HETEROSCEDASTICITE

L'hypothèse H_1^3 (les variances sont égales) n'est plus vérifiée.
Les variances des erreurs peuvent être constantes sur tout l'échantillon, ($Var(\epsilon_t)$=σ^2constante) on dira que les erreurs sont homoscédastiques ou bien varier ($Var(\epsilon_i) \neq Var(\epsilon_j)$) il y a alors hétéroscédasticité des erreurs. Supposons que les erreurs ne soient pas autocorrélées, dans ce cas la matrice de variance-covariance des erreurs est une matrice diagonale avec les variances de chaque erreur sur la diagonale et des 0 ailleurs (non autocorrélation), cette matrice est carrée et symétrique.

$$V_{\overrightarrow{\epsilon}} = \begin{pmatrix} V(\epsilon_1) & 0 & ... & 0 \\ 0 & V(\epsilon_2) & ... & 0 \\ ... & ... & ... & 0 \\ 0 & 0 & .. & V(\epsilon_n) \end{pmatrix}$$

Si les erreurs sont homoscédastiques, toutes ces variances sont égales à une constante notée σ^2 et $V_{\overrightarrow{\epsilon}} = \sigma^2 I$. Dans le cas d'hétéroscédasticité on note cette matrice $V_{\overrightarrow{\epsilon}} = \sigma^2\Omega$ où σ^2 est la moyenne des variances et les coefficients de Ω (n,n) sont les $\frac{V(\epsilon_t)}{\sigma^2}$. On remarque que dans le cas d'homoscédasticité on trouve $\Omega = I$.

1 Les MCO sous hétéroscédasticité

Soit le modèle $\overrightarrow{Y} = X\overrightarrow{a} + \overrightarrow{\epsilon}$ avec hétéroscédasticité $V_{\overrightarrow{\epsilon}} = \sigma^2\Omega$. Les hypothèses des MCO H1,H2 et H4 ne sont pas modifiées, seule l'hypothèse H3 l'est. Quelles sont sous ces hypothèses les propriétés des MCO ?

- Espérance : on constate que $\overrightarrow{\widehat{a}} = ({}^tXX)^{-1}\,{}^tX\overrightarrow{Y} = ({}^tXX)\,{}^tX(X\overrightarrow{a} + \overrightarrow{\epsilon}) = \overrightarrow{a} + ({}^tXX)^{-1}\,{}^tX\overrightarrow{\epsilon}$ or si on fait les deux hypothèses de base H1 : les variables explicatives ne sont pas aléatoires et H2: L'espérance des erreurs est nulle on obtient

  $$E(\overrightarrow{\widehat{a}}) = \overrightarrow{a} + E(\overrightarrow{a} + ({}^tXX)^{-1t}X\overrightarrow{\epsilon}) = \overrightarrow{a} + (\overrightarrow{a} + ({}^tXX)^{-1t}XE(\overrightarrow{\epsilon}) = \overrightarrow{a} + \overrightarrow{0} = \overrightarrow{a}$$

  L'estimateur des MCO est donc toujours sans biais, on constate que seules les hypothèses H1 et H2 sont utilisées et non l'hypothèse H3 sur les variances

- Matrice de variances-covariances de l'estimateur des MCO : on a constaté dans le calcul de l'espérance que $\overrightarrow{\widehat{a}} - \overrightarrow{a} = ({}^tXX)^{-1}\,{}^tX\overrightarrow{\epsilon}$

  $$V_{\widehat{a}} = E((\overrightarrow{\widehat{a}} - \overrightarrow{a})\,{}^t(\overrightarrow{\widehat{a}} - \overrightarrow{a})) = E(({}^tXX)^{-1}\,{}^tX\overrightarrow{\epsilon}\,{}^t\overrightarrow{\epsilon}X({}^tXX)^{-1})$$

  or la matrice X des variables explicatives est supposée être non aléatoire, on peut donc sortir de l'espérance tout ce qui n'est pas aléatoire

  $$V_{\widehat{a}} = ({}^tXX)^{-1}\,{}^tXE(\overrightarrow{\epsilon}\,{}^t\overrightarrow{\epsilon})X({}^tXX)^{-1}$$

on constate que dans le cas d'hétéroscédasticité $V_{\overrightarrow{\epsilon}} = E(\overrightarrow{\epsilon}\ {}^t\overrightarrow{\epsilon}) = \sigma^2\Omega$ donc

$$V_{\widehat{a}} = \sigma^2({}^tXX)^{-1}\ {}^tX\Omega X({}^tXX)^{-1}$$

Cette matrice est différente du cas d'homoscédasticité indice (0) où $V^0_{\widehat{a}} = \sigma^2({}^tXX)^{-1}$. On montre que la matrice $V_{\widehat{a}} - V^0_{\widehat{a}}$ est une matrice définie >0, donc que les variances sous hétéroscédasticité sont plus grandes que sous homoscédasticité rendant les estimateurs moins bons.

Il existe dans la littérature économétrique de nombreux tests d'hétéroscédasticité dépendant chacun de la forme que l'on donne à l'hypothèse H_1. En effet, si l'hypothèse H_0 est toujours l'absence d'hétéroscédasticité, l'hypothèse H_1 peut être "il y a hétéroscédasticité" ou bien "l'hétéroscédasticité prend une forme précise" et dans ce dernier cas il est évident que pour chaque forme d'hétéroscédasticité il y a un test différent. En pratique trois tests sont les plus utilisés, le test de Goldfeld et Quandt, le test de Breusch-Pagan et le test de White.
Ces trois tests sont basés sur le fait que l'estimateur des MCO reste convergent même dans le cas d'hétéroscédasticité des erreurs. On peut penser alors que les résidus approcheront assez correctement les erreurs et qu'ainsi une utilisation des résidus pourra détecter correctement une éventuelle hétéroscédasticité. Si le test de Goldfeld et Quandt nécessité la normalité des erreurs, les deux derniers tests sont plus généraux et donc plus utilisés.

2 Le test de Goldfeld et Quandt

2.1 Théorie

Ce test est basé sur l'hypothèse H_1: les variances des erreurs sont des fonctions monotones d'une variable X_i, en général cette variable est l'une des variables explicatives du modèle. Le test prend donc la forme

$$\begin{aligned} H_0 &: V(\epsilon_t) = \sigma^2 \text{ constante} \\ H_1 &: V(\epsilon_t) = f(X_t) \text{ avec } f \text{ monotone} \end{aligned}$$

Ce test est basé sur l'hypothèse: les erreurs suivent une loi Normale.

Si le test déduit H_0, on en conclut qu'il y a homoscédasticité ou que l'hétéroscédasticité est d'une autre forme. Si le test déduit H_1, on en conclut que la variance est une fonction de X_i. Dans les deux cas il faut poursuivre le test sur les autres variables du modèle car bien sûr plusieurs variables peuvent être responsables de l'hétéroscédasticité.
Pour avoir une fonction monotone (car comme on va le voir tout repose sur cette construction), il faut ordonner les indices t en fonction monotone (dans la programmation on prendra croissante). On classe donc X_{it} en croissant et bien sur les autre variables avec le même indice que X_i.Par exemple:
Données de base

t=1 :Y_1 =200 $X_1 = 10$ $Z_1 = 20$
t=2 : Y_2 =250 $X_2 = 8$ $Z_2 = 17$
t=3 : Y_3 =170 $X_3 = 12$ $Z_3 = 12$
Données classées en fonction de X
t=1 : Y_1 =250 $X_1 = 8$ $Z_1 = 17$
t=2 :Y_2 =200 $X_2 = 10$ $Z_2 = 20$
t=3 : Y_3 =170 $X_3 = 12$ $Z_3 = 12$
La variable X_i étant croissante, alors si H_1 est la bonne hypothèse les variances des erreurs vont aussi être croissantes. Goldfeld et Quandt proposent de partager l'échantillon en trois parties en théorie égales.

- échantillon des N1=N/3 premières valeurs : sur cet échantillon, on effectue la méthode des MCO et on calcule S_1^2 la variance estimée des erreurs, sous l'hypothèse H_0 vraie les variances des erreurs sont constantes. La somme des carrés des résidus sur σ^2 suit alors la loi du χ^2 a N1-k degrés de liberté. $SCR_1/\sigma^2 = \chi^2_{(N_1-k)}$

- échantillon des N3=N/3 données suivantes : cet échantillon n'est pas pris en compte afin de permettre l'indépendance des erreurs des deux autres sous-échantillons.

- échantillon des N2=N/3 dernières valeurs : on effectue les MCO et on calcule S_2^2 la variance estimée des erreurs, sous l'hypothèse H_0 vraie les variances des erreurs sont constantes. La somme des carrés des résidus sur σ^2 suit alors la loi du χ^2 a N2-k degrés de liberté. $SCR_2/\sigma^2 = \chi^2_{(N_2-k)}$

- Pour effectuer le test, on prend le rapport des deux variances . En effet, le rapport de deux χ^2 indépendants (cette indépendance étant garantie par le tiers des données non utilisées au milieu de l'échantillon) divisés par leur degré de liberté suit une loi de Fisher.

$$\frac{\frac{SCR_2}{\sigma^2}/(N_2-k)}{\frac{SCR_1}{\sigma^2}/(N_1-k)} = F_{(N2-1,N1-1)} = \frac{S_2^2}{S_1^2}$$

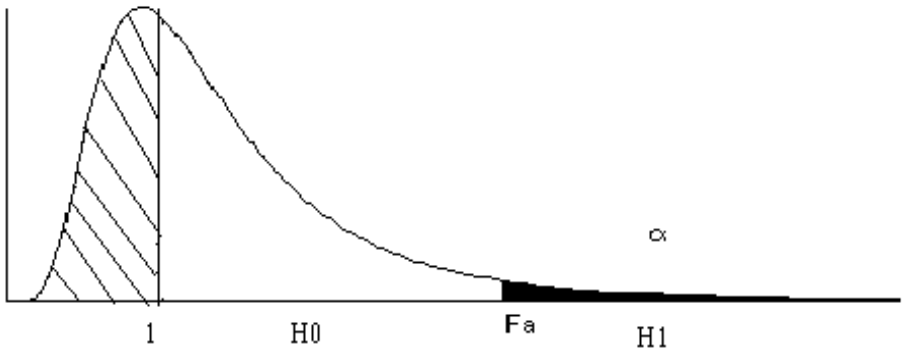
  En simplifiant par σ^2 commun dans les deux sous-échantillons sous l'hypothèse H_0, on montre que

$$\frac{SCR_2/(N_2-k)}{SCR_1/(N_1-k)} = F_{(N2-k,N1-k)} = \frac{S_2^2}{S_1^2}$$

  Plusieurs résultats possibles: Ce rapport peut être supérieur ou inférieur à 1

  Si $\frac{S_2^2}{S_1^2}$ est proche de 1 (qu'il soit >1 ou <1) on décidera H_0 , s'il est trop éloigné de 1 (soit nettement supérieur à 1 soit nettement inférieur à 1) on décidera H_1. Ce test est donc bilatéral. Afin de faciliter la présentation, on a coutume de transformer ce test bilatéral en test unilatéral en mettent systématiquement au numérateur la variance empirique la plus grande.

  IMPORTANT: on mets au numérateur l'indice 1 ou 2 qui correspond à la plus grande variance calculée , supposons ici que ce soit la variance correspondant au dernier



échantillon

$\frac{S_2^2}{S_1^2}$ = F suit un Fisher à $N2-k, N1-k$ degrés de liberté si la valeur calculée $s_2^2 > s_1^2$

$\frac{S_1^2}{S_2^2}$ = F suit un Fisher à $N1-k, N2-k$ degrés de liberté si la valeur calculée $s_1^2 > s_2^2$

Ce rapport F est donc toujours supérieur à 1 ; si ce rapport est très nettement supérieur à 1 alors les variances des deux sous-échantillons sont admises comme différentes et on décide H_1.

EN PRATIQUE : si $s_2^2 > s_1^2$ on calcule le rapport $f = s_2^2/s_1^2$ qui est >1 et si ce rapport est supérieur à la borne Fisher(N2-k,N1-k) on décide H_1 sinon on décide H_0

2.2 Pratique

2.2.1 La procédure GOLDFELD_QUANDT.SRC

Il faut indiquer la variable que l'on prend pour responsable de l'hétéroscédasticité. En général on refait ce test pour toutes les variables explicatives. Si l'on souhaite que cette variable soit une variable retardée X{i} il faut d'abord définir ce retard comme une variable (set XI=X{i}) et utiliser XI. Les deux sous-échantillons sont pris de taille égale N1=N2 afin d'avoir des variances estimées comparables. Sans l'option de choix de N3, celui-ci est pris (partie entière de N/3).
Les exemples ci-dessous seront retrouvés dans le programme CMUS.PRG de la consommation des ménages aux USA , dans la partie ' étude de l'hétéroscédasticité ' .

Sans option La procédure sans option est utilisée pour un modèle sans retards sur les variables explicatives . De plus elle prend comme valeur de N3 le tiers des données ou du moins la partie entière de N3 et en répartissant en deux échantillons égaux N1 et N2.
source goldfeld_quandt.src
@goldfeld_quandt explic start end X
liste des variables explicatives

* X est la variable correspondant à H_1, les données seront classées en fonction croissante de X
Exemple : on teste si la variable RD est responsable d'une hétéroscédasticité dans le modèle CMUS
CM est la consommation des ménages US , RD le revenu disponible, SP un indice de richesse, TCHO le taux de chômage.

$$CM = a_0 + a_1 RD + a_2 SP + a_3 TCHO + \epsilon$$

*On appelle la procédure
source goldfeld_quandt.src
*On donne le nom de la variable endogène, le début et la fin de l'échantillon et le nom de la variable éventuellement responsable de l'hétéroscédasticité.
@goldfeld_quandt cm start end rd
*On donne la liste des explicatives précédées de #
 # constant rd sp tcho

```
*****************************************************************
                  TEST DE GOLDFELD et QUANDT
variable eventuellement responsable de l heteroscedasticite  RD
*****************************************************************
Linear Regression - Estimation by Least Squares
Dependent Variable CMGQ
Quarterly Data From 1955:01 To 1970:04
Usable Observations     64      Degrees of Freedom    60
Centered R**2     0.998451      R Bar **2   0.998374
Uncentered R**2   0.999945      T x R**2      63.996
Mean of Dependent Variable      1736.1953125
Std Error of Dependent Variable  336.7655293
Standard Error of Estimate        13.5799837
Sum of Squared Residuals        11064.957433
Regression F(3,60)                12894.4543
Significance Level of F           0.00000000
Log Likelihood                    -255.69704
Durbin-Watson Statistic             0.817819

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  constant                 92.96741591  16.62071136      5.59347  0.00000058
2.  RDGQ                      0.83818950   0.01167609     71.78682  0.00000000
3.  SPGQ                      0.46872795   0.22682206      2.06650  0.04310537
4.  TCHOGQ                   -3.17718882   1.83873518     -1.72792  0.08914654

 Linear Regression - Estimation by Least Squares
Dependent Variable CMGQ
Quarterly Data From 1987:01 To 2002:04
Usable Observations     64      Degrees of Freedom    60
Centered R**2     0.996360      R Bar **2   0.996178
Uncentered R**2   0.999920      T x R**2      63.995
Mean of Dependent Variable      5160.0375000
Std Error of Dependent Variable  781.2105759
VStandard Error of Estimate        48.2938590
Sum of Squared Residuals        139937.80899
Regression F(3,60)                 5475.0507
Significance Level of F           0.00000000
Log Likelihood                    -336.89432
Durbin-Watson Statistic             1.612418

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  constant               -349.9651712   95.7444592     -3.65520  0.00054237
2.  RDGQ                      0.9875425    0.0213271     46.30453  0.00000000
```

```
3.  SPGQ                      0.1024657    0.0484439      2.11514  0.03857949
4.  TCHOGQ                  -22.5660547    9.6596325     -2.33612  0.02284413

F(60,60)=     12.64694 with Significance Level 0.00000000
```

On remarque que les variables portent leur nom suivi de GQ, ceci afin de ne pas changer l'ordre dans les variables de base. Ces variables sont donc les variables de bases mais avec un classement correspondant à l'ordre de la variable RD .
Dans cet exemple on a partagé l'échantillon en trois parties égales, on donne le résultat des MCO pour les parties 1 et 3. La valeur calculée du rapport des variances estimées suit sous l'hypothèse de non hétéroscédasticité une loi de Fisher (60,60), la valeur calculée est 12.65 nettement supérieure à la borne du Fisher car le niveau de significativité ns=prob(F>12.65)=0.0000, on en déduit donc que la variable RD est responsable d'une hétéroscédasticité.

avec option Trois options sont proposées:

- **OPTION LAGS**: Si dans la liste des variables explicatives figurent des variables retardées indiquées par z{i} il est impératif de mettre l'option lags. Le nom des variables ne pouvant pas être récupéré (problème de RATS), le numéro des variables est celui de la liste des variables. Ici VARGQ2 est RD , VARGQ3 est RD{1}....

$$CM_t = a_0 + a_1 RD_t + a_2 RD_{t-1} + a_3 SP + a_4 TCHO + \epsilon$$

source goldfeld_quandt.src
@goldfeld_quandt(lags) cm start end rd
 # constant rd{0 to 1} sp tcho

```
*****************************************************************
                  TEST DE GOLDFELD et QUANDT
variable eventuellement responsable de l heteroscedasticite  RD
*****************************************************************
Linear Regression - Estimation by Least Squares
Dependent Variable CMGQ
Quarterly Data From 1955:01 To 1970:04
Usable Observations     64      Degrees of Freedom    59
Centered R**2     0.998643      R Bar **2   0.998551
Uncentered R**2   0.999952      T x R**2      63.997
Mean of Dependent Variable      1736.1953125
Std Error of Dependent Variable  336.7655293
Standard Error of Estimate        12.8209998
Sum of Squared Residuals        9698.3040852
Regression F(4,59)                10851.8088
Significance Level of F           0.00000000
Log Likelihood                    -251.47841
Durbin-Watson Statistic             0.727446

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  constant                 94.91380485  15.70629396      6.04304  0.00000011
2.  vargq2                    0.49873793   0.11824048      4.21800  0.00008600
3.  vargq3                    0.33760973   0.11708673      2.88342  0.00548190
4.  vargq4                    0.59109085   0.21830932      2.70758  0.00885249
5.  vargq5                   -3.21211754   1.73601072     -1.85029  0.06928135

Linear Regression - Estimation by Least Squares
```

```
Dependent Variable CMGQ
Quarterly Data From 1987:01 To 2002:04
Usable Observations     64      Degrees of Freedom    59
Centered R**2     0.997207      R Bar **2   0.997017
Uncentered R**2   0.999938      T x R**2      63.996
Mean of Dependent Variable      5160.0375000
Std Error of Dependent Variable  781.2105759
Standard Error of Estimate        42.6650123
Sum of Squared Residuals        107397.89305
Regression F(4,59)                 5265.7256
Significance Level of F            0.00000000
Log Likelihood                     -328.42527
Durbin-Watson Statistic              1.139523

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  constant                -362.3808038   84.6360088     -4.28164  0.00006921
2.  vargq2                     0.5553388    0.1039457      5.34259  0.00000155
3.  vargq3                     0.4416982    0.1044694      4.22801  0.00008312
4.  vargq4                     0.0944007    0.0428401      2.20356  0.03147111
5.  vargq5                   -25.6936655    8.5657642     -2.99958  0.00395531

F(59,59)=     11.07388 with Significance Level 0.00000000
```

On constate toujours hétéroscédasticité.

- **PARTIE DU MILIEU NON EGALE A n/3 MAIS A NGQ A FIXER**. Il est possible de prendre pour N3 une autre valeur que N/3, dans ce cas on prend l'option NGQ=la valeur choisie pour nouvelle valeur de N3. (Le programme prenant bien automatiquement N1 et N2 pour qu'ils soient des entiers). Ce cas est fréquent dans les petits échantillons où on laisse au milieu moins du tiers des données ce qui permet d'augmenter N1=N2. Dans le modèle :

$$CM = a_0 + a_1 RD + a_2 SP + a_3 TCHO + \epsilon$$

source goldfeld_quandt.src
@goldfeld_quandt(ngq=50) cm start end tcho
 # constant rd sp tcho
Ici on teste la variable TCHO comme éventuelle responsable de l'hétéroscédasticité.

```
*****************************************************************
                  TEST DE GOLDFELD et QUANDT
variable eventuellement responsable de l heteroscedasticite  TCHO
*****************************************************************
 Linear Regression - Estimation by Least Squares
Dependent Variable CMGQ
Quarterly Data From 1955:01 To 1972:03
Usable Observations     71      Degrees of Freedom    67
Centered R**2     0.999680      R Bar **2   0.999665
Uncentered R**2   0.999919      T x R**2      70.994
Mean of Dependent Variable      3137.9690141
Std Error of Dependent Variable 1842.5124863
Standard Error of Estimate        33.7136501
Sum of Squared Residuals        76152.883662
Regression F(3,67)                69670.1779
Significance Level of F           0.00000000
Log Likelihood                    -348.45719
Durbin-Watson Statistic             1.740199
    Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  constant                  36.98636448  31.78603052      1.16360  0.24871157
```

```
2.  RDGQ                      0.86104493   0.00708634    121.50770  0.00000000
3.  SPGQ                      0.35759005   0.02723570     13.12946  0.00000000
4.  TCHOGQ                   -0.10681633   7.70645348     -0.01386  0.98898236

Linear Regression - Estimation by Least Squares
Dependent Variable CMGQ
Quarterly Data From 1985:02 To 2002:04
Usable Observations     71      Degrees of Freedom    67
Centered R**2     0.998477      R Bar **2   0.998409
Uncentered R**2   0.999882      T x R**2      70.992
Mean of Dependent Variable      3376.8154930
Std Error of Dependent Variable  985.7809885
Standard Error of Estimate        39.3224448
Sum of Squared Residuals        103599.06256
Regression F(3,67)                14641.8080
Significance Level of F           0.00000000
Log Likelihood                    -359.38357
Durbin-Watson Statistic             1.922328
    Variable                    Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  constant                  147.3268087   37.7389637      3.90384  0.00022246
2.  RDGQ                        0.8381808    0.0092466     90.64712  0.00000000
3.  SPGQ                        0.5668268    0.0777449      7.29085  0.00000000
4.  TCHOGQ                    -10.8258449    5.0284958     -2.15290  0.03493086

     F(67,67)=      1.36041 with Significance Level 0.10519022
********************************************************************************
```

On constate dans cet exemple que TCHO n'est pas responsable de l'hétéroscédasticité..

- **OPTION NOPRINT**. On peut choisir de ne pas sortir les résultats des deux régressions : option NOPRINT

$$CM = a_0 + a_1 RD + a_2 SP + a_3 TCHO + \epsilon$$

source goldfeld_quandt.src
@goldfeld_quandt(noprint) cm start end rd
 # constant rd sp tcho

```
****************************************************************
                TEST DE GOLDFELD et QUANDT
variable eventuellement responsable de l heteroscedasticite  RD
****************************************************************
F(60,60)=     12.64694 with Significance Level 0.00000000

****************************************************************
```

@goldfeld_quandt(noprint) cm start end tcho
 # constant rd sp tcho

```
****************************************************************
                TEST DE GOLDFELD et QUANDT
variable eventuellement responsable de l heteroscedasticite  TCHO
****************************************************************
F(60,60)=      1.54080 with Significance Level 0.04836289

****************************************************************
```

@goldfeld_quandt(noprint) cm start end sp
 # constant rd sp tcho

```
*****************************************************************
                  TEST DE GOLDFELD et QUANDT
variable eventuellement responsable de l heteroscedasticite  SP
*****************************************************************
F(60,60)=      9.53446 with Significance Level 0.00000000

*****************************************************************
```

On constate que SP et RD sont responsables de l'hétéroscédasticité.
Il est bien sur possible de mettre plusieurs options en les séparant par des virgules

@goldfeld_quandt(noprint,lags) cm start end sp
 # constant rd sp tcho cm{1}

```
*****************************************************************
                  TEST DE GOLDFELD et QUANDT
variable eventuellement responsable de l heteroscedasticite  SP
*****************************************************************
F(59,59)=      3.76341 with Significance Level 0.00000048

*****************************************************************
```

Conclusion Sur l'exemple $CM = a_0 + a_1 RD + a_2 SP + a_3 TCHO + \epsilon$ on constate que pour ce modèle la variable TCHO n'est pas responsable de l'hétéroscédasticité (ns=.0855), ce qui ne veut pas dire qu'il n'y a pas d'hétéroscédasticité, mais simplement que s'il y a hétéroscédasticité, celle-ci n'est pas due à TCHO. Par contre RD (ns=0.00) et SP ns=0.00) sont responsables. Il y a donc hétéroscédasticité.
ATTENTION : POUVONS-NOUS VRAIMENT ICI UTILISER CE TEST ? SOMMES-NOUS DANS LE CAS DE NORMALITE DES ERREURS ? (voir le résultat dans CMUS.PRG). IL FAUT TOUJOURS COMMENCER PAR VERIFIER LA NORMALITE. Nous ne sommes pas ici dans le cas de normalité des erreurs et le test de Fisher n'est ici qu'une approximation. On lui préférera alors d'autres tests d'hétéroscédasticité que l'on verra ci-dessous.

2.2.2 exemple sur le modèle Y1 en fonction de X1 et X2

Cet exemple porte sur le modèle

$$Y1 = a_0 + a_1 X1 + a_2 X2 + \epsilon$$

Les données sont dans hetero1.rat et le programme dans hetero1.prg.
Contrairement au cas précédent nous allons commencer par bien vérifier que l'on est dans le cas d'utilisation de ce test c'est-à-dire qu'il y a normalité des erreurs. Pour cela on utilise les tests de normalité (voir le chapitre correspondant pour les détails).
cal 1952 1 4

all 1986:4
open data hetero1.rat
data(for=rats) / Y1 X1 X2
***** aller dans Wizards+show series windows pour voir la liste des variables**
com start = 1952:1
com end = 1986:4
smpl start end
lin Y1 / res
 # constant X1 X2
stat res

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y1
Usable Observations    140      Degrees of Freedom   137
Centered R**2     0.970857      R Bar **2   0.970432
Uncentered R**2   0.997557      T x R**2     139.658
Mean of Dependent Variable      702505.89295
Std Error of Dependent Variable 213255.81299
Standard Error of Estimate       36670.18838
Sum of Squared Residuals         1.84224e+11
Regression F(2,137)                2282.0001
Significance Level of F           0.00000000
Log Likelihood                   -1668.49581
Durbin-Watson Statistic             2.018234

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                 -4570.41278  12901.72696     -0.35425  0.72369758
2.  X1                           0.93619      0.36250      2.58255  0.01085645
3.  X2                           0.50147      0.02601     19.28315  0.00000000

Statistics on Series RES
Observations                  140
Sample Mean             -0.000000      Variance            1325354475.402771
Standard Error       36405.418215      of Sample Mean      3076.819410
t-Statistic (Mean=0)    -0.000000      Signif Level        1.000000
Skewness                -0.141123      Signif Level (Sk=0) 0.500070
Kurtosis (excess)       -0.679835      Signif Level (Ku=0) 0.109344
Jarque-Bera              3.160725      Signif Level (JB=0) 0.205900
```

TEST DE NORMALITE

test de H0 : $\alpha_3 = 0$

L'estimateur $\widehat{\alpha_3}$ suit sous l'hypothèse H0 asymptotiquement une loi Normale, soit $S = \sqrt{\frac{n}{6}}(\widehat{\alpha_3} - 0)$ suit asymptotiquement une loi $N(0,1)$

$\widehat{\alpha_3} = -0,1411$ donc $\sqrt{\frac{n}{6}}(\widehat{\alpha_3} - 0) = -0,68$ valeur comprise entre -1.96 et 1.96 on décide donc H0 $\alpha_3 = 0$ loi symétrique. On peut dire aussi que la P VALUE=0,5 est très supérieure à 0,05 on décide alors H0: la loi est symétrique.

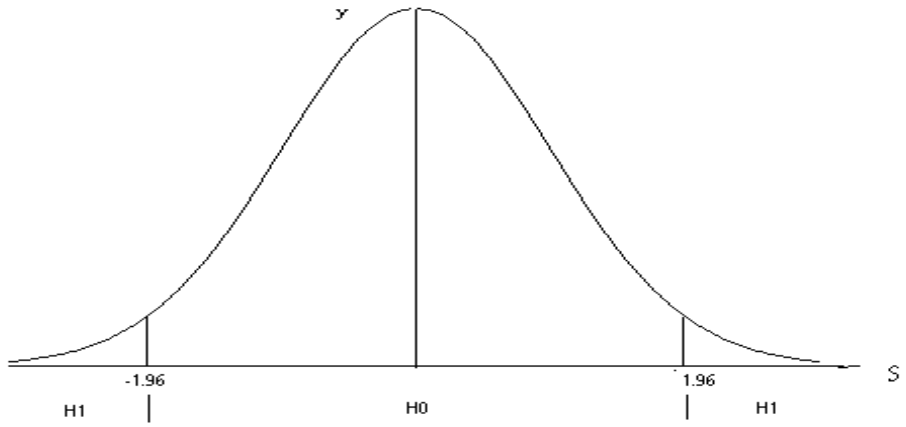
test de H0: $\alpha_4 = 3$ ou $\alpha_4 - 3 = 0$

L'estimateur $\widehat{\alpha_4}$ suit sous l'hypothèse H0 asymptotiquement une loi Normale soit $S = \sqrt{\frac{n}{24}}(\widehat{\alpha_4} - 3)$ suit asymptotiquement une loi $N(0,1)$

Ici $(\widehat{\alpha_4} - 3)$ =-0,6798 et $\sqrt{\frac{n}{24}}(\widehat{\alpha_4} - 3)$ =-1,64 valeur comprise entre -1.96 et 1.96 on décide donc H0 $\alpha_4 = 3$ loi de coefficient d'aplatissement égal à 3. On peut dire aussi directement que la P VALUE=0,109 étant supérieure à 0,05 on décide donc H0.

Les deux tests donnant H0 on en déduit la normalité des erreurs. Le test de Goldfeld et Quandt peut donc être utilisé.

source goldfeld_quandt.src



@goldfeld_quandt Y1 start end X1
 # constant X1 X2
Comme il n'y a aucune option, on laisse au milieu le tiers des valeurs non utilisées. L'ordre en fonction de X1 est fait automatiquement dans le sous-programme.

```
****************************************************************
                 TEST DE GOLDFELD et QUANDT
variable eventuellement responsable de l heteroscedasticite  X1
****************************************************************
Linear Regression - Estimation by Least Squares
Dependent Variable Y1GQ
Usable Observations     47      Degrees of Freedom    44
Centered R**2     0.842087      R Bar **2   0.834909
Uncentered R**2   0.997269      T x R**2      46.872
Mean of Dependent Variable      463063.74165
Std Error of Dependent Variable  62093.38074
Standard Error of Estimate       25229.36852
Sum of Squared Residuals         28006925572
Regression F(2,44)                  117.3175
Significance Level of F           0.00000000
Log Likelihood                    -541.52101
Durbin-Watson Statistic             2.581698

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  constant                -41295.01904  37118.51245     -1.11252  0.27195693
2.  X1GQ                         1.54107      1.14999      1.34006  0.18710381
3.  X2GQ                         0.49276      0.08799      5.60045  0.00000130

Linear Regression - Estimation by Least Squares
Dependent Variable Y1GQ
Usable Observations     47      Degrees of Freedom    44
Centered R**2     0.862351      R Bar **2   0.856094
Uncentered R**2   0.997677      T x R**2      46.891
Mean of Dependent Variable      924790.75277
Std Error of Dependent Variable 122477.03300
Standard Error of Estimate       46461.63482
Sum of Squared Residuals         94982074460
Regression F(2,44)                  137.8263
Significance Level of F           0.00000000
Log Likelihood                    -570.22006
Durbin-Watson Statistic             1.868573

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  constant                -54745.77289  68516.38458     -0.79902  0.42857262
```

```
2.  X1GQ                        0.42384      0.79995      0.52983  0.59889580
3.  X2GQ                        0.57308      0.05619     10.19826  0.00000000

F(44,44)=      3.39138 with Significance Level 0.00004568

*****************************************************************
```

Le Fisher théorique avec $\alpha = 5\%$ étant d'environ 1,65 on constate que 3,39 est nettement supérieur on décidera donc H1 il y a de l'hétéroscédasticité due à la variable X1. On peut également constater que la P Value = 0,000045 est nettement inférieure à 0,05 donc on décide H1.

@goldfeld_quandt Y1 start end X2
 # constant X1 X2

```
*****************************************************************
                  TEST DE GOLDFELD et QUANDT
variable eventuellement responsable de l heteroscedasticite  X2
*****************************************************************
Linear Regression - Estimation by Least Squares
Dependent Variable Y1GQ
Usable Observations     47      Degrees of Freedom    44
Centered R**2     0.849708      R Bar **2   0.842877
Uncentered R**2   0.997277      T x R**2      46.872
Mean of Dependent Variable      464129.34602
Std Error of Dependent Variable  63730.25434
Standard Error of Estimate       25261.88222
Sum of Squared Residuals         28079158500
Regression F(2,44)                  124.3820
Significance Level of F            0.00000000
Log Likelihood                     -541.58154
Durbin-Watson Statistic               1.970607

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  constant                -63233.10732  37019.69617     -1.70809  0.09466608
2.  X1GQ                         1.75167      1.14141      1.53466  0.13202750
3.  X2GQ                         0.50312      0.08866      5.67454  0.00000101

Linear Regression - Estimation by Least Squares
Dependent Variable Y1GQ
Usable Observations     47      Degrees of Freedom    44
Centered R**2     0.798436      R Bar **2   0.789274
Uncentered R**2   0.997851      T x R**2      46.899
Mean of Dependent Variable      942007.86622
Std Error of Dependent Variable  98854.17388
Standard Error of Estimate       45378.94082
Sum of Squared Residuals         90606923887
Regression F(2,44)                  87.1463
Significance Level of F            0.00000000
Log Likelihood                     -569.11186
Durbin-Watson Statistic               2.138877

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  constant                -139314.3672   84382.7861     -1.65098  0.10586334
2.  X1GQ                         0.0800       0.6231      0.12843  0.89839623
3.  X2GQ                         0.6494       0.0780      8.32038  0.00000000

F(44,44)=      3.22684 with Significance Level 0.00008450

*****************************************************************
```

Pour les mêmes raisons la variable X2 est également responsable d'une hétéroscédasticité.

3 Le test de BREUSCH-PAGAN

3.1 Théorie

On trouve plusieurs présentations de ce test. Il est basé sur la connaissance de la forme de l'hétéroscédasticité comme le test de Goldfeld et Quandt mais la différence entre eux est que a normalité des erreurs n'est pas nécessaire et que l'on peut tester plusieurs variables. Il faut cependant que la taille de l'échantillon ne soit pas trop petite car c'est un test asymptotique. La variance des erreurs est une fonction non linéaire de r variables Z_i en théorie indépendantes sous l'hypothèse H_1 soit $V(\epsilon_t) = \sigma^2 h(\alpha_0 + \sum \alpha_i Z_{it})$

$$H_0 \ : \ V(\epsilon_t) = \sigma^2 h(\alpha_0) = cte \implies \alpha_0 \neq 0 \text{ et } \alpha_1 = ... = \alpha_r = 0$$
$$H_1 \ : \ V(\epsilon_t) = \sigma^2 h(\alpha_0 + \sum_1^r \alpha_i Z_{it}) \implies \alpha_0 \neq 0 \text{ et l'un au moins des } \alpha_i \neq 0$$

On montre comme $V(\epsilon_t) = E(\epsilon_t^2)$ est approximée par $\epsilon_t^2 = \sigma^2 h(\alpha_0 + \sum_1^r \alpha_i Z_{it}) + terme\ erreur$. De plus si on fait un développement limité, si n est grand on peut finalement faire l'approximation

$$\epsilon_t^2 = \sigma^2(\alpha_0 + \sum_1^r \alpha_i Z_{it}) + terme\ erreur$$

et en remplaçant les erreurs ϵ_t par les résidus e_t:

$$e_t^2 = \sigma^2(\alpha_0 + \sum_1^r \alpha_i Z_{it}) + terme\ erreur$$

Dans ce dernier modèle l'hypothèse H_0 revient à tester la nullité des coefficients des variables Z_i pour que la variance des erreurs approximée par les résidus au carré soit une constante. On teste ainsi $\alpha_1 = ... = \alpha_r = 0$. Dans ce cas il est impossible d'effectuer le test de Fisher de nullité des coefficients car même si les erreurs suivent une loi normale, leur carré ne suit pas une loi normale, donc le terme erreur ne suit pas non plus une loi normale, le Fisher n'est donc pas utilisable. On utilise alors le test de Wald sur le modèle précédent divisé par s^2 l'estimation de σ^2. Le modèle s'écrit alors, en notant u_t le terme erreur du modèle:

$$\frac{e_t^2}{s^2} = \alpha_0 + \sum_1^r \alpha_i Z_{it} + u_t$$

Sous l'hypothèse H_0 : tous les coefficients nuls sauf la constante (car si la constante était nulle, la variance des erreurs serait également nulle ce qui est impossible, ou trop beau car alors les erreurs seraient nulles), le test de Wald montre que le nR^2 suit asymptotiquement une loi du χ^2 à r degrés de liberté (autant de degrés de liberté que de variables dans cette équation autres que la constante.

3.2 Procédure : breuschpagan.src

3.2.1 Sur le modèle Y1 en fonction de X1 et X2

On effectue la régression de base par exemple $Y1 = a_0 + a_1 X_1 + a_2 X_2 + \epsilon$ et on récupère les résidus e_t de ce modèle une fois estimé par les MCO.

Les données sont les mêmes que dans l'exemple précédent et la programmation est également dans heteros1.prg

lin Y1 start end res

 # constant X1 X2

On récupère les résidus du modèle dans le vecteur RES

Si l'on souhaite tester la forme particulière d'hétéroscédasticité $V(\epsilon_t) = \sigma^2 h(\alpha_0 + \alpha_1 X_{1t})$, on appelle la procédure breuschpagan.src

source breuschpagan.src

puis donne dans la procédure le nom du vecteur des résidus (ici RES) ainsi que le début et la fin de l'échantillon

@breuschpagan res start end

enfin on donne la ou les variables Z_i qui peuvent être responsables de l'hétéroscédasticité SANS METTRE LA CONSTANTE qui sera ajoutée automatiquement (ici la seule variable X1 dans l'hypothèse H_1 ci-dessus)

 # X1

```
Test de Breusch Pagan sur  RES
 Linear Regression - Estimation by Least Squares
Dependent Variable RES^2/S2
Quarterly Data From 1952:01 To 1986:04
Usable Observations    140      Degrees of Freedom   138
Centered R**2     0.185219      R Bar **2   0.179315
Uncentered R**2   0.539218      T x R**2      75.490
Mean of Dependent Variable      1.0000000000
Std Error of Dependent Variable 1.1449952873
Standard Error of Estimate      1.0372706012
Sum of Squared Residuals        148.47838140
Regression F(1,138)                  31.3706
Significance Level of F           0.00000011
Log Likelihood                    -202.76718
Durbin-Watson Statistic             2.181347

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                     -0.9002       0.3504     -2.56907  0.01125940
2.  X1                         1.8485e-05   3.3004e-06      5.60095  0.00000011

Chi-Squared(1)=     25.930625 with Significance Level 0.00000035
```

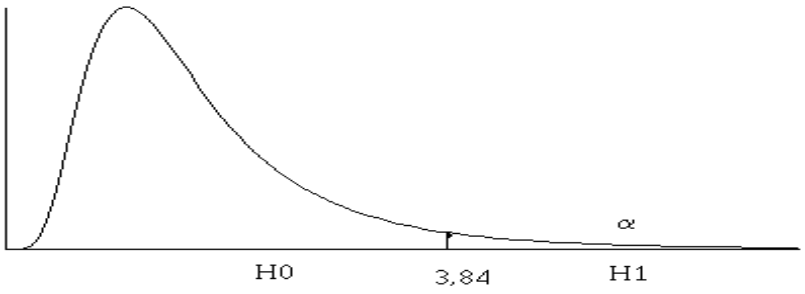
nR2=140x0.185=25,93 taille de l'échantillon par le R2 centré (car le modèle a obligatoirement un terme constant). Sous l'hypothèse H0 la variable X1 n'est pas responsable de l'hétéroscédasticité nR2 suit asymptotiquement une loi du χ^2 à un degré de liberté (sous H0 le seul coefficient de X1 doit être nul d'où un seul degré de liberté). Décision du test: Comme nR2>3.84 on décide H1: les erreurs sont hétéroscédastiques, la variance des erreurs est une fonction de X1. On peut également constater que la P-value=.000000035 est très nettement inférieure à .05 donc décision H1.

Si on veut tester H1: $V(\epsilon_t) = \sigma^2 h(\alpha_0 + \alpha_1 X2_t)$

@breuschpagan res start end

 # X2

```
Test de Breusch Pagan sur  RES
Linear Regression - Estimation by Least Squares
Dependent Variable RES^2/S2
Quarterly Data From 1952:01 To 1986:04
Usable Observations    140      Degrees of Freedom   138
```




```
Centered R**2     0.183735      R Bar **2   0.177820
Uncentered R**2   0.538378      T x R**2      75.373
Mean of Dependent Variable      1.0000000000
Std Error of Dependent Variable 1.1449952873
Standard Error of Estimate      1.0382146632
Sum of Squared Residuals        148.74877679
Regression F(1,138)                  31.0627
Significance Level of F           0.00000013
Log Likelihood                    -202.89454
Durbin-Watson Statistic             2.169671

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                      -0.6088       0.3017     -2.01799  0.04553082
2.  X2                          1.3208e-06   2.3698e-07      5.57339  0.00000013

Chi-Squared(1)=     25.722892 with Significance Level 0.00000039
```

Les conclusions sont les mêmes on décide H1
Si on veut tester H1: $V(\epsilon_t) = \sigma^2 h(\alpha_0 + \alpha_1 X1_t + \alpha_2 X2_t)$

```
  Test de Breusch Pagan sur  RES
 Linear Regression - Estimation by Least Squares
Dependent Variable RES^2/S2
Quarterly Data From 1952:01 To 1986:04
Usable Observations    140      Degrees of Freedom   137
Centered R**2     0.189547      R Bar **2   0.177715
Uncentered R**2   0.541665      T x R**2      75.833
Mean of Dependent Variable      1.0000000000
Std Error of Dependent Variable 1.1449952873
Standard Error of Estimate      1.0382808373
Sum of Squared Residuals        147.68971230
Regression F(2,137)                  16.0206
Significance Level of F           0.00000056
Log Likelihood                    -202.39437
Durbin-Watson Statistic             2.188723

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                      -0.8129       0.3653     -2.22542  0.02768756
2.  X1                          1.0173e-05   1.0264e-05      0.99117  0.32335192
3.  X2                          6.2980e-07   7.3632e-07      0.85533  0.39386309

Chi-Squared(2)=     26.536525 with Significance Level 0.00000173
```

Sous l'hypothèse H0 nR^2 suit ici un χ^2_2
Les conclusions sont bien sur les mêmes. X1 et X2 sont responsables de l'hétéroscédasticité. On remarque ici que les t de Student de X1 et X2 sont très faibles. Cela est du à la

colinéarité entre X1 et X2. On rappelle que regarder la valeur du t de Student n'est pas utilisable dans cette régression car les erreurs au carré donc également les u_t ne suivant pas une loi normale on ne peut utiliser ni le Student ni les Fisher. Le dernier test proposé n'est pas très valable en théorie car X1 et X2 sont corrélées dans cet exemple pour le voir on va estimer l'une des variables en fonction de l'autre:

lin X1 start end
 # constant X2

```
Linear Regression - Estimation by Least Squares
Dependent Variable X1
Quarterly Data From 1952:01 To 1986:04
Usable Observations    140      Degrees of Freedom   138
Centered R**2     0.896405      R Bar **2   0.895654
Uncentered R**2   0.993516      T x R**2     139.092
Mean of Dependent Variable      102797.72143
Std Error of Dependent Variable  26657.66370
Standard Error of Estimate        8611.12312
Sum of Squared Residuals         10232898913
Regression F(1,138)                1194.1078
Significance Level of F            0.00000000
Log Likelihood                    -1466.15759
Durbin-Watson Statistic              0.134979

   Variable                     Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                20063.198473  2502.392354      8.01761  0.00000000
2.  X2                          0.067921     0.001966     34.55586  0.00000000
```

On voit bien la colinéarité entre les deux est forte (coefficient de X2 très significatif et R2 fort).
Si on veut savoir si $X1^2$ ou $X2^2$ joue un rôle dans l'hétéroscédasticité:

set X12 = X12**
@breuschpagan res start end
 # X12

```
  Test de Breusch Pagan sur  RES
Linear Regression - Estimation by Least Squares
Dependent Variable RES^2/S2
Quarterly Data From 1952:01 To 1986:04
Usable Observations    140      Degrees of Freedom   138
Centered R**2     0.191492      R Bar **2   0.185634
Uncentered R**2   0.542765      T x R**2      75.987
Mean of Dependent Variable      1.0000000000
Std Error of Dependent Variable 1.1449952873
Standard Error of Estimate      1.0332695633
Sum of Squared Residuals        147.33514668
Regression F(1,138)                  32.6848
Significance Level of F           0.00000006
Log Likelihood                    -202.22612
Durbin-Watson Statistic             2.199853

   Variable                     Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                 -9.1980e-03       0.1969     -0.04670  0.96281701
2.  X12                       8.9524e-11   1.5659e-11      5.71706  0.00000006

Chi-Squared(1)=     26.808922 with Significance Level 0.00000022
```

nR2= 140x0,19149 = 26,8089 la P-value ou niveau de significativité étant très inférieure à 0,05 on décide H1 , la variable $X1^2$ est responsable de l'hétéroscédasticité.

3.2.2 Sur le modèle de consommation

Pour ce test, comme nous l'avons vu, l'hypothèse de Normalité des erreurs n'est pas nécessaire. C'est pour cela que ce test et le suivant ont un si gros succès au dépend de Goldfeld et Quandt. On reprend donc cmus.prg
lin cm start end residus
 # constant rd sp tcho
Sur ce modèle test H1 : $V(\epsilon_t) = \sigma^2 h(\alpha_0 + \alpha_1 RD)$
source breuschpagan.src
@breuschpagan residus start end
 # rd

```
   Test de Breusch Pagan sur  RESIDUS
Linear Regression - Estimation by Least Squares
Dependent Variable RESIDUS^2/S2
Quarterly Data From 1955:01 To 2002:04
Usable Observations    192      Degrees of Freedom   190
Centered R**2     0.159957      R Bar **2   0.155535
Uncentered R**2   0.305789      T x R**2      58.712
Mean of Dependent Variable      1.0000000000
Std Error of Dependent Variable 2.1875206351
Standard Error of Estimate      2.0102168976
Sum of Squared Residuals        767.78467532
Regression F(1,190)                  36.1788
Significance Level of F           0.00000001
Log Likelihood                    -405.49354
Durbin-Watson Statistic             1.851287

   Variable                     Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                -1.004889553  0.363524490     -2.76430  0.00626640
2.  RD                       0.000540782  0.000089907      6.01488  0.00000001

Chi-Squared(1)=     30.711656 with Significance Level 0.00000003
```

On constate que le niveau de significativité est très inférieur à 0,05 donc on décide H1, RD est responsable de l'hétéroscédasticité.
Test de l'hypothèse H0 : $\alpha_1 = \alpha_2 = \alpha_3 = 0$
contre H1 : $V(\epsilon_t) = \sigma^2 h(\alpha_0 + \alpha_1 RD + \alpha_2 SP + \alpha_3 TCHO)$
@breuschpagan residus start end
 # rd sp tcho

```
  Test de Breusch Pagan sur  RESIDUS
Linear Regression - Estimation by Least Squares
Dependent Variable RESIDUS^2/S2
Quarterly Data From 1955:01 To 2002:04
Usable Observations    192      Degrees of Freedom   188
Centered R**2     0.179913      R Bar **2   0.166827
Uncentered R**2   0.322282      T x R**2      61.878
Mean of Dependent Variable      1.0000000000
Std Error of Dependent Variable 2.1875206351
Standard Error of Estimate      1.9967320353
Sum of Squared Residuals        749.54449832
Regression F(3,188)                  13.7480
Significance Level of F           0.00000004
Log Likelihood                    -403.18535
Durbin-Watson Statistic             1.887052

   Variable                     Coeff       Std Error      T-Stat     Signif
```

```
*****************************************************************************
1.  Constant                -1.076488911  0.650509017     -1.65484  0.09962525
2.  RD                       0.000155998  0.000200893      0.77652  0.43841528
3.  SP                       0.001970849  0.000921681      2.13832  0.03378163
4.  TCHO                     0.157369926  0.121390717      1.29639  0.19642961

Chi-Squared(3)=     34.543365 with Significance Level 0.00000015
```

On déduit toujours hétéroscédasticité due à au moins l'une des variables RD SP ou TCHO

3.3 Conclusion:

Ce test de Breusch-Pagan remplace le test de Goldfeld et Quandt quand il n'y a pas normalité des erreurs et quand il peut y avoir plusieurs variables responsables de l'hétéroscédasticité. Il faut définir les variables éventuellement responsables de l'hétéroscédasticité, on peut prendre des variables du modèle, des transformations de ces variables (par exemple leur carré ...) ou des variables non prises en compte dans le modèle. Le principal problème de ce test est que l'on teste une certaine forme d'hétoroscédasticité (par exemple pour le dernier résultat on test la responsabilité des variables RD, SP, TCHO), si d'autres variables sont responsables on ne le voit pas. C'est pour cela qu'il faut être méfiant pour tous ces tests d'hétoroscédasticité, et bien préciser que si vous trouvez H0 il fait bien préciser non pas il y a homoscédasticité mais les variables incluses ne sont pas responsables d'une éventuelle hétéroscédasticité.

4 Le test de WHITE

4.1 Théorie

C'est le test le plus utilisé. Il fonctionne comme le test précédent en pratique. C'est un test asymptotique, on ne peut donc l'utiliser que si n'est assez grand (>50) . White compare le vecteur $V_{\widehat{a}}$ (voir la théorie plus haut) dans le cas d'hétéroscédasticité et d'homoscédasticité .Il montre que dans le cas d'hétoroscédasticité la variance des erreurs est fonction des variables du modèle, de leurs carrés et des doubles produits et cela **indépendamment de la forme de l'hétéroscédasticité**. On en a déduit le test suivant : dans le cas d'hétéroscédasticité les résidus au carré du modèle estimé sont fonction des variables explicatives, de leurs carrés et des produits deux à deux.
H_0 : les erreurs sont homoscédastiques
H_1 : les erreurs sont hétéroscédastiques sans que l'on connaisse la forme exacte de l'hétéroscédasticité.
Mise en place du test : on effectue la méthode des MCO sur le modèle et on en déduit les résidus. On construit alors la régression sur les résidus au carrés

$$e_t^2 = \alpha_0 + \sum \alpha_i X_{it} + \sum \beta_i X_{it}^2 + \sum \gamma_{ij} X_{iy} X_{jt} + u_t$$

u_t étant l'erreur de ce modèle. On utilise encore ici le test de Wald , puisque comme e_t^2 la variable u_t ne suit pas une loi normale. On estime donc ce dernier modèle et sous l'hypothèse H0 (tous les α_i, β_i et γ_i sont nuls sauf α_0) le nR2 de ce modèle estimé suit asymptotiquement une loi du χ^2 à autant de degrés de liberté qu'il y a de coefficients α_i, β_i et γ_i (c'est-à-dire à autant de degrés de liberté que de variables expliquant e_t^2 hors la constante.

4.2 Applications: utilisation de la procédure White.src

Cette procédure a été faite par l'équipe d'Estima. On reprend le programme hetero1.prg

4.2.1 Sur le modèle Y1 en fonction de X1 et X2

lin Y1 start end res
 # constant X1 X2
On récupère les residus du modèle dans le vecteur RES
source white.src
@white res start end
 # X1 X2
On met donc la liste des variables explicatives sauf la constante qui sera mise automatiquement ainsi que les carrés et les produits deux à deux.
Attention en général on ne met pas certaines variables comme explicatives de l'hétéroscédasticité:

- les variables muettes car bien sur elles ne jouent aucun rôle dans l'hétéroscédasticité.
- Très souvent on ne met pas non plus les retards sur les variables car d'une part ces retards sont souvent très colinéaires aux variables de base et d'autre part elles augmentent alors beaucoup le nombre de variables dans la régression et l'on sait que plus on teste de variables simultanément moins celui-ci sera puissant.

Exemple H0: il y a homoscédasticité contre H1 : il y a hétéroscédasticité
Le test de White étudie donc les résidus du modèle au carré en fonction des variables du modèle, leur carré, et les produits deux à deux.

```
Linear Regression - Estimation by Least Squares
Dependent Variable RES^2
Quarterly Data From 1952:01 To 1986:04
Usable Observations    140      Degrees of Freedom   134
Centered R**2     0.199420      R Bar **2   0.169548
Uncentered R**2   0.547249      T x R**2      76.615
Mean of Dependent Variable      1315887657.7
Std Error of Dependent Variable 1506685166.7
Standard Error of Estimate      1373029391.2
Sum of Squared Residuals         2.52618e+20
Regression F(5,134)                   6.6757
Significance Level of F           0.00001390
Log Likelihood                   -3141.22516
Durbin-Watson Statistic             2.206746

   Variable                    Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                1321577797.136185 1929715270.879267      0.68486  0.49461746
2.  X1                      -64951.471589 101372.874575     -0.64072  0.52279983
3.  X2                       3357.328923  7142.846709      0.47003  0.63909997
4.  X1^2                        0.810818     1.413473      0.57364  0.56717601
5.  X2^2                        0.002225     0.006648      0.33467  0.73840085
6.  X1*X2                      -0.075558     0.183899     -0.41087  0.68182697

Chi-Squared(5)=     27.918825 with Significance Level 0.00003775
```

On constate que le sous-programme a estimé les résidus au carré en fonction des variables, de leur carré et des doubles produits.

nR2=140x0,19942=27,9188 Sous l'hypothèse H0 la variable nR2 suit asymptotiquement une loi du χ^2 à 5 degrés de liberté (nombre de variables explicatives hors la constante) au risque $\alpha = 5\%$ la borne du χ^2_5 est 11,07 <27,9188 on décide donc H1. On peut également dire que le niveau de significativité (p-value)=0,0000377 très inférieur à $\alpha = 5\%$, on décide donc H1, il y a hétéroscédasticité mais on n'en connaît pas la forme.

4.2.2 Sur le modèle de consommation

lin cm start end residus
 # constant rd sp tcho
source white.src
@white residus start end
 # rd sp tcho

```
Linear Regression - Estimation by Least Squares
Dependent Variable RESIDUS^2
Quarterly Data From 1955:01 To 2002:04
Usable Observations    192      Degrees of Freedom   182
Centered R**2     0.323380      R Bar **2   0.289921
Uncentered R**2   0.440842      T x R**2      84.642
Mean of Dependent Variable      1805.5005176
Std Error of Dependent Variable 3949.5696391
Standard Error of Estimate      3328.1518723
Sum of Squared Residuals        2015940269.1
Regression F(9,182)                   9.6649
Significance Level of F          0.00000000
Log Likelihood                  -1824.45439
Durbin-Watson Statistic            2.216642

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                -3292.995428  5099.673400     -0.64573  0.51926964
2.  RD                          4.494379     3.039844      1.47849  0.14100520
3.  SP                        -76.526236    18.488834     -4.13905  0.00005327
4.  TCHO                      367.877833  1136.514423      0.32369  0.74654474
5.  RD^2                       -0.000417     0.000491     -0.84967  0.39662483
6.  SP^2                        0.005859     0.011424      0.51282  0.60870180
7.  TCHO^2                     -3.782402   125.126374     -0.03023  0.97591782
8.  RD*SP                       0.005863     0.004584      1.27881  0.20259312
9.  RD*TCHO                    -0.346501     0.471070     -0.73556  0.46294414
10. SP*TCHO                     8.033894     3.394110      2.36701  0.01898246

Chi-Squared(9)=     62.088987 with Significance Level 0.00000000
```

Le sous-programme White.src étudie les résidus du modèle au carré en fonction des variables, de leurs carrés et des produits deux à deux. La p-value étant nulle donc nettement inférieure à α on en déduit très nettement hétéroscédasticité, mais toujours sans donner la forme exacte de cette hétéroscédasticité.

5 Conclusion générale

Il existe de très nombreux tests d'hétéroscédasticité. Les trois proposés ci-dessus sont les plus utilisés. Le plus puissant est le test de Goldfeld et Quandt mais il nécessite la normalité et la connaissance de la variable responsable de l'hétéroscédasticité. Le test de White indique la présence d'hétéroscédasticité mais ne fournit pas la forme de l'hétéroscédasticité, notons aussi que plus il y a de variables dans le modèle de base moins ce test sera puissant,

c'est cependant le plus utilisé. Enfin le test de Breusch et Pagan indique si les variables que l'on a proposées (hypothèse H1) sont responsables de l'hétéroscédasticité mais l'on est jamais sur de ne pas en oublier.

6 Que faire en cas d'hétéroscédasticité?

- La théorie indique de faire la méthode des QMCG ou plutôt dans le cas de seule hétéroscédasticité, la régression pondérée. Si par exemple dans le modèle $Y1 = a_0 + a_1X_1 + a_2X_2 + \epsilon$ on décide de prendre comme forme d'hétéroscédasticité $V(\epsilon_t) = \sigma^2 X1^2$, la régression pondérée consiste à transformer le modèle afin que l'on puisse estimer par les MCO le modèle transformé. Dans cet exemple simple on remarque qu'il suffit de diviser l'équation par X1 pour que le nouveau modèle ne soit plus hétéroscédastique. en effet si on divise par X1 on obtient:

$$\frac{Y1}{X1} = a_0\frac{1}{X1} + a_1 + a_2\frac{X_2}{X1} + \frac{\epsilon}{X1}$$

  On pose YR = $\frac{Y1}{X1}$, CR=$\frac{1}{X1}$ X2R=$\frac{X_2}{X1}$ et u=$\frac{\epsilon}{X1}$. Dans ce modèle transformé on constate que si l'hypothèse d'hétéroscédasticité est vraie alors V(u)=V($\frac{\epsilon}{X1}$) = $\frac{1}{X1^2}V(\epsilon) = \sigma^2$ il y a donc homoscédasticité et on peut appliquer les MCO.

  Modèle transformé sur lequel on applique les MCO:

$$YR = a_0CR + a_1 + a_2X2R + u$$

  On récupère ainsi les coefficients estimés α_i . Pour être sur que l'on a bien travaillé il faut bien sur vérifier avec White que l'on n'a plus d'hétéroscédasticité.

```
Linear Regression - Estimation by Least Squares
Dependent Variable YR
Quarterly Data From 1952:01 To 1986:04
Usable Observations    140      Degrees of Freedom   137
Centered R**2     0.745651      R Bar **2   0.741938
Uncentered R**2   0.997469      T x R**2     139.646
Mean of Dependent Variable      6.7757679364
Std Error of Dependent Variable 0.6816957120
Standard Error of Estimate      0.3463000235
Sum of Squared Residuals        16.429547761
Regression F(2,137)                 200.8152
Significance Level of F           0.00000000
Log Likelihood                     -48.67212
Durbin-Watson Statistic              1.990794

   Variable                     Coeff       Std Error      T-Stat     Signif
*******************************************************************************
1.  Constant                      1.09907      0.37426      2.93660  0.00389366
2.  CR                       -10080.41753  10817.48972     -0.93186  0.35304666
3.  X2R                           0.49239      0.02694     18.27423  0.00000000

Linear Regression - Estimation by Least Squares
Dependent Variable RESIDUS^2
Quarterly Data From 1952:01 To 1986:04
Usable Observations    140      Degrees of Freedom   134
Centered R**2     0.022661      R Bar **2  -0.013807
Uncentered R**2   0.498055      T x R**2      69.728
```

```
Mean of Dependent Variable      0.1173539126
Std Error of Dependent Variable 0.1210195367
Standard Error of Estimate      0.1218521438
Sum of Squared Residuals        1.9896246218
Regression F(5,134)                   0.6214
Significance Level of F           0.68370438
Log Likelihood                      99.10736
Durbin-Watson Statistic             2.098042

   Variable                     Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                     1.333563     1.507117      0.88484  0.37782580
2.  CR                       -93608.863088 83016.755368     -1.12759  0.26150811
3.  X2R                         -0.143322     0.202632     -0.70730  0.48060628
4.  CR^2                    688476265.491639 1349465248.721366      0.51018  0.61076155
5.  X2R^2                        0.003636     0.007150      0.50851  0.61193443
6.  CR*X2R                    7061.303506  6052.330520      1.16671  0.24540037

Chi-Squared(5)=      3.172494 with Significance Level 0.67341196
```

On constate que le niveau de significativité est très nettement supérieur à 5% on décide donc H0 pas d'hétéroscédasticité.
On remarque que l'on récupère l'estimation du modèle de base en précisant par la méthode des MCG

$$\widehat{Y1} = -10080.4175 + 1.09907\ X1 + 0.49239\ X2$$

- **REMARQUE IMPORTANTE**: Dans le cas général il est indispensable de **ne pas utiliser les MCG** car on obtient souvent de bons modèles homoscédastiques en rendant dynamique le modèle statique hétéroscédastique. On va prendre comme exemple hetero1.prg.

  Prenons par exemple le modèle autorégressif (qui contient des retards sur la variable endogène ici Y1) et à retards échelonnés (qui contient des retards sur une ou plusieurs variables explicatives ici X1 et X2). Ces modèles qui seront vus en M1 sont dits MODELES DYNAMIQUES (un modèle seulement autorégressif ou seulement à retards échelonnés est aussi un MODELE DYNAMIQUE). Les retards sont déterminés par la procédure retards.src

lin Y1 start end res
 # constant Y1{1 to 5} X1{0 to 6} X2{0 to 10}
@white res start end
 # X1 X2

```
Linear Regression - Estimation by Least Squares
Dependent Variable Y1
Quarterly Data From 1952:01 To 1986:04
Usable Observations    130      Degrees of Freedom   106
 Total Observations    140      Skipped/Missing       10
Centered R**2     0.979633      R Bar **2   0.975214
Uncentered R**2   0.998569      T x R**2     129.814
Mean of Dependent Variable      727313.40897
Std Error of Dependent Variable 200693.64506
Standard Error of Estimate       31596.54233
Sum of Squared Residuals         1.05824e+11
Regression F(23,106)                221.6733
Significance Level of F           0.00000000
Log Likelihood                   -1518.10020
```

```
Durbin-Watson Statistic             2.028950

   Variable                      Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                  -2304.75212  14503.67269     -0.15891  0.87404359
2.  Y1{1}                        0.05478      0.08446      0.64857  0.51802070
3.  Y1{2}                       -0.02174      0.08398     -0.25881  0.79628565
4.  Y1{3}                        0.00860      0.08533      0.10077  0.91992117
5.  Y1{4}                        0.03465      0.08675      0.39949  0.69033792
6.  Y1{5}                        0.24718      0.08483      2.91366  0.00435776
7.  X1                           1.02883      1.49119      0.68994  0.49173960
8.  X1{1}                        1.21382      1.94082      0.62541  0.53304299
9.  X1{2}                       -2.59602      1.89791     -1.36783  0.17425699
10. X1{3}                        0.06228      1.92647      0.03233  0.97427046
11. X1{4}                        1.82990      1.96703      0.93029  0.35433652
12. X1{5}                       -5.02134      1.97284     -2.54524  0.01235963
13. X1{6}                        6.60528      1.60231      4.12234  0.00007468
14. X2                          -5.94731      4.25588     -1.39743  0.16520180
15. X2{1}                        6.79060      7.70958      0.88080  0.38041847
16. X2{2}                       13.34559      8.25674      1.61633  0.10899617
17. X2{3}                      -18.87873      8.30605     -2.27289  0.02504977
18. X2{4}                        5.07974      8.51078      0.59686  0.55187398
19. X2{5}                       -5.15133      8.90015     -0.57879  0.56395779
20. X2{6}                        8.85925      8.72100      1.01585  0.31201268
21. X2{7}                       -5.77231      8.54297     -0.67568  0.50071555
22. X2{8}                      -23.44746      8.59284     -2.72872  0.00744490
23. X2{9}                       42.88159      8.96651      4.78242  0.00000562
24. X2{10}                     -17.56738      4.33205     -4.05521  0.00009587

Performing White's Test for Heteroskedasticity on RES
  using the regressors, their squares, and non-redundant cross-products

Linear Regression - Estimation by Least Squares
Dependent Variable RES^2
Quarterly Data From 1952:01 To 1986:04
Usable Observations    130      Degrees of Freedom   124
 Total Observations    140      Skipped/Missing       10
Centered R**2     0.033173      R Bar **2  -0.005812
Uncentered R**2   0.398698      T x R**2      51.831
Mean of Dependent Variable       814032289.8
Std Error of Dependent Variable 1048107313.5
Standard Error of Estimate      1051148745.1
Sum of Squared Residuals         1.37009e+20
Regression F(5,124)                   0.8509
Significance Level of F           0.51631936
Log Likelihood                   -2881.90000
Durbin-Watson Statistic             2.335324

   Variable                      Coeff       Std Error      T-Stat     Signif
********************************************************************************
1.  Constant                 72178067.137656 1865606156.352617      0.03869  0.96920072
2.  X1                      -92387.185643 83329.449970     -1.10870  0.26970626
3.  X2                       9233.102504  5588.720091      1.65210  0.10104503
4.  X1^2                        1.399216     1.108802      1.26192  0.20934727
5.  X2^2                        0.003415     0.005114      0.66782  0.50549084
6.  X1*X2                      -0.165993     0.142089     -1.16823  0.24495360

Chi-Squared(5)=      4.312474 with Significance Level 0.50535855
```

Au passage on remarque que RATS indique qu'il ne peut pas utiliser tout l'échantillon car il y a 10 retards au maximum donc l'échantillon ne peut commencer qu'à la valeur 11 de l'échantillon , il ne reste donc plus qu'une taille d'échantillon de 130. On constate que pour faire le test de White les retards n'ont pas été pris (le nombre de variables aurait été beaucoup trop important et donc le test moins puissant). Si on utilise le niveau de significativité, on constate que l'on est dans la zone d'acceptation de l'hypothèse H0 car 0.505 est supérieur à 0.05. On a donc réglé avec des retards le problème d'hétéroscédasticité.