Luı́s Filipe Martins	Dpt. de Métodos Quantitativos	ISCTE - EG
luis.martins@iscte.pt	http://home.iscte.pt/~lfsm	Lisboa, 12/09/2005

HETEROCEDASTICIDADE

Wooldridge §8

0.1 Introdução e Definição

De entre as hipóteses clássicas do MRLM, $M1 - M5$,

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i = x_i^{'} \beta + u_i, i = 1, ..., n \tag{1}$$

consideremos a de homocedasticidade dos erros $M4.1$ em que estes têm uma variância condicional que é idêntica (constante) para todas as observações[1]:

$$E(u_i^2 | x_i) = E(u_i^2 | x_{i2}, ..., x_{ik}) = \sigma^2 \in (0, \infty), i = 1, ..., n. \tag{2}$$

Esta hipótese pode no entanto não ser válida para um vasto conjunto de modelos, nomeadamente os que fazem uso de dados seccionais. Vejamos o seguinte exemplo:

Example 1 *Retornos na educação:*

$$S_i = \beta_1 + \beta_2 A_i + u_i,$$

em que S_i é o salário auferecido pelo trabalhador i (num dado periodo de tempo e numa dada unidade de medida) e A_i é o numero de anos de escolaridade.

Existe uma forte evidência empı́rica de que, em média, um trabalhador com mais anos de escolaridade aufere um salário mais elevado, isto é, $\beta_2 > 0$. Por outro lado, trabalhadores com um baixo nı́vel de escolaridade têm salários relativamente semelhantes enquanto que os que apresentam altos nı́veis de escolaridade as disparidades de ordenados estão mais acentuadas. Neste cenário, a variabilidade no valor dos salários para um numero reduzido de anos de escolaridade é diferente e menor do que para um elevado numero de anos de escolaridade.

Como o segundo momento (variância) condicionado da variável dependente S_i coincide com o de u_i (todos os factores que influenciam os salários e que não estão no modelo como, por exemplo, anos de experiência, idade, sector de actividade, sexo, QI, habilidade ...),

$$E(u_{baixo}^2 | A_{baixo}) = \sigma_{baixo}^2 < \sigma_{alto}^2 = E(u_{alto}^2 | A_{alto}).$$

[1] $E(u_1^2 | x_1) = \sigma^2, a.s.$, porque assume-se que os dados são identicamente distribuidos (M1) e X é aleatório.

- (Plots)

Generalizando esta ideia, sob a hipótese de heterocedasticidade nos erros (não pertence às hipóteses clássicas do MRLM),

$$\sigma_i^2 = \sigma^2(x_i) = E(u_i^2|x_i) = V(y_i|x_i), i = 1, ..., n. \tag{3}$$

Sob (3) (e ausência de autocorrelação dos erros),

$$\Omega_{n\times n} = E(\mathrm{u}\,\mathrm{u}'|X) = \begin{pmatrix} \sigma_1^2 & ... & 0 \\ ... & & ... \\ 0 & ... & \sigma_n^2 \end{pmatrix} = diag\left(\sigma_1^2, ..., \sigma_n^2\right) = f(x_1, ..., x_n) \neq \sigma^2 I_n. \tag{4}$$

A forma funcional de $\sigma^2(x_i)$ pode ser muito diversa. Note-se que se os regressores x_i são estocásticos então σ_i^2 e a matriz Ω são estocásticos também.

Example 2 $\sigma_i^2 = \sigma^2(A_i) = A_i^2; \sigma_i^2 = \sigma^2(x_i) = x_{1i}^2 + x_{2i}; \sigma_i^2 = \frac{1}{2}\exp(x_{4i}); ...$

Note-se que σ_i^2 é conceptualmente diferente de σ^2 e que no modelo com heterocedasticidade y_i não é identicamente distribuido. É importante notar ainda que uma incorrecta especificação do modelo (1) pode introduzir heterocedasticidade no mesmo. Portanto, em primeiro lugar convém testar a forma funcional do modelo de regressão.

Como é obvio, se a hipótese de homocedasticidade não se verifica em deterimento de uma dada especificação de heterocedasticidade dos erros (3), o estimador OLS perderá algumas das suas qualidades estatísticas.

0.2 Consequências para o OLS

Suponhamos que as hipóteses clássicas do MRLM se verificam mas sob erros heterocedasticos (3). Porque $M4.2$ não é utilizada para derivar o primeiro momento do estimador, o OLS mantém as propriedades de centrado e consistente, $E\left(\widehat{\beta}|X\right) = \beta$ e $p\lim_{n\to\infty}\widehat{\beta} = \beta$. No entanto, a fórmula para $V\left(\widehat{\beta}|X\right)$ não é válida sob (3) e o seu estimador habitual $\widehat{\Sigma}$ é enviesado. Consequentemente, a inferência torna-se errónea. Por outro lado, o teorema de Gauss-Markov não é válido e o OLS deixa de ser BLUE, mesmo assimptóticamente. O OLS não sendo BLUE, significa que existe um outro estimador LU para β que é mais efficiente - "menor" variância.

A expressão para $V\left(\widehat{\beta}|X\right)$ sob (3), onde $\widehat{\beta}$ é o estimador OLS, é apresentada em seguida. No MRLS e sob (3),

$$V\left(\widehat{\beta}_2|\mathbf{x}\right) = \frac{\sum_{i=1}^n \sigma_i^2 (x_i - \overline{x})^2}{\left[\sum_{i=1}^n (x_i - \overline{x})^2\right]^2}, \tag{5}$$

que apenas coincide com a expressão standard $\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$ se $\sigma_i^2=\sigma^2, i=1,...,n-$ homocedasticidade! No MRLM,

$$\Sigma_{k\times k}=V\left(\widehat{\beta}|X\right)=\left(X^{'}X\right)^{-1}X^{'}\Omega X\left(X^{'}X\right)^{-1}=\left(\sum_{i=1}^{n}x_ix_i^{'}\right)^{-1}\left(\sum_{i=1}^{n}\sigma_i^2x_ix_i^{'}\right)\left(\sum_{i=1}^{n}x_ix_i^{'}\right)^{-1} \tag{6}$$

que, em geral, é diferente de $\sigma^2\left(X^{'}X\right)^{-1}=\sigma^2\left(\sum_{i=1}^{n}x_ix_i^{'}\right)^{-1}$ e onde

$$p\lim_{n\to\infty}\left(\frac{1}{n}X^{'}\Omega X\right)=p\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2x_ix_i^{'}\right)=D,$$

matriz positiva definida.

- (Monte Carlo)

O estudo do MRLM sob (3) requer portanto uma de duas alternativas: *(i)* manter o OLS mas usar (6); ou *(ii)* utilizar o estimador (F)GLS, do qual o MQP é um caso particular. Estas duas abordagens serão desenvolvidas mais à frente. O (F)GLS será estudado no capítulo de introdução à regressão com time series e autocorrelação dos erros. Entretanto, propomos testes cujo objectivo é inferir se os erros são homocedasticos ou não. Se os testes apontam para a hipótese de homocedasticidade, pode-se utilizar o OLS. Caso contrário, aplicar *(i)* ou *(ii)*.

0.3 Testes para a sua Detecção

O objectivo é inferir sobre $\sigma_i^2=E(u_i^2|x_i)=E(u_i^2|x_{i2},...,x_{ik})$. Como sabemos, os erros u_i não sendo observados podem ser estimados pelos resíduos OLS

$$\widehat{u}_i=y_i-\widehat{y}_i=y_i-x_i^{'}\widehat{\beta}, i=1,...,n; \widehat{u}_{n\times 1}=y-X\widehat{\beta}. \tag{7}$$

Um estimador centrado para $\sigma_i^2, i=1,...,n$ quando $\sigma_i^2\neq\sigma^2$ é $\widehat{u}_i^2$. Por outras palavras, $\widehat{u}_i^2$ é uma proxy/medida de $\sigma_i^2=E(u_i^2|x_i)$. Se $\widehat{u}_i^2$ é relativamente constante para todo i então há indícios de homocedasticidade. Caso contrário, assume-se heterocedasticidade e pelo comportamento de $\widehat{u}_i^2$ pode-se induzir de quais as variáveis responsáveis e que forma funcional tem σ_i^2 - linear, quadrática, ...

A hipótese nula é a de erros homocedasticos,

$$H_0:\sigma_i^2=V\left(u_i|x_i\right)=E(u_i^2|x_i)=E(u_i^2)=E(u_1^2)=\sigma^2.$$

A estatística de teste (e sua distribuição) basea-se numa regressão auxiliar[2] que é construída à custa de uma forma funcional específica para $\sigma_i^2=E(u_i^2|x_i)=f\left(x_i\right)$ sob a hipótese alternativa de heterocedasticidade.

[2] A regressão auxiliar não é o modelo original que pretendemos estudar. É auxiliar no sentido de que auxilia-nos a testar a hipótese de homocedasticidade versus heterocedasticidade.

Case 3 *$f(x_i)$ linear: Breusch-Pagan.*

$$\begin{aligned}
u_i^2 &= \delta_1 + \delta_2 x_{i2} + ... + \delta_k x_{ik} + v_i = x_i^{'}\delta + v_i, i = 1, ..., n \\
H_0 &: \delta_2 = ... = \delta_k = 0; H_1 : \textit{Pelo menos um deles} \neq 0 \\
F &= \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{k-1,n-k}; LM = nR^2 \xrightarrow{d} \chi^2_{k-1} \\
R^2 &: \widehat{u}_i^2 = \delta_1 + \delta_2 x_{i2} + ... + \delta_k x_{ik} + v_i = x_i^{'}\delta + v_i, i = 1, ..., n
\end{aligned}$$

Case 4 *$f(x_i)$ não-linear (quadrados e produtos cruzados): White. Exemplo para $k = 3$:*

$$\begin{aligned}
u_i^2 &= \delta_1 + \delta_2 x_{i2} + \delta_3 x_{i3} + \delta_4 x_{i2}^2 + \delta_5 x_{i3}^2 + \delta_6 x_{i2} x_{i3} + v_i, \\
H_0 &: \delta_2 = ... = \delta_6 = 0; H_1 : \textit{Pelo menos um deles} \neq 0 \\
F &= \frac{R^2/5}{(1-R^2)/(n-6)} \sim F_{5,n-6}; LM = nR^2 \xrightarrow{d} \chi^2_5 \\
R^2 &: \widehat{u}_i^2 = \delta_1 + \delta_2 x_{i2} + \delta_3 x_{i3} + \delta_4 x_{i2}^2 + \delta_5 x_{i3}^2 + \delta_6 x_{i2} x_{i3} + v_i,
\end{aligned}$$

Se k elevado. Exemplo para potência até dois:

$$\begin{aligned}
u_i^2 &= \textit{linear, quadrados e produtos cruzados } x_{ij}, j = 2, ..., k, \\
H_0 &: \delta_2 = \delta_3 = 0; H_1 : \textit{Pelo menos um deles} \neq 0 \\
F &= \frac{R^2/2}{(1-R^2)/(n-3)} \sim F_{2,n-3}; LM = nR^2 \xrightarrow{d} \chi^2_2 \\
R^2 &: \widehat{u}_i^2 = \delta_1 + \delta_2 \widehat{y}_i + \delta_3 \widehat{y}_i^2 + v_i,
\end{aligned}$$

Existem outros testes na literature. Por exemplo, Gleijser considera

$$|\widehat{u}_i| = \delta_1 + \delta_2 x_{i2} + ... + \delta_k x_{ik} + v_i = x_i^{'}\delta + v_i, i = 1, ..., n \tag{8}$$

e Park assume

$$\ln\left(\widehat{u}_i^2\right) = \delta_1 + \delta_2 \ln(x_{i2}) + ... + \delta_k \ln(x_{ik}) + v_i, i = 1, ..., n. \tag{9}$$

O teste de Goldfeld-Quandt divide a amostra em dois sub-periodos e testa a diferença entre as duas variâncias dos erros. De salientar que para todos os testes descritos anteriormente os regressores da regressão auxiliar podem ser apenas uma parte (sub-conjunto) dos apresentados e que podem mesmo incluir variáveis que **não** são regressores do modelo original (por exemplo, análises per capita.) Devido à sua natureza, o teste de White também pode ser interpretado como um teste de especificação.

Como é sabido, a região crítica do teste F encontra-se na cauda (lado direito) da distribuição. Se o valor observado da estatística F é próximo de zero, então não se rejeita a hipótese nula que neste caso é equivalente a homocedasticidade dos erros. Caso F_{obs} seja maior do que o valor crítico, para um dado nível de significância, rejeita-se H_0, ou seja, admite-se como válida a presença de erros heterocedásticos. Neste cenário, é imperioso solucionar estatisticamente o "problema" - o que será objecto nas próximas duas secções.

0.4 Estimação Robusta de $V(\widehat{\beta}_{OLS})$

Se o modelo tem erros heterocedasticos, uma forma de resolver o "problema" da inferência é, mantendo o estimador consistente e ineficiente OLS, usar estimadores para (5) e (6). Como vimos anteriormente, σ_i^2 pode ser estimado por $\widehat{u}_i^2$ e portanto,

$$V\widehat{\left(\widehat{\beta}_2|\mathbf{x}\right)} = \frac{\sum_{i=1}^n \widehat{u}_i^2 (x_i - \overline{x})^2}{\left[\sum_{i=1}^n (x_i - \overline{x})^2\right]^2}; \tag{10}$$

$$\widehat{\Sigma} = \left(X'X\right)^{-1} X'\widehat{\Omega}X \left(X'X\right)^{-1} = \left(\sum_{i=1}^n x_i x_i'\right)^{-1} \left(\sum_{i=1}^n \widehat{u}_i^2 x_i x_i'\right) \left(\sum_{i=1}^n x_i x_i'\right)^{-1}. \tag{11}$$

A raíz quadrada de $V\widehat{\left(\widehat{\beta}_2|\mathbf{x}\right)}$ e $\widehat{\Sigma}_{jj}$ denomina-se desvio padrão de $\widehat{\beta}_j$ *robusto* a heterocedasticidade. A matriz $\widehat{\Sigma}$ é o estimador consistente para Σ robusto a heterocedasticidade de White. Os testes t e os intervalos de confiança baseam-se na expressões anteriores.

Como as propriedades de $\widehat{\Sigma}$ foram derivados assimptoticamente, este procedimento é recomendado só para amostras de grande dimensão. O estimador BLUE que representa uma alternativa ao OLS sob heterocedasticidade é apresentado em seguida.

0.5 Método dos MQP

Se o modelo tem erros heterocedasticos, existe um estimador alternativo ao OLS que é eficiente (BLUE, de menor variância condicional). Este denomina-se mínimos quadrados ponderados (MQP) (ou também WLS) porque pondera de uma forma diferente os erros e pertence à classe de estimadores GLS ou FGLS. A ideia é a seguinte.

O MRLM (1) é

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i = x_i'\beta + u_i, i = 1, ..., n$$

em que

$$E(u_i^2|x_i) = \sigma_i^2 = \sigma^2(x_i) = \sigma^2 h(x_i) = \sigma^2 h_i \tag{12}$$

para uma dada função positiva $h(\cdot)$ de x_i. Ver exemplos 2. Por agora, suponhamos que $\sigma_i^2 = \sigma^2 h_i$ é conhecido. Então, é possível transformar o modelo de tal modo que o "novo" modelo tem erros homocedasticos. Qual é a transformação (única) apropriada? É dividir ambos os membros da equação por $\sqrt{h_i}$ porque desta maneira

$$\frac{y_i}{\sqrt{h_i}} = \frac{1}{\sqrt{h_i}}\beta_1 + \beta_2\frac{x_{i2}}{\sqrt{h_i}} + ... + \beta_k\frac{x_{ik}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}} = \frac{x_i'}{\sqrt{h_i}}\beta + \frac{u_i}{\sqrt{h_i}}, i = 1, ..., n \tag{13}$$

com

$$\frac{y_i}{\sqrt{h_i}} = y_i^*; \frac{x_i'}{\sqrt{h_i}} = x_i^{*'}; \frac{u_i}{\sqrt{h_i}} = u_i^*; E(u_i^*|x_i) = 0; E(u_i^{*^2}|x_i) = \sigma^2, constante. \tag{14}$$

O MQP (pertence à classe GLS) é o OLS aplicado ao modelo transformado (13) onde dar-se-à mais peso às observações i com variância σ_i^2 (ou h_i) mais reduzida:

$$\begin{aligned}
\widehat{\beta} &= \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i^* - x_i^{*'}\beta\right)^2 = \arg\min_{\beta} \sum_{i=1}^{n} \frac{1}{h_i}\left(y_i - x_i'\beta\right)^2 = \arg\min_{\beta} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left(y_i - x_i'\beta\right)^2 && (15)\\
&= \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}y = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i x_i'\right)^{-1} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i y_i && (16)
\end{aligned}$$

onde

$$\begin{aligned}
V\left(\widehat{\beta}|X\right) &= \left(X'\Omega^{-1}X\right)^{-1} = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} x_i x_i'\right)^{-1}\\
\widehat{\sigma}^2 &= \widehat{u}^{*\prime}\widehat{u}^* = \frac{1}{n-k}\sum_{i=1}^{n} \widehat{u}_i^{*2} = \frac{1}{n-k}\sum_{i=1}^{n} \frac{\widehat{u}_i^2}{h_i}.
\end{aligned}$$

E se $\sigma_i^2 = \sigma^2 h_i$ não é conhecido e a teoria económica não aponta numa dada direcção, como é usual em qualquer aplicação? Quando o último estimador não é viável porque $\sigma_i^2 = \sigma^2 h_i$ não é conhecido então procede-se à estimação de $\sigma_i^2 = \sigma^2 h_i$ e substitui-se essa quantidade no procedimento descrito anteriormente. Esse estimador pertence à classe FGLS pois passa a ser Feasible - viável. Suponhamos que

$$E(u_i^2|x_i) = \sigma_i^2 = \sigma^2 h_i = \sigma^2 exp\left(\delta_1 + \delta_2 x_{i2} + ... + \delta_k x_{ik}\right) = \sigma^2 exp\left(x_i'\delta\right) \tag{17}$$

onde σ^2, δ são desconhecidos. Então,

$$\begin{aligned}
u_i^2 &= \sigma^2 exp\left(\delta_1 + \delta_2 x_{i2} + ... + \delta_k x_{ik}\right) v_i = \sigma^2 exp\left(x_i'\delta\right) v_i; E\left(v_i|x_i\right) = 1\\
\ln\left(u_i^2\right) &= \delta_1^* + \delta_2 x_{i2} + ... + \delta_k x_{ik} + e_i = x_i'\delta^* + e_i; \delta_1^* = \delta_1 + 2\ln\sigma; E\left(e_i|x_i\right) = 0, e_i \perp x_i,
\end{aligned}$$

e os parâmetros podem ser estimáveis através de

$$\ln\left(\widehat{u}_i^2\right) = \delta_1^* + \delta_2 x_{i2} + ... + \delta_k x_{ik} + e_i = x_i'\delta^* + e_i, \tag{18}$$

onde $\widehat{u}_i$ são os resíduos OLS. Desta forma,

$$\widehat{h}_i = exp\left(\widehat{\delta}_1^* + \widehat{\delta}_2 x_{i2} + ... + \widehat{\delta}_k x_{ik}\right) = exp\left(x_i'\widehat{\delta}^*\right), \tag{19}$$

e o estimador é

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(\widehat{y_i^*} - \widehat{x_i^{*'}}\beta\right)^2 = \arg\min_{\beta} \sum_{i=1}^{n} \frac{1}{\widehat{h}_i}\left(y_i - x_i'\beta\right)^2 = \arg\min_{\beta} \sum_{i=1}^{n} \frac{1}{\widehat{\sigma}_i^2}\left(y_i - x_i'\beta\right)^2. \tag{20}$$

A forma geral do estimador GLS e FGLS será alvo de discussão no capítulo de introdução à regressão com time series e autocorrelação dos erros. No entanto, importa referir o seguinte. Se a forma funcional de σ_i^2 está correctamente especificada, pode-se demostrar que a distribuição assimptótica do FGLS coincide com a do GLS. Caso contrário, o FGLS é ineficiente e pode mesmo ser mais ineficiente do que o OLS, mesmo assimptóticamente. Na verdade, mesmo se a forma funcional é correctamente especificada, em amostras de dimensão finita o OLS é normalmente mais preciso do que o FGLS pois este último requer a estimação de Ω. Em jeito de conclusão pode-se afirmar que se estamos perante um modelo com erros heterocedásticos (hipótese a ser testada) é de bom senso utilizar-se o OLS para estimar o modelo e a matrix de variâncias-covariâncias de White robusta a heterocedasticidade para os testes t e intervalos de confiança.

- (Monte Carlo; Vcov WLS $\leq$ Vcov OLS)

Aplicações

(...)

Exercícios

1. Considere o modelo

$$C_i = \beta_1 + \beta_2 D_i + u_i,$$

em que C_i é a taxa de crescimento do lucro da empresa i para um dado ano e D_i é a dimensão da empresa (medida por exemplo pelo numero de trabalhadores). Discuta a hipótese de homocedasticidade para este modelo.

2. Porque

$$\arg\min_{\beta} \sum_{i=1}^{n} \frac{1}{h_i} \left(y_i - x_i' \beta \right)^2 = \arg\min_{\beta} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \left(y_i - x_i' \beta \right)^2 ?$$

3. ...