NOTES DE STATISTIQUE ET D'INFORMATIQUE

94/4

LA REGRESSION LINEAIRE PONDEREE: PRINCIPES ET APPLICATION

R. PALM

Faculté des Sciences agronomiques

Centre de Recherches agronomiques

GEMBLOUX

(Belgique)

LA RÉGRESSION LINÉAIRE PONDÉRÉE : PRINCIPES ET APPLICATIONS

R. PALM[1]

RÉSUMÉ

Cette note présente brièvement les principes de la régression linéaire pondérée et donne deux exemples numériques réels d'application, traités de façon détaillée.

SUMMARY

This note briefly describes the principles of weighted linear regression and gives two completely worked out examples from the real world.

1. INTRODUCTION

Une des conditions d'application des méthodes classiques de l'inférence statistique en relation avec les problèmes de régression linéaire est l'homoscédasticité et lorsque cette condition n'est pas vérifiée, l'utilisation de la régression pondérée doit être envisagée.

L'objectif de cette note est de présenter brièvement les principes de la régression pondérée et d'illustrer la méthode par deux exemples concrets.

Nous consacrons tout d'abord un paragraphe à la description de la régression pondérée proprement dite (paragraphe 2). Ensuite, nous examinons comment mettre en évidence la non constance de la variance résiduelle (paragraphe 3) et comment déterminer une relation exprimant l'évolution de cette variance en fonction d'une ou de plusieurs caractéristiques (paragraphe 4). Les principes sont alors appliqués à deux exemples concrets (paragraphes 5 et 6). Enfin, nous terminons par quelques informations complémentaires (paragraphe 7).

[1] Chef de travaux et Maître de conférences à la Faculté des Sciences agronomiques de Gembloux.

2. MOINDRES CARRÉS PONDÉRÉS

Supposons que, pour tout individu d'une population donnée, le modèle suivant :

$$y = \boldsymbol{x}\,\beta + \epsilon\,,$$

soit applicable. Dans cette relation, y est la variable dépendante, $\boldsymbol{x}$ est le vecteur, de dimensions $1 \times p$, relatif aux variables explicatives, β est le vecteur des p coefficients de régression théoriques et ϵ est le résidu. D'autre part, soit $\boldsymbol{y}$ le vecteur, de dimensions $n \times 1$, des valeurs observées de la variable à expliquer et $\boldsymbol{X}$ la matrice, de dimensions $n \times p$, des valeurs observées des variables explicatives, les observations étant réalisées sur un échantillon aléatoire et simple de n individus, prélevé dans la population. On a alors :

$$\boldsymbol{y} = \boldsymbol{X}\,\hat{\beta} + \boldsymbol{e}\,,$$

$\hat{\beta}$ étant le vecteur des coefficients de régression estimés, et $\boldsymbol{e}$ le vecteur, de dimensions $n \times 1$, des résidus observés.

Pour tenir compte d'un éventuel terme indépendant, il suffit de considérer qu'une variable explicative est constante et égale à l'unité : le vecteur $\boldsymbol{x}$ du modèle théorique comporte alors un élément de plus que de variables explicatives et la matrice $\boldsymbol{X}$ une colonne de plus que de variables explicatives.

Sous les conditions d'application classiques des méthodes relatives à la régression [DAGNELIE, 1982, DRAPER et SMITH, 1966, PALM, 1988, WEISBERG, 1985, entre autres], l'estimateur au sens des moindres carrés :

$$\hat{\beta} = (\boldsymbol{X}'\,\boldsymbol{X})^{-1}\,\boldsymbol{X}'\,\boldsymbol{y}\,,$$

est l'estimateur linéaire non biaisé de variance minimum et la matrice des variances et covariances estimées de $\hat{\beta}$ est égale à :

$$\hat{\boldsymbol{V}}\,(\hat{\beta}) = \hat{\sigma}^2(\boldsymbol{X}\,\boldsymbol{X})^{-1}\,,$$

avec :

$$\hat{\sigma}^2 = \boldsymbol{e}'\,\boldsymbol{e}/(n-p)\,.$$

Une des conditions d'application des méthodes classiques de l'inférence statistique en relation avec les problèmes de régression est l'homoscédasticité, c'est-à-dire la constance de la variance résiduelle, qui implique, si le modèle est correct, la constance des variances conditionnelles :

$$\sigma^2_{y|\boldsymbol{x}} = \sigma^2\,.$$

Si cette condition n'est pas remplie, l'estimateur au sens des moindres carrés ordinaires n'est plus de variance minimum et les différentes procédures d'inférence statistique doivent être modifiées en conséquence.

Soit $\sigma^2_{y|\boldsymbol{x}_i}$ $(i = 1, \ldots, n)$, les variances conditionnelles relatives aux différentes observations de l'échantillon et soit $\boldsymbol{\Sigma}$ la matrice diagonale, de dimensions

$n \times n$, dont les éléments diagonaux sont les $\sigma^2_{y|\boldsymbol{x}_i}$. On peut montrer que la meilleure estimation du vecteur $\boldsymbol{\beta}$ est donnée par :

$$\hat{\boldsymbol{\beta}}_w = (\boldsymbol{X}' \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{\Sigma}^{-1} \boldsymbol{y} .$$

Le vecteur $\hat{\boldsymbol{\beta}}_w$ ainsi obtenu est le vecteur qui minimise la somme des carrés des écarts entre les valeurs observées et les valeurs estimées de la variable à expliquer, ces écarts étant pondérés par l'inverse de leur variance :

$$\mathrm{SCE}_r = (\boldsymbol{y} - \boldsymbol{X}\,\hat{\boldsymbol{\beta}}_w)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\,\hat{\boldsymbol{\beta}}_w) = \sum_{i=1}^{n} \left[(y_i - \hat{y}_i)^2 / \sigma^2_{y|\boldsymbol{x}_i} \right] .$$

L'utilisation de facteurs de pondérations :

$$w_i = 1/\sigma^2_{y|\boldsymbol{x}_i} ,$$

justifie l'appellation d'estimation par les moindres carrés pondérés[2] donnée à la technique.

La somme des carrés des écarts qui est minimisée peut encore s'écrire :

$$\mathrm{SCE}_r = \sum_{i=1}^{n} (z_i - \boldsymbol{u}_i\,\hat{\boldsymbol{\beta}}_w)^2 ,$$

avec : $\quad z_i = y_i/\sigma_{y|\boldsymbol{x}_i} \quad$ et $\quad \boldsymbol{u}_i = \boldsymbol{x}_i/\sigma_{y|\boldsymbol{x}_i} \quad (i = 1, \ldots, n)$,

$\boldsymbol{x}_i$ étant le vecteur des valeurs observées des p variables explicatives pour le $i^{\text{ème}}$ individu.

On constate donc que le vecteur $\hat{\boldsymbol{\beta}}_w$ peut être obtenu par la méthode des moindres carrés non pondérés, à condition de diviser, pour chaque individu, les valeurs de la variable à expliquer et des variables explicatives par l'écart-type conditionnel relatif à l'individu en question. Pour le modèle ainsi transformé, les méthodes classiques d'inférence pourront être appliquées, puisque, dans ce cas, la condition d'homoscédasticité est remplie.

En particulier, la variance d'une valeur estimée pourra être obtenue à partir de la variance de la variable transformée. Ainsi la variance de la valeur moyenne $\hat{y}_{\boldsymbol{x}_0}$ correspondant au vecteur $\boldsymbol{x}_0$ est égale à la variance $v(\hat{z}_{\boldsymbol{u}_0})$, multipliée par la variance conditionnelle $\sigma^2_{y|\boldsymbol{x}_0}$. La notation $v\,(\hat{z}_{\boldsymbol{u}_0})$ représente la variance de la valeur moyenne estimée $\hat{z}_{\boldsymbol{u}_0}$, $\boldsymbol{u}_0$ étant le vecteur des variables explicatives, $\boldsymbol{x}_0$, divisé par l'écart-type conditionnel, $\sigma_{y|\boldsymbol{x}_0}$. Pour obtenir la variance d'une valeur individuelle, il suffit de rajouter, dans cette expression, le terme $\sigma^2_{y|\boldsymbol{x}_0}$.

En pratique cependant, les variances conditionnelles théoriques, $\sigma^2_{y|\boldsymbol{x}_i}$, ne sont généralement pas connues mais sont remplacées par des estimations. Le paragraphe 4 sera consacré à ce sujet.

Enfin, notons encore que la formule donnée ci-dessus pour $\hat{\boldsymbol{\beta}}_w$ reste valable également lorsque la matrice des variances et covariances des résidus, $\boldsymbol{\Sigma}$,

[2] En anglais : *weighted least squares*.

n'est plus, à une constant près, une matrice diagonale, mais bien une matrice symétrique définie positive. Il s'agit alors de la méthode des moindres carrés généralisés[3] [DRAPER et SMITH, 1981 ; THEIL, 1971 ; WEISBERG, 1985].

3. MISE EN ÉVIDENCE DE L'INÉGALITÉ DES VARIANCES CONDITIONNELLES

Dans certaines situations, l'inégalité des variances conditionnelles est tellement évidente que tout test statistique est superflu. Dans d'autres cas, par contre, la situation est moins nette et il peut être utile de tester l'hypothèse d'homoscédasticité.

Un tel test peut être utile également afin de s'assurer que les poids utilisés lors du calcul d'une équation de régression pondérée sont adéquats. Les tests viseront alors à vérifier l'égalité des variances conditionnelles pour la variable transformée, définie au paragraphe précédent :

$$z_i = y_i / \hat{\sigma}_{y|x_i} .$$

On ne perdra pas de vue non plus que la mise en évidence de l'hétérogénéité des variances n'est qu'une première étape dans la résolution du problème. En effet, s'il y a hétérogénéité, il faut encore définir un modèle permettant de décrire l'évolution des variances conditionnelles (paragraphe 4). Au-delà de la simple décision d'accepter ou de rejeter l'hypothèse d'homoscédasticité, les statistiques définies dans le cadre des tests permettent également, du moins dans certains cas, d'orienter le choix d'une fonction décrivant l'évolution des variances. Dans cette dernière optique, nous présentons aussi des procédures qui sont purement descriptives et qui ne correspondent pas à des tests proprement dits.

Dans le cas particulier où plusieurs observations sont caractérisées par un même vecteur x, les tests classiques de BARTLETT ou de HARTLEY [DAGNELIE, 1975] peuvent être réalisés. Ces tests présentent cependant l'inconvénient de ne pas tenir compte de l'ordre dans lequel les sous-échantillons relatifs aux différents vecteurs x_i sont logiquement rangés.

En pratique, les vecteurs x_i sont le plus souvent différents pour chacun des individus et les tests de BARTLETT ou de HARTLEY sont inapplicables ; l'utilisateur doit alors se tourner vers d'autres solutions. Nous envisageons d'abord le cas où on ne dispose que d'une seule variable explicative et nous verrons ensuite la généralisation au cas de plusieurs variables explicatives.

L'inégalité des variances conditionnelles peut être visualisée par le diagramme de régression de y en fonction de x, du moins si cette inégalité est suffisamment accentuée. Une solution plus souvent retenue consiste cependant à établir des diagrammes dans lesquels on porte, en ordonnée, non pas les valeurs y_i, mais les résidus, ou une fonction des résidus, qui ont été obtenus à la

[3] En anglais : *generalized least squares.*

suite de l'ajustement de l'équation de régression non pondérée. Parmi les transformations des résidus proposées dans la littérature, on peut citer, notamment, la valeur absolue des résidus, le logarithme de la valeur absolue des résidus, le carré des résidus et la racine cubique du carré des résidus. Une discussion des avantages et des inconvénients de ces diverses solutions est donnée par CARROLL et RUPPERT [1988]. Ces derniers proposent la représentation du logarithme des résidus absolus en fonction du logarithme de x. Ce type de diagramme permet en effet de tirer des informations concernant la nature de la relation liant l'écart-type conditionnel à la valeur de x, car il s'apparente à un graphique donnant le logarithme de la variance conditionnelle en fonction du logarithme de la variable explicative, dont nous parlerons au paragraphe 4.

Il faut noter aussi que l'appréciation de l'hétérogénéité des variances conditionnelles par l'examen des résidus ne peut se justifier que si le modèle retenu est satisfaisant. Nous reviendrons sur ce point au paragraphe 7.

Des approches numériques permettent également d'apprécier l'inégalité des variances conditionnelles. Ainsi, si on soupçonne l'existence d'une relation croissante ou décroissante entre $\sigma_{y|x}$ et x, on peut classer les observations par ordre croissant de x et calculer la somme des carrés des résidus sur les m premières observations et sur les m dernières observations et effectuer le rapport des ces deux quantités :

$$\sum_{i=1}^{m} e_i^2 \Big/ \sum_{i=n-m+1}^{n} e_i^2 .$$

Un rapport trop différent de l'unité indique une inégalité des variances conditionnelles. La valeur de m dépend de l'effectif total : m sera égal, par exemple, à $n/2$ (ou à $n/2 - 0{,}5$, si n est impair) lorsque le nombre total d'observations est inférieur ou égal à 20 et m sera de l'ordre de $n/3$ pour des effectifs plus importants.

Le rapport ainsi déterminé est analogue à la valeur F_{obs}, calculée lors de la réalisation du test d'égalité de deux variances. Il n'est cependant pas distribué selon une variable F de FISHER-SNEDECOR, même si la variance résiduelle est constante, car le numérateur et le dénominateur ne sont pas indépendants, les différents résidus estimés, e_i, étant corrélés. On peut toutefois assurer l'indépendance des deux quantités en effectuant deux ajustements séparés, l'un pour les faibles valeurs de x et l'autre pour les valeurs élevées [GOLDFELD et QUANDT, 1965].

CARROLL et RUPPERT [1988] proposent de mesurer l'inégalité des variances en calculant le coefficient de corrélation de rang de SPEARMAN des valeurs x_i et des résidus standardisés. Le niveau de signification de ce coefficient peut être pris comme un indicateur de la non-constance des variances, bien qu'il ne s'agisse pas d'un test rigoureux. Cette approche repose également sur l'hypothèse d'une relation croissante ou décroissante entre la variance conditionnelle et la variable explicative.

Se basant sur la même hypothèse et postulant le modèle suivant :

$$\sigma_{y|x_i}^2 = \sigma^2[\exp(\lambda x_i)] ,$$

BREUSCH et PAGAN [1979], d'une part, et COOK et WEISBERG [1983], d'autre part, proposent une solution pour tester la nullité de λ. A partir des résidus de la régression non pondérée, e_i, on définit une nouvelle variable explicative r_i :

$$r_i = e_i^2/\tilde{\sigma}^2 \quad \text{avec} \quad \tilde{\sigma}^2 = \sum_{i=1}^{n} e_i^2/n\,.$$

On calcule alors la somme des carrés des écarts, SCE_{reg}, liée à la régression de r en fonction de x, ainsi que la quantité :

$$\chi^2_{obs} = \mathrm{SCE}_{reg}/2\,,$$

qui, comme l'ont montré les auteurs cités ci-dessus, suit une distribution χ^2 à un degré de liberté, lorsque l'hypothèse de nullité de λ est vérifiée.

Quand il y a inégalité des variances conditionnelles et que les paramètres de l'équation sont obtenus par les moindres carrés non pondérés, la formule classique donnant la matrice des variances et covariances des paramètres est inadéquate. Pour ce cas, WHITE [1980] propose un estimateur consistant de cette matrice. Sur la base de la comparaison de cet estimateur consistant et de l'estimateur usuel, il définit également un test relatif à la spécification du modèle. Le rejet de l'hypothèse nulle signifie que la forme de l'équation retenue n'est pas adéquate ou bien qu'il y a inégalité des variances conditionnelles. L'estimation consistante de la matrice des variances et covariances et les résultats du test sont donnés par la procédure PROC REG du logiciel SAS [X, 1990], en faisant appel aux options ACOV et SPEC.

Les approches graphiques et numériques décrites ci-dessus peuvent être généralisées au cas de plusieurs variables explicatives. Pour les représentations graphiques et le calcul du coefficient de corrélation de rang, on choisit *a priori* l'une ou l'autre variable explicative dont on pense qu'elle pourrait être pertinente pour expliquer la non-constance des variances conditionnelles. On peut également remplacer la variable explicative par les valeurs estimées de la variable à expliquer. Pour le test de BREUSCH et PAGAN [1979] ou de COOK et WEISBERG [1983], on peut également remplacer le calcul de la régression de r en fonction de x par la régression de r en fonction de $\hat{y}$, ou calculer une équation de régression multiple en fonction de toutes ou d'un sous-ensemble de variables explicatives. Dans ce cas, et sous l'hypothèse d'homoscédasticité, la quantité χ^2_{obs} suit une distribution χ^2 à k degrés de liberté, k étant le nombre de paramètres de l'équation, à l'exclusion du terme indépendant.

4. ESTIMATION DES VARIANCES CONDITIONNELLES

Nous avons vu, au paragraphe 2, que le calcul de la régression pondérée nécessite une connaissance préalable des variances conditionnelles $\sigma^2_{y|x_i}$. Par ailleurs, si on souhaite déterminer l'erreur-standard de l'estimation relative à un individu ne faisant pas partie de l'échantillon, il faut également connaître l'écart-type conditionnel relatif à cet individu.

Ces variances sont, en pratique, rarement connues *a priori* et il est le plus souvent nécessaire de déterminer une relation exprimant les variances conditionnelles en fonction d'une ou de plusieurs caractéristiques des individus. Ces caractéristiques sont fréquemment les variables explicatives du modèle de régression ou des transformations de ces variables.

Une solution préconisée pour estimer les variances conditionnelles consiste à répartir les résidus de la régression non pondérée en classes, en fonction d'une variable explicative particulière. On estime ensuite les variances des résidus dans chacune de ces classes, $\tilde{\sigma}^2_{y|x}$, et on détermine alors, par régression, la relation existant entre ces variances conditionnelles et les points moyens des classes, $\bar{x}$, et, dans ce but, on utilise fréquemment un modèle doublement logarithmique :

$$\log\tilde{\sigma}^2_{y|x} = a + b\log\bar{x}\,,$$

qui a notamment l'avantage de toujours conduire à des variances estimées positives.

La procédure peut encore être étendue au cas où la répartition des résidus en classes se fait en fonction de plus d'une variable explicative.

Une telle méthode d'estimation des variances conditionnelles suppose que les observations sont assez nombreuses pour être réparties en un nombre suffisant de classes dont l'effectif permet une estimation valable de la variabilité résiduelle. D'autre part, même dans le cas d'un effectif suffisant, la méthode soulève le problème du choix du nombre de classes à retenir. En effet, plus ce nombre est élevé, plus on aura de couples de points pour calculer la régression mais moins on aura de valeurs pour calculer la variance par classe. A la limite, la procédure revient à établir la relation entre le logarithme du carré des résidus et le logarithme des valeurs x_i, chaque résidu élevé au carré constituant une estimation de la variance résiduelle.

Une autre solution consiste à fixer *a priori* un petit nombre de modèles du type :

$$\sigma^2_{y|}\boldsymbol{x} = f(\boldsymbol{x})\,,$$

et à choisir, parmi ces modèles, celui qui s'avère le plus adéquat.

Parmi les modèles simples, on considérera, par exemple, que la variance conditionnelle est proportionnelle à une variable explicative donnée ou au carré ou encore à toute autre puissance de cette variable.

Pour choisir le modèle de variance le plus adéquat parmi les candidats, on calcule, pour chacun de ces modèles, la régression non pondérée :

$$\boldsymbol{z} = \boldsymbol{U}\,\boldsymbol{\beta}_w + \boldsymbol{d}\,,$$

le vecteur $\boldsymbol{z}$ et la matrice $\boldsymbol{U}$ étant obtenus, comme nous l'avons expliqué au paragraphe 2, en divisant les valeurs de la variable à expliquer et des variables explicatives initiales par l'écart-type résiduel donné par le modèle de variance testé, et $\boldsymbol{d}$ étant le vecteur des résidus observés pour la régression sur les variables ainsi transformées. On retient alors le modèle de variance pour lequel la variance

résiduelle $\sigma^2_{z|u}$ est la plus constante. On évite donc le problème de l'ajustement des modèles de variance sur base des résidus de la régression non pondérée. Cet avantage peut être appréciable, surtout si l'ajustement de la régression non pondérée conduit des résidus observés fort différents de ceux qu'on obtiendrait par la régression pondérée, comme nous le verrons au paragraphe 6.

On notera que, dans le cas où plusieurs observations sont caractérisées par un même vecteur x, on dispose immédiatement d'estimations des variances conditionnelles. L'utilisation directe de l'inverse de ces estimations comme facteurs de pondération est cependant le plus souvent à proscrire pour les deux raisons suivantes. Tout d'abord, ces estimations sont de mauvaise qualité, à moins que les nombres de répétitions pour les différents vecteurs de variables explicatives ne soient assez grands. Ensuite, ces estimations sont généralement insuffisantes puisque, ne connaissant pas la relation permettant de modéliser la variance, on ne dispose pas d'une estimation de la variance conditionnelle pour des observations relatives à des vecteurs de variables explicatives différents de ceux de l'échantillon. Or, nous avons vu, au paragraphe 2, que cette variance conditionnelle est indispensable pour le calcul de la variance de la valeur estimée par l'équation de régression.

5. EXEMPLE 1 : CAS D'UNE VARIABLE EXPLICATIVE

Le premier exemple traité concerne une situation particulièrement favorable, puisqu'il s'agit d'un problème de régression en fonction d'une seule variable explicative, dans le cas où on dispose de plusieurs observations pour les différentes valeurs de x.

Des échantillons contenant de l'acide ascorbique à six doses différentes (x, en μg/l) ont été soumis à un polarographe, à raison de cinq échantillons par dose, et l'intensité du courant (y, en nA) a été lue sur le polarographe. A partir des résultats obtenus et repris dans le tableau 1, on se propose de déterminer l'équation de régression exprimant l'intensité du courant en fonction de la concentration en acide ascorbique [RODRIGUEZ ZEVALLOS, 1992].

Le tableau 2 donne les moyennes, $\bar{x}$, et les écarts-types, $\bar{\sigma}_{y|x}$, des cinq répétitions en fonction des doses. Ce tableau donne aussi les écarts-types conditionnels de deux autres variables, z_1 et z_2, dont il sera question par la suite. On remarque immédiatement que les écarts-types augmentent très nettement avec les doses et tout test statistique d'homogénéité des variances est superflu.

Du fait de l'inégalité de ces variances, la droite de régression au sens des moindres carrés exprimant l'intensité du courant en fonction de la dose conduit à des résultats peu satisfaisants, comme le montre le tableau 3, qui donne l'estimation ponctuelle, l'erreur-standard et les limites de confiance de cette estimation. Ce tableau donne également ces mêmes caractéristiques pour trois autres modèles qui seront examinés ci-dessous.

L'examen du tableau 3 montre, par exemple, que l'intensité du courant déduite de l'équation de régression non pondérée, pour une dose de 1 μg/l,

Tableau 1. Doses d'acide ascorbique (x, en μg/l) et intensités du courant données par le polarographe (y, en nA) : valeurs pour 30 échantillons.

x	y	x	y
1	25,0	50	1.828
1	25,5	50	1.848
1	24,0	50	1.854
1	23,7	50	1.826
1	25,5	50	1.814
10	351,0	70	2.604
10	343,9	70	2.564
10	357,0	70	2.570
10	357,8	70	2.616
10	359,5	70	2.567
25	936,0	100	3.808
25	920,0	100	3.780
25	896,0	100	3.700
25	904,0	100	3.774
25	907,0	100	3.804

Tableau 2. Intensités moyennes, $\bar{y}$, et écarts-types conditionnels des variables y, z_1 et z_2, en fonction de la dose ($z_1 = y/x^{0,82}$ et $z_2 = y/x$).

x	$\bar{y}$	$\tilde{\sigma}_{y\|x}$	$\tilde{\sigma}_{z_1\|x}$	$\tilde{\sigma}_{z_2\|x}$
1	24,7	0,8	0,84	0,84
10	353,8	6,4	0,97	0,64
25	912,6	15,7	1,12	0,63
50	1.834,0	16,6	0,67	0,33
70	2.584,2	24,0	0,74	0,34
100	3.773,2	43,5	1,00	0,44

est égale à 8,8, soit une valeur environ trois fois plus faible que la moyenne observée, qui est de 24,7. Il montre également que les limites de confiance de cette estimation, considérée comme une estimation individuelle, valent respectivement - 54,2 et 71,8, ce qui ne correspond manifestement pas à la réalité.

La régression du logarithme des écarts-types en fonction du logarithme des doses conduit à l'équation suivante :

$$\log\tilde{\sigma}_{y|x} = -0,143 + 0,820\log x \quad (r^2 = 0,98)\,.$$

Les écarts-types conditionnels sont donc proportionnels à $x^{0,82}$:

$$\hat{\sigma}_{y|x} = k\,x^{0,82}\,.$$

Tableau 3. Estimations ponctuelles ($\hat{y}$), erreurs-standards des estimations (E.S.), limites de confiance, au niveau 0,95, pour une valeur moyenne (LIM et LSM) et limites de confiance, au niveau 0,95, pour une valeur individuelle (LII et LSI) ; résultats pour les modèles suivants :

(1) $y = a + bx$ et $\hat{\sigma}_{y|x} = k$
(2) $z_1 = a/x^{0,82} + bx^{0,18}$ et $\hat{\sigma}_{y|x} = kx^{0,82}$
(3) $z_2 = a/x + b$ et $\hat{\sigma}_{y|x} = kx$
(4) $z_2 = a/x + b + cx$ et $\hat{\sigma}_{y|x} = kx$

Modèles	x	$\hat{y}$	E.S.	LIM	LSM	LII	LSI
(1)	1	8,8	8,44	– 8,5	26,1	– 54,2	71,8
	10	348,3	7,42	333,1	363,5	285,8	410,7
	25	914,1	6,06	901,6	926,5	852,2	975,9
	50	1.857,0	5,52	1.845,7	1.868,3	1.795,4	1.918,7
	70	2.611,4	6,88	2.597,3	2.625,5	2.549,2	2.673,6
	100	3.743,0	10,44	3.721,6	3.764,4	3.678,7	3.807,2
(2)	1	24,6	0,51	23,6	25,6	22,0	27,2
	10	359,5	1,14	357,1	361,8	343,8	375,2
	25	917,6	2,89	911,7	923,5	884,1	951,0
	50	1.847,8	5,88	1.835,8	1.859,8	1.788,4	1.907,2
	70	2.592,0	8,28	2.575,0	2.608,9	2.513,5	2.670,4
	100	3.708,2	11,92	3.683,8	3.732,6	3.602,7	3.813,7
(3)	1	24,7	0,29	24,1	25,3	23,2	26,2
	10	358,7	1,23	356,2	361,2	345,1	372,3
	25	915,4	3,25	908,7	922,0	881,3	949,5
	50	1.843,2	6,65	1.829,5	1.856,8	1.774,9	1.911,4
	70	2.585,4	9,38	2.566,2	2.604,6	2.489,8	2.680,9
	100	3.698,7	13,40	3.671,2	3.726,2	3.562,1	3.835,2
(4)	1	24,7	0,25	24,2	25,3	23,5	26,0
	10	354,0	1,76	350,4	357,7	342,0	366,1
	25	907,3	3,73	899,6	914,9	877,5	937,0
	50	1.841,4	5,75	1.829,7	1.853,2	1.782,7	1.900,1
	70	2.599,7	9,10	2.581,0	2.618,5	2.517,1	2.682,4
	100	3.755,4	20,60	3.713,0	3.797,84	3.632,8	3.878,0

Par conséquent aussi, les écarts-types conditionnels de la variable :

$$z_1 = y/x^{0,82}$$

sont à peu près constants et, en tout cas, non liés à x, comme le montre la colonne intitulée $\tilde{\sigma}_{z_1|x}$ du tableau 2. Ces écarts-types s'obtiennent très facilement, en divisant directement les écarts-types de y par $x^{0,82}$.

La valeur de l'exposant étant relativement proche de l'unité, on pourrait, pour simplifier le calcul de la régression pondérée, considérer que les écarts-types conditionnels de y sont à peu près proportionnels à x et donc aussi que les

écarts-types conditionnels de :

$$z_2 = y/x\,,$$

désignés par $\tilde{\sigma}_{z_2|x}$, sont à peu près constants.

Le tableau 2 montre que les écarts-types de z_2 sont moins constants que ceux de z_1. Toutefois, pour z_2, le rapport entre le maximum et le minimum n'est que de 2,5. Le rapport des variances est donc de l'ordre de 6,5, soit une valeur nettement inférieure à 29,5, qui est la valeur critique du test de HARTLEY, au niveau 0,05, pour 6 populations et 4 degrés de liberté.

Si on admet que $\sigma_{y|x}$ est proportionnel à x, l'ajustement du modèle :

$$y = a + b\,x\,,$$

obtenu en pondérant les écarts résiduels par l'inverse des variances conditionnelles, est équivalent à l'ajustement non pondéré de :

$$z_2 = \frac{a}{x} + b\,.$$

Les résultats de cet ajustement réalisé avec le logiciel Minitab sont donnés dans la figure 1, en même temps que les résultats des prédictions pour les six doses considérées. Le passage des valeurs de z_2 aux valeurs de y se fait simplement en multipliant les nombres par x. Les résultats sont donnés dans la troisième partie du tableau 3.

A titre de comparaison, le tableau 3 donne aussi les résultats provenant de la régression suivante :

$$z_1 = a/x^{0,82} + b\,x^{0,18}\,.$$

On constate que la simplification apportée dans le modèle exprimant les écarts-types conditionnels en fonction de x est pratiquement sans incidence sur les valeurs estimées par la régression mais qu'elle se marque davantage sur les limites de confiance des estimations des valeurs individuelles. Par ailleurs, on remarque aussi que, quel que soit le type de pondération retenu, les limites de confiance conduisent à des résultats bien plus raisonnables que la régression non pondérée.

Dans la mesure où on dispose de plusieurs répétitions pour chacune des doses, on peut tester la linéarité de la relation, en décomposant la somme des carrés des écarts résiduelle en une somme des carrés des écarts liée à la non-linéarité et une somme des carrés des écarts résiduelle pure et en calculant la valeur F_{obs} résultant du rapport des carrés moyens correspondants [DAGNELIE, 1975]. Cette décomposition doit se faire sur le modèle transformé, pour lequel les conditions d'application sont remplies. La figure 1 montre que la probabilité associée à ce test est égale à 0,03 et on peut donc conclure que la composante non linéaire est significative. On peut dès lors envisager le recours à une équation du second degré, qui, après transformation s'écrit :

$$z_2 = \frac{a}{x} + b + c\,x\,.$$

Nous n'avons pas repris, sous forme de figure, les résultats fournis par Minitab pour ce modèle. On a cependant constaté que les trois coefficients de

```
The regression equation is
Y/X = 37.1 - 12.4 1/X

Predictor       Coef      Stdev    t-ratio       p
Constant     37.1111     0.1359     273.09   0.000
1/X         -12.4204     0.3308     -37.54   0.000

s = 0.6528      R-sq = 98.1%     R-sq(adj) = 98.0%

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1      600.52      600.52    1409.39   0.000
Error        28       11.93        0.43
Total        29      612.45

Unusual Observations
Obs.      1/X         Y/X         Fit  Stdev.Fit    Residual    St.Resid
  7      0.10      34.390      35.869      0.123      -1.479      -2.31R

R denotes an obs. with a large st. resid.

    Fit  Stdev.Fit          95% C.I.              95% P.I.
 24.691      0.291   (  24.094,  25.287)   (  23.226,  26.155)
 35.869      0.123   (  35.616,  36.122)   (  34.508,  37.230)
 36.614      0.130   (  36.348,  36.881)   (  35.251,  37.978)
 36.863      0.133   (  36.590,  37.135)   (  35.498,  38.228)
 36.934      0.134   (  36.660,  37.208)   (  35.568,  38.299)
 36.987      0.134   (  36.712,  37.262)   (  35.621,  38.352)

Pure error test - F = 3.25  P = 0.0289  DF(pure error) = 24
```

Figure 1. Régression non pondérée de $z_2 = y/x$ en fonction de $1/x$.

régression sont significatifs et que la variance résiduelle est égale à 0,31. Cette dernière valeur est à comparer à la valeur 0,43 de la figure 1.

La réduction de la variance résiduelle par l'introduction d'un terme supplémentaire dans le modèle n'est pas donc négligeable. Par ailleurs, le test F basé sur l'erreur résiduelle pure a montré que les composantes d'ordre supérieur à deux (composante cubique, composante du $4^{ième}$ degré, etc.) sont non significatives ($F = 0{,}77$).

En conclusion, nous retiendrons donc ce dernier modèle, qui peut encore s'écrire :

$$y = -11{,}72 + 36{,}45x + 0{,}01217x^2\,,$$

l'écart-type conditionnel étant égal à :

$$\hat{\sigma}_{y|x} = \sqrt{0{,}31}x = 0{,}557\,x\,.$$

Les valeurs estimées et les limites de confiance correspondantes sont reprises, pour ce modèle, dans la quatrième partie du tableau 3.

La régression pondérée peut aussi être réalisée directement par certains logiciels, sans qu'il soit nécessaire d'effectuer les transformations de variables. C'est le cas, notamment, pour les logiciels Minitab et SAS, qui proposent l'option "WEIGHTS", les poids étant définis au préalable et enregistrés dans une variable particulière. [X, 1989, X, 1990].

```
The regression equation is
Y = - 12.4 + 37.1 X

Predictor       Coef      Stdev    t-ratio       p
Constant     -12.4204     0.3308     -37.54   0.000
X             37.1111     0.1359     273.09   0.000

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1       31776       31776   74577.02   0.000
Error        28          12           0
Total        29       31788

Unusual Observations
Obs.        X           Y         Fit  Stdev.Fit   Residual    St.Resid
  7        10      343.90      358.69       1.23      -14.79      -2.31R

R denotes an obs. with a large st. resid.
* WARNING * Use caution in interpreting prediction intervals with
          * weighted regression.  See HELP REGRESS PREDICT.

    Fit  Stdev.Fit          95% C.I.              95% P.I.
  24.69       0.29  (   24.09,   25.29)  (   23.23,   26.15)
 358.69       1.23  (  356.16,  361.22)  (  355.83,  361.55) XX
 915.36       3.25  (  908.69,  922.02)  (  908.56,  922.15) XX
1843.13       6.64  ( 1829.52, 1856.74)  ( 1829.46, 1856.81) XX
2585.36       9.36  ( 2566.18, 2604.53)  ( 2566.14, 2604.58) XX
3698.69      13.43  ( 3671.16, 3726.21)  ( 3671.13, 3726.24) XX

X  denotes a row with X values away from the center
XX denotes a row with very extreme X values


Pure error test - F = 2.06   P = 0.1176   DF(pure error) = 24
```

Figure 2. Régression pondérée de y en x, avec poids **égaux à** $1/x^2$.

A titre d'illustration, la figure 2 reprend les résultats obtenus par Minitab, dans le cas de la régression linéaire de y en x et pour des poids, w_i, égaux à $1/x_i^2$. Des résultats tout à fait similaires auraient d'ailleurs été obtenus par SAS.

On constate que l'équation de régression est identique à l'équation obtenue à partir de la régression sur z_2 (figure 1), mais que le tableau d'analyse de la variance est différent. Cette discordance provient du mode de calcul des sommes des carrés des écarts. Ainsi, la somme des carrés des écarts totale, pour la variable transformée, est égale à :

$$\mathrm{SCE}_{z_2} = \sum_{i=1}^{n} \left(\frac{y_i}{x_i}\right)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} \frac{y_i}{x_i}\right)^2 ,$$

et le calcul direct par Minitab se fait de la façon suivante :

$$\begin{aligned}(\mathrm{SCE}_y)_w &= \sum_{i=1}^{n} w_i (y_i - \bar{y}_w)^2 = \sum_{i=1}^{n} w_i y_i^2 - \left(\sum_{i=1}^{n} w_i y_i\right)^2 \Big/ \sum_{i=1}^{n} w_i \\ &= \sum_{i=1}^{n} \left(\frac{y_i}{x_i}\right)^2 - \left[\sum_{i=1}^{n} \left(\frac{y_i}{x_i^2}\right)\right]^2 \Big/ \sum_{i=1}^{n} \left(\frac{1}{x_i^2}\right) .\end{aligned}$$

Pour la somme des carrés des écarts résiduelle, les deux tableaux d'analyse de la variance donnent les mêmes résultats. En effet :

$$\mathrm{SCE}_{z_2.1/x} = \sum_{i=1}^{n} \left(\frac{y_i}{x_i} - \frac{\hat{y}_i}{x_i} \right)^2 = \sum_{i=1}^{n} \left(\frac{y_i}{x_i} - \frac{a}{x_i} - b \right)^2$$

et $$(\mathrm{SCE}_{y.x})_w = \sum_{i=1}^{n} w_i (y_i - a - b\,x_i)^2 = \sum_{i=1}^{n} \left(\frac{y_i}{x_i} - \frac{a}{x_i} - b \right)^2 .$$

Quant aux sommes des carrés des écarts liées aux régressions, qui s'obtiennent par différence, elles sont forcément différentes.

Enfin, on constate encore que les limites de confiance d'une valeur estimée individuelle données par Minitab sont incorrectes. En désignant par $v\,(\hat{y}_{x_0})$ la variance de l'estimation d'une valeur moyenne, les limites de confiance d'une valeur individuelle sont égales à :

$$\hat{y}_{x_0} \pm t_{1-\alpha/2} \sqrt{v\,(\hat{y}_{x_0})) + \hat{\sigma}^2_{y|x_0}}\,,$$

Pour $x_0 = 50$, par exemple, la variance de l'estimation de la valeur moyenne est égale au carré de l'erreur-standard donnée dans la figure 2 soit :

$$v\,(\hat{y}_{50}) = 6{,}64^2$$

et la variance résiduelle vaut :

$$\hat{\sigma}^2_{y|x} = (0{,}43)(50^2)\,,$$

la valeur 0,43 étant la variance résiduelle déduite du tableau d'analyse de la variance de la régression non pondérée de z_2 en fonction de $1/x$ (figure 1) ou de la régression pondérée de y en fonction de x (figure 2). Les limites de confiance sont donc égales à :

$$1.843{,}2 \pm 2{,}045 \sqrt{6{,}64^2 + (0{,}43)(50^2)}\,,$$

soit 1.774,7 et 1.911,6. On retrouve bien, aux erreurs d'arrondis près, les limites données dans le tableau 3.

6. EXEMPLE 2 : CAS DE DEUX VARIABLES EXPLICATIVES

Une application classique de la régression pondérée concerne l'établissement d'équations de cubage d'arbres. Ces équations expriment généralement le volume d'un arbre en fonction du diamètre, mesuré à 1,30 m du sol, et éventuellement de la hauteur de cet arbre. Elles sont obtenues par régression multiple, à partir de mesures précises de volumes faites, le plus souvent, sur des arbres abattus.

Lorsque l'amplitude des diamètres des arbres de l'échantillon disponible est grande, les variances conditionnelles des volumes peuvent varier dans des

rapports allant, par exemple, de 1 à 10.000 [PALM, 1981a, 1986]. Dans de telles conditions, le calcul d'équations de régression non pondérées peut fournir, pour les petits arbres, des estimations parfois complètement aberrantes, comme, par exemple, des volumes négatifs ou des volumes croissant lorsque le diamètre diminue. L'utilisation de la régression pondérée permet de supprimer ces inconvénients [PALM, 1981b].

Nous allons considérer le calcul d'une équation de cubage pour l'*Eucalyptus globulus*, à partir des mesures des volumes (v, en m^3), des diamètres pris à 1,30 m du sol (d, en m) et des hauteurs (h, en m). Les données relatives à 72 arbres sont fournies par VILLARROEL [1993].

L'équation de régression suivante a tout d'abord été ajustée par les moindres carrés ordinaires :

$$v = a + b_1 d + b_2 d^2 + b_3 h + b_4 d^2 h \,.$$

Ce modèle a été retenu, *a priori*, sur la base d'études antérieures relatives à d'autres espèces forestières [PALM, 1981a, 1981b]. Cependant, à cause des corrélations très importantes entre les variables explicatives, les valeurs t_{obs} associées aux tests de signification des coefficients de régression partielle sont toutes très nettement inférieures à 2, sauf pour la variable d^2h. Le modèle comporte certainement trop de paramètres. Toutefois, dans un premier temps du moins, il est préférable de n'éliminer aucune variable car, dans la mesure où on soupçonne une inégalité des variances conditionnelles, les tests classiques de signification des coefficients de régression partielle sont incorrects.

D'autre part, l'expérience montre que, pour ce type de problème, la variance conditionnelle est une fonction croissante de la dimension des arbres et on peut considérer que la variable d^2h est une bonne mesure globale de la dimension, puisqu'elle tient compte à la fois de la grosseur et de la hauteur de ceux-ci.

La figure 3 donne le diagramme de dispersion du logarithme du carré des résidus en fonction du logarithme de la variable d^2h. D'après ce diagramme, il semble bien que les écarts les plus importants soient associés aux arbres les plus grands, comme on pouvait s'y attendre. L'hétéroscédasticité n'est cependant pas très manifeste, à cause de la faible amplitude des grosseurs, les diamètres étant, en effet, tous compris entre 0,10 et 0,26 m.

La régression des résidus, élevés au carré et divisés par la moyenne des carrés, en fonction de d^2h conduit à une somme des carrés des écarts due à la régression de 75,3, soit une valeur χ^2_{obs} du test d'homoscédasticité de COOK et WEISBERG (paragraphe 3) égale à 37,6, qui est très hautement significative. L'hypothèse d'homoscédasticité ne peut donc être acceptée.

Si on calcule la droite de régression, pour le diagramme de dispersion donné à la figure 3, on obtient un coefficient de régression égal à 1,25, ce qui conduirait à un écart-type conditionnel proportionnel à la quantité d^2h élevée à la puissance 0,625. Il ne nous paraît cependant pas opportun de retenir ce modèle de variance, car l'examen de la figure 3 montre que le résidu relatif à l'arbre le plus petit c'est-à-dire correspondant à la plus faible valeur de d^2h est anormalement élevé,

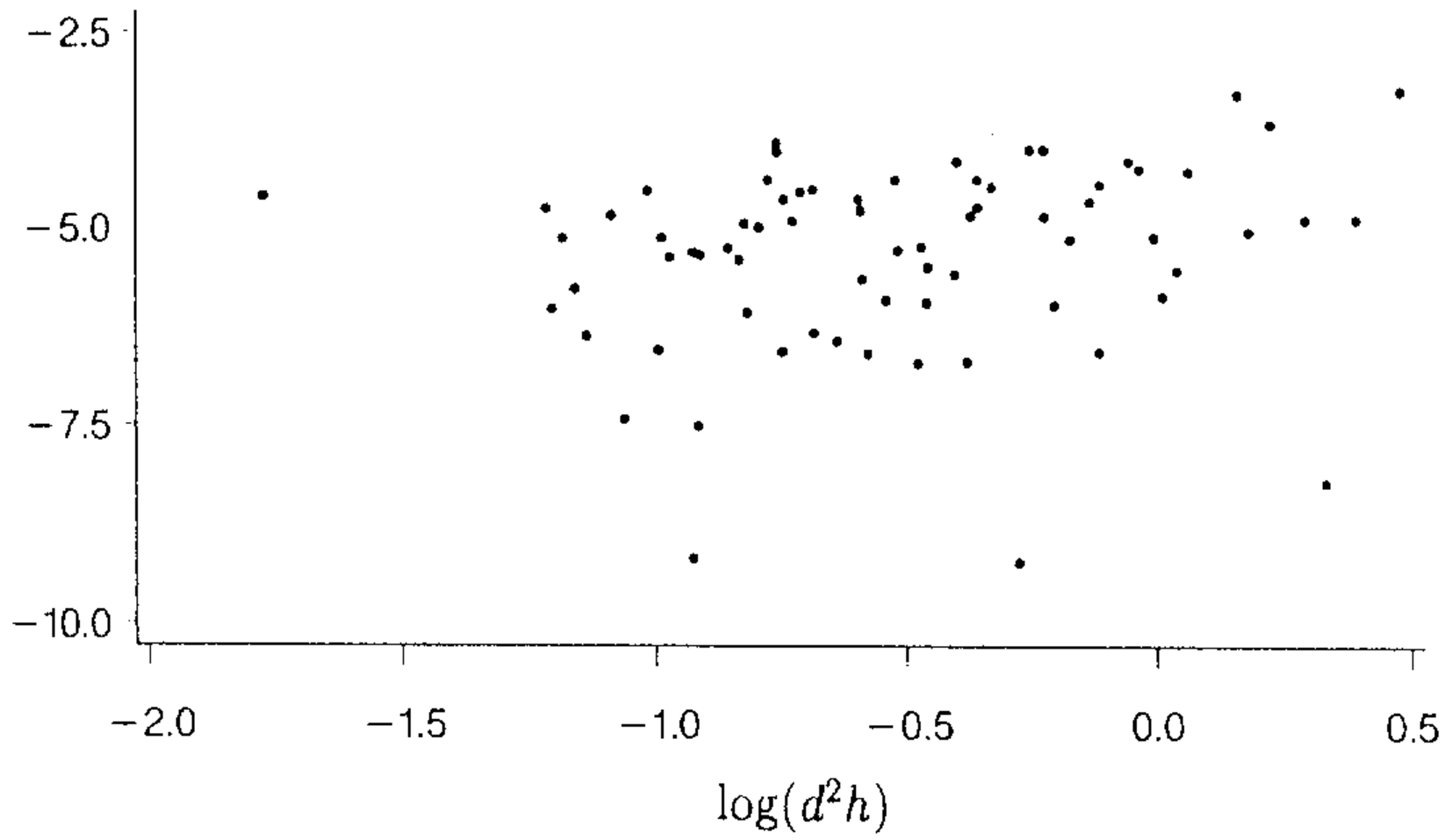

Figure 3. Diagramme de dispersion des logarithmes des carrés des résidus en fonction des logarithmes de d^2h.

à cause du mauvais ajustement de l'équation de cubage par la méthode des moindres carrés non pondérés. Il nous semble au contraire préférable, et les calculs ont montré le bien fondé de ce choix, de retenir un modèle a *priori*.

Parmi les modèles simples d'évolution de l'écart-type résiduel en fonction de la grosseur des arbres, on peut penser aux deux modèles suivants :

$$\sigma_{v|d,h} = k_1 d^2 h \quad \text{et} \quad \sigma_{v|d,h} = k_2 d^2 ,$$

qui traduisent, en première approximation, que l'écart-type augmente de façon linéaire avec le volume. En effet, l'expérience montre que les relations entre le volume et d^2h, d'une part, et entre le volume et d^2, d'autre part, sont approximativement linéaires, surtout si l'amplitude des grosseurs est assez faible.

Partant de ces deux hypothèses, nous avons calculé les deux équations de régression suivantes :

$$\frac{v}{d^2h} = \frac{a}{d^2h} + \frac{b_1}{dh} + \frac{b_2}{h} + \frac{b_3}{d^2} + b_4$$

et

$$\frac{v}{d^2} = \frac{a}{d^2} + \frac{b_1}{d} + b_2 + b_3\frac{h}{d^2} + \frac{b_4}{h} .$$

Les valeurs χ^2_{obs} du test d'homoscédasticité sont, pour ces deux équations, respectivement égales à 0,025 et 1,28, aucune des deux valeurs n'étant significa-

Tableau 4. Estimations ponctuelles ($\hat{v}$), erreurs-standards des estimations (E.S.), limites de confiance, au niveau 0,95, pour une valeur moyenne (LIM et LSM) et limites de confiance pour une valeur individuelle (LII et LSI).

d	h	$\hat{v}$	E.S.	LIM	LSM	LII	LSI
0,10	14	0,055	0,0014	0,052	0,058	0,049	0,060
0,10	16	0,060	0,0012	0,058	0,063	0,055	0,066
0,10	18	0,067	0,0011	0,064	0,069	0,060	0,073
0,15	18	0,139	0,0010	0,137	0,141	0,126	0,152
0,15	20	0,153	0,0009	0,151	0,154	0,138	0,167
0,15	22	0,166	0,0009	0,164	0,167	0,150	0,182
0,20	20	0,262	0,0018	0,258	0,266	0,236	0,288
0,20	22	0,285	0,0022	0,281	0,290	0,257	0,314
0,20	24	0,309	0,0027	0,303	0,314	0,277	0,340
0,25	22	0,437	0,0045	0,428	0,447	0,392	0,483
0,25	24	0,474	0,0054	0,464	0,485	0,425	0,524
0,25	26	0,510	0,0062	0,497	0,523	0,457	0,564

tive. Et puisque la première transformation conduit à une valeur χ^2_{obs} inférieure à la seconde, nous admettrons que l'écart-type est proportionnel à d^2h.

Pour la transformation retenue, on peut maintenant procéder à une sélection de variables, afin d'éliminer les variables redondantes. La sélection pas à pas conduit au modèle suivant :

$$\frac{v}{d^2h} = 0{,}2928 + \frac{0{,}138}{dh},$$

l'écart-type résiduel étant égal à 0,0160.

L'équation de cubage s'écrit donc finalement :

$$v = 0{,}2928d^2h + 0{,}138d,$$

et l'écart-type conditionnel est égal à :

$$\hat{\sigma}_{v|d,h} = 0{,}016d^2h.$$

A titre d'illustration, le tableau 4 donne les informations relatives aux estimations des volumes pour quelques valeurs de d et de h.

On notera que lors de la procédure de sélection des variables, on a supposé, implicitement, que la suppression de l'une ou l'autre variable non significative ne modifie pas de façon sensible l'ordre de grandeur des résidus et ne remet pas en cause l'hypothèse d'homoscédasticité. Cette supposition, qui a d'ailleurs été vérifiée dans le cas de l'exemple, semble a *priori* assez raisonnable et pourra le plus souvent être acceptée en pratique.

Enfin, cet exemple met clairement en évidence un phénomène qui est tout à fait général, bien qu'il ne soit pas toujours aussi marqué que dans le cas présent. Il s'agit de la modification du coefficient de détermination multiple, R^2, du fait de la transformation de la variable dépendante. Ainsi, pour l'équation de régression non pondérée établie en premier lieu et dont la variable à expliquer est v, le coefficient de détermination multiple vaut 0,99. Pour les deux régressions non pondérées dont les variables dépendantes sont v/d^2h et v/d^2, les valeurs de R^2 sont respectivement égales à 0,35 et 0,02.

Par définition, dans le cas de la régression linéaire avec ordonnée à l'origine, le coefficient de détermination exprime la part de la variance de la variable dépendante qui est expliquée par la régression.Cette interprétation reste valable pour les trois équations mais la comparaison des trois valeurs entre elles n'a aucun sens, car les variables dépendantes sont différentes. En particulier, le choix d'une transformation sur base de ce coefficient serait tout à fait injustifié.

On notera, par ailleurs, que lorsqu'on calcule directement l'équation de régression pondérée avec le logiciel Minitab, aucune valeur de R^2 n'est indiquée. Par contre, le logiciel SAS donne un coefficient de détermination multiple, qui est obtenu en faisant simplement le rapport entre la somme des carrés des écarts liée à la régression et la somme des carrés des écarts totale, ces sommes de carrés des écarts étant calculées à l'aide des formules décrites au paragraphe 5.

7. QUELQUES INFORMATIONS COMPLÉMENTAIRES

Les exemples analysés montrent clairement qu'en présence d'une hétéroscédasticité très marquée, l'utilisation de la régression non pondérée peut conduire à des résultats incohérents alors que la régression pondérée permet de traiter ce problème de façon tout à fait satisfaisante.

La régression pondérée n'est cependant pas la seule méthode utilisable dans ce contexte. En effet, dans certains cas, la transformation de la variable à expliquer permet de stabiliser les variances conditionnelles. Dans cette optique, la transformation logarithmique est fréquemment proposée. Il faut cependant remarquer que ce type de transformation altère la nature de la relation qui existe entre la variable à expliquer et la ou les variables explicatives. Ainsi, pour le premier exemple traité, la relation entre le logarithme de y et x est très nettement non linéaire, alors que la relation entre y et x n'est que très légèrement non linéaire. Pour cet exemple particulier, la transformation logarithmique ne stabilise d'ailleurs pas très bien les variances conditionnelles.

La transformation logarithmique peut, par contre, être avantageusement retenue si elle stabilise les variances et si, en même temps, elle linéarise le modèle. Un exemple concret d'une telle situation est donné par DAGNELIE [1992].

Si on dispose d'une relation satisfaisante entre y et les différentes variables explicatives, une transformation de variable adéquate affectant les deux membres de la relation peut être également envisagée. Des informations à ce sujet sont données par CARROLL et RUPPERT [1988].

Il faut noter également que, dans certains cas particuliers, la démarche présentée dans les paragraphes précédents se simplifie nettement. C'est le cas, lorsque, sur la base de considérations théoriques, on sait *a priori* qu'il y a une inégalité de variances conditionnelles et que, de plus, ces variances conditionnelles sont connues, éventuellement à une constante près. C'est, par exemple, le cas si les y_i sont les moyennes de n_i valeurs individuelles correspondant au même vecteur $\boldsymbol{x}_i$:

$$y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \quad (i = 1, \ldots, n) .$$

Si les variances conditionnelles des y_{ij} ne dépendent pas de $\boldsymbol{x}_i$, alors les variances des moyennes conditionnelles y_i sont égales à σ^2/n_i. De même, si la relation entre une variable y et une variable x est linéaire et si les valeurs de y_i sont distribuées selon une loi de POISSON, on a :

$$\sigma^2_{y|x} = a + b\,x .$$

L'utilisation de la régression pondérée avec pondérations connues peut aussi être préconisée lors du calcul de l'équation de régression exprimant les variances conditionnelles estimées par classes, $\tilde{\sigma}_{y|\boldsymbol{x}_i}$, en fonction des moyennes des classes $\bar{x}_i$, à laquelle nous avons fait allusion au paragraphe 4. Dans ce cas, les facteurs de pondération seront les degrés de liberté associés aux variances estimées des différentes classes.

D'une façon générale cependant, la loi donnant l'évolution de la variance conditionnelle n'est pas connue *a priori* mais est estimée à partir du comportement des résidus obtenus par une régression non pondérée. Nous avons par ailleurs déjà signalé qu'une très forte inégalité des variances conditionnelles pouvait conduire à des valeurs estimées de y qui sont aberrantes. Il en résulte que la loi donnant l'évolution de la variance conditionnelle établie à partir des résidus obtenus par un mauvais ajustement peut être inadéquate. Pour pallier cet inconvénient, on peut utiliser la méthode itérative des moindres carrés repondérés[4] [CARROLL et RUPPERT, 1988, WEISBERG, 1985]. Cette méthode consiste à répéter plusieurs fois le calcul de l'équation de régression pondérée en utilisant, comme estimations des variances conditionnelles, les résultats déduits des résidus de l'équation obtenue à l'itération précédente, le processus s'arrêtant lorsque les variances conditionnelles ne se modifient plus de façon sensible. On notera cependant que la théorie élémentaire qui a été présentée au paragraphe 2 suppose que la matrice $\boldsymbol{\Sigma}$ est connue. Dans le cas d'échantillons de petit taille, le remplacement de $\boldsymbol{\Sigma}$ par $\hat{\boldsymbol{\Sigma}}$ peut conduire à une sous-estimation des erreurs-standards des paramètres de l'équation de régression, surtout si cette estimation résulte d'une utilisation répétée des données de l'échantillon. Pour cette raison, le recours à des modèles donnant l'évolution des variances conditionnelles qui sont simples et choisis *a priori* nous semble préférable.

Enfin, signalons encore que nous nous sommes limités au problème de la régression linéaire. Il est toutefois possible d'étendre les principes exposés au

[4] En anglais : *iteratively reweighted least squares.*

cas de la régression non linéaire. Des informations à ce sujet sont données par CARROLL et RUPPERT [1988].

8. BIBLIOGRAPHIE

BREUSCH T.S., PAGAN A.R. [1979]. A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47, 1287-1294.

CARROLL R.J., RUPPERT D. [1988]. *Transformation and weighting in regression.* New York, Chapman and Hall, 249 p.

COOK R.D., WEISBERG S. [1983]. Diagnostics for heteroscedasticity in regression. *Biometrika* 70, 1-10.

DAGNELIE P. [1975]. *Théorie et méthodes statistiques : applications agronomiques* (vol. 2). Gembloux, Presses Agronomiques, 463 p.

DAGNELIE P. [1982]. *Analyse statistique à plusieurs variables.* Gembloux, Presses Agronomiques, 362 p.

DAGNELIE P. [1992]. *Statistique théorique et appliquée* (tome 1). Gembloux, Presses Agronomiques, 492 p.

DRAPER N.R., SMITH H. [1981]. *Applied regression analysis.* New York, Wiley, 709 p.

GOLDFELD S.M., QUANDT R.E. [1965]. Some tests for homoscedasticity. *J. Amer. Stat. Assoc.* 60, 535-547.

PALM R. [1981a]. *Contribution méthodologique au cubage des arbres et à la construction de tables de cubage et d'assortiments* (Thèse de doctorat). Gembloux, Faculté des Sciences agronomiques, 295 p.

PALM R. [1981b]. Calcul et choix des équations de cubage d'arbres. *Bull. Rech. Agron. Gembloux* 16, 351-370.

PALM R. [1986]. Etude des résidus de régression : principes et application. *Notes Stat. Inform.* (Gembloux) 86/1, 13 p.

PALM R. [1988]. Les critères de validation des équations de régression linéaire. *Notes Stat. Inform.* (Gembloux) 88/1, 27 p.

RODRIGUEZ ZEVALLOS A.R. [1992]. *Comparaison des méthodes d'analyse de l'acide ascorbique et de ses principaux produits de dégradation en vue de l'étude de la stabilité de purées de fraises et de compotes de pommes* (Thèse de doctorat), Gembloux, Faculté des Sciences agronomiques, 224 p.

THEIL H. [1971]. *Principles of econometrics.* New York, Wiley, 736 p.

VILLARROEL L.A. [1993]. *Traitement statistique des données de parcelles semi-permanentes d'*Eucalyptus globulus L. *à Cochabamba* (Bolivie) (Travail de fin d'étude). Gembloux, Faculté des Sciences agronomiques, 89 p. + annexes.

WEISBERG S. [1985]. *Applied linear regression.* New York, Wiley, 324 p.

WHITE H. [1980]. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* 48, 817-838.

X [1989]. *Minitab reference manual.* Valley Forge, Data Tech. Industries, 349 p.

X [1990]. *SAS/STAT user's guide, version 6* (2 vol.). Cary, SAS Institute, 1686 p.