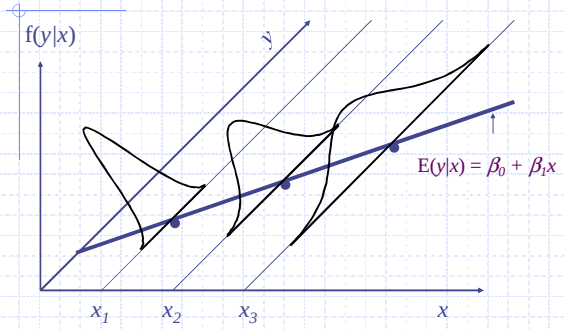
Heterocedasticidade

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$

Outras formas de detectar heterocedasticidade

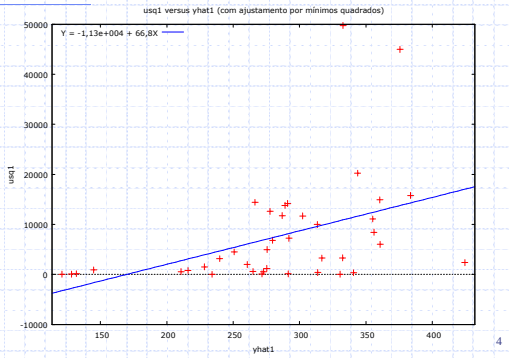
Exemplo de heterocedasticidade



Gráficos residuais

- Estimar o modelo usando MQO e fazer os gráficos dos resíduos relacionando-os com as variáveis explicativas ou com o valor estimado de *y*.
- Se erros são heterocedásticos – relação sistemática.

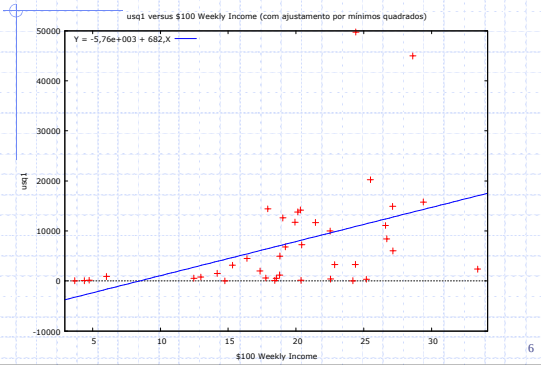
Gráficos residuais



Gráficos residuais

- No exemplo, o exame dos resíduos ao quadrado (proxies dos termos de erro), é feito quando eles são plotados junto com o *y* estimado.
- Verifica-se que o valor médio estimado de y se relaciona de forma sistemática com os resíduos elevados ao quadrado.

Gráficos residuais



Gráficos residuais

- No exemplo acima, o exame dos resíduos ao quadrado é feito quando eles são plotados junto com a variável explicativa x.
- Verifica-se que o valor médio estimado de y se relaciona de forma sistemática com os resíduos elevados ao quadrado.
- Para RLS, o gráfico é idêntico ao anterior. Para RLM, plota-se para cada uma das variáveis explicativas.

7

Teste Goldfeld-Quandt

- Separar a amostra em duas subamostras aproximadamente iguais.
- Uma amostra com variância grande σ_1 e outra com variância pequena σ_2.
- Calcular a variância estimada do erro para cada subamostra.
- Se a hipótese nula de variâncias iguais não é verdade, esperamos que a razão de $\sigma_{1/} \sigma_2$ seja grande.

8

Teste GQ

$$GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

$$H_0 : \sigma_t^2 = \sigma^2$$

$$H_1 : \sigma_t^2 = \sigma^2 x_t$$

- Rejeito a hipótese nula quando GQ>Fc.
- Fc é o valor crítico da distribuição F com (T1-K) e (T2-K) graus de liberdade.

- **Exemplo**: GQ = 2285,9/682,46 = 3,35 > Fc=2,22.
- Rejeitamos a hipótese nula e concluímos que há heterocedasticidade.

9

Mínimos quadrados ponderados

- O estimador de *White* para os erros padrão ajuda a melhorar o problema de inferências incorretas das estimativas de MQO com heterocedasticidade.
- Podemos obter um estimador melhor do que o MQO, mas depende de como modelamos a heterocedasticidade.

10

Mínimos quadrados ponderados (cont.)

- A idéia é minimizar a soma dos quadrados ponderados por $1/h_i$.
- Dá-se menos peso para as observações com maior variância.
- **No gretl temos que informar esta função que pondera as observações**.

11

Exemplo

- Banco de dados: food.gdt
- Relacionar gastos em alimentos com a renda mensal
- Definir a variável peso: 1/x (inverso da renda).
- Cada variável, inclusive a constante, é multiplicada pela raiz quadrada do peso.
- Modelo: outros modelos lineares\mínimos quadrados ponderados

12

Exemplo

Estimativas WLS
Variável dependente: y

	Modelo 1
const	78,68**
	(23,79)
x	10,45**
	(1,386)
n	40
R^2	0,5994
lnL	-172,980

Erros padrão entre parênteses
* indica significância num nível de 10 por cento
** indica significância num nível de 5 por cento

MQGF

- Heterocedasticidade tem uma forma específica.
- *Heterocedasticidade multiplicativa:*

$$\sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_i)$$

- *A variável explicativa z determina como a variância muda para cada observação. Defina* $z_i = ln(x_i)$
- Estima os parâmetros por MQO e usa o valor predito para calcular o peso.

MQGF (cont.)

- A estimativa de *h* é obtida por $\hat{h} = \exp(\hat{g})$; o peso será o inverso dessa estimativa.
- Resumindo:
- Faça a regressão por MQO da equação original, salve os resíduos, $\hat{u}$, eleve-os ao quadrado e tire o log.
- Faça a regressão de $\ln(\hat{u}^2)$ em todas as variáveis independentes o obtenha o valor ajustado $\hat{g}$.
- Faça a regressão por MQP utilizando $1/\exp(\hat{g})$ como ponderador.

Exemplo

$$\sigma_i^2 = \exp(\alpha_1 + \alpha_2 z_i)$$

$$z_i = ln(x_i)$$

$$\ln(\sigma_i^2) = \alpha_1 + \alpha_2 z_i + v_i$$

$$\ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 z_i + v_i$$

Exemplo

Modelo 1:Mínimos Quadrados (OLS), usando as observações 1-40
Variável dependente: y

	Coeficiente	*Erro Padrão*	*razão-t*	*p-valor*	
const	83,416	43,4102	1,9216	0,06218	*
x	10,2096	2,09326	4,8774	0,00002	***

Média var. dependente	283,5735	D.P. var. dependente	112,6752
Soma resíd. quadrados	304505,2	E.P. da regressão	89,51700
R-quadrado	0,385002	R-quadrado ajustado	0,368818
F(1, 38)	23,78884	P-valor(F)	0,000019
Log da verossimilhança	-235,5088	Critério de Akaike	475,0176
Critério de Schwarz	478,3954	Critério Hannan-Quinn	476,2389

Salva os resíduos ao quadrado e tira o log destes.

Exemplo

$$\ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 z_i + v_i$$

- Tira exponencial dos valores preditos (acha h).
- Ache o peso (w=1/h)

Modelo 2:Mínimos Quadrados (OLS), usando as observações 1-40
Variável dependente: l_usq2

	Coeficiente	*Erro Padrão*	*razão-t*	*p-valor*	
const	0,937796	1,58311	0,5924	0,55711	
l_x	2,32924	0,541336	4,3028	0,00011	***

Média var. dependente	7,648159	D.P. var. dependente	2,071519
Soma resíd. quadrados	112,5310	E.P. da regressão	1,720855
R-quadrado	0,327597	R-quadrado ajustado	0,309903
F(1, 38)	18,51375	P-valor(F)	0,000114
Log da verossimilhança	-77,44452	Critério de Akaike	158,8890
Critério de Schwarz	162,2668	Critério Hannan-Quinn	160,1103

Exemplo

$$\hat{y} = \underset{(9.7135)}{76.0538} + \underset{(0.97151)}{10.6335}\, x$$

$$T = 40 \quad \bar{R}^2 = 0.7529 \quad F(1,38) = 119.8 \quad \hat{\sigma} = 1.5467$$

- Rode MQP

Modelo 3: WLS, usando as observações 1-40
Variável dependente: y
Variável usada como peso: w

	Coeficiente	*Erro Padrão*	*razão-t*	*p-valor*	
const	76,0538	9,71349	7,8297	<0,00001	***
x	10,6335	0,971514	10,9453	<0,00001	***

Estatísticas baseadas nos dados ponderados:

Soma resíd. quadrados	90,91135	E.P. da regressão	1,546740
R-quadrado	0,759187	R-quadrado ajustado	0,752850
F(1, 38)	119,7991	P-valor(F)	2,62e-13
Log da verossimilhança	-73,17765	Critério de Akaike	150,3553
Critério de Schwarz	153,7331	Critério Hannan-Quinn	151,5766

Estatísticas baseadas nos dados originais:

Média var. dependente	283,5735	D.P. var. dependente	112,6752
Soma resíd. quadrados	304869,6	E.P. da regressão	89,57055

Exemplo

Variável dependente: y

	Modelo 1 WLS h=x	Modelo 2 WLS h=exp(g)	Modelo 3 Mínimos Quadrados (OLS)
const	78,68**	76,05**	83,42*
	(23,79)	(9,713)	(43,41)
x	10,45**	10,63**	10,21**
	(1,386)	(0,9715)	(2,093)
n	40	40	40
R^2	0,5994	0,7592	0,3850
lnL	-172,980	-73,178	-235,509

Erros padrão entre parênteses
* indica significância num nível de 10 por cento
** indica significância num nível de 5 por cento

Observações sobre MQP

- Lembre-se que utiliza-se MQP apenas por eficiência, pois MQO continua não tendencioso e consistente.
- As estimativas serão diferentes devido a erros amostrais, mas se forem muito diferentes, então alguma outra hipótese de Gauss-Markov também deve estar sendo violada.

Multicolinearidade

Multicolinearidade

- Quando existem relação linear exata entre as variáveis independentes será impossível calcular os estimadores de MQO.
- O procedimento MQO utilizado para estimação não será efetivado.
- Mensagem: "matriz quase singular" (uma matriz quase singular X´X não pode ser invertida) ou "a variável x_k dropped".
- ***Relacionamento linear exato***: só quando os dados foram construídos pelo pesquisador, pe., no caso de inclusão de dummies.
- ***Relacionamento linear aproximado*** entre as variáveis independentes: comuns em economia.
- O procedimento de estimação não é rompido quando as variáveis são bastante correlacionadas, contudo, surgem problemas de estimação.

Multicolinearidade

- Multicolinearidade: nome dado ao fenômeno de presença de relação linear aproximada entre os regressores.
- Problema de estimação causado por uma amostra particular. Não é um problema teórico.
- Multicolinearidade nos dados pode existir por diferentes motivos:
  - Regressores possuem a mesma tendência temporal.
  - Algumas variáveis caminham na mesma direção porque os dados não foram coletados de uma base grande.
  - Pode existir realmente algum tipo de relacionamento aproximado entre os regressores.

Variância do estimador de MQO

A variância estimada de b_k é
Var[b_2/**X**] =

$$\frac{s^2}{\left[(1-R_{2.}^2)\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2\right]} = \frac{s^2}{(1-R_{2.}^2)S_{22}}.$$

Quanto maior o fit da regressão de $\mathbf{x}_2$ em $\mathbf{X}_1$, maior a variância. No limite, um ajuste perfeito produz uma variância infinita.

Variância do estimador de MQO Forma mais geral

Defina a matriz **X** que contém uma constante e K-1 variáveis explicativas

A variância estimada de b_k é
Var[b_k/**X**] =

$$\frac{s^2}{\left(1-R_{k.}^2\right)\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}$$

Ingrediente para existência de multicolinearidade:
- Quanto maior a correlação entre xk e as outras variáveis (R^2_k).

Consequências da Multicolinearidade

- O estimador de MQO permanece **não viesado** e **BLUE.**
- O grau de ajuste não é afetado.
- Problemas práticos:
  - Pequenas mudanças nos dados produzem grandes variações nas estimativas dos parâmetros.
  - Os coeficientes estimados apresentam erros padrão muito elevados e baixos níveis de significância, mesmo que sejam conjuntamente significativos e com o grau de ajuste da regressão elevado (R^2).
  - Os coeficientes podem ter o sinal "errado" e magnitudes irreais.

Consequências da Multicolinearidade

- Na presença de multicolinearidade, o procedimento de estimação MQO não recebe variação independente suficiente de uma variável para realizar o cálculo com confiança do efeito que esta tem sobre a variável dependente.
- Quando os regressores são altamente correlacionados, a maior parte da sua variação é comum às duas variáveis, deixando pouca variação exclusiva a cada variável.
- MQO tem pouca informação para usar ao fazer as estimativas do coeficiente (similar a um problema de amostra pequena ou que a variável não mudasse muito).

Consequências da Multicolinearidade

- As variâncias dos estimadores MQO dos parâmetros são muito grandes – **Imprecisão dos estimadores dos parâmetros.**
- **Erros de especificação :** não sabemos qual variável é mais ou menos importante para explicar a variação da variável dependente.

Como detectar?

- Controvérsia: muitos métodos inadequados.
- Sinais hipotéticos não são encontrados.
- Variáveis consideradas *a priori* importantes não são significativas individualmente, mas estatística F (significância coletiva) é alta.
- Resultados alterados quando uma variável independente é excluída ou quando uma observação é retirada.
- Matriz de correlação (0,8 a 0,9 são valores absolutos altos): detecta colinearidade de duas variáveis, mas não de mais de duas.

Como detectar?

- Índice de condição dos dados (IC):
  - Raiz quadrada da razão da maior para a menor raiz característica de X´X

$$\gamma = \left[\frac{raizmáxima}{raizmínima}\right]^{1/2}$$

  - Medida de sensibilidade das estimativas a pequenas pertubações dos dados.
  - Medida de proximidade de X´X da singularidade (multicolinearidade perfeita): quanto maior o IC maior dificuldade em inverter a matriz.
  - Índice maior que 20 indica colinearidade forte: mudança de 1% nos dados faz surgir uma mudança de IC% nos estimadores.

Como detectar?

- Inverso da matriz de correlação:
  - Elementos na diagonal: Fatores de inflação da variância (VIF).

$$VIF = \frac{1}{(1 - R_{k.}^2)}$$

R^2 da regressão da k-ésima variável independente em todas demais variáveis independentes.

  - Quanto maior VIF, mais o R^2_k está perto da unidade.
  - Medida da quantidade pela qual a variância da k-ésima estimativa do coeficiente é aumentada devido a associação linear com as outras variáveis explicativas.
  - Se VIF > 10: presença de colinearidade

No stata:

```
. reg ln_sal_hora  filho idade idade2 sexo educa

      Source |       SS       df       MS              Number of obs =   14537
-------------+------------------------------           F(  5, 14531) = 1939.23
       Model |   5434.065     5    1086.813            Prob > F      =  0.0000
    Residual | 8143.68463 14531  .560435251            R-squared     =  0.4002
-------------+------------------------------           Adj R-squared =  0.4000
       Total | 13577.7496 14536  .934077438            Root MSE      =  .74862

------------------------------------------------------------------------------
 ln_sal_hora |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       filho |   -.209508   .0202922   -10.32   0.000    -.2492833   -.1697328
       idade |   .0604502   .0028366    21.31   0.000       .05489    .0660103
      idade2 |  -.0005105   .0000332   -15.37   0.000    -.0005756   -.0004454
        sexo |   -.346604   .0129488   -26.77   0.000    -.3719854   -.3212227
       educa |   .1304724   .0014665    88.97   0.000     .1275979    .1333469
       _cons |  -.4814204    .061482    -7.83   0.000     -.601933   -.3609078
------------------------------------------------------------------------------

. vif

    Variable |       VIF       1/VIF
-------------+----------------------
       idade |     33.37    0.029969
      idade2 |     30.63    0.032650
       filho |      1.63    0.611927
       educa |      1.11    0.901243
        sexo |      1.04    0.961969
-------------+----------------------
    Mean VIF |     13.56
```

No stata:

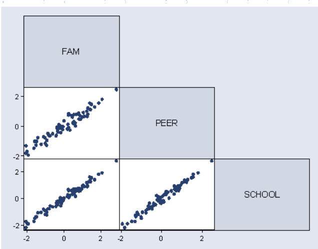
```
. collin  idade idade2 sexo educa filho

  Collinearity Diagnostics

                        SQRT                   R-
  Variable      VIF     VIF    Tolerance    Squared
----------------------------------------------------
     idade     27.65    5.26    0.0362      0.9638
    idade2     19.99    4.47    0.0500      0.9500
      sexo      1.02    1.01    0.9843      0.0157
     educa      1.55    1.25    0.6449      0.3551
     filho      3.11    1.76    0.3218      0.6782
----------------------------------------------------
  Mean VIF     10.66

                           Cond
        Eigenval          Index
---------------------------------
    1     4.3513          1.0000
    2     1.0883          1.9996
    3     0.3723          3.4187
    4     0.1424          5.5283
    5     0.0395         10.4905
    6     0.0063         26.3514
---------------------------------
 Condition Number        26.3514
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.0194
```

No stata: graph matrix fam peer school



Multicolinearidade

Não existe "cura" para a colinearidade.

1. Exclusão de variáveis: eliminar as variáveis que causam o problema – impor na regressão a hipótese de que a variável problemática não deve aparecer no modelo. **Possível problema de especificação.**
2. Obtenção de mais dados: dados adicionais e tamanho da amostra.
3. Formalizar os relacionamentos entre os regressores: equações simultâneas.
4. Especificar o relacionamento entre alguns parâmetros: dois parâmetros iguais ou que a soma das elasticidades deve ser igual a um, etc.
5. Análise componente principal: as variáveis colineares poderiam ser agrupadas para formar um índice composto capaz de representar este conjunto de variáveis. Variável só pode ser criada se tiver uma interpretação econômica.