# FOAD
# COURS D' ECONOMETRIE 1
# CHAPITRE 2 : Hétéroscédasicité des erreurs.
# 23 mars 2012.

Christine Maurel
Maître de conférences en Sciences Economiques
Université de Toulouse 1 - Capitole
Toulouse School of Economics-ARQADE

# Table des matières



# I Introduction

Dans le chapitre précédent nous avons supposé que $Var(u_i) = \sigma^2$ pour tout $i$ c'est à dire que les erreurs sont homoscédastiques. Sur des données individuelles, l'hypothèse d'homoscédasticité peut poser problème. Il y a en pratique deux stratégies possibles. Nous choisirons la première stratégie si on soupçonne un problème d'hétéroscédasticité sans en connaître la forme c'est à dire si l'on pense que la variance des erreurs n'est pas constante pour tous les individus, ou encore que la variance des erreurs dépend de $i$ mais nous n'avons aucune information sur la forme de cette dépendance : la variance de $u_i$ dépend de $i$ mais on ne sait pas "comment". Cette stratégie utilise les propriétés asymptotiques des estimateurs et est donc valable en grand échantillon. En pratique c'est la stratégie la plus utilisée quand l'échantillon est de taille suffisante.
La seconde stratégie sera utilisée si nous disposons d' une information sur la forme de l'hétéroscédasticité. Cette information auxiliaire permet d'utiliser les Moindres Carrés Généralisés, noté MCG (GLS en anglais). Dans ce cas nous pourrons utiliser les propriétés exactes des estimateurs c'est à dire en échantillon fini.
Avant de présenter ces deux stratégies, nous étudierons les conséquences de l'hétéroscédasticité sur les propriétés des estimateurs des MCO dans un premier paragraphe.

# II Conséquences de l'hétéroscédasticité sur les MCO (OLS en anglais)

Si la variance des erreurs n'est plus constante, on montre que les estimateurs des MCO sont toujours sans biais et consistants mais ils ne sont plus les meilleurs

estimateurs (linéaires et sans biais). En effet la variance des MCO, c'est à dire, $\sigma^2(X'X)^{-1}$ n'est plus valide. Ainsi tous les tests présentés dans le chapitre précédent ne sont plus valides. Deux stratégies sont alors possibles : la première stratégie consiste à conserver les estimateurs des MCO et à corriger la variance des estimateurs des MCO. La variance corrigée (ou variance robuste à l'hétéroscédasticité) est une estimation robuste à l'hétéroscédasticité (par défaut et dans la suite de ce chapitre toute mention de robustesse correspond à la robustesse à l'hétéroscédasticité) valable en grand échantillon. Cette prémière stratégie très utilisée en pratique sera présentée dans le premièr paragraphe de ce chapitre. Les paragraphes suivants présenteront la seconde stratégie qui consiste à abandonner les MCO et à faire une hypothèse sur la spécification de la variance des erreurs afin d'appliquer une autre méthode d'estimation, les Moindres Carrés Généralisés (MCG ou GLS en anglais).

# III Inférence robuste à l'hétéroscédasticité

La première stratégie est utilisée si on n'a aucune information sur la forme de l'hétéroscédasticité et permet d'éviter de spécifier la variance des erreurs. White(1980) a proposé une matrice de variance-covariance asymptotique des paramètres estimés qui est la suivante : $\hat{\Sigma} = (X'X)^{-1}(X'diag(\hat{u}_i^2)X)(X'X)^{-1}$ où diag($\lambda_i$) est la matrice diagonale qui contient $\lambda_i$ sur la diagonale.
SAS dispose d'une option, "Acov" pour asymptotic covariance matrix, dans la proc reg qui permet d'afficher cette matrice :

Fichier de données : wage1
Source : Wooldridge, 2006 .
On dispose d'un échantillon de 526 salariés pour lesquels on observe :
Wage : average hourly earnings
Lwage : log(wage)
Married : =1 if married
Female : =1 if female
Educ : years of education
Exper : years potential experience
Expersq : Exper ** 2
Tenure : years with current employer
Tenursq : tenure ** 2
Marrmale =1 si le salarié est un homme marrié, 0 sinon
Marrfem =1 si le salarié est une femme marriée, 0 sinon
Singfem =1 si le salarié est une femme célibataire, 0 sinon

Nous estimons le modèle suivant :

$$Lwage_i = \alpha_0 + \alpha_1 Marrmale_i + \alpha_2 Marrfem_i + \alpha_3 Singfem_i$$

$+\alpha_4 Educ_i + \alpha_5 Exper_i + \alpha_6 Expersq_i + \alpha_7 Tenure_i + \alpha_8 Tenursq_i + u_i$ avec $i = 1, \ldots, 526$.

Nous supposons que toutes les hypothèses des MCO sont vérifiées sauf une : la variance des erreurs n'est pas constante dans l'échantillon. Ainsi nous supposons toujours que la covariance entre les erreurs est nulle et que, hors diagonale la matrice de Variance-Covariance des erreurs ne contient que des 0. Une seule hypothèse est donc levée sur cette matrice : les termes sur la diagonale ne sont pas constants, ils dépendent de $i$. En pratique pour savoir si cette hypothèse d'homoscédasticité doit être levée sur l'échantillon étudié, on procède à des tests d'hétéroscédasticité sur les erreurs. Nous présenterons ces tests après l'étude graphique des résidus.
Pour répresenter les résidus il faut d'abord les enregistrer ; ensuite on les représente en fonction des variables quantitatives du fichier :lotsize et sqrft.

```
proc reg data=tpfoad.wage1 ;
model lwage=marrmale marrfem singfem educ exper expersq tenure tenursq / acov ;
run ;
```

Avec le logiciel R :

```
mod1 <- lm(LWAGE~MARRMALE+MARRFEM+SINGFEM+EDUC+EXPER+EXPERSQ
+TENURE+TENURSQ,data=wage1)
coeftest(mod1,vcov=sandwich)
```

| Valeurs estimées des paramètres | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Cohérent avec l'hétéroscédasticité | | |
| Variable | Libellé | DDL | Valeur estimée des paramètres | Erreur type | Valeur du test t | Pr > \|t\| | Erreur type | Valeur du test t | Pr > \|t\| |
| Intercept | Intercept | 1 | 0.32138 | 0.10001 | 3.21 | 0.0014 | 0.10853 | 2.96 | 0.0032 |
| marrmale | | 1 | 0.21268 | 0.05536 | 3.84 | 0.0001 | 0.05665 | 3.75 | 0.0002 |
| marrfem | | 1 | -0.19827 | 0.05784 | -3.43 | 0.0007 | 0.05827 | -3.40 | 0.0007 |
| singfem | | 1 | -0.11035 | 0.05574 | -1.98 | 0.0483 | 0.05663 | -1.95 | 0.0519 |
| EDUC | | 1 | 0.07891 | 0.00669 | 11.79 | <.0001 | 0.00735 | 10.73 | <.0001 |
| EXPER | | 1 | 0.02680 | 0.00524 | 5.11 | <.0001 | 0.00509 | 5.26 | <.0001 |
| EXPERSQ | | 1 | -0.00053525 | 0.00011043 | -4.85 | <.0001 | 0.00010543 | -5.08 | <.0001 |
| TENURE | | 1 | 0.02909 | 0.00676 | 4.30 | <.0001 | 0.00688 | 4.23 | <.0001 |
| TENURSQ | | 1 | -0.00053314 | 0.00023124 | -2.31 | 0.0215 | 0.00024159 | -2.21 | 0.0278 |

FIGURE 1 – inférence robuste sur wage1

Dans la fenêtre de résultats de SAS 9.3 et dans le tableau "valeurs estimées des paramètres" (voir Fugure 1 - inférence robuste sur wage1), la première colone contient les estimateurs par MCO. La colone suivante contient les écart-types des MCO qui ne sont plus valables en présence d'hétéroscédasticité. Les deux colones suivantes contiennent les tests de significativité qui utilisent les écart-tyoes précédents et qui ne sont donc plus valables non plus en présence d'hétéroscédasticité.

Les trois dernière colonnes de " cohérent avec l'hétéroscédasticité" contiennent les écart-types robustes et les tests asymptotiques correspondant.
Les commentaires sont les suivants :
- La taille de l'échantillon est de 526 qui est suffisament grande pour raisonner asymptotiquement.
- En général les écart-types robustes sont plus grands que les écart-types des MCO. C'est le cas ici sauf pour les variables Exper et Expersq.
- Les écart-types robustes sont du même ordre de grandeur que les écart-types des MCO sauf pour la variable Educ. Ainsi les seuils de signification sont sensiblement les mêmes avec les deux écart-types.
- En ce qui concerne l'hétéroscédasticité étant donné que les écart-types ont le même ordre de grandeur on peut penser que l'hétéroscédasticité n'est pas très présente.

Pour terminer il est recommandé de lire dans la documentation de SAS (ou dans la documentation du package 'sandwich' de R, page 18) les différentes matrices asympotiques qui peuvent être utilisées avec l'option HCCMETHOD où 3 matrices sont proposées (par défaut il s'agit de la matrice de White présentée dans ce paragraphe) ; en particuiler un article de Long and Ervin (2000) recommende d'utiliser la troisième matrice si la taille d'échantillon est inférieure à 250.

Nous allons maintenant présenter la seconde stratégie qui consiste à abandonner les estimateurs des MCO et à estimer le modèle par MCG. La première étape de cette stratégie consiste à procéder à des tests d'hétéroscédasticité. En cas de rejet de l'homoscédasticité des erreurs, la seconde étape consistera à appliquer les MCG en faisant une hypothèse sur la forme de l'hétéroscédasticité.

# IV Tests d'hétéroscédasticité

Nous allons présenter dans ce paragraphe deux tests d'hétéroscédasticité des erreurs qui sont des tests asymptotiques très utilisés en pratique :le test de Breusch-Pagan (1979)et celui de White (1980). Nous présenterons brièvement le principe d'un test asymptotique du multiplicateur de Lagrange avant de présenter les tests d'hétéroscédasticité (source : Wooldridge, 2006).

1. Rappel : Test asympotique du multiplicateur de Lagrange (ML) :
   Soit le modèle de regression partitionnée [1] suivant :
   $Y = X_1\beta_1 + X_2\beta_2 + u$ où $X_1$ est de dimension (1,$k_1$) en incluant une constante et $X_2$ de dimension (1,$k_2$) . Nous allons présenter une autre approche pour tester $H_0$ : $\beta_2 = 0$. Cette approche utilise le modèle contraint. Notons $\hat{\beta}_{1c}$ [2] l'estimateur de $\beta_1$ sous $H_0$ . Le résidu du modèle contraint est $\hat{u}_c = Y - X_1\hat{\beta}_{1c}$.
   Sous $H_0$, $\hat{u}_c$ ne doit pas être corrélé avec $X_2$. Le test du multiplicateur de

1. Un modèle partitionné est une écriture différente d'un modèle, écriture qui est simple à utiliser pour présenter le test de ML
2. où "c" signifie "Contraint" comem dans le chapitre 1

Lagrange ou test du Score, repose sur cette observation sur le modèle contraint. On procède de la manière suivante : on régresse le résidu du modèle contraint $\hat{u}_c$ sur $X_1$ et $X_2$. Soit le $R^2$ de cette régression[3]. La statistique de test est $N \times R^2$ qui suit, sous $H_0$, une loi de $\chi^2$ à $k_2$ degré de liberté. Si $N \times R^2$ est "grand", le résidu est corrélé avec $X_2$ et on rejette $H_0$. Attention : il est important de faire figurer $X_1$ dans la regression auxiliaire des résidus même si ce résidu est toujours orthogonal à $X_1$. Si $X_1$ est exclue , la statistique ne suit généralement pas une loi de $\chi^2$ quand $X_1$ et $X_2$ sont corrélées, ce qui est très souvent le cas en pratique.

Nous pouvons maintenant présenter les tests de Breusch-Pagan puis celui de White qui utilisent un test du multiplicateur de Lagrange.

2. Test de Breusch-Pagan (1979)

   Ce test sera noté BP dans la suite du cours.

   Il s'agit de tester l'hypothèse selon laquelle la variance des erreurs ne dépend pas des variables explicatives du modèle. Soit le modèle de regression multiple $Y = X\beta + u$ .L'hypothèse nulle d'homoscédasticité est $H_0 : Var(u/X) = \sigma^2$ ou encore $H_0 : E(u^2/X) = \sigma^2$. Pour tester la violation de cette hypothèse nous allons tester si $u_i^2$ dépend d'une ou de plusieurs variables explicatives. Si $H_0$ est rejettée alors l'espérance (conditionelle) de $u_i^2$ est une fonction des variables explicatives. La spécification la plus simple de cette fonction est $u_i^2 = \delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k + v$ où $v$ vérifie toutes les hypothèses des MCO. L'hypothèse nulle d'homoscédasticité devient $H_0 : \delta_1 = \delta_2 = \ldots = \delta_k = 0$. Si les paramètres $\delta$ sont tous nuls, alors la dispersion des erreus ne dépend pas des variables explicatives ; cette dispersion devient une constante et nous retrouvons l'hypothèse d'homoscédasticité. Nous procédons ensuite à un test du multiplicateur de Lagrange pour tester la nullité des $\delta$ :

   - On régresse Y sur l'ensemble des variables explicatives par MCO ;on sauve les résidus et on calcule $\hat{u}_i^2$
   - On regresse $\hat{u}_i^2$ sur toutes les variables explicatives . On obtient le $R^2$ de cette régression auxiliaire.
   - On calcule la statistique du multiplicateur de Lagrange :$LM = n \times R^2$
   - On compare cette valeur observée à la valeur critique d'une table de $\chi^2$ à k-1 dégré de liberté où $k$ est le nombre de paramètre de la regression auxiliaire (ou on calcule la probabilité de dépasser la valeur observée). On procède de la manière habituelle pour rejeter ou pas $H_0$.

   APPLICATION :

   Le fichier de données, Hprice, fera l'objet d'un TP de ce chapitre.

   Fichier de données : Hprice

   Source : Wooldridge, 2006 p 281.

   On dispose d'un échantillon de 88 observations (88 maisons) pour lesquelles on observe :

   Price : prix de la maison (house price), en milliers de dollars

3. on nomme ce type de régression "régression auxiliaire" car elle sert un objectif statistique mais elle n'a pas d'intérêt économique.

Bdrms : nombre de chambre (number of bedrooms)
Lotsize : superficie du terrain (size of lot in square feet)
Sqrft : superficie de la maison (size of house in square feet)

On régresse les résidus au carré sur l'ensemble des variables explicatives du prix :

```
*test de BP ;
proc reg data=tpfoad.hprice ;
model price=lotsize sqrft bdrms ;
output out=t r=res ;
run ;
data t1 ;set t ;
res2=res*res ;run ;
proc reg data=t1 ;
model res2=lotsize sqrft bdrms ;run ;
data ct1 ;n=88 ;r2= 0.1601 ;ddl= 3 ;
obs=n*r2 ;p=1-probchi(obs,ddl) ;run ;
proc print data=ct1 ;run ;
*BP=14.0899 p= .0028 ;
```

Avec R :

```
ols <- lm(PRICE   LOTSIZE+SQRFT+BDRMS,data=hprice)
summary(ols)
pbp <- lm(I(residuals(ols)^2) LOTSIZE+SQRFT+BDRMS,data=hprice)
summary(pbp)
jour_pbp <- summary(pbp)
class(jour_pbp)
names(jour_pbp)
r2aux <- jour_pbp$r.square
r2aux
obs <- r2aux*88
obs
pvalue <- 1-pchisq(obs,3)
pvalue
```

Avec le test de BP, on rejette l'hypothèse d'homoscédasticité. Avec SAS, la procédure "proc model" calcule le test de BP. Je tiens tout de même à ce que les étudiants sachent retrouver, en programmant, la valeur observée des tests de SAS car il est parfois difficile de comprendre quel est le test utilsé par une procédure SAS et que de plus la programmation du test permet de mieux le comprendre. Le programme est le suivant :

```
*Test de BP avec proc model;
proc model data=tpfoad.hprice;
parms b0 b1 b2 b3;
price = b0 + b1* lotsize + b2 *sqrft +b3*bdrms;
fit price / OLS breusch=( 1 lotsize sqrft bdrms);
run;
```

Avec R :

```
library(AER)
ols <- lm(PRICE~LOTSIZE+SQRFT+BDRMS,data=hprice)
summary(ols)
bptest(ols,data=hprice)
```

On retrouve les mêmes valeurs que dans notre programme : obs= 14.09 et p = 0.0028 avec ddl= 3.

3. Test de White (1980)
Le test de White repose sur le même principe que le test de BP. La seule différence réside dans le fait que la liste des variables explicatives de la regression auxiliaire est plus grande. Pour obtenir cette liste on effectue le "produit" des variables du modèle de la manière suivante :
Nous avons $X = (1, lotsize, sqrft, bdrms)$. Nous effectuons le produit de ces variables par elle-même :
$(1, lotsize, sqrft, bdrms) \times (1, lotsize, sqrft, bdrms)$ ce qui nous fournit la liste des variables de la régression auxiliaire suivante :
$(1, lotsize, sqrft, bdrms, lotsize^2, lotsize \times sqrft, lotsize \times bdrms, sqrft^2, sqrft \times bdrms, bdrms^2)$

Remarquons que pour le test de White la liste des variables explicatives de la régression auxiliaire contient toutes les variables du modèle (comme le test de BP) ainsi que tous les doubles-produits des variables. Ainsi si le modèle de régression contient les variables $(1, X_1, X_2, X_3)$ alors la régression auxiliaire est

$$u_i^2 = \delta_0 + \delta_1 X_1 + \delta_2 X_2 \delta_3 X_3 + \delta_4 X_1^2 \delta_5 X_2^2 + \delta_6 X_3^2 + \delta_7 X_1 X_2 + \delta_8 X_1 X_3 + \delta_9 X_2 X_3 + erreur$$

. On procède ensuite à un test du multiplicateur de Lagrange comme précédement. Nous remarquons que le nombre de paramètres de la régression auxiliaire est assez élevé même avec 3 paramètres dans le modèle initial donc la perte de degré de liberté peut poser des problèmes en pratique. Une seconde version de ce test est la suivante : dans la regression auxiliaire on regresse $\hat{u}_i^2$ sur $\hat{Y}_i$ et $\hat{Y}_i^2$. Dans ce cas le degré de liberté de la loi du $\chi^2$ est égal à 2.
Le test de White est plus général que le test de BP mais la puissance de ce test est plus faible ; en effet le test de White peut rejetter l'homoscédasticité en cas d'erreur de spécification comme l'omission d'une variable explicative au carré dans le modèle. Nous reviendrons sur ce point lors du TP qui concerne le même fichier "hprice" avec une autre spécification du modèle.
Programme de calcul avec SAS :

```
*test de White ;
data verifw ;set t1 ;
*creation des var de la reg aux ;
lot2=lotsize*lotsize ;sq2=sqrft*sqrft ;
lotsq=lotsize*sqrft ;lotb=lotsize*bdrms ;
sqb=sqrft*bdrms ;bdrms2=bdrms*bdrms ;run ;
*reg aux ;
proc reg data=verifw ;
model res2= lotsize sqrft bdrms
lot2 lotsq lotb sq2 sqb bdrms2 ;run ;
data ww ;n=88 ;r2= 0.3833 ;obs=n*r2 ;ddl= 9 ;
p=1-probchi(obs,ddl) ;run ;
proc print ;run ;
```

Avec R je vous laisse faire le programme.
On obtient une statistique observée pour le test de White égale à 33.7304 , avec ddl= 9 et p=.000099.
Avec la proc model :

```
*Test de W avec proc model ;
proc model data=tpfoad.hprice ;
parms b0 b1 b2 b3 ;
price = b0 + b1* lotsize + b2 *sqrft +b3*bdrms ;
fit price / OLS white ;
run ;
```

Avec R :

```
bptest(ols, LOTSIZE+SQRFT+BDRMS+I(LOTSIZE^2)+(LOTSIZE*SQRFT)+
(LOTSIZE*BDRMS)+I(SQRFT^2)+(SQRFT*BDRMS)+I(BDRMS^2) , data=hprice)
```

On retrouve les mêmes valeurs que dans notre programme.
Dans la proc model on peut demander le test de BP et celui de White simultanément en écrivant : “fit price / OLS breusch=( 1 lotsize sqrft bdrms) white ;"
La proc model ne calcule pas la statistique de White de la seconde version du test de White et donc il faut la programmer :

```
*test de White seconde version ;
proc reg data=tpm1.hprice ;
model price=lotsize sqrft bdrms ;
output out=f r=res p=pricehat ;
quit ;run ;
data w2 ;set f ;res2=res*res ;pricehat2=pricehat*pricehat ;run ;
proc reg data=w2 ;model res2 =pricehat pricehat2 ;run ;
data calculw2 ;n=88 ;r2= 0.1849 ;obs=n*r2 ;ddl= 2 ;
p=1-probchi(obs,ddl) ;run ;
proc print ;run ;
```

On obtient une statistique observée égale à 16.2712 et une probabilité de 0.000292923

avec un degré de liberté de 2. On rejette aussi l'homoscédasticité avec cette seconde version du test de WHite.
Conclusion : En pratique il suffit qu'un test au moins sur les 2 tests présentés, rejette l'homoscédasticité des erreurs pour que l'on soupçonne un problème d'hétéroscédasticité des erreurs. Nous sommes donc amenés à penser que les erreurs de l'équation du prix des maisons sont hétéroscédastiques. En présence d'hétéroscédasticité nous savons que les MCO sont toujours sans biais ce qui signifie que l'élasticité du prix par rapport à la taille du terrain par exemple, est aussi sans biais. Par contre les estimateurs des MCO ne sont pas les meilleurs estimateurs ce qui signifie que leur variance n'est pas minimale. Ainsi les tests auxquels nous avons procédé ne sont pas valables en présence d'hétéroscédasticité, comme par exemple les tests de significativité des variables. Pour obtenir des estimateurs efficaces il faut estimer le modèle par Moindres Carrés Généralisés. Cette méthode est présentée dans le paragraphe qui suit.

# V Correction de l'hétéroscédasticité : les Moindres Carrés Généralisés (Generalized Least Squares)

Avant d'appliquer les MCG sur notre échantillon nous allons brièvement donner des éléments théoriques.

1. Eléments théoriques :
   Nous avons vu dans le chapitre 1 et dans le paragraphe sur les MCO, le modèle de vrégression multiple suivant : $Y = X\beta + u$ avec $E(u/X) = 0$ et $Var(u/X) = \sigma^2 I$. En présence d'hétéroscédasticité , la diagonale de la matrice $Var(u/X)$ n'est plus constante ; elle dépend de $i$ et on a

$$Var(u/X) = \sigma^2\Omega = \sigma^2 \begin{pmatrix} \omega_1 & 0 & \dots & 0 \\ 0 & \omega_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \omega_N \end{pmatrix}$$

   L'estimateur des MCG en présence d'hétérosédasticité est :
   $\hat{\beta}_{MCG} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$ avec $Var\hat{\beta}_{MCG} = \sigma^2(X'\Omega^{-1}X)^{-1}$ où $\sigma^2$ est estimé de la manière suivante :
   On construit le résidu des MCG :$\hat{u}_{MCG} = Y - X\hat{\beta}_{MCG}$ puis on calcule $\hat{\sigma^2} = \frac{\hat{u}'_{MCG}\Omega^{-1}\hat{u}_{MCG}}{N-k}$
   On montre que $\Omega$ étant une matrice symétrique définie positive, $\Omega^{-1}$ existe et elle est aussi symétrique définie positive donc il existe une matrice $T$ telle que $\Omega^{-1} = T'T$.
   $T$ est la matrice de transformation du modèle $Y = X\beta + u$. Le modèle transformé par $T$ est $TY = TX\beta + Tu$. On montre que $Var(Tu/X) = \sigma^2 I$ et donc les MCO sur le modèle transformé sont BLUE puisque les erreurs du modèle transformé sont homoscédastiques. Nous utiliserons ceci dans les applications des MCG sur le fichier “hprice".

Remarque : Dans ce chapitre nous levons l'hypothèse d'homoscédasticité mais nous conservons l'hypothèse de non corrélation des erreurs. Ce qui rend très simple le calcul de la matrice $T$. Ainsi , $\Omega = \begin{pmatrix} \omega_1 & 0 & \dots & 0 \\ 0 & \omega_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \omega_N \end{pmatrix}$ et

$$\Omega^{-1} = \begin{pmatrix} \frac{1}{\omega_1} & 0 & \dots & 0 \\ 0 & \omega_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{1}{\omega_N} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{\omega_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\omega_2}} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\omega_N}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{\omega_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\omega_2}} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\omega_N}} \end{pmatrix}$$

donc

$$T = \begin{pmatrix} \frac{1}{\sqrt{\omega_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{\omega_2}} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{\omega_N}} \end{pmatrix}$$

Il suffit donc de multiplier le modèle par $\frac{1}{\sqrt{\omega_i}}$ pour obtenir le modèle transfomé. Nous reviendrons sur la transformation du modèle dans les applications sur le fichier "hprice".
Dans la méthode des MCG, deux cas peuvent se présenter : le premier cas concerne les situations où la matrice $\Omega$ est connue ; le second concerne des situations où cette matrice $\Omega$ doit être estimée. En fait le cas où "$\Omega$ est connue" signifie que l'on n'a pas besoin d'estimer $\Omega$.

2. Premier Cas : $\Omega$ est connue
Supposons que $Var(u_i) = \sigma^2 Lotsize^2$, noté exemple1.

   (a) Comment appliquer les MCO sur le modèle transformé ? :
   Si $Var(u_i) = \sigma^2 Lotsize^2$ alors la matrice $\Omega$ est la suivante :

$$\Omega = \begin{pmatrix} Lotsize_1^2 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & Lotsize_N^2 \end{pmatrix}$$

   Nous sommes donc bien dans le cas où $\Omega$ est connue puisque nous connaissons la diagonale de la matrice $\Omega$ car la variable Lotsize de chaque individu est connue.
   La transformation du modèle consiste toujours à diviser l'équation initiale par la racine carrée de la diagonale de la matrice $\Omega^{-1}$. Procédons ainsi sur notre modèle qui est le suivant :
   $Pirce_i = \beta_0 + \beta_1 Lotsize_i + \beta_2 Sqrft_i + \beta_3 Bdrms_i + u_i$
   On transforme ce modèle en divisant par Lotsize et on obtient :

$$\frac{Price_i}{Lotsize_i} = \beta_0 \frac{1}{Lotsize_i} + \beta_1 \frac{Lotsize_i}{Lotsize_i} + \beta_2 \frac{Sqrft_i}{Lotsize_i} + \beta_3 \frac{Bdrms_i}{Lotsize_i} + \frac{u_i}{Lotsize_i}$$

ou encore

$$\frac{Price_i}{Lotsize_i} = \beta_0 \frac{1}{Lotsize_i} + \beta_1 + \beta_2 \frac{Sqrft_i}{Lotsize_i} + \beta_3 \frac{Bdrms_i}{Lotsize_i} + \frac{u_i}{Lotsize_i}$$

Notons $v_i$ l'erreur de ce modèle transformé et calculons sa variance.

$Var(v_i/X) = Var(\frac{u_i}{Lotsize_i}/X) = \frac{1}{Lotsize_i^2} Var(u_i/X) = \frac{1}{Lotsize_i^2} \sigma^2 Lotsize_i^2 = \sigma^2$.

Donc en appliquant la transformation adéquate sur le modèle initial on obtient un modèle transformé dont les erreurs, $v_i$, sont homoscédastiques. Nous pouvons donc appliquer les OLS sur le modèle transformé : ils seront BLUE si $Var(u_i) = \sigma^2 Lotsize_i^2$ est vraie (nous supposons que toutes les autres hypothèses des MCO sont vérifiées).
Pour estimoer par MCO le modèle transformé il faut d'abord créer les variables dont nous avons besoin dans une étape data puis utiliser la proc reg :

```
*OMEGA CONNUE;
*EXEMPLE 1 var u = sigma 2 lotsize2;
data mt1;set tpm1.hprice;
pl=price/lotsize;
ilot=1/lotsize;
sl=sqrft/lotsize;
bl=bdrms/lotsize;
run;
proc reg data=mt1;
model pl=ilot sl bl;run;
```

Les résultats sont donnés dans la figure 2.
Commentaires :
- Attention à la lecture des paramètres estimés dans le modèle transformé : la constante $\beta_0$ est le paramètre de ilot dans le modèle transformé mais c'est la constante dans notre modèle initial et $\beta_1$ est le paramètre de lotsize du modèle initial et ce paramètre est devenu la constante dans le modèle transformé. Pour éviter ce problème de lecture nous utiliserons les Moindres Carrés Pondérés (MCP ou WLS, Weighted Lesat Squares) : Au lieu de transformer le modèle et de minimiser $\sum v_i^2$ c'est à dire $\sum (\frac{u_i}{Lotsize_i})^2$ on peut minimiser la somme pondérée des $u_i^2$ c'est à dire $\sum \frac{1}{Lotsize^2} u_i^2$ où $\frac{1}{Lotsize^2}$ est la pondération. Avec SAS et avec la proc reg , il faut d'abord créer la variable de pondération dans une étape data puis utiliser l'instruction “weight" de la proc reg.
Le programme est le suivant :

Dependent Variable: pl

| | |
|---|---|
| Number of Observations Read | 88 |
| Number of Observations Used | 88 |

Analyse de variance

| Source | DF | Somme des carrés | Carré moyen | Valeur F | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 0.08304 | 0.02768 | 480.77 | <.0001 |
| Error | 84 | 0.00484 | 0.00005757 | | |
| Corrected Total | 87 | 0.08787 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.00759 | R-Square | 0.9450 |
| Dependent Mean | 0.04266 | Adj R-Sq | 0.9430 |
| Coeff Var | 17.78757 | | |

Résultats estimés des paramètres

| Variable | DF | Résultat estimé des paramètres | Erreur std | Valeur du test t | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 0.00736 | 0.00172 | 4.27 | <.0001 |
| ilot | 1 | 21.90458 | 30.41595 | 0.72 | 0.4734 |
| sl | 1 | 0.09729 | 0.00892 | 10.91 | <.0001 |
| bl | 1 | 3.83713 | 7.04293 | 0.54 | 0.5873 |

FIGURE 2 – MCO sur le modèle transformé : Exemple 1

```
*MCP ou WLS en anglais;
data wls; set tpfoad.hprice;
poids=1/(lotsize*lotsize);
run;
proc reg data=wls;
model price=sqrft lotsize bdrms;
weight poids;
quit;run;
```

Avec R :

```
wls1 <- lm(PRICE~LOTSIZE+SQRFT+BDRMS,weights= 1/LOTSIZE^2,
data=hprice)
summary(wls1)
```

Quand nous comparons ces estimations (MCP ou WLS) à celles obtenues par MCO dans le modèle transformé nous constatons que les paramètres WLS obtenus dans le listing de SAS sont directement les paramètres du modèle du prix qui nous intéressent. Il est donc inutile avec les WLS de revenir au modèle initial comme nous l'avons fait pour les MCO sur le modèle transformé.

liste des var significatives ? ou faire inf robuste pares MCG pour voir si hypo sur la variance n'est pas rejettée.

(b) Second cas : $\Omega$ est inconnue

$\Omega$ est inconnue quand cette matrice dépend de paramètres inconnus. Supposons que, par exemple, cette matrice soit une fonction des 3 variables du modèle et dépende de 4 paramètres inconnus de la manière suivante : EXEMPLE 2 : $Var(u_i) = \sigma^2(\theta_0 + \theta_1 Lotsize_i + \theta_2 Sqrft + \theta_3 Bdrms) = \sigma_i^2 = \sigma^2 w_i$.
Dans ce cas on procède en deux étapes pour estimer le modèle de manière efficace : la première étape consiste à estimer $\omega_i$ et à obtenir $\hat{\Omega}$ ; la seconde consiste à utiliser $\hat{\Omega}$ pour calculer un estimateur des MCG : $\hat{\beta}_{MCG} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}Y$. Cette méthode d'estimation des MCG en deux étapes est appelée Moindres Carrés Quasi Généralisés, MCQG, (ou Feasible GLS).
Dans la première étape on régresse les résidus au carré sur les variables spécifiées dans l'équation de la variance de l'exemple 2 c'est à dire sur lotsize, sqrft et bdrms . Grâce à cette régression on obtient une estimation de la variance des $u_i$ en calculant la variable endogène estimée de l'équation des résidus. En effet étant donné que $u_i$ a une espérance nulle sa variance est égale à $E(u_i^2)$. Dans l'exemple 2 nous allons donc supposer que $E(u_i^2) = \sigma_i^2 = \sigma^2(\theta_0 + \theta_1 Lotsize_i + \theta_2 Sqrft + \theta_3 Bdrms)$.
Ainsi en régressant $u_i^2$ sur les variables explicatives on obtient les $\hat{w}_i$, et on dispose donc d'une estimation de $\Omega$.
Appliquons les MCQG sur le modèle de prix :

```
*Exemple 2 :MCQG ou FGLS en anglais ;
*etape 1 ;
proc reg data=tpm1.hprice ;
model price=lotsize sqrft bdrms ;
output out=f r=res ;
quit ;run ;
data a ;set f ;res2=res*res ;run ;
proc reg data=a ;
model res2=lotsize sqrft bdrms ;
output out=sortie2 p=omega ;
quit ;run ;
```

A la fin de cette première étape nous disposons d'une variance estimée des erreurs $u_i$ qui est la variable "omega" dans le programme précédent. Nous utilisons ensuite cette variable "omega" pour tranformer le modèle et appliquer les MCO sur le modèle transformé et nous nous ramenons au cas où $\Omega$ est connue , ce qui donne :

```
*etape 2 ;
data mt3 ;set sortie2 ;
is=1/sqrt(omega) ;
pricet=price/sqrt(omega) ;
lotsizet=lotsize/sqrt(omega) ;
sqrftt=sqrft/sqrt(omega) ;
bdrmst=bdrms/sqrt(omega) ;run ;
proc reg data=mt3 ;
model pricet=is lotsizet sqrftt bdrmst /noint ;
run ;
```

Remarquons qu'au lieu de transformer le modèle et d'appliquer les MCO sur le modèle transformé on peut aussi directement estimer le modèle par MCP avec comme poids la variable "oméga".

- Inconvénient des MCQG (FGLS en anglais) : il est possible que la variance estimée des erreurs, la variable "oméga", soit négative ; en effet rien n'assure que la variable estimée de la regression auxiliaire soit positive [4]. Nous notons ici que la régression du prix contient 87 observations car l'estimation de la variance est négative pour 1 obs qui est donc éliminée. On peut vérifier ceci avec la condition "if omega<0" dans une étape data par exemple. Ce problème peut être aussi identifié dans le journal de SAS qui a détecté une erreur dans la création de la table "mt3" ; SAS indique "argument invalide pour la fonction sqrt" en précisant le nombre de fois où se produit ce problème ici 1 fois. Si on choisit d'utiliser les MCP , le journal de SAS nous met en garde sur un poids invalide pour 1 observation.
Avec R :

```
auxreg <- lm(residuals(ols)^2~LOTSIZE+SQRFT+BDRMS , data=hprice)
fgls <- lm(PRICE~LOTSIZE+SQRFT+BDRMS,
weights= 1/fitted(auxreg)[fitted(auxreg)>0],
data=hprice[fitted(auxreg)>0, ])
summary(fgls)
```

En pratique le cas où $\Omega$ est connue se produit au moins dans les deux situations suivantes :
- quand on utilise des données temporelles par exemple quand le terme d'erreur est AR(1) [5].

4. pour s'assurer que la variable "oméga" soit toujours positive on peut choisir de spécifier la regression auxiliaire de la manière suivant :$Var(u_i) = \sigma^2 exp(\theta_0 + \theta_1 Lotsize_i + \theta_2 Sqrft + \theta_3 Bdrms)$. voir par exemple "Applied Econometrics with R" p77

5. Un type d'erreur AR(1) fait partie du vocabulaire de l'économétrie des séries temporelles : parmi les hypothèses habituelles de base, il y a l'hypothèse que les erreurs ne sont pas corrélées entre elles (ou la covariance entre deux erreurs est nulle). Dans les séries temporelles cela est très souvent non vérifiée pour des données économiques c'est à dire qu'il existe un lien entre les erreurs ;par exemple pour les modèles AR(1) l'erreur à un moment donné dépend de l'erreur

- le cas de données groupées : Supposons pour simplifier que nous nous intéressions à la relation entre le profit, noté $\Pi$ et la dépense en Recherche-Développement, notée $RD$. Si nous disposions de données individuelles d'entreprise le modèle serait : $\Pi_i = \alpha + \beta RD_i + u_i$ pour $i = 1, \ldots, N$.
Le problème est le suivant : on ne dispose pas de données individuelles d'entreprise , peut être pour des raisons de confidentialité, mais seulement de moyennes par groupe d'entreprises ; c'est ce que l'on appelle des données groupées.
Supposons que les deux premières entreprises aient été groupées : on observe seulement la moyenne de la variable $RD$ et celle du profit pour ce groupe de deux entreprises. On obtient donc $\frac{\Pi_1+\Pi_2}{2} = \alpha + \beta\frac{RD_1+RD_2}{2} + \frac{u_1+u_2}{2}$. Notons le modèle sur des données groupées de la manière suivante :$\Pi_g = \alpha + \beta RD_g + v_g$ où $v_g$ est le terme d'erreur du groupe g avec $g = 1, \ldots, G$. Sur notre exemple $v_1$ représente le terme d'erreur du groupe 1 c'est à dire la moyenne des erreurs des deux premiers individus. On calcule facilemnt $Var(v_g) = \frac{\sigma_u^2}{n_g}$ où $n_g$ est l'effectif du groupe g. On obtient facilement la matrice ,

$$\Omega = \begin{pmatrix} \frac{1}{n_1} & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & \frac{1}{n_G} \end{pmatrix}$$

Attention la matrice $\Omega$ est de taille (G,G).
Donc pour transformer le modèle et obtenir des erreurs homoscédastiques il suffit de multiplier l'équation de profit par $\sqrt{n_g}$.

# VI Conclusion du Chapitre 2

Je voudrais d'abord insister sur l'importance de la spécification du modèle. Comme nous le verrons en TP, il est possible que les tests d'hétéroscédasticité , en particulier celui de White, rejette l'hypothèse d'homoscédastcité quand le modèle est mal spécifié. Ce sera le cas sur le fichier hprice que nous avons utilisé dans ce chapitre ; nous en discuterons en TP. Il faut donc régler le problème de la spécification du modèle avant de procéder à des tests d'hétéroscédasticité si on souhaite estimer le modèle par MCG.

L'inconvénient des MCG est qu'elle suppose une spécification de la variance des erreurs. Cette spécification est une hypothèse qui peut s'avérer inexacte. Wooldridge (2006) suggère de procéder à des tests robustes après avoir estimé le modèle par MCG. C'est une première solution. En pratique la solution la plus commune est d'utiliser directement l'inférence robuste du premier paragraphe de ce cours sans utiliser les MCG.

commise à la période précédente. c'est un processus Auto Régressif ( la variable erreur dépend de ses propres valeurs passées ) d'ordre 1 ( une période précédente).

# VII Références

Breusch T. and A. Pagan “A Simple Test for Heteroscedasticity and Random Coefficient Variation" *Econometrica*, 47, 1979.

White H. “A Heteroscedasticity-Consistent Covariance Matrix Estimator and a direct test for Heteroscedasticity" *Econometrica*, 48, 1980.