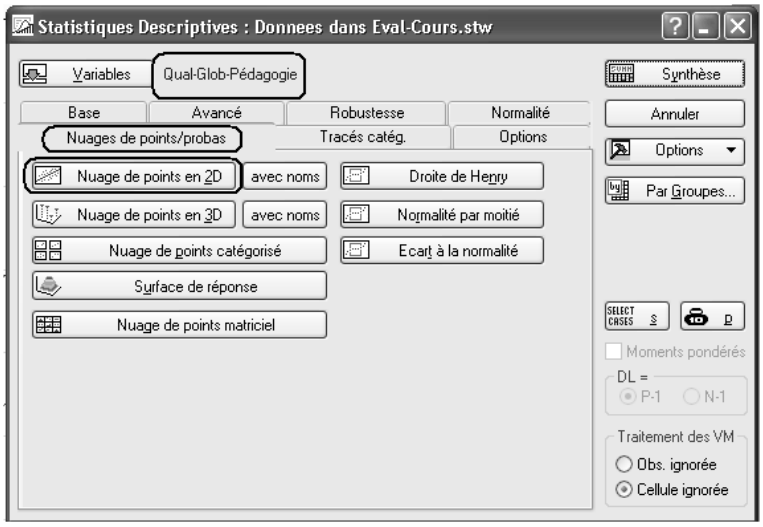
N=50	Synthèse de la Régression; Variable Dép. : Qual-Glob R= ,80386295 R²= ,64619564 R² Ajusté = ,63882472 F(1,48)=87,668 p<,00000 Err-Type de l'Estim.: ,36872					
	b*	Err-Type de b*	b	Err-Type de b	t(48)	valeur p
OrdOrig.			0,1541	0,3664	0,4205	0,6760
Pédagogie	0,8039	0,0859	0,9268	0,0990	9,3631	0,0000

On obtient ainsi comme équation pour la régression :

Qual-Glob = 0,1541 + 0,9268 * Pédagogie.

19.1.2 Nuage de points et droite de régression

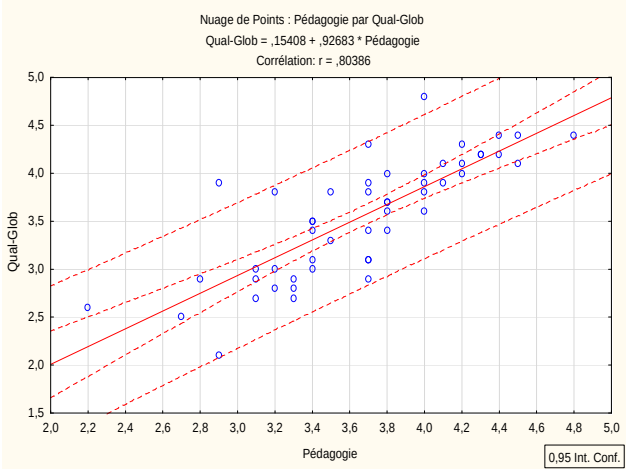
Le plus simple est d'utiliser ici le menu Statistiques - Statistiques Elémentaires - Statistiques Descriptives et l'onglet "Nuages de points/probas" :



Statistica nous affiche le nuage de points, la droite de régression, et les "bandes" donnant l'intervalle de confiance pour la droite de régression, au degré de confiance de 95%. Cet intervalle de confiance correspond aux différentes positions que la droite serait susceptible d'occuper si on recommençait les calculs à partir d'un autre échantillon.

En cliquant sur le graphique à l'aide du bouton droit de la souris, on a accès au menu Propriétés du Graphique (Toutes les Options). L'onglet "Bandes de Régr" permet alors de supprimer les bandes donnant l'intervalle de confiance, ou de leur substituer les représentations graphiques de l'intervalle de détermination, c'est-à-dire la bande du plan qui devrait rassembler 95% des couples (x, y) observés sur la population.

On peut aussi (comme ci-dessous), représenter les deux types de bandes en introduisant un deuxième jeu de bandes à l'aide du bouton "Nouvelles Bandes".



19.2. Régression linéaire à plusieurs variables : recherche d'un modèle explicatif

19.2.1 Présentation de l'exemple

Exercice adapté à partir de "Les disparités géographiques des dépenses de santé: deux modèles explicatifs pour le secteur libéral", de Roquefeuil, L., Solidarité Santé, N° 4, 1996.

Des variations dans le niveau des dépenses de santé allant du simple au double ont été observées entre les départements. Plusieurs variables peuvent expliquer ce phénomène : la densité des médecins libéraux et la

densité de leur clientèle, la morbidité de la population, la proportion de personnes âgées ou l'influence du tiers-payant sur la dépense. Sont étudiées ici :

- l'IDRS ou indicateur des dépenses de remboursement de soins du secteur libéral
- la densité de médecins libéraux dans l'unité géographique concernée
- la mobilité de la clientèle des médecins libéraux : un indicateur de mobilité positif signifie que la valeur des soins "produits" par les médecins de l'unité géographique est supérieure à la valeur des soins "consommés" par la population de l'unité ; un indicateur négatif au contraire, signifie qu'une partie de la population de l'unité va se faire soigner à l'extérieur de celle-ci.
- la mobilité de la clientèle des médecins spécialistes
- le taux de mortalité, corrigé de la structure par âge de la population totale
- la proportion de personnes âgées de 70 ans et plus
- la part (en %) de dépenses de santé réglées en tiers payant.

Deux niveaux d'unités géographiques sont considérés : les données sont fournies par département et par région.

N.B. Les données figurant dans le fichier sont celles indiquées par l'auteur en annexe de son article, et non des données recréées artificiellement.

19.2.2 Etude au niveau départemental

Ouvrez le classeur IDRS.stw et activez la feuille IDRS-Dept.
Affichez les statistiques descriptives relatives aux données présentées. Vous devriez obtenir :

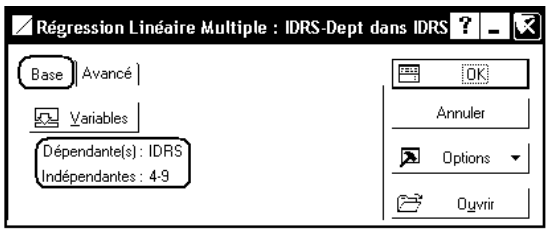
	Statistiques Descriptives (IDRS-Dept dans IDRS.stw)				
Variable	N Actifs	Moyenne	Minimum	Maximum	Ecart-type
IDRS	89	3262,76	2237,40	4959,20	495,39
Densité Médecins	89	178,45	123,50	309,10	37,54
Mobilité omnipraticiens	89	4,49	-10,80	31,20	6,18
Mobilité spécialistes	89	-8,52	-61,60	31,00	19,16
Taux de mortalité	89	9,16	8,10	11,00	0,63
Plus de 70 ans	89	11,01	7,00	17,60	2,30
Tiers payant	89	25,58	11,70	56,80	6,60

Affichez la matrice des corrélations entre les variables :

	Corrélations (IDRS-Dept dans IDRS.stw) Corrélations significatives marquées à p < ,05000 N=89 (Observations à VM ignorées)						
Variable	IDRS	Densité Médecins	Mobilité nipraticien:	Mobilité pécialistes	Taux de mortalité	Plus de 70 ans	Tiers payant
IDRS	1,00	0,67	-0,32	-0,05	-0,13	0,51	0,69
Densité Médecins	0,67	1,00	0,21	0,48	-0,31	0,14	0,36
Mobilité omnipraticiens	-0,32	0,21	1,00	0,39	-0,34	-0,13	-0,28
Mobilité spécialistes	-0,05	0,48	0,39	1,00	-0,11	-0,33	-0,05
Taux de mortalité	-0,13	-0,31	-0,34	-0,11	1,00	-0,38	0,04
Plus de 70 ans	0,51	0,14	-0,13	-0,33	-0,38	1,00	0,15
Tiers payant	0,69	0,36	-0,28	-0,05	0,04	0,15	1,00

Effectuez ensuite une régression linéaire multiple de la variable IDRS sur les autres variables numériques.

Utilisez ensuite le menu Statistiques - Régression Multiple
Sous l'onglet "Base", spécifiez IDRS comme variable dépendante, les 6 autres variables numériques comme variables indépendantes.



Statistica nous affiche alors l'essentiel des résultats de la régression. On peut notamment 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 IDRS et les autres variables :

	Analyse de Variance (IDRS-Dept dans IDRS.stw)				
Effet	Sommes Carrés	dl	Moyennes Carrés	F	niveau p
Régress.	19632968	6	3272161	136,6935	0,0000
Résidus	1962912	82	23938		
Total	21595880				

On peut cliquer sur le bouton OK pour avoir accès à d'autres résultats.

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

N=89	Synthèse de la Régression; Variable Dép. : IDRS (IDRS-Dept dans R= ,95347109 R²= ,90910713 R² Ajusté = ,90245643 F(6,82)=136,69 p<0,0000 Err-Type de l'Estim.: 154,72					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(82)	niveau p
OrdOrig.			-302,094	367,7148	-0,8215	0,4137
Densité Médecins	0,6518	0,0460	8,601	0,6072	14,1659	0,0000
Mobilité omnipraticiens	-0,2355	0,0404	-18,895	3,2444	-5,8239	0,0000
Mobilité spécialistes	-0,1381	0,0450	-3,570	1,1625	-3,0712	0,0029
Taux de mortalité	0,0921	0,0407	72,162	31,9124	2,2613	0,0264
Plus de 70 ans	0,3358	0,0412	72,172	8,8481	8,1567	0,0000
Tiers payant	0,3276	0,0394	24,586	2,9580	8,3117	0,0000

La colonne "B" donne les coefficients de l'équation de régression linéaire. Le modèle fourni par la régression linéaire est le suivant :

IDRS = -302 + 8,6 *Dens. Méd - 18,9 * Mobi Géné - 3,57 * Mobi Spéc + 72.2 * Mort. + 72,2 Part Agées + 24,6 * Tiers-P

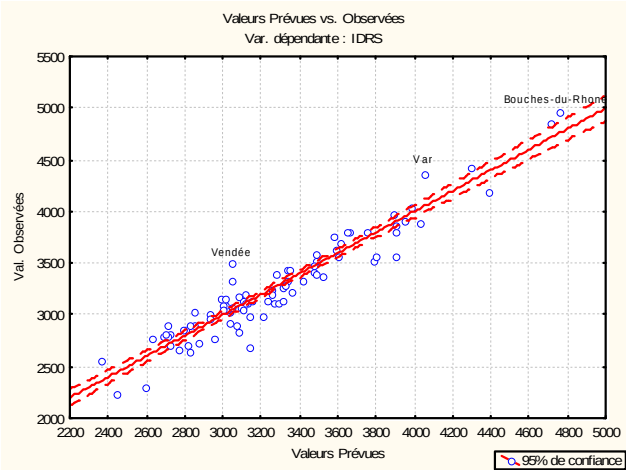
La valeur de R^2 est de 0,91 : 91% de la variance de la variable IDRS est expliquée par le modèle.

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 "Densité Médecins" augmente d'un écart type, la variable "IDRS" estimée augmente de 0,65 écart type, lorsque la variable "Mobilité omnipraticiens" augmente d'un écart type, "IDRS" diminue de 0,23 écart type.
Par exemple, on pourra vérifier que

$$Beta(Densité) = \frac{Ecart\ type(Densité)}{Ecart\ type(IDRS)} \times B(Densité) = \frac{37,54}{495,39} \times 8,601 = 0,6518$$

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.

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 :



19.3. Liens entre les prédicteurs, tolérance, coefficients de corrélation partielle

On reprend les données Eval-Cours.stw. On veut estimer la variable Qual-Glob en utilisant comme prédicteurs les 5 autres variables.
Déterminer l'équation de régression et le coefficient de corrélation. Vous devriez obtenir :

Qual-Glob = - 1.19 + 0.763 Péda + 0.132 Exam + 0.489 Connai - 0.184 Rés +0.000525 Inscr

Mais, est-il bien nécessaire d'utiliser un modèle à 5 prédicteurs ? Un modèle comportant moins de prédicteurs ne serait-il pas tout aussi pertinent ?

On peut déjà noter que seuls les coefficients relatifs à Pédagogie et à Connaissance sont significativement différents de 0 :

N=50	Synthèse de la Régression; Variable Dép. : Qual-Glob R= ,86916574 R²= ,75544908 R² Ajusté = ,72765920 F(5,44)=27,184 p<,00000 Err-Type de l'Estim.: ,32018					
	b*	Err-Type de b*	b	Err-Type de b	t(44)	valeur p
OrdOrig.			-1,1948	0,6312	-1,8931	0,0649
Pédagogie	0,6620	0,1153	0,7632	0,1329	5,7421	0,0000
Examen	0,1061	0,1309	0,1320	0,1628	0,8107	0,4219
Connaissance	0,3251	0,0908	0,4890	0,1365	3,5813	0,0008
Résultat	-0,1055	0,0947	-0,1843	0,1655	-1,1137	0,2715
Inscription	0,1242	0,0922	0,0005	0,0004	1,3472	0,1848

D'autres éléments de réponse peuvent être obtenus à partir du bouton "Corrélations partielles" du dialogue "Résultats".

La colonne "Corrél Partiel." donne les coefficients de corrélation partielle entre la variable Qual-Glob et chacun des prédicteurs, c'est à dire les coefficients observés lorsque les autres variables sont "contrôlées". On voit ici que seuls Pédagogie et Connaissance semblent avoir un effet significatif sur Qual-Glob, résultat qu'on retrouverait en faisant une régression linéaire pas à pas (cf. exemple suivant).

La colonne R^2 donne le carré du coefficient de régression multiple de chacune des variables prédictrices sur les autres prédicteurs. La tolérance est simplement la quantité $1-R^2$. Un R^2 très proche de 1 (par exemple une tolérance inférieure à 0,1) indique que la variable concernée est "presque" une combinaison linéaire des autres variables. Il est alors préférable d'éliminer des prédicteurs dans le modèle.

Variable	Variables dans l'équation ; VD: Qual-Glob (Donnees dans Eval-Cours.stw)						
	Bêta ds	Corrél. Partiel.	Corrél. Semipart	Tolérance	R^2	t(44)	Niveau p
Pédagogie	0,6620	0,6545	0,4281	0,4182	0,5818	5,7421	0,0000
Examen	0,1061	0,1213	0,0604	0,3246	0,6754	0,8107	0,4219
Connaissance	0,3251	0,4751	0,2670	0,6746	0,3254	3,5813	0,0008
Résultat	-0,1055	-0,1656	-0,0830	0,6197	0,3803	-1,1137	0,2715
Inscription	0,1242	0,1990	0,1004	0,6534	0,3466	1,3472	0,1848

Finalement, le modèle le plus pertinent semble être celui ne faisant intervenir que les deux prédicteurs Pédagogie et Connaissance :

N=50	Synthèse de la Régression; Variable Dép. : Qual-Glob R= ,85953033 R^2= ,73879238 R^2 Ajusté = ,72767716 F(2,47)=66,467 p<,00000 Err-Type de l'Estim.: ,32017					
	b*	Err-Type de b*	b	Err-Type de b	t(47)	valeur p
OrdOrig.			-1,2984	0,4773	-2,7200	0,0091
Pédagogie	0,6155	0,0877	0,7097	0,1011	7,0208	0,0000
Connaissance	0,3579	0,0877	0,5383	0,1319	4,0818	0,0002

Qual-Glob = - 1.2984 + 0.7097 Péda + 0.5383 Connai

19.3.1 Calcul "à la main" des coefficients de corrélation partielle

On veut calculer le coefficient de corrélation partielle entre la variable Qual-Glob et la variable Pédagogie. Nous allons procéder en trois étapes :

- Déterminez les résidus de la régression de la variable Qual-Glob par rapport aux 4 autres variables (Examen, Connaissance, Résultat, Inscription).

- Déterminez de même les résidus de la régression de la variable Pédagogie par rapport aux 4 autres variables.

- Créez une nouvelle feuille de données et collez dans les deux premières colonnes de cette feuille les colonnes "Résidus" des feuilles de résultats précédentes.

- Supprimez les 4 dernières observations qui viennent d'être collées (il s'agit de paramètres descriptifs des résidus, sans intérêt ici).

- Calculez enfin le coefficient de corrélation entre les deux variables ainsi définies. Vous devriez retrouver le résultat, à savoir : r=0,65.

19.3.2 Corrélations partielles et neutralisation de l'effet d'un facteur

Ouvrez le fichier Coping.stw du répertoire Corrélations-Partielles.
On a relevé les valeurs de deux variables numériques, RSS et DI, sur 12 sujets, 6 hommes et 6 femmes :

	Sujet	Sexe	RSS	DI	RSS centrée par sexe	DI centrée par sexe
1	s1	F	5	0	-4	-1,33
2	s2	F	8	0	-1	-1,33
3	s3	F	9	2	1	-0,33
4	s4	F	10	1	0	0,67
5	s5	F	10	3	3	0,67
6	s6	F	12	2	1	1,67
7	s7	H	2	0	-3,33	-0,17
8	s8	H	4	0	-1,33	-0,17
9	s9	H	6	0	0,67	-0,17
10	s10	H	6	0	0,67	-0,17

11	s11	H	6	0	0,67	-0,17
12	s12	H	8	1	2,67	0,83

On constate un effet important du facteur sexe : pour les deux variables, les scores des hommes et ceux des femmes sont notablement différents.

Calculer le coefficient de corrélation des variables RSS et DI. On obtient : r = 0,77. Ce coefficient est difficile à interpréter, car il est dû à la fois au lien éventuel entre RSS et DI et à l'effet du facteur Sexe.

Comment neutraliser l'effet du sexe dans le calcul de l'intensité du lien entre RSS et DI ?

1) Faites une régression multiple de DI sur les variables Sexe et RSS, afin d'évaluer les coefficients de corrélation partielle. Vous devriez obtenir :

	Variables dans l'équation ; VD: DI (Donnees dans Classeur1)						
Variable	Bêta ds	Corrél. Partiel.	Corrél. Semipart	Tolérance	R^2	t(9)	Niveau p
Sexe	0,111458	0,129428	0,082673	0,550186	0,449814	0,391578	0,704480
RSS	0,694655	0,631059	0,515257	0,550186	0,449814	2,440493	0,037334

Ainsi, après neutralisation de l'effet du sexe, la corrélation entre RSS et DI n'est que de r'=0,63. Elle reste cependant significative.

2) De manière équivalente, on peut remplacer chaque valeur observée de RSS et DI par son écart algébrique à la moyenne par sexe correspondante. C'est ce qui a été fait dans les colonnes "RSS centrée par sexe" et "DI centrée par sexe".

Par exemple, RSS vaut 5 sur le sujet féminin s1, et vaut 9 pour les femmes. La valeur de "RSS centrée par sexe" sur le sujet s1 est donc : 5 - 9 = -4.

Le coefficient de corrélation des deux variables centrées par sexe est alors exactement le coefficient de corrélation partielle précédent :

	Corrélations (Donnees dans Classeur1) Corrélations significatives marquées à p < ,0500 N=12 (Observations à VM ignorées)	
Variable	RSS centrée par sexe	DI centrée par sexe
RSS centrée par sexe	1,000000	0,631050
DI centrée par sexe	0,631050	1,000000

20. Régression linéaire pas à pas

20.1. Principe de la méthode

Les données sont formées par une VD Y et plusieurs variables explicatives X_1, X_2, ..., X_p.
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 X_1.
On calcule l'équation de régression linéaire de Y sur X_1 : $Y = b_1 X_1 + b_0$.
On calcule alors les résidus : $R_1 = Y - b_1 X_1 - b_0$
On choisit, parmi les variables explicatives restantes, celle qui est le mieux corrélée à R_1. Nous supposons ici qu'il s'agit de la variable X_2.
On calcule l'équation de régression linéaire de Y sur X_1 et X_2 : $Y = b'_1 X_1 + b_2 X_2 + b'_0$.
On calcule les nouveaux résidus : $R_2 = Y - (b'_1 X_1 + b_2 X_2 + b'_0)$ et on poursuit la méthode jusqu'à ce que les variables explicatives restantes ne soient plus significativement corrélées aux résidus.

20.2. Présentation de l'exemple

Exercice adapté à partir de "Intelligence pratique ou traditionnelle : Que mesure l'entrevue structurée situationnelle ?", Durivage A., St-Martin J., Barette J., Revue européenne de Psychologie Appliquée, 1995, vol. 45 n° 3, pp. 171-178.

L'objectif de l'étude consiste à explorer le construit sous-jacent à l'entrevue structurée situationnelle lors de la sélection du personnel. Constitue-t-elle une mesure de l'intelligence traditionnelle (QI) ou de connaissances tacites associées théoriquement à de l'intelligence pratique.

Méthodologie : l'entrevue structurée situationnelle et les tests ont été administrés à 48 candidats potentiels à un poste de responsable des bénévoles dans un centre hospitalier psychiatrique du Québec. Les variables suivantes ont été recueillies :

- Score à l'entrevue structurée : échelle de 0 à 40

- Score au BGTA : batterie générale de tests d'aptitude mesurant l'intelligence traditionnelle. Ce score est ici donné sous la forme d'une variable centrée et réduite

- Scores sur les dimensions "Organisation", "Impulsivité", "Compréhension", "Altruisme" des tests de personnalité de Jackson (échelles de 0 à 20)

- Age (de 20 à 41 ans dans l'expérience originale)

- Le nombre d'années d'expérience de travail à temps plein (de 0 à 21 ans dans l'expérience originale).

20.3. Régression linéaire pas à pas de Entrevue sur les autres variables

Chargez le classeur Entrevue-structuree.stw, qui contient des données générées artificiellement, conformes aux résultats indiqués par les auteurs.

Affichez les statistiques descriptives relatives aux données présentées. Vous devriez obtenir :

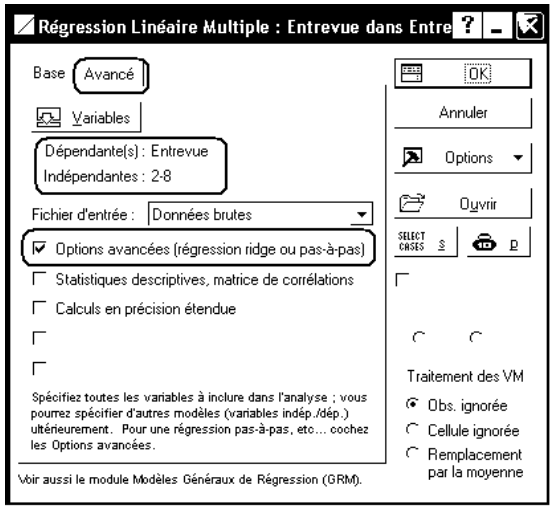
	Statistiques Descriptives (Entrevue dans Entrevue-structuree.stw				
Variable	N Actifs	Moyenne	Minimum	Maximum	Ecart-type
Entrevue	48	30,09	20,77	38,58	4,43
Comprehension	48	12,18	5,09	19,02	3,31
Organisation	48	13,06	6,05	20,70	3,69
Altruisme	48	12,00	2,49	19,58	4,25
Impulsivite	48	9,51	2,66	16,48	3,41
BGTA	48	-0,00	-2,08	2,31	1,01
Age	48	25,00	16,11	38,06	4,75
Ancienne te	48	8,60	0,00	18,62	4,02

Affichez la matrice des corrélations entre les différentes variables. Quelles sont les corrélations qui apparaissent significatives ?

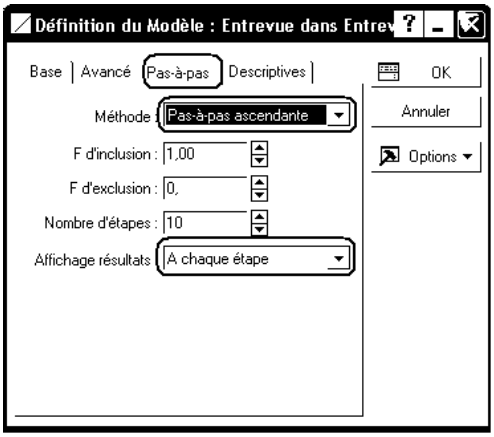
	Corrélations (Entrevue dans Entrevue-structuree.stw) Corrélations significatives marquées à p < ,05000 N=48 (Observations à VM ignorées)							
Variable	Entrevue	ıprehensi	rganisation	Altruisme	ıpulsivite	BGTA	Age	ıncienne te
Entrevue	1,00	0,48	-0,14	0,02	0,03	-0,06	0,40	0,33
Comprehension	0,48	1,00	-0,18	0,00	0,12	0,07	0,01	0,04
Organisation	-0,14	-0,18	1,00	0,18	-0,51	-0,12	0,02	0,06
Altruisme	0,02	0,00	0,18	1,00	-0,19	-0,22	-0,05	0,00
Impulsivite	0,03	0,12	-0,51	-0,19	1,00	-0,22	-0,17	-0,28
BGTA	-0,06	0,07	-0,12	-0,22	-0,22	1,00	0,03	0,01
Age	0,40	0,01	0,02	-0,05	-0,17	0,03	1,00	0,83
Ancienne te	0,33	0,04	0,06	0,00	-0,28	0,01	0,83	1,00

20.3.1 Exécution de la procédure

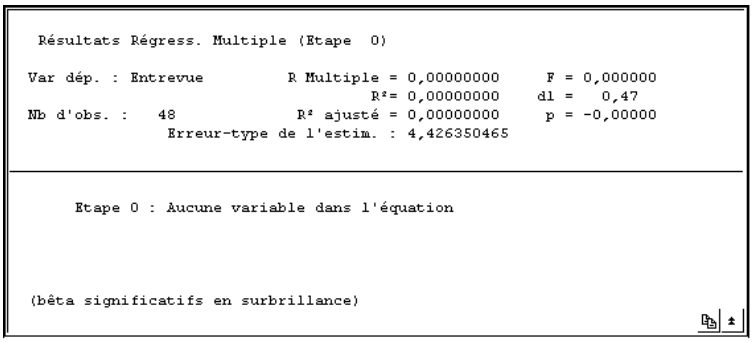
Utilisez ensuite le menu Statistiques - Régression Multiple
Sous l'onglet "Avancé", spécifiez Entrevue comme variable dépendante, les 7 autres variables comme variables indépendantes. Cochez l'option "régression ridge ou pas-à-pas".



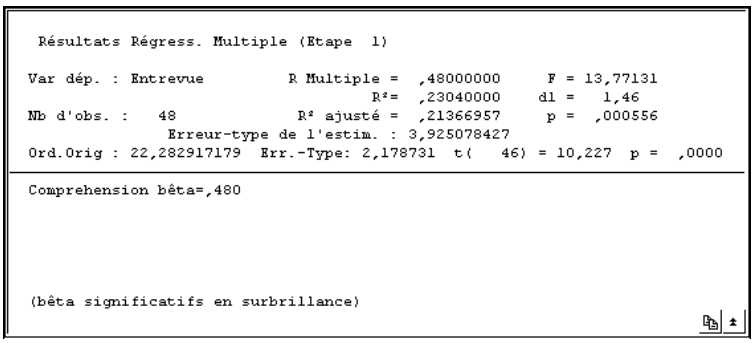
Dans le dialogue suivant, activez l'onglet "pas-à-pas" et sélectionnez la méthode "pas à pas ascendante", et l'affichage des résultats à chaque étape :



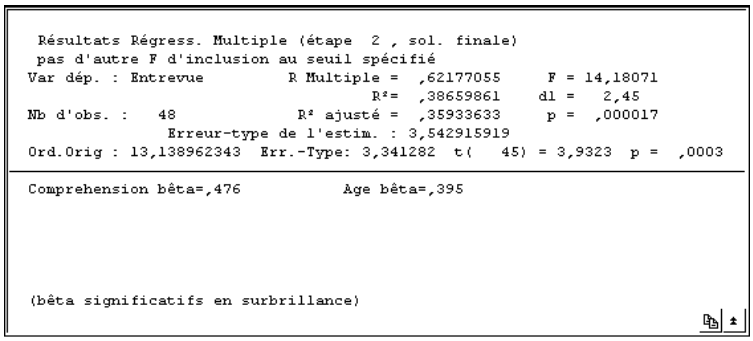
A la première étape, Statistica affiche les résultats suivants :



Cliquez sur "suivant". On obtient :



puis :



Il ne reste plus de variable significativement corrélée aux résidus, et Statistica substitue le bouton "OK" au bouton "Suivant". Cliquez sur ce bouton.

20.3.2 Analyse des résultats

Sous l'onglet "Avancé", le bouton "Synthèse de la régression" permet d'obtenir les résultats suivants :

	Synthèse de la Régression; Variable Dép. : Entrevue (Entrevue d R= ,62177055 R²= ,38659861 R² Ajusté = ,35933633 F(2,45)=14,181 p<,00002 Err-Type de l'Estim.: 3,5429					
N=48	Bêta	Err-Type de Bêta	B	Err-Type de B	t(45)	niveau p
OrdOrig.			13,14	3,34	3,93	0,0003
Comprehension	0,48	0,12	0,64	0,16	4,08	0,0002
Age	0,40	0,12	0,37	0,11	3,39	0,0015

Ainsi, l'équation de régression obtenue par ce modèle est :

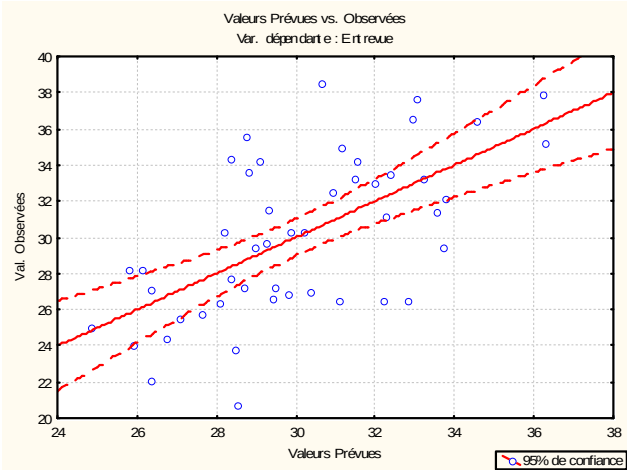
Entrevue = 0,64 * Comprehension + 0,37 * Age + 13,14

Ce modèle explique 38% de la variance de la variable Entrevue. 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 "Comprehension" augmente d'un écart type, la variable "Entrevue" estimée augmente de 0,48 écart type, lorsque la variable "Age" augmente d'un écart type, "Entrevue" augmente de 0,4 écart type.
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 également obtenir le tableau des valeurs observées et des valeurs estimées de la variable Entrevue (Onglet "Avancé", bouton Synthèse : Résidus et prévisions)

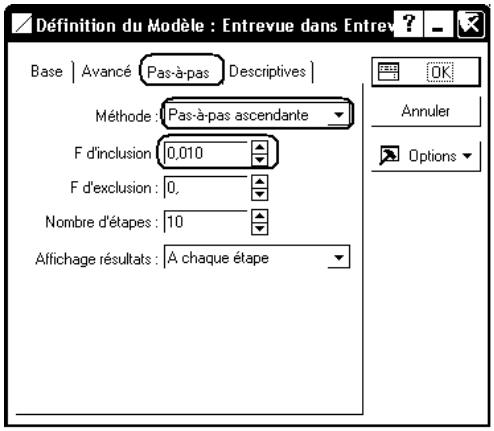
	Valeurs Prévues & Résidus (Entre Var. dépendante : Entrevue		
N° d'Obs.	Valeur Observée	Valeur Prévue	Résidus
s1	32,47	30,92	1,54
s2	28,22	26,14	2,09
s3	35,16	36,28	-1,12

Diverses représentations graphiques peuvent être obtenues. Un nuage de points tri-dimensionnel Entrevue - Comprehension - Age serait peu lisible. En revanche, on pourra construire un graphique comparant les valeurs observées aux valeurs estimées par le modèle (Onglet "Nuages", bouton "Prévues v/s observées") :



20.3.3 Variantes

On peut souhaiter recueillir également des informations concernant les variables qui ont été écartées du modèle. Reprenez par exemple l'étude, en indiquant cette fois 0,01 comme valeur limite de F pour inclure une variable :



La régression pas à pas est alors faite sur toutes les variables, avec les résultats suivants. On notera que l'ajout des 5 variables restantes ne permet pas vraiment d'augmenter la part de variance expliquée (40% au lieu de 38%). On notera que, lorsqu'est introduite la variable "Ancienneté", fortement corrélée à l'âge, ni "Ancienneté" ni "Age" ne restent significatifs.

N=48	Synthèse de la Régression; Variable Dép. : Entrevue (Entrevue da R= ,63694958 R²= ,40570477 R² Ajusté = ,30170311 F(7,40)=3,9009 p<,00250 Err-Type de l'Estim.: 3,6988					
	Bêta	Err-Type de Bêta	B	Err-Type de B	t(40)	niveau p
OrdOrig.			14,3833	6,0676	2,3705	0,0227
Comprehension	0,4728	0,1249	0,6314	0,1668	3,7852	0,0005
Age	0,4415	0,2203	0,4115	0,2053	2,0038	0,0519
BGTA	-0,1208	0,1360	-0,5293	0,5957	-0,8886	0,3795
Organisation	-0,1001	0,1496	-0,1201	0,1795	-0,6692	0,5072
Altruisme	0,0259	0,1298	0,0269	0,1351	0,1992	0,8431
Impulsivite	-0,0402	0,1629	-0,0523	0,2117	-0,2468	0,8064
Anciennete	-0,0557	0,2279	-0,0614	0,2509	-0,2446	0,8080

20.4. Exercice à rendre par mail

Réalisez l'étude demandée dans l'exercice ci-dessous. Faites parvenir votre travail (classeur Statistica contenant les traitements demandés, commentaire saisi dans un rapport Statistica ou un fichier Word) par mail à votre enseignant (adresse : Francois.Carpentier@univ-brest.fr).

Source des données : Source : http://www.amstat.org/publications/jse/jse_data_archive.html

Dans les débats autour des réformes du système éducatif, il a été fréquemment affirmé que la dépense d'éducation par élève n'était pas un gage de réussite. A l'appui de cette affirmation, certains ont souligné que parmi les dix états américains dont la dépense moyenne par élève était la plus basse en 1994/95, quatre se trouvaient parmi les dix états qui avaient les meilleurs résultats au SAT.

Pour étudier cette question, des statisticiens américains ont extrait du *Digest of Education Statistics* les données suivantes :

Nom de l'état

Dépense par élève moyenne dans les écoles publiques élémentaires et secondaires, en milliers de dollars (1994-95)

Taux d'encadrement : ratio élève par enseignant

Salaire annuel moyen estimé des enseignants des écoles publiques élémentaires et secondaires

Pourcentage d'inscrits au SAT parmi les élèves satisfaisant les conditions d'inscription

Score moyen observé au SAT verbal

Score moyen observé au SAT mathématique

Score moyen observé global au SAT global.

Les données correspondantes se trouvent dans le classeur Statistica SATdata.stw du serveur des salles de TD.

1) Etudier la corrélation entre la dépense moyenne par élève et le score moyen observé au SAT. La corrélation est-elle significative ? Interpréter le signe du coefficient de corrélation.
En effectuant une régression linéaire du score sur la dépense moyenne, retrouver le résultat indiqué dans la source :

> *"every $1,000 increase in spending per student per year is associated with a decline of nearly 21 points in the average statewide SAT score, an estimate that easily reaches conventional levels of statistical significance ($p < .01$)."*

Représenter le nuage de points et la droite de régression.

2) Etudier la corrélation entre les variables "Pourcentage Inscrits au SAT" et "Score global au SAT". Représenter le nuage de points correspondant. Que peut-on en conclure ?

3) Etudier la régression linéaire multiple de la variable "Score global au SAT" sur les deux variables "Pourcentage Inscrits au SAT" et "Dépense par élève". Analyser les résultats obtenus. Compléter cette étude par le calcul et l'interprétation des coefficients de corrélation partiels.
Retrouver ainsi la conclusion :

> *"With a robust R^2 and slope coefficients that are both highly statistically significant ($p < .01$), it is now clear that while the bulk of variation in statewide SAT scores is attributable to the percentage of students taking the exam, increased spending on public education is in fact associated with better academic performance."*