



HAL
open science

Comparaison de plusieurs droites de regression. Programme BASIC d'application

Frederic Lardeux

► **To cite this version:**

Frederic Lardeux. Comparaison de plusieurs droites de regression. Programme BASIC d'application. Rapport ITRMLM n°38/88/ITRM/Doc.Ent, Institut Louis Malardé, Tahiti, French Polynesia. 1988, pp.1-22. hal-04709098

HAL Id: hal-04709098

<https://hal.science/hal-04709098v1>

Submitted on 25 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

I.T.R.M.L.M.
B.P. 30
Papeete - Tahiti

O.R.S.T.O.M
B.P. 529
Papeete - Tahiti

COMPARAISON DE PLUSIEURS DROITES DE REGRESSION

PROGRAMME BASIC D'APPLICATION

F. LARDEUX

Décembre 1988

Référence ITRMLM : 38 / 88 / ITRM / Doc-Ent

PLAN

INTRODUCTION

1 - COMPARAISON DE K DROITES DE REGRESSION

1.1 - Généralités

1.2 - Test d'égalité des variances résiduelles

1.3 - Analyse de covariance

1.3.1 - Test d'égalité des pentes

1.3.2 - Test d'égalité des ordonnées à l'origine

2 - LE PROGRAMME INFORMATIQUE

2.1 - Description générale

2.2 - Fichiers des données

2.3 - Mise en route du programme

2.4 - Spécificités

CONCLUSION

ANNEXE : LISTING DU PROGRAMME

INTRODUCTION

Le programme présenté ici est issu de l'analyse donnée par SCHERRER (1984).

Avant de décrire le programme informatique, on reprendra brièvement la présentation théorique du problème, exposée dans le livre de SCHERRER, afin de mieux comprendre et interpréter les résultats du programme.

Le programme présenté ici ne tient compte que du cas où les diverses droites dérivent du modèle $Y = a + b.X$, où X est une variable contrôlée. Le cas où les deux variables X et Y sont aléatoires (d'où une estimation de la droite de régression par l'axe majeur réduit), n'est pas étudié.

Le programme informatique proprement dit, se veut simple dans son installation et son utilisation. Ecrit en TURBO BASIC, il est facilement modifiable en fonction des besoins spécifiques d'autres programmeurs. Une version compilée est disponible au laboratoire, pour les personnes ne disposant pas de TURBO BASIC.

1 - COMPARAISON DE K DROITES DE REGRESSION

1.1 - Généralités

Soit le modèle linéaire $Y = a + b.X$.

Il arrive fréquemment que la relation entre Y et X soit obtenue sur plusieurs échantillons (par exemple sur une population de mâles, de femelles, d'immatures) et que l'on désire comparer les droites de régression pour savoir si on peut les considérer comme identiques ou non.

La méthode utilisée consiste, dans un premier temps, à tester le parallélisme des droites par comparaison statistique de leurs pentes. Dans un deuxième temps, et à condition que les droites soient parallèles, on peut tester si les ordonnées à l'origine sont identiques ou non. Si les ordonnées sont identiques, les droites peuvent alors être considérées comme confondues.

Si il n'y a que deux droites de régression, un simple test t permet de comparer les deux pentes et les deux ordonnées à l'origine (ZAR, 1974). Si le nombre de droites dépasse deux, une analyse plus complexe est nécessaire. Il s'agit alors de soumettre les échantillons à une analyse de covariance. Cette méthode reste cependant valide dans le cas de deux échantillons.

L'analyse de covariance teste l'hypothèse nulle H_0 : "les pentes des k populations d'origine sont égales" contre l'hypothèse alternative H_1 : "toutes les pentes ne sont pas égales" (c'est à dire, qu'il existe au moins une des pentes qui diffère des autres). Si H_0 est acceptée, on peut continuer l'analyse pour comparer les ordonnées à l'origine : H_0 "Toutes les ordonnées sont égales" contre H_1 "une des ordonnées au moins diffère des autres".

L'analyse de covariance repose sur l'hypothèse d'égalité des variances résiduelles (c'est à dire, des variances des résidus d'estimation). Il faut donc tout d'abord vérifier cette condition.

1.2 - Test d'égalité des variances résiduelles

L'égalité des variances résiduelles peut être vérifiée par un test de Bartlett (BARTLETT, 1937) ou un test Log-anova (MARTIN et GAMES, 1977), ou tout autre test équivalent.

En général les hypothèses à tester sont du type $H_0 =$ "toutes les variances sont égales" contre $H_1 =$ "une au moins des variances est différente".

Le test de Bartlett est très sensible aux données non normales, ce qui implique que le rejet de H_0 peut indiquer un manque de normalité comme un manque d'homogénéité des variances. La confusion entre ces deux interprétations peut constituer un handicap. Toutefois, si la régression a été correctement analysée, les résidus sont distribués selon une loi normale et le test de Bartlett ne mettra alors en évidence qu'une inégalité des variances, si elle existe. Ce test consiste à calculer une statistique (notée B) qui suit approximativement une loi du Chi-2 à $k-1$ degrés de liberté. La conformité à la loi du chi-2 a pu être améliorée grâce à un facteur de correction. Ce facteur a été inclus dans le programme. En pratique, on rejette H_0 (c'est à dire l'homogénéité des variances) si la valeur du test est supérieure au Chi-2 tabulé. La statistique de Bartlett est calculée par le programme.

Le test Log-anova est moins sensible à la non normalité des données. Aussi, comme on vient de le voir, il peut paraître luxueux d'utiliser un test robuste sur des données (les résidus) supposées suivre une loi normale. Ce test est néanmoins calculé par le programme.

Le test Log-anova consiste à subdiviser les éléments de chacun des k groupes en un certain nombre de sous-groupes (en fait, \sqrt{n} sous-groupes). Les éléments sont affectés aléatoirement dans les sous-groupes issus de leur groupe d'origine, en prenant soin de constituer, dans la mesure du possible, des sous-groupes de taille égale. La variance de chaque sous-groupe est calculée et une transformation logarithmique du résultat est effectuée afin de normaliser au mieux la distribution des variances.

On vérifie alors, par analyse de variance, si la variation inter-groupe est supérieure à la variation intra-groupe qui résulte uniquement des fluctuations aléatoires d'échantillonnage. Si tel est le cas, on retient l'hypothèse d'inégalité des variances. Sinon, on accepte l'hypothèse d'égalité des variances des k groupes.

Le tableau d'analyse de variance a l'allure suivante :

Source de variation	Somme des carrés des écarts	degrés de liberté	variance
inter groupe	SCI	$k - 1$	$V_i = SCI / (k-1)$
intra groupe	SCE	ui	$V_e = SCE / ui$
Totale	SCT		

Le test F est $F_c = V_i / V_e$. On rejette H_0 si F_c est supérieur ou égal au F tabulé, au seuil de signification α , pour $k-1$ et ui degrés de liberté. Le test est unilatéral, de par sa conception.

Ce tableau d'analyse de variance est calculé par le programme informatique.

1.3 - Analyse de covariance.

1.3.1 - Test de l'égalité des pentes.

Les étapes de calcul du test sont les suivantes :

- On calcule la somme des carrés des écarts (erreurs résiduelles) de chaque droite de régression.
 - On additionne ces sommes pour obtenir l'erreur résiduelle totale (SCEET)
 - On calcule l'erreur résiduelle par rapport à une série de k droites parallèles dont la pente est égale à la pente pondérée des droites observées (SCEEDP). Le nombre de degrés de liberté associé est noté DDLe_{dp}.
- Les erreurs d'estimation obtenues avec des droites de même pente seront plus élevées qu'avec des droites dont la pente est ajustée pour chaque échantillon. Cette erreur sera aussi d'autant plus grande que les différences de pente seront élevées. Par conséquent, la différence entre SCEEDP et SCEET rend compte des divergences entre les pentes.
- On calcule donc SCEEVP = SCEEDP - SCEET (avec DDLe_{vp} degrés de liberté)
 - Le test F correspondant à H_0 indique si l'erreur résiduelle due à la variation des pentes est trop grande pour être explicable par les fluctuations fortuites d'échantillonnage. On le calcule à l'aide de SCEEVP et SCEEDP :

$$F_{vp} = (SCEEVP / DDLe_{vp}) / (SCEEDP / DDLe_{dp})$$

Si F_{vp} est supérieur ou égal à la valeur critique du F tabulé, on rejette l'hypothèse d'égalité des pentes.

1.3.2 - Test d'égalité des ordonnées à l'origine

Si l'hypothèse d'égalité des pentes est acceptée, et uniquement dans ce cas, il est possible de procéder à une comparaison des ordonnées à l'origine.

Les hypothèses deviennent alors :

- H_0 : toutes les ordonnées à l'origine sont égales.
- H_1 : une au moins des ordonnées à l'origine est différente.

Si toutes les pentes sont égales, il est alors possible de combiner les données de tous les échantillons pour calculer une droite de régression commune à l'ensemble des n observations.

L'erreur résiduelle par rapport à cette droite, symbolisée par SCEEC s'obtient selon la démarche habituelle appliquée à un échantillon unique.

Les étapes des calculs sont donc :

- calcul de l'erreur résiduelle SCEEC (droite "commune")
- calcul de l'erreur résiduelle due aux différences d'ordonnées à l'origine $SCEEVOO = SCEEC - SCEEDP$, avec $k - 1$ degrés de liberté.
- On teste alors si l'erreur résiduelle due aux différences de position est trop grande pour être le fruit des fluctuations fortuites d'échantillonnage par le test F suivant :

$$F_{voo} = (SCEEVOO / (k-1)) / (SCEEDP / DDLeDP)$$

remarque : DDLeDP vaut $n - k - 1$.

L'hypothèse d'égalité des ordonnées à l'origine sera acceptée si F_{voo} est inférieur à la valeur du F tabulé pour $k-1$ et $n-k-1$ degrés de liberté.

Le tableau général d'analyse de covariance des droites de régression de k échantillons et les règles de décision relatives à cette analyse peuvent se résumer selon les tableaux 1 et 2.

2 - LE PROGRAMME INFORMATIQUE

2.1 - Description générale

Le programme proposé ici a été écrit en TURBO BASIC sur un compatible IBM-PC (Goupil G4).

Il effectue les calculs suivants :

- Calcul du test de Bartlett et du test Log-anova d'homogénéité des variances résiduelles.
- Calcul de chaque droite de régression : pente (et son écart type), ordonnée à l'origine, coefficient de corrélation.
- Calcul de la régression "commune".
- Calcul du tableau d'analyse de covariance tel qu'il a été présenté précédemment.
- Calcul des tests d'égalité des pentes et des ordonnées à l'origine.

Les résultats apparaissent à l'écran et leur sortie sur imprimante est possible.

Le programme a été écrit dans l'optique d'une utilisation la plus conviviale possible. L'utilisateur n'a pas besoin de connaissances en programmation : il doit simplement répondre aux questions posées en clair par le programme.

Nature de la variation	Calculs préliminaires	Erreurs résiduelles	ν	Variances résiduelles
Échantillon 1	$(n_1 - 1) s_{x_1}^2, (n_1 - 1) s_{xy_1}, (n_1 - 1) s_{y_1}^2$	$SCEE_1 = \sum (y_1 - \hat{y}_1)^2$	$n_1 - 2$	$s_{e_1}^2 = \frac{SCEE_1}{n_1 - 2}$
Échantillon j	$(n_j - 1) s_{x_j}^2, (n_j - 1) s_{xy_j}, (n_j - 1) s_{y_j}^2$	$SCEE_j = \sum (y_j - \hat{y}_j)^2$	$n_j - 2$	$s_{e_j}^2 = \frac{SCEE_j}{n_j - 2}$
Échantillon k	$(n_k - 1) s_{x_k}^2, (n_k - 1) s_{xy_k}, (n_k - 1) s_{y_k}^2$	$SCEE_k = \sum (y_k - \hat{y}_k)^2$	$n_k - 2$	$s_{e_k}^2 = \frac{SCEE_k}{n_k - 2}$
k droites différentes		$SCEET = \sum_{j=1}^k SCEE_j$	$\sum_{j=1}^k (n_j - 2)$	$s_{eT}^2 = \frac{SCEET}{n - 2k}$
k droites parallèles	$\underbrace{\sum_{j=1}^k (n_j - 1) s_{x_j}^2}_A, \underbrace{\sum_{j=1}^k (n_j - 1) s_{xy_j}}_B, \underbrace{\sum_{j=1}^k (n_j - 1) s_{y_j}^2}_C$	$SCEEDP = C - \frac{B^2}{A}$	$n - k - 1$	$s_{eDP}^2 = \frac{SCEEDP}{n - k - 1}$
Différence de pentes		$SCEEVP = SCEEDP - SCEET$	$k - 1$	$s_{eVP}^2 = \frac{SCEEVP}{k - 1}$
Une seule droite commune	$(n - 1) s_x^2, (n - 1) s_{xy}, (n - 1) s_y^2$	$SCEEC = \sum (y - \hat{y})^2$	$n - 2$	$s_{eC}^2 = \frac{SCEEC}{n - 2}$
Différence de position		$SCEEVOO = SCEEC - SCEEDP$	$k - 1$	$s_{eVOO}^2 = \frac{SCEEVOO}{k - 1}$

Tableau 1 : Tableau d'analyse de covariance pour comparer k droites de régression. (d'après SCHERRER, 1984)

Hypothèse principale	Hypothèse alternative	Rejet de H_0 si	Lecture de la valeur critique	Remarque
$H_0 : \alpha_1 = \alpha_j = \alpha_k$ Pentes égales	$H_1 : \text{tous les } \alpha \text{ ne sont pas égaux}$	$F_{VP} > F_\alpha$	F_α $\nu_1 = k - 1$ et $\nu_2 = n - k - 1$ degrés de liberté	$F_{VP} = \frac{s_{eVP}^2}{s_{eDP}^2}$
$H_0 : \beta_1 = \beta_j = \beta_k$ Positions identiques	$H_1 : \text{toutes les ordonnées à l'origine ne sont pas identiques}$	$F_{VOO} > F_\alpha$	F_α $\nu_1 = k - 1$ et $\nu_2 = n - k - 1$ degrés de liberté	$F_{VOO} = \frac{s_{eVOO}^2}{s_{eDP}^2}$ Le test F_{VOO} ne peut être réalisé que si $F_{VP} < F_\alpha$

Tableau 2 : Tableau récapitulatif des règles de décision relatives à une analyse de covariance pour comparer les pentes et les ordonnées à l'origine de k droites de régression. (d'après SCHERRER, 1984)

2.2 - Fichiers des données

La mise en route du programme suppose, au préalable, que les données brutes correspondant aux diverses droites de régression soit stockées sur un ou plusieurs fichiers.

Ces fichiers doivent être séquentiels, en format ASCII. Il ne doit pas y avoir de noms de variables : seules les données chiffrées doivent être inscrites sur ces fichiers.

Ces fichiers sont supposés correspondre à une saisie des données représentables sous forme de matrice de i lignes et c colonnes. Chaque ligne représente un individu et chaque colonne représente une variable (fig. 1).

Ainsi, un fichier minimal sera composé de i lignes et 2 colonnes, l'une correspondant à la variable Y de la régression, l'autre à la variable X . Autrement, la variable Y et la variable X peuvent être situées à n'importe quel rang dans le fichier, le programme informatique demandant à quel rang dans le fichier se situe la variable Y et la variable X .

Les variables Y et X d'une même régression doivent être sur le même fichier. Par contre, il n'est pas nécessaire que chaque régression soit sur un fichier différent : par exemple, si on désire comparer trois régressions, il se peut que les deux premières soit dans un même fichier (avec, par exemple, rang de $Y1 = 3$, rang de $X1 = 6$, rang de $Y2 = 5$, rang de $X2 = 2$) et la troisième dans un autre fichier.

Pour un même fichier, le nombre i de lignes est limité, dans le programme, à 2000 et le nombre de colonnes à 100.

2.3 - Mise en route du programme

Il suffit de charger TURBO BASIC puis le programme COMPREG et de l'exécuter.

Pour la version compilée, il suffit de se placer dans le répertoire où se trouve le programme et de taper COMPREG.

Lors de l'exécution du programme, un fichier temporaire (nommé \$\$\$ dans le programme et détruit automatiquement à la fin de l'exécution) est créé. Ce fichier correspond à la concaténation des variables X et Y des diverses droites de régression, nécessaire pour le calcul de la droite "commune". Il faut donc s'assurer qu'il existe suffisamment de place disponible sur le disque (ou la disquette) où se trouve COMPREG.

Une fois le programme lancé, il suffit de suivre les indications portées à l'écran : un premier écran apparaît (fig. 2), indiquant le but du programme, la publication source, le nom et l'adresse du programmeur ainsi qu'une question : "nombre de droites à comparer". Il suffit d'indiquer ce nombre. Le programme est limité à la comparaison de 30 droites de régression.

Ensuite, le programme demande pour chaque droite, quatre spécifications :

- Le nom du fichier (avec son chemin d'accès complet). Par exemple, si les données correspondant à une régression sont stockées dans un fichier ayant pour nom TIKE1, situé dans le même répertoire que le programme, il suffit d'indiquer : TIKE1. Si ce fichier n'est pas dans le même répertoire, il faut

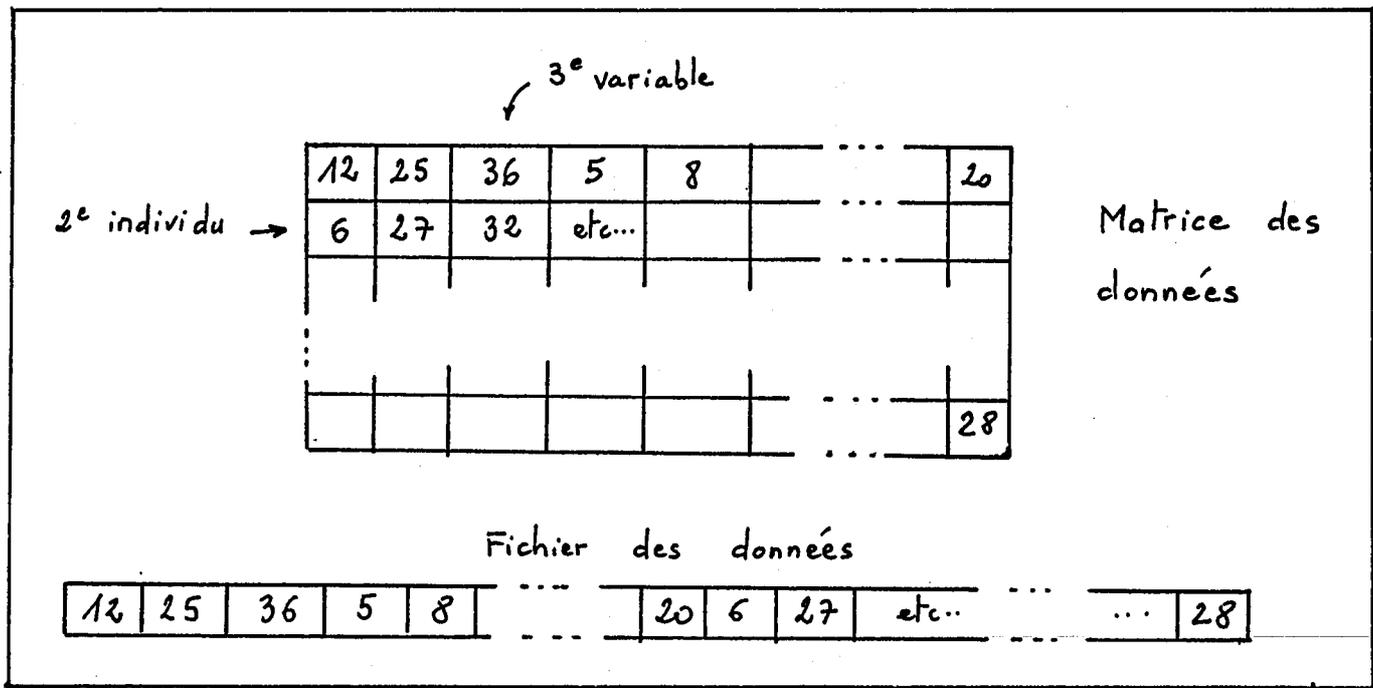


Figure 1 : Exemple de structure de fichier des données.

COMPARAISON DE K DROITES DE REGRESSION

Source : SCHERRER - 1984 - Biostatistiques. Gaëtan morin Ed.
pp. 386 et seq. et 676 et seq.

Programme F. LARDEUX - Centre ORSTOM de Tahiti. B.P. 529

Nombre de droites à comparer :

Figure 2 : Premier écran.

Droite n° 1
 Nom du fichier (avec son chemin d'accès) : TIKE1
 Nombre total de variables dans ce fichier : 14
 Rang de la variable Y : 2
 Rang de la variable X : 8

Droite n° 2
 Nom du fichier (avec son chemin d'accès) : TIKE1
 Nombre total de variables dans ce fichier : 14
 Rang de la variable Y : 2
 Rang de la variable X : 12

Droite n° 3
 Nom du fichier (avec son chemin d'accès) : C:\DYN\RANGI
 Nombre total de variables dans ce fichier : 3
 Rang de la variable Y : 1
 Rang de la variable X : 2

Figure 3 : Saisie des caractéristiques des fichiers des données.

indiquer le chemin d'accès complet. Par exemple, C:\DYN\TIKE1 si le fichier se trouve sur un disque dur (C:) dans le répertoire DYN. La figure 3 donne un exemple des diverses possibilités : deux régressions dans un même fichier TIKE1 sur le même répertoire que le programme et une troisième dans un autre fichier sur un autre répertoire du disque dur : C:\DYN\RANGI.

- Le nombre total de variables dans le fichier
- Le rang de la variable Y
- Le rang de la variable X

Pour ces trois questions, il suffit de répondre simplement, comme le montre l'exemple de la figure 3.

Une fois ces données initiales saisies, le programme effectue tous les calculs. Les exemples de calculs illustrant les possibilités du programme sont issus de LARDEUX et TETUANUI (1988), qui ont comparé les droites de régression de la largeur de la capsule céphalique en fonction du stade de développement chez les larves de deux espèces de moustiques : Aedes aegypti et Aedes polynesiensis.

Un écran apparaît avec le calcul du test de Bartlett et le tableau d'analyse de variance du test Log-anova (Fig. 4). Le programme demande par O (=Oui) ou N (= Non) si on désire continuer l'exécution. La réponse dépend de la validité du test de Bartlett (ou du test Log-anova). En cas de réponse négative, l'exécution du programme est arrêtée.

Si les calculs sont poursuivis, le programme affiche successivement les écrans correspondant aux calculs de chaque droite de régression (Fig. 5), en rappelant pour chaque droite le nom du fichier, le rang de la variable X et celui de la variable Y. Ceci permet de vérifier qu'il n'y a pas eu d'erreur de saisie lors du lancement du programme.

Chaque écran comporte les calculs relatifs à deux droites de régression. Pour afficher les résultats suivants, il suffit de faire un 'Retour-chariot', comme l'indique le programme. Le dernier écran, correspondant aux droites de régression, contient les calculs de la régression "commune" (c'est à dire, avec tous les couples X,Y de toutes les droites, mélangés) (Fig. 6).

L'écran suivant affiche les résultats de l'analyse de covariance comme indiqué au tableau 1 précédent (fig. 7).

Le dernier écran redonne le résultat du test de Bartlett et du test Log-anova et affiche les tests F d'égalité des pentes et des ordonnés à l'origine (Fig. 8).

Enfin, une possibilité de sortie de ces résultats sur l'imprimante est offerte. Il suffit de répondre O (oui) pour obtenir les résultats sous la forme de la figure 9. Sinon (réponse N pour "non"), l'exécution du programme s'arrête.

2.4 - Spécificités

Les calculs correspondant aux sommes des carrés des écarts ont été effectués en double précision, afin de conserver un plus grand nombre de chiffres significatifs. Les risques d'erreurs sont donc diminués, en particulier celui d'obtenir des variances négatives!

TEST D'HOMOGENEITE DES VARIANCES RESIDUELLES

TEST DE BARTLETT

Test = 0.8614 avec 1 degrés de liberté.

Si la valeur du test est inférieure au CHI-2 tabulé, on accepte l'hypothèse d'homogénéité des variances résiduelles

TEST LOG-ANOVA

Tableau d'analyse de variance :

Source de variation	Somme des carrés	Degrés de liberté	Variance	Test F
Inter groupe	0.034356	1	0.034356	0.048
Intra groupe	29.450648	41	0.718308	
Totale	29.485003			

Si le test F est inférieur à la valeur du F tabulé à 1 et 41 degrés de liberté on accepte l'hypothèse d'homogénéité des variances.

Voulez vous continuer (O/N) :

Figure 4 : Résultats des tests d'homogénéité des variances résiduelles.

COMPARAISON DE 2 DROITES DE REGRESSION

REGRESSIONS POUR CHAQUE DROITE

Droite n° 1 Fichier : FLTAAR Rang de X : 2 Rang de Y : 1
 Nbr. points = 282
 Pente = 0.197 Ecart-type = 0.00190
 Origine = 0.103
 Coeff. corr = 0.9872

Droite n° 2 Fichier : FLTAPR Rang de X : 1 Rang de Y : 2
 Nbr. points = 697
 Pente = 0.198 Ecart-type = 0.00115
 Origine = 0.030
 Coeff. corr = 0.9884

Faites un 'retour-chariot' pour continuer ...

Figure 5 : Calcul des équations de régression pour chaque droite.

REGRESSION TOUTES DROITES CONFONDUES

Nbr. points = 979
 Pente = 0.198 Ecart-type = 0.00133
 Origine = 0.050
 Coeff. corr = 0.9787

Faites un 'retour-chariot' pour continuer ...

Figure 6 : Calcul de l'équation de la droite de régression "commune".

TABLEAU D'ANALYSE DE COVARIANCE

Source	SCE	DDL	Variance
Droite 1	0.3530740	280	0.0012610
Droite 2	0.7974590	695	0.0011474
2 droites différentes	1.1505330	975	0.0011800
2 droites parallèles	1.1512091	976	0.0011795
Différences de pentes	0.0006761	1	0.0006761
Une droite commune	2.0914223	977	0.0021407
Différence de position	0.9402132	1	0.9402132

Faites un 'retour chariot' pour continuer ...

Figure 7 : Tableau d'analyse de covariance.

Test de Bartlett d'homogénéité des variances : 0.861
avec 1 degrés de liberté.

Test Log-anova d'homogénéité des variances : 0.048.
avec 1 et 41 degrés de liberté.

Test F d'égalité des pentes : 0.573
avec 1 et 976 degrés de liberté.

Test F d'égalité des origines : 797.117
avec 1 et 976 degrés de liberté.

Si les F calculés sont supérieurs aux F tabulés,
on refuse les hypothèses d'égalité.

Voulez vous les résultats sur imprimante (O/N) : ?

Figure 8 : Calculs des tests : homogénéité des variances résiduelles (rappel) et tests de parallélisme des droites et d'égalité des ordonnées à l'origine.

COMPARAISON DE 2 DROITES DE REGRESSION

REGRESSIONS POUR CHAQUE DROITE

Droite n° 1 Fichier : FLTAAR Rang de X : 2 Rang de Y : 1
 Nbr. points = 282
 Pente = 0.197 Ecart-type = 0.00190
 Origine = 0.103
 Coeff. corr = 0.9872

Droite n° 2 Fichier : FLTAPR Rang de X : 1 Rang de Y : 2
 Nbr. points = 697
 Pente = 0.198 Ecart-type = 0.00115
 Origine = 0.030
 Coeff. corr = 0.9884

REGRESSION TOUTES DROITES CONFONDUES

Nbr. points = 979
 Pente = 0.198 Ecart-type = 0.00133
 Origine = 0.050
 Coeff. corr = 0.9787

TABLEAU D'ANALYSE DE COVARIANCE

Source	SCE	DDL	Variance
Droite 1	0.3530740	280	0.0012610
Droite 2	0.7974590	695	0.0011474
2 droites différentes	1.1505330	975	0.0011800
2 droites parallèles	1.1512091	976	0.0011795
Différences de pentes	0.0006761	1	0.0006761
Une droite commune	2.0914223	977	0.0021407
Différence de position	0.9402132	1	0.9402132

Test de Bartlett d'homogénéité des variances : 0.861
 avec 1 degrés de liberté.

Test Log-anova d'homogénéité des variances : 0.048
 avec 1 et 41 degrés de liberté.

Test F d'égalité des pentes : 0.573
 avec 1 et 976 degrés de liberté.

Test F d'égalité des origines : 797.117
 avec 1 et 976 degrés de liberté.

Si les F calculés sont supérieurs aux F tabulés,
 on refuse les hypothèses d'égalité.

Figure 9 : Sortie des résultats sur imprimante.

Le programme est limité à la comparaison simultanée de 30 droites de régression, chaque droite ne pouvant pas être construite par plus de 2000 couples.

Par ailleurs, les fichiers des données sont limités à 100 variables (et 2000 individus).

CONCLUSION

Le programme informatique présenté ici calcule simplement les tests nécessaires à la comparaison statistique de plusieurs droites de régression. Aussi, il ne fournit pas d'analyse détaillée des droites de régression, bien que celles-ci soient calculées. Pour cela, l'utilisateur est renvoyé à des programmes spécifiques.

On peut contacter l'auteur de ce rapport pour tout renseignement complémentaire ou la fourniture sur disquette du programme.

BIBLIOGRAPHIE

- BARTLETT M.S. - 1937 - Some example of statistical methods of research in agriculture and applied biology. J. Royal Soc. Suppl. 8 : 128 - 138.
- LARDEUX F., TETUANUI A. - 1988 - Etude biométrique des capsules céphaliques des larves d'Aedes aegypti et d'Aedes polynesiensis : recherche d'un critère de différenciation sexuelle précoce. Rapport ITRMLM n° 34 / 88 / ITRM / Doc-Ent : 38 p.
- MARTIN C.G., GAMES P.A. - 1977 - Anova tests for homogeneity of variances : non normality and unequal samples. J. Educ. Stat. 2 : 187 - 206.
- SCHERRER B. - 1984 - Biostatistiques. Gaëtan Morin Ed. 850 p.
- ZAR J. H. - 1974 - Biostatistical analysis. Prentice-Hall, Englewood Cliffs, N.J.

LISTING DU PROGRAMME

```

PROGRAMME COMPREG
comparaison de k droites de regression
F. Lardeux Novembre 1988
d'après SCHERRER - 1984 - Biostatistique - Gaëtan morin ed.

```

```

DIM FICH$(30),TOTAVAR(30), RX(30), RY(30)
DIM T(100) , SOMX(30), SOMY(30), SOMX2(30), SOMY2(30), SOMXY(30),N(30)
DIM S2X(30), S2Y(30), SXY(30), A(30), B(30), C(30), SCE(30)
DIM NSG(30), NEL(30), RESIDU(2000), Z(30), MOYZ(30), SOMZ(30), SOMDL(30)
DIM SOMZ2(30), PENTE(30), ORIGINE(30)

```

```

DEFBL A - C, S

```

```

CLS
PRINT
PRINT
PRINT TAB(20) "COMPARAISON DE K DROITES DE REGRESSION"
PRINT : PRINT : PRINT
PRINT
PRINT TAB(10) "Source : SCHERRER - 1984 - Biostatistiques. Gaëtan morin Ed."
PRINT TAB(10) " pp. 386 et seq. et 676 et seq."
PRINT : PRINT
PRINT TAB(10) "Programme F. LARDEUX - Centre ORSTOM de Tahiti. B.P. 529"
PRINT: PRINT : PRINT

```

```

INPUT "Nombre de droites à comparer : ", ND

```

```

CLS
PRINT : PRINT
FOR I = 1 TO ND
PRINT "Droite n° ";I
INPUT " Nom du fichier (avec son chemin d'accès) : ", FICH$(I)
INPUT " Nombre total de variables dans ce fichier : ", TOTVAR(I)
INPUT " Rang de la variable Y : ",RY(I)
INPUT " Rang de la variable X : ",RX(I)
PRINT
PRINT
NEXT I

```

```

'Overture d'un fichier temporaire pour stocher les X et Y de tous
' les fichiers
OPEN "O",#2, "$$$"

```

```

'Calculs pour chaque fichier séparé

```

```

FOR I = 1 TO ND
'mise à zéro des sommes
SOMX(I) = 0
SOMY(I) = 0
SOMX2(I) = 0
SOMY2 (I) = 0
SOMXY(I) = 0
N(I) = 0

'lecture des variables du fichier I
OPEN "I",#1,FICH$(I)

WHILE NOT EOF(1)

FOR J = 1 TO TOTVAR(I)
INPUT #1,T(J)
NEXT J

```

```

PRINT #2, T(RX(I)), T(RY(I))
SOMX(I) = SOMX(I) + T(RX(I))
SOMY(I) = SOMY(I) + T(RY(I))
SOMX2(I) = SOMX2(I) + T(RX(I))^2
SOMY2(I) = SOMY2(I) + T(RY(I))^2
SOMXY(I) = SOMXY(I) + T(RX(I))*T(RY(I))
N(I) = N(I) + 1

```

```
WEND
```

```
CLOSE #1
```

```

'calcul des variances et covariances de X et Y pour le fichier I
S2X(I) = (SOMX2(I) - SOMX(I)*SOMX(I)/N(I)) / (N(I) - 1)
S2Y(I) = (SOMY2(I) - SOMY(I)*SOMY(I)/N(I)) / (N(I) - 1)
SXY(I) = (N(I)*SOMXY(I) - SOMX(I)*SOMY(I)) / (N(I)*(N(I)-1))
'calculs des coeff de la regression pour la droite I
PENDE (I) = SXY(I) / S2X(I)
ORIGINE (I) = (SOMY(I) / N(I)) - (SOMX(I) / N(I)) * (SXY(I) / S2X(I))

```

```

'calculs rapides des SCE pour le fichier I
A(I) = (N(I) - 1) * S2X(I)
B(I) = (N(I) - 1) * SXY(I)
C(I) = (N(I) - 1) * S2Y(I)

```

```

SCE(I) = C(I) - (B(I)*B(I) / A(I))
DDL(I) = N(I) - 2
S2E(I) = SCE(I) / DDL(I)

```

```

'Calculs des résidus pour la droite I
OPEN "I", #1, FICH(I)
K = 1

```

```

SOR = 0
SOR2 = 0
WHILE NOT EOF(1)
FOR J = 1 TO TOTVAR(I)
INPUT #1, T(J)
NEXT J
RESIDU(K) = T(RY(I)) - (ORIGINE(I) + PENDE(I) * T(RX(I)))
SOR = SOR + RESIDU(K)
SOR2 = SOR2 + RESIDU(K) * RESIDU(K)
K = K + 1
WEND

```

```

' calculs préliminaires pour test de Bartlett
BVAR(I) = (SOR2 - SOR*SOR/N(I)) / (N(I) - 1)
BS = BS + BVAR(I) * (N(I) - 1)
BN = BN + (N(I) - 1)
BL = BL + (N(I) - 1) * LOG(BVAR(I))
B1N = B1N + 1 / (N(I) - 1)

```

```
CLOSE #1
```

```

'calculs du test log anova d'égalité des variances résiduelles entre les
'fichiers I

```

```

'nombre de sous groupes pour la droite I
NSG(I) = CINT (SQR(N(I)))
MU2 = MU2 + NSG(I) - 1 : 'Calcul du nbre de ddl pour sce du test log-anova

```

```

'nombre d'éléments par sous groupe
SNEL = 0
FOR J = 1 TO NSG(I) - 1
NEL (J) = CINT(N(I) / NSG(I))
SNEL = SNEL + NEL(J)
NEXT J
NEL(NSG(I)) = N(I) - SNEL

```

```

'Affectation aléatoire des éléments résidu(i) dans chaque sous groupe
'Calcul des variances de chaque sous groupe et de la stat Z
FOR J = 1 TO NSG(I)
  SR = 0
  SR2 = 0
  FOR K = 1 TO NEL(J)
    WHILE RESIDU(X) = 1 E+30 OR X = 0
      RANDOMIZE TIMER
      X = INT(RND * N(I)) + 1
    WEND
    SR = SR + RESIDU(X)
    SR2 = SR2 + RESIDU(X) * RESIDU(X)
    RESIDU(X) = 1 E+30
  NEXT K

  Z(J) = LOG( (SR2 - (SR*SR / NEL(J))) / (NEL(J) - 1) ) * 0.4342945
  SOMZ(I) = SOMZ(I) + Z(J) * (NEL(J) - 1)
  SOMDL(I) = SOMDL(I) + (NEL(J) - 1)
  SOMZ2(I) = SOMZ2(I) + Z(J) * Z(J) * (NEL(J) - 1)
NEXT J

```

```

'calcul pour chaque droite I : moyenne pondérée, SCT , SCI
MOYZ(I) = SOMZ(I) / SOMDL(I)
SOM1 = SOM1 + MOYZ(I) * (N(I) - NSG(I))
SOM2 = SOM2 + ((N(I) - NSG(I)) * MOYZ(I))
SOM3 = SOM3 + (N(I) - NSG(I))
SOM4 = SOM4 + SOMZ(I)
SOM5 = SOM5 + SOMZ2(I)

```

NEXT I

```

' Fin du calcul du test log anova
SCI = SOM1 - SOM2 * SOM2 / SOM3
SCT = SOM5 - SOM4 * SOM4 / SOM3
SCE = SCT - SCI

```

LOGANOVA = (SCI / (ND-1)) / (SCE / MU2)

```

'fin du calcul du test de Bartlett
S2XP = BS / BN
B = ((LOG(S2XP)) * BN - BL)
C = 1 + (1/(3*(ND - 1))) * (B1N - 1/BN)
BC = B / C

```

' sortie des résultats du test de Bartlett et test log anova

```

CLS
PRINT TAB(18) "TEST D'HOMOGENEITE DES VARIANCES RESIDUELLES"
PRINT "TEST DE BARTLETT"
PRINT
PRINT "      Test = ";
PRINT USING "#####.####";BC;
PRINT " avec ";ND - 1;" degrés de liberté."
PRINT
PRINT "      Si la valeur du test est inférieure au CHI-2 tabulé,";
PRINT "      on accepte l'hypothèse d'homogénéité des variances résiduelles"
PRINT " _____"
PRINT "TEST LOG-ANOVA"
PRINT
PRINT "Tableau d'analyse de variance : "
PRINT
PRINT " Source de          Somme          Degrés de          Variance          Test"
PRINT " variation          des carrés        liberté           F"
PRINT

```

```

PRINT " Inter groupe";
PRINT USING "#####.#####";SCI;
PRINT USING "#####";ND - 1;
PRINT USING "#####.#####";SCI / (ND-1);
PRINT USING "#####.###";LOGANOVA
PRINT " Intra groupe";
PRINT USING "#####.#####";SCE;
PRINT USING "#####";MU2;
PRINT USING "#####.#####";SCE/MU2
PRINT " Totale ";
PRINT USING "#####.#####";SCT
PRINT
PRINT "Si le test F est inférieur à la valeur du F tabulé à ";ND-1;" et ";MU2;" degrés de Liberté,"
PRINT " on accepte l'hypothèse d'homogénéité des variances."
PRINT
WHILE Y$ <> "O" AND Y$ <> "o" AND Y$ <> "n" AND Y$ <> "N"
  INPUT " Voulez vous continuer (O/N) : ",Y$
WEND
IF Y$ ="N" OR Y$ = "n" THEN END
CLOSE #2

```

'Calculs pour le fichier SOMME = tous les X et les Y mélangés
'C'est à dire une seule droite commune

```

OPEN "I",#1, "$$$"
  SOX = 0
  SOY = 0
  SOXY = 0
  SOX2 = 0
  SOY2 = 0
  N = 0

  WHILE NOT EOF(1)
    INPUT #1,X,Y
    N = N + 1
    SOX = SOX + X
    SOY = SOY + Y
    SOXY = SOXY + X * Y
    SOX2 = SOX2 + X * X
    SOY2 = SOY2 + Y * Y
  WEND
CLOSE #1

  SX2 = (SOX2 - SOX * SOX / N) / (N-1)
  SY2 = (SOY2 - SOY * SOY / N) / (N-1)
  SYX = (N * SOXY - SOX * SOY) / (N * (N-1))

  A = (N-1) * SX2
  B = (N-1) * SYX
  C = (N-1) * SY2

  SCEEC = C - B*B / A
  DDLEC = N - 2
  S2EC = SCEEC / DDLEC

```

! Calculs pour ND droites parallèles

```

FOR I = 1 TO ND
  AT = AT + A(I)
  BT = BT + B(I)
  CT = CT + C(I)
  SCEET = SCEET + SCE(I)
  DDLET = DDLET + DDL(I)
NEXT I

```

```

SCEEDP = CT - ((BT * BT) / AT)
DOLEDP = N - ND - 1
S2EOP = SCEEDP / DOLEDP

```

```

'ND droites différentes
S2ET = SCEET / DOLET

```

```

'Différence de pentes
SCEEVP = SCEEDP - SCEET
DDLEVP = ND - 1
S2EVP = SCEEVP / DDLEVP

```

```

'différence d'ordonnée à l'origine
SCEEV00 = SCEEC - SCEEDP
DDLEV00 = ND - 1
S2EV00 = SCEEV00 / DDLEV00

```

```

KILL "$$$"

```

```

'SORTIE DES RESULTATS

```

```

CLS
SORTIE$ = "SCRN:"
IMPR = 1
WHILE IMPR = 1
  OPEN SORTIE$ FOR OUTPUT AS #1
  PRINT #1,
  PRINT #1, "COMPARAISON DE ";ND;" DROITES DE REGRESSION"
  PRINT #1,
  PRINT #1, "REGRESSIONS POUR CHAQUE DROITE":PRINT #1,
  FOR I = 1 TO ND
    PRINT #1, "Droite n° ";I; : PRINT #1, " Fichier : ";FICH$(I);
    PRINT #1, " Rang de X : ";RX(I);:PRINT #1, " Rang de Y : ";RY(I)
    PRINT #1, " Nbr. points = ";
    PRINT #1, USING "#####";N(I)
    PRINT #1, " Pente = ";
    PRINT #1, USING "#####.###";PENTE (I);
    PRINT #1, " Ecart-type = ";
    PRINT #1, USING "#####.####"; SQR((SCE(I)/(N(I)-2))/((N(I)-1) * S2X(I)))
    PRINT #1, " Origine = ";
    PRINT #1, USING "#####.###";ORIGINE(I)
    PRINT #1, " Coeff. corr = ";
    PRINT #1, USING "###.###";SXY(I) / (SQR(S2X(I)) * SQR(S2Y(I)))
    PRINT #1,
    IF (SORTIE$ = "SCRN:" AND I>1 AND INT(I/2) - (I/2) = 0) THEN PRINT : PRINT :INPUT " Faites un 'retour-chariot' pour con
    continuer ... ",Y$: CLS : FOR H=1 TO 6: PRINT:NEXT H
  NEXT I

```

```

PRINT #1, "REGRESSION TOUTES DROITES CONFONDUES" : PRINT
PRINT #1, " Nbr. points = ";
PRINT #1, USING "#####";N
PRINT #1, " Pente = ";
PRINT #1, USING "#####.###";SYX / SX2;
PRINT #1, " Ecart-type = ";
PRINT #1, USING "#####.####";SQR((SCEEC/(N-2))/((N-1) * SX2))
PRINT #1, " Origine = ";
PRINT #1, USING "#####.###";(SOY/N) - ((SOX/N) * (SYX/SX2))
PRINT #1, " Coeff. corr = ";
PRINT #1, USING "###.###"; SYX / (SQR(SX2) * SQR(SY2))
PRINT #1,
IF SORTIE$ = "SCRN:" THEN PRINT #1, :INPUT " Faites un 'retour-chariot pour continuer ... ", Y$: CLS
PRINT #1, "TABLEAU D'ANALYSE DE COVARIANCE"
PRINT #1,
PRINT #1, "Source SCE DDL Variance
PRINT #1,

```

```

FOR I = 1 TO ND
  PRINT #1,"Droite ";I;" ";
  PRINT #1, USING "#####.#####";SCE(I);
  PRINT #1, USING "#####";DOL(I);
  PRINT #1, USING "#####.#####";S2E(I)
NEXT I

```

```

PRINT #1,
PRINT #1,ND;" droites"
PRINT #1,"différentes ";
PRINT #1, USING "#####.#####";SCEET;
PRINT #1, USING "#####";DLET;
PRINT #1, USING "#####.#####";S2ET
PRINT #1,

```

```

PRINT #1,ND;" droites"
PRINT #1,"parallèles ";
PRINT #1, USING "#####.#####";SCEEDP;
PRINT #1, USING "#####";DDLEDP;
PRINT #1, USING "#####.#####";S2EDP
PRINT #1,

```

```

PRINT #1,"Différences
print #1,"de pentes ";
PRINT #1, USING "#####.#####"; SCEEVP;
PRINT #1, USING "#####";DDLEVP;
PRINT #1, USING "#####.#####";S2EVP
PRINT #1,

```

```

PRINT #1, "Une droite
PRINT #1,"commune ";
PRINT #1, USING "#####.#####"; SCEEC;
PRINT #1, USING "#####";DDLEC;
PRINT #1, USING "#####.#####";S2EC
PRINT #1,

```

```

PRINT #1,"Différence"
PRINT #1,"de position ";
PRINT #1, USING "#####.#####"; SCEEVOO;
PRINT #1, USING "#####";DDLEVOO;
PRINT #1, USING "#####.#####";S2EVOO
PRINT #1,

```

```

IF SORTIES$ = "SCRN:" THEN INPUT "      Faites un 'retour chariot' pour continuer ...",Y$: CLS
PRINT #1, : PRINT : PRINT :PRINT
PRINT #1,"Test de Bartlett d'homogénéité des variances : ";
PRINT #1, USING "#####.###";BC
PRINT #1,"      avec ";ND - 1;" degrés de liberté."
PRINT #1,
PRINT #1,"Test Log-anova d'homogénéité des variances : ";
PRINT #1, USING "#####.###";LOGANOVA
PRINT #1,"      avec ";ND-1; " et ";MU2;" degrés de liberté."
PRINT #1,
PRINT #1,"_____ "
PRINT #1,
PRINT #1,"Test F d'égalité des pentes : ";
PRINT #1, USING "#####.###";S2EVP/S2EDP
PRINT #1,"      avec ";DDLEVP; " et ";DDLEDP;" degrés de liberté."
PRINT #1,
PRINT #1,"Test F d'égalité des origines : ";
PRINT #1, USING "#####.###";S2EVOO/S2EODP
PRINT #1,"      avec ";DDLEVOO; " et ";DDLEDP;" degrés de liberté."
PRINT #1,
PRINT #1,"Si les F calculés sont supérieurs aux F tabulés,";
PRINT #1," on refuse les hypothèses d'égalité."
PRINT #1, : PRINT #1, : PRINT #1,
  WHILE R$ <>"0" AND R$ <>"o" AND R$ <>"N" AND R$ <>"n"
    INPUT "      Voulez vous les résultats sur imprimante (O/N) : ";R$
  WEND
IF R$ = "0" OR R$ = "o" THEN R$ = "A" :CLOSE #1 : SORTIES$ = "LPT1:" ELSE IMPR = 0
WEND

```

END