AULA 15 - Heterocedasticidade

Econometria I

Área quantitativa - IE/UFRJ

Heterocedasticidade

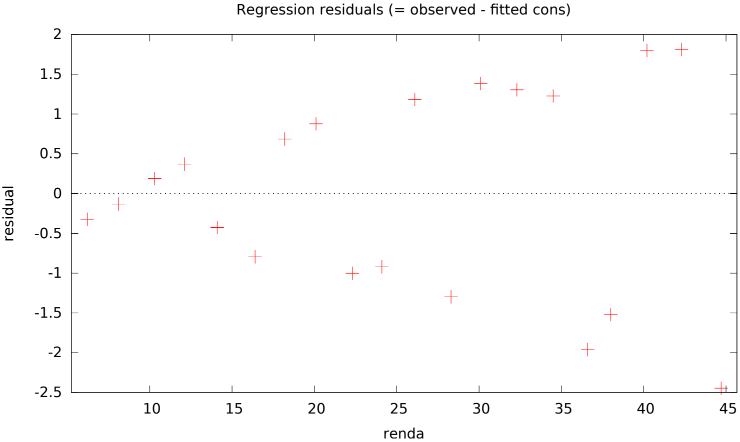
- Questões que iremos responder:
  - Como **detectamos** este problema?
  - Quais são as **consequências** nas propriedades dos estimadores por mínimos quadrados?
  - Quais são as **soluções** para este problema?
- A heteroscedasticidade aparece em várias **aplicações**, tanto em *cross-section* quanto em séries temporais.
- Observação: em séries de tempo exite uma forma de heterocedasticidade em que o distúrbio depende de próprio distúrbio defasado, sendo usado um método específico para a estimativa, chamado Autoregressivo Condicional a heterocedasticidade (ARCH).

Heterocedasticidade

Example (Gastos com consumo e renda para 20 famílias)

Estimando a equação por mímos quadrados, obtemos:

$$\text{consumo} = \underset{(0.703)}{0.847} + \underset{(0.025)}{0.899}\text{renda} + u$$



Heterocedasticidade

- No modelo de regressão **heterocedástico**:

$$Var[u_i|x_i] = \sigma_i^2, \quad i = 1, ..., n$$

- Assumindo que os erros são **não correlacionados**

$$E[uu'|X] =$$

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

- Para facilitar a notação escreveremos $\sigma_i^2 = \sigma^2\omega_i$

- A regressão clássica com erros **homocedásticos** pode ser visto como um **caso especial** em que $\omega_i = 1, \, i = 1, ...n$

Teste para heterocedasticidade

O teste geral considera a seguite hipótese:

$$\begin{aligned} H_0 &: \sigma_i^2 = \sigma^2 \quad \text{para todo}\, i \\ H_1 &: \text{Não}\, H_0. \end{aligned}$$

- É feita a regressão da seguinte forma: $\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + \epsilon$
- A estatística do teste é o $nR^2_{\hat{u}^2}$ (teste LM multiplicador de Lagrange) em que n são as observações e $R^2_{\hat{u}^2}$ é o poder de explicação da regressão acima, distribuída assintoticamente por uma $\chi^2(k)$.
- A versão LM do teste é geralmente chamada teste de BREUSCH-PAGAN da heterocedasticidade.

Teste para heterocedasticidade

Example (Gastos com consumo e renda para 20 famílias)

```
 model_hetero <- lm((resid(model))^2 ~ renda, data = tutoria2_2019_1)
> summary(model_hetero)

Call:
lm(formula = (resid(model))^2 ~ renda, data = tutoria2_2019_1)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1212 -0.2607 -0.1955  0.1335  2.1811 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.37033    0.38860  -3.526  0.00241 ** 
renda        0.11580    0.01398   8.282 1.49e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7259 on 18 degrees of freedom
Multiple R-squared:  0.7921,	Adjusted R-squared:  0.7806 
F-statistic: 68.59 on 1 and 18 DF,  p-value: 1.492e-07

O teste LM=n*R-squared
> 20*0.7921
[1] 15.842

A distribuição \chi^2(1) para 95% de confiança é 3.84.

O teste foi 15.842>3.84$, então rejeitamos a hipótese nula de homocedasticidade.
```

Outros testes para heterocedasticidade

- No **teste de White** regredimos $\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_1^2 + \epsilon$ no caso de regressão simples e incluimos o produto cruzado se for uma regressão múltipla.
- **Teste Goldfeld-Quandt**
  - Assume que as observações podem ser divididas em **dois grupos**, tal que a **hipótese nula de homocedasticidade** significa que a **variância dos erros serão as mesmas** nos dois grupos.
  - A ideia é separar as observações em dois grupos com **alta e baixa variância**.
  - É estimado separadamente os dois conjuntos de observações.
  - A **estatística** é
    $$F[n_1 - K, n_2 - K] = \frac{u_1' u_1/(n_1 - K)}{u_2' u_2/(n_2 - K)}$$
    em que assumimos que a variância do erro é maior na primeira amostra. Sob a hipótese nula de homocedasticidade, esta estatística tem distribuição F com $n_1 - K$ e $n_2 - K$ graus de liberdade.

Ineficiência na estimativa por mínimos quadrados

- Quando **falha a hipótese** de **homocedasticidade**, temos que o estimador por mínimos quadrados, apesar de ser não viesado ele é ineficiente.
- Em um modelo de regressão simples temos que a inclinação do estimador por mínimos quadrados é escrito como:
  $\hat{\beta} = \beta + \frac{\sum(x_i - \overline{x})u_i}{\sum(x_i - \overline{x})^2} = \beta + \sum w_i u_i$
  $E[\hat{\beta}] = E[\beta + \sum w_i u_i] = \beta + \sum w_i E(u_i) = \beta$
  $V[\hat{\beta}] = V[\sum w_i u_i] = \sum w_i^2 V[u_i] + \sum\sum w_i w_j Cov[u_i, u_j] = \sum w_i^2 \sigma_i^2$
  Se σ_i^2 é constante temos a fórmula usual $\sigma^2 / \sum(x_i - \overline{x})^2$

Soluções para o problema de heterocedasticidade

- **Casos simples**: transformar os dados em logaritmo ou fazer o uso de deflatores.
- Se o tipo de **hetorocedasticidade é conhecido** podemos usar o mínimo quadrado generalizado.
- Se a precisa forma de **heterocedasticidade não é conhecida**, devemos usar técnicas de estimação apropriadas que requerem uma formulação de Ω.
- O **estimador de White** mostra que é possível obter um estimador adequado para a variância usando o resíduo dos mínimos quadrados na estimativa da variância assintótica.

Equação do Log dos de Preços dos Imóveis

$$\widehat{preço} = -21{,}77 + 0{,}00207\ tamterr + 0{,}123\ arquad + 13{,}85\ qtdorm$$
$$(29{,}48)\quad (0{,}00064)\qquad\qquad (0{,}013)\qquad\qquad (9{,}01)$$
$$n = 88,\ R^2 = 0{,}672.$$

- O R-quadrado da regressão $\hat{u}^2$ sobre tamterr, arquad e qtdorm é $0,1601$ e temos n=88 e k=3
- Fazendo o teste para heterocedasticidade $LM = 88(0,1601) = 14,09 > 7,815$ temos heterocedasticidade.

$$\log(\widehat{preço}) = -1{,}30 + 0{,}168 \log(tamterr) + 0{,}700 \log(arquad) + 0{,}037\ qtdorm$$
$$(0{,}65)\quad (0{,}038)\qquad\qquad (0{,}093)\qquad\qquad (0{,}028)$$
$$n = 88,\ R^2 = 0{,}643.$$

- Fazendo o teste para heterocedasticidade $LM = 88(0,048) = 4,22 < 7,815$ não temos mais heterocedasticidade.

Quando o Ω é conhecido - MQG

- Sendo o estimador generalizado: $\hat{\beta} = \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}$ e considere $Var[u_i|x_i] = \sigma_i^2 = \sigma^2\omega_i$.
- Então Ω^{-1} é uma matriz diagonal com o ith elemento da diagonal $1/\omega_i$.
- Aplicando mínimos quadrados no modelo transformado obtemos o estimador por **mínimos quadrados ponderados**
  - $\hat{\beta} = \left[\sum_{i=1}^{n} \frac{1}{\omega_i}\mathbf{x}_i\mathbf{x}_i'\right]^{-1}\left[\sum_{i=1}^{n} \frac{1}{\omega_i}\mathbf{x}_i y_i\right]$
- Ideia: observações com menor variância tem um maior peso.

Exemplo Gastos com consumo usando MQG

Example (Gastos com consumo e renda para 20 famílias)

Assumindo que $\sigma_i^2 = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2$ podemos fazer o WLS em dois passos

Regredindo o u_i^2 com $renda_i$ e $renda_i^2$

$$\omega_i^2 = 0.4875 - 0.0701\, renda_i + 0.0037\, renda_i^2$$

Regredindo $cons_i/\omega_i$ com $1/\omega_i$ e $renda_i/\omega_i$

$$cons_i/\omega_i = \underset{(0.3299)}{0.7287}\, 1/\omega_i + \underset{(0.0199)}{0.9052}\, renda_i/\omega_i + e$$

Temos que o β estimado é maior do que o MQO, a inclinação menor e os erros padrão menores.

Quando o Ω não é conhecido - MQGE

- Podemos obter uma estimativa consistente de Ω, digamos $\widehat{\Omega}$, usando um MQG implementável conhecido como estimador mínimos quadrados generalizados estimado (MQGE).
- O modelo é feito em dois estágios:
  - (1) o modelo é estimado por MQO ou outro estimador consistente (mas ineficiente), e os resíduos são usados para construir um estimador consistente da matriz de covariância de erros;
  - (2) usando o estimador consistente da matriz de covariância dos erros, pode-se implementar a estrutura do MQG.

Quando o Ω não é conhecido - MQGE

- O estimador MQO é calculado de forma usual
$$\hat{\beta}_{MQO} = (X'X)^{-1}X'y$$
and estimates of the residuals $u_j = (Y - X\mathbf{b})_j$ are constructed.
- Cada entrada da diagonal deve ser estimada pelo resíduo e_j tal que
$$\widehat{\Omega}_{MQO} = \mathrm{diag}(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \ldots, \widehat{\sigma}_n^2)$$
- Estima β_{MQGE1} usando mínimos quadrados ponderados
$$\hat{\beta}_{MQGE1} = (X'\widehat{\Omega}_{MQO}^{-1}X)^{-1}X'\widehat{\Omega}_{MQO}^{-1}y$$

Quando o Ω não é conhecido - MQGE

- O procedimento pode ser iterado. A primeira iteração é dada por

$$e_{MQGE1} = Y - X\hat{\beta}_{MQGE1}$$

$$\widehat{\Omega}_{MQGE1} = \operatorname{diag}(\widehat{\sigma}^2_{MQGE1,1}, \widehat{\sigma}^2_{MQGE1,2}, \ldots, \widehat{\sigma}^2_{MQGE1,n})$$

$$\hat{\beta}_{MQGE2} = (X'\widehat{\Omega}^{-1}_{MQGE1}X)^{-1}X'\widehat{\Omega}^{-1}_{MQGE1}y$$

- Esta estimativa de $\widehat{\Omega}$ pode ser iterada para convergência.
- As propriedades assintóticas do estimador iterado são as mesmas que as do estimador de dois passos.

Inferência Robusta em relação à heterocedasticidade após a estimação MQO

- Sabemos que a inferência habitual dos estimadores MQO é viesada na presença de heterocedasticidade.
- Com esta informação temos que decidir se abandonaremos de vez o método MQO.
- Felizmente, ele ainda é útil.
- É possível ajustar erros-padrão, estatísticas t, F e LM de forma a torná-la válidas na presença de heterocedasticidade de forma desconhecida.
- Esses métodos são conhecidos como **procedimentos robustos em relação à heterocedasticidade**.
- A ideia aqui é como calculamos a $Var(\hat{\beta}_j)$, na presença de heterocedasticidade.

Inferência Robusta em relação à heterocedasticidade após a estimação MQO

- Considere um modelo de regressão simples:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- Assumindo que as hipóteses clássicas são satisfeitas, exceto que: $Var(u_i|x_i) = \sigma_i^2$
- Vimos que

$$Var(\hat{\beta}_j) = \frac{\sum (x_i - \overline{x})^2 \sigma_i^2}{SQT_x^2}$$

onde $SQT_x = \sum (x_i - \overline{x})^2$ é a soma dos quadrados totais de x_i.
- Quando $\sigma_i^2 = \sigma^2$ essa fórmula se reduz em σ^2 / SQT_x.

Inferência Robusta em relação à heterocedasticidade após a estimação MQO

- Na presença de heterocedasticidade temos que achar uma forma de estimar

$$Var(\hat{\beta}_j) = \frac{\sum(x_i - \overline{x})^2 \hat{u}_i^2}{SQT_x^2}$$

- Uma forma de estimadar a equação acima é:

$$\hat{Var}(\hat{\beta}_j) = \frac{\sum \hat{r}_{ij}^2 \hat{u}_i^2}{SQT_j^2}$$

  onde $\hat{r}_{ij}$ representa o i-ésimo resíduo da regressão de x_j sobre todas as outras variáveis independentes e SQR_j é a soma dos resíduos quadrados dessa regressão.

- A raiz quadrada da estimativa acima é chamada de **erro-padrão robusto em relação a heterocedasticidade** de $\hat{\beta}_j$

Inferência Robusta em relação à heterocedasticidade após a estimação MQO

- Com os erros-padrão robustos em relação a heterocedasticidade podemos construir a estatística t como anteriormente.

  $$t = \frac{\text{estimativa-valor hipotético}}{\text{erro-padrão}}$$

  assim a única diferença é o erro-padrão calculado.

Equação do Log dos Salários com erros-padrão robustos em relação à heterocedasticidade

$$
\begin{aligned}
\widehat{\log(salário)} = \underset{\substack{(0{,}100)\\ [0{,}109]}}{0{,}321} &+ \underset{\substack{(0{,}055)\\ [0{,}057]}}{0{,}213}\, hcasados - \underset{\substack{(0{,}058)\\ [0{,}058]}}{0{,}198}\, mcasadas - \underset{\substack{(0{,}056)\\ [0{,}057]}}{0{,}110}\, msolteiras \\
&+ \underset{\substack{(0{,}0067)\\ [0{,}0074]}}{0{,}0789}\, educ + \underset{\substack{(0{,}0055)\\ [0{,}0051]}}{0{,}0268}\, exper - \underset{\substack{(0{,}00011)\\ [0{,}00011]}}{0{,}00054}\, exper^2 \\
&+ \underset{\substack{(0{,}0068)\\ [0{,}0069]}}{0{,}0291}\, perm - \underset{\substack{(0{,}00023)\\ [0{,}00024]}}{0{,}00053}\, perm^2 \\
&n = 526,\ R^2 = 0{,}461.
\end{aligned}
$$

- Em parênteses () erro-padrão usual do MQO e em colchetes [] o erro-padrão robusto em relação à heterocedasticidade
- Neste caso as variáveis eram significativas e continuarem sendo, pois não houve muita mudança.
- Erros-padrão robustos podem ser maiores ou menores do que os usuais.

Equação do Log dos Salários com erros-padrão robustos em relação à heterocedasticidade

- Não sabemos se este exemplo existe problema de heterocedasticidade.
- Se a diferença for muito grande a suspeita é que haja problema.
- Também podemos nos perguntar se dada que é feita a correção por que não utilizamos esses erros-padrão robustos em vez dos usuais?
- A responta é que se os erros são homocedásticos e normalmente distribuídos as estatísticas t usuais têm distribuições t exatas, independente do tamanho da amostra.
- O erros robustos são justificados somente quando o tamanho da amostra se torna grande.
- Com amostras pequenas as estatísticas t robustas podem ter distorções que não sejam muito próximas da distribuição t, e isso pode ofuscar nossa inferência.

Equação do número de prisões no ano atual

- Exemplo 8.3 do Wooldridge sobre do efeito da média de tempo das penas cumpridas de condenações passadas afeta o número de prisões no anos atual.

$$
\begin{aligned}
\widehat{npre86} = {} & \underset{\substack{(0{,}036)\\ [0{,}040]}}{0{,}567} - \underset{\substack{(0{,}040)\\ [0{,}034]}}{0{,}136}\, pcond + \underset{\substack{(0{,}0097)\\ [0{,}0101]}}{0{,}0178}\, sentmed - \underset{\substack{(0{,}00030)\\ [0{,}00021]}}{0{,}00052}\, sentmed^2 \\
& - \underset{\substack{(0{,}0087)\\ [0{,}0062]}}{0{,}0394}\, ptemp86 - \underset{\substack{(0{,}0144)\\ [0{,}0142]}}{0{,}0505}\, empr86 - \underset{\substack{(0{,}00034)\\ [0{,}00023]}}{0{,}00148}\, rend86 \\
& + \underset{\substack{(0{,}045)\\ [0{,}058]}}{0{,}325}\, negro + \underset{\substack{(0{,}040)\\ [0{,}040]}}{0{,}193}\, hispan \\
& n = 2{,}725,\ R^2 = 0{,}0728.
\end{aligned}
$$

- A estatística t habitual de $sentmed^2$ é aproximadamente $-1,73$ (não significativa a 5%) e a estatística robusta é $-2,48$ (signficativo).