ESTATÍSTICA II Ec/Fi

Econometria – parte 2

Introdução

Para fazer inferência sobre o MRLM, foi necessário introduzir a hipótese MRL 6 referente à distribuição normal.

Na 1ª parte da UC, em Estatística, quando se passou de populações normais para "grandes amostras" recorreu-se ao Teorema do Limite Central.

Agora, no quadro do MRLM vamos fazer algo semelhante e levantar a hipótese MRL 6 no quadro das grandes amostras.

Aproveitaremos o tratamento de grandes amostras para apresentar algumas propriedades assintóticas dos estimadores OLS, nomeadamente a consistência.

Em termos da UC de Estatística 2 apenas cobriremos uma versão "light" do capítulo 5.

Consistência do estimador OLS

Assumindo as hipóteses MRL 1 a MRL 4, mostra-se que $\hat{\beta}$, estimador OLS de β **é consistente**, isto é, tende em probabilidade para o verdadeiro valor β:

$$\forall \varepsilon > 0, \lim_{n \to \infty} Pr\left(\left|\hat{\beta}_j - \beta_j\right| < \varepsilon\right) = 1 \quad \text{para } j = 0,1,\dots,k$$

o que se resume escrevendo

$$plim\left(\hat{\beta}\right) = \beta$$

Demonstração no Apêndice E do Wooldridge.

Tecnicamente, tem de se garantir que $var\left(x_j\right) < \infty$.

Omissão de variáveis relevantes e estimador OLS

A existência de correlação entre o termo de erro, u, e qualquer das variáveis x_j incluídas no modelo origina a inconsistência de todos os $\hat{\beta}_j$ para além do enviesamento do estimador.

Este ponto é de grande importância já que mostra que o enviesamento não irá desaparecer com o aumento da amostra.

A hipótese MRL 4 vai para além de garantir esta não correlação.

Não se verificando a hipótese fala-se de endogeneídade por oposição a exogeneídade (quando a hipótese se verifica).

Omissão de variáveis relevantes e estimador OLS

A omissão de uma variável relevante no modelo – violação de MRL 1 – (o seu efeito passará a estar incluído em u), se correlacionada com alguma das variáveis incluídas, originará assim a inconsistência do estimador OLS o que é bastante mais grave do que a inclusão de variáveis irrelevantes (que só originam a perda de eficiência do estimador). Esta é uma razão para manter variáveis de controle que não se mostram estatisticamente significantes.

Variáveis explicativas omitidas e variáveis explicativas irrelevantes

Aquando da escolha dos regressores a incluir no modelo, deve-se ter em conta que:

- Regressores em excesso geram redução de eficiência
- Omissão de regressores (tendo em conta que os regressores omitidos estão incluídos no erro u) gera:
  - Inconsistência se $E(U|X) \neq 0 \rightarrow$ endogeneidade
  - Consistência se $E(U|X) = 0 \rightarrow$ exogeneidade

Teste assintótico para um β_j

Em termos da distribuição por amostragem do estimador, podendo-se aplicar o TLC, não existem grandes problemas a não ser que as distribuições passam a ser aproximadas.

Nalguns casos uma transformação da variável y pode ajudar.

Teste a coeficientes individuais:

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \xrightarrow{a} N(0,1)$$

Embora também se possa utilizar $t_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \xrightarrow{a} t(n-k-1)$

Observação: Esta abordagem também se aplica, sem dificuldade quando se quer fazer inferência sobre uma combinação linear de β_j.

$$t_j = \frac{\hat{\delta} - \delta}{\hat{\sigma}_{\hat{\delta}}} \xrightarrow{a} N(0,1) \quad \text{ou} \quad t_j = \frac{\hat{\delta} - \delta}{\hat{\sigma}_{\hat{\delta}}} \xrightarrow{a} t(n-k-1)$$

Testes assintóticos para um conjunto de coeficientes

- Pode utilizar-se o teste F, apesar da distribuição ser apenas aproximada (neste caso a aproximação é mais lenta do que com a *t-Student*)
- Pode-se aplicar o **teste LM** (*Lagrange multiplier*). Este teste apenas envolve o modelo restrito. Supondo q restrições
  - Estimar a regressão assumindo H_0 verdadeira e guardar os resíduos desta regressão que se designarão por $\tilde{u}$ (para os diferenciar dos resíduos do modelo sem restrição)
  - Estimar a regressão de $\tilde{u}$ em **todas** as variáveis x_j e obter o coeficiente de determinação desta regressão, R_u^2.
  - A estatística de teste será $LM = nR_u^2 \xrightarrow{a} X_q^2$
  - **Região de rejeição:** aba direita da distribuição *Qui-quadrado*

Teste LM - exemplo

Exemplo: Retoma-se o exemplo dos salários dos jogadores de baseball que se fez no tópico de "teste à significância conjunta de *q* coeficientes") e vai testar-se $H_0: \beta_3 = \beta_4 = \beta_5 = 0$ *vs* H_1: H_0 falsa (pelo menos um dos $\beta \neq 0$)

Estimação do modelo com a restrição

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.772704072					
R Square	0.597071582					
Adjusted R Square	0.594769134					
Standard Error	0.752731258					
Observations	353					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	293.8640431	146.9320216	259.3203218	8.2202E-70	
Residual	350	198.3115214	0.566604347			
Total	352	492.1755645				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	11.22380399	0.108312013	103.6247386	9.966E-265	11.01077971	11.43682826
years	0.071317962	0.012505011	5.703150453	2.50422E-08	0.046723543	0.095912381
gamesyr	0.02017448	0.00134287	15.02340897	1.01913E-39	0.017533371	0.022815589

Seguidamente estima-se a regressão dos resíduos deste modelo no conjunto de todas as variáveis

Teste LM - exemplo

Cujo output foi

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.276169921					
R Square	0.076269825					
Adjusted R Square	0.062959592					
Standard Error	0.726577259					
Observations	353					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	5	15.12518511	3.025037023	5.730164532	4.26556E-05	
Residual	347	183.1863363	0.527914514			
Total	352	198.3115214				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.031383232	0.288822864	-0.1086591	0.913535688	-0.599446973	0.536680508
years	-0.002455338	0.012114544	-0.202676916	0.839506351	-0.026282514	0.021371837
gamesyr	-0.007622368	0.002646763	-2.879882894	0.004225253	-0.012828086	-0.002416651
bavg	0.000978594	0.001103509	0.886802204	0.375799595	-0.001191814	0.003149002
hrunsyr	0.014429519	0.01605698	0.898644642	0.369465108	-0.017151734	0.046010772
rbisyr	0.01076574	0.007174962	1.50045958	0.134404707	-0.003346147	0.024877626

$LM_{obs} = 353 \times 0.076269825 = 26.92 \quad q_{0.05} = 7.814$ (3 g.l)

$p - value = P(Q \geq 26.92) = 6.11E - 06$ Rejeita-se $H_0: \beta_3 = \beta_4 = \beta_5 = 0$

A mesma conclusão que se tinha tirado com o teste F

Teste LM para os k declives

Neste caso o teste é quase imediato (alguns softwares reportam mesmo o seu resultado)

$$LM = nR^2 \sim X_k^2$$

Em que R^2 é o coeficiente de determinação do modelo. Como é evidente, a região de rejeição mantém-se na cauda direita da distribuição *Qui-quadrado*

Exemplo:

$LM.obs = 353 \times 0.6278$

$= 221.6$

$q_{0.05} = 11.07$

$p - value \approx 0$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.792340118
R Square	0.627802862
Adjusted R Square	0.622439791
Standard Error	0.726577259
Observations	353

ANOVA

	df	SS	MS	F	Significance F
Regression	5	308.9892282	61.79784565	117.0603271	2.93802E-72
Residual	347	183.1863363	0.527914514		
Total	352	492.1755645			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	11.19242075	0.288822864	38.75185162	4.1871E-128	10.62435701	11.76048449
years	0.068862623	0.012114544	5.684293499	2.7876E-08	0.045035448	0.092689799
gamesyr	0.012552112	0.002646763	4.742438528	3.08865E-06	0.007346394	0.017757829
bavg	0.000978594	0.001103509	0.886802204	0.375799595	-0.001191814	0.003149002
hrunsyr	0.014429519	0.01605698	0.898644642	0.369465108	-0.017151734	0.046010772
rbisyr	0.01076574	0.007174962	1.50045958	0.134404707	-0.003346147	0.024877626

Propriedades do estimador OLS - Síntese

Hipóteses:

1. Modelo linear nos parâmetros: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$
2. Amostra aleatória
3. Ausência de colinearidade perfeita
4. Exogeneidade: $E(u|\boldsymbol{x}) = 0$
5. Homoscedasticidade: $var(u|\boldsymbol{x}) = \sigma^2$
6. Normalidade do erro: $u \sim Normal(0, \sigma^2)$

Propriedades dos Estimadores

Pequenas amostras	Propriedades assimptóticas
1-4: estimadores centrados	1-4: estimadores consistentes
1-5: estimadores centrados e eficientes	1-5: estimadores consistentes, eficientes e com dist aprox normal
1-6: estimadores centrados, eficientes e normalmente distribuídos	

Complementos sobre a forma funcional

Este tópico cobre o essencial das secções 6.1, 6.2 e 6.3 e também a secção 9.1, juntando vários pontos que se podem mostrar interessante em termos da modelação de determinado fenómeno. Destacam-se 5 pontos:

- Efeitos da alteração de escala (numa ou em mais variáveis) no MRLM
- Uma segunda leitura das alterações logarítmicas
- Introdução de termos quadráticos na regressão
- Introdução de termos de interação no MRLM
- Teste RESET

Alteração de escala

Modelo: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

Modelo estimado: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1\, x_1 + \cdots + \hat{\beta}_k\, x_k$

A pergunta a que se irá responder é muito simples: O que acontece aos coeficientes do modelo quando se alteram as unidades (alteração linear) em que são medidas uma ou mais das variáveis?

Exemplo: Regresse-se ao exemplo referente ao preço de um imóvel onde se tinha

$$\widehat{preço} = -19.286 + 1.3836\ area + 15.121\ quartos$$

Qual o efeito nos resultados do modelo (coeficientes e demais estatísticas de interesse) de se medir a área em pés-quadrados (em vez de m2) ou o preço em dólares (em vez de milhares de dólares)?

Alteração em y

Alteração em y → $y^* = c\,y$

Modelo inicial: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

Novo modelo: $y^* = c\,y = c\,(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u)$

$$= c\,\beta_0 + c\,\beta_1 x_1 + \cdots + c\,\beta_k x_k + c\,u$$

$$= \beta_0^* + \beta_1^* x_1 + \cdots + \beta_k^* x_k + u^*$$

em que $\beta_j^* = c\,\beta_j$ $(j = 0,1,\ldots,k)$ e $u^* = c\,u$

Repare-se que $E(u^*) = E(c\,u) = cE(u) = 0$ MRL 4 mantém-se

$$var(u^*) = var(c\,u) = c^2\,var(u) = c^2\,\sigma^2$$

a variância vem multiplicada por c^2 mas a essência de MRL 5 mantém-se.

Tal como seria de esperar ir-se-á obter $\hat{\beta}_j^* = c\,\hat{\beta}_j$ e $\hat{u}_i^* = c\,\hat{u}_i$ e portanto $\hat{\sigma}_{\hat{\beta}_j}^* = c\,\hat{\sigma}_{\hat{\beta}_j}$, $SST^* = c^2 SST$, $SSE^* = c^2 SSE$ e $SSR^* = c^2 SSR$, mantendo-se inalterados quer o coeficiente de correlação quer os coeficientes de determinação.

Alteração em y

Exemplo: Regressem ao exemplo e meçam o preço dos imóveis em dólares (isto é, multipliquem por 1000 os valores de y_i) e voltem a estimar o modelo para verificarem os resultados apresentados, nomeadamente que $R^2 = 0.6319$ e que $\hat{\beta}_1 = 1383.6$.

Suponham agora que a alteração em y era dada por $y^* = y - c$.

O que aconteceria?

Alteração em y - exemplo

Regressão original (preço em 10^3dólares, área em m^2)

$$\widehat{preço} = -19.2855 + \mathbf{1.3836}\ area + 15.1213\ quartos$$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.794906945					
R Square	0.63187705					
Adjusted R Square	0.623215334					
Standard Error	63.04837628					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	579971.1994	289985.5997	72.95055817	3.58672E-19	
Residual	85	337883.3088	3975.097751			
Total	87	917854.5083				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-19.28550028	31.0475285	-0.621160563	0.536156232	-81.0163048	42.44530424
area(m2)	1.38360615	0.148943494	9.289470184	1.40049E-14	1.08746658	1.67974572
quartos	15.12133684	9.488597692	1.593632413	0.114730383	-3.744537442	33.98721112

Alteração em y - exemplo

Alteração 1 (preço em dólares, área em m^2)

$$\widehat{preço} = -19285.5 + 1383.6\ area + 15121.3\ quartos$$

SUMMARY OUTPUT Preço em dolares

Regression Statistics	
Multiple R	0.794906945
R Square	0.63187705
Adjusted R Square	0.623215334
Standard Error	63048.37628
Observations	88

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	5.79971E+11	2.89986E+11	72.95055817	3.58672E-19
Residual	85	3.37883E+11	3975097751		
Total	87	9.17855E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-19285.50028	31047.5285	-0.621160563	0.536156232	-81016.3048	42445.30424
area	1383.60615	148.9434943	9.289470184	1.40049E-14	1087.46658	1679.74572
quartos	15121.33684	9488.597692	1.593632413	0.114730383	-3744.537442	33987.21112

Alteração em y - exemplo

Alteração 2 (preço em 10^3dólares descontando 100 (mil dólares), área em m^2)

$$\widehat{preço} = -119.2855 + 1.3836\ area + 15.1213\ quartos$$

SUMMARY OUTPUT Preço em milhares de dolares (descontando 100)

Regression Statistics	
Multiple R	0.794906945
R Square	0.63187705
Adjusted R Square	0.623215334
Standard Error	63.04837628
Observations	88

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	579971.1994	289985.5997	72.95055817	3.58672E-19
Residual	85	337883.3088	3975.097751		
Total	87	917854.5083			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-119.2855003	31.0475285	-3.842028852	0.000234516	-181.0163048	-57.5546958
area	1.38360615	0.148943494	9.289470184	1.40049E-14	1.08746658	1.67974572
quartos	15.12133684	9.488597692	1.593632413	0.114730383	-3.744537442	33.98721112

Alteração em x_j

Alteração em $x_j \rightarrow x_j^* = c\ x_j$ e façamos, sem perda de generalidade, $j = 1$

Modelo inicial: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

Novo modelo: $y = \beta_0 + \beta_1 \left(\frac{x_1^*}{c}\right) + \cdots + \beta_k x_k + u$

$$= \beta_0 + (\beta_1/c) x_1^* + \cdots + \beta_k x_k + u$$

$$= \beta_0 + \beta_1^*\ x_1^* + \cdots + \beta_k x_k + u$$

em que $\beta_1^* = \frac{\beta_1}{c}$.

Apenas se altera o parâmetro β_1.

Em termos dos parâmetros estimados teremos apenas 2 alterações: a primeira, mais evidente, $\hat{\beta}_1^* = \frac{\hat{\beta}_1}{c}$ mas também

$\hat{\sigma}_{\hat{\beta}_1}^* = \hat{\sigma}_{\hat{\beta}_1}/c$ não se alterando o rácio t. Uma alteração de escala deste tipo nunca altera a significância estatística da variável.

Alteração em x_j exemplo

Exemplo: sabendo que um m2 equivale a cerca de 10.764 pés quadrados multiplique x_i por este valor (deve encontrar a menos de erros de arredondamento os dados originais do Wooldridge) e comprove que apenas $\hat{\beta}_1$ e $\hat{\sigma}_{\hat{\beta}_1}$ se alteram.

SUMMARY OUTPUT

Preço em milhares de dolares

Área em pés quadrados

Regression Statistics	
Multiple R	0.794906945
R Square	0.63187705
Adjusted R Square	0.623215334
Standard Error	63.04837628
Observations	88

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	579971.1994	289985.5997	72.95055817	3.58672E-19
Residual	85	337883.3088	3975.097751		
Total	87	917854.5083			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-19.28550028	31.0475285	-0.621160563	0.536156232	-81.0163048	42.44530424
area(pé^2)	0.128540148	0.013837188	9.289470184	1.40049E-14	0.10102811	0.156052185
quartos	15.12133684	9.488597692	1.593632413	0.114730383	-3.744537442	33.98721112

Coeficientes beta

O Wooldridge aborda ainda os chamados coeficientes beta que consistem em estimar a regressão depois de estandardizar todas as variáveis envolvidas (y e x_j para $j = 1,2,\dots,k$) em termos amostrais.

Estandardização: $x_{ij}^{*} = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}}$

O modelo transformado não inclui termo independente já que a estandardização da variável associada com o termo independente origina o valor 0 para cada i.

Estes modelos são utilizado quando é mais interessantes interpretar as variações das variáveis em termos dos seus desvios padrões do que nas unidades em que são medidas.

Este ponto não nos vai interessar diretamente e portanto iremos abandoná-lo mas podem ler a secção no livro. Não envolve nada de complicado.

Complemento sobre o uso de logaritmos

No capítulo 3 interpretaram-se os parâmetros do modelo (embora de forma aproximada) no quadro dos modelos log-log e log-lin. Para recordar a situação, considere-se o modelo

$$\widehat{\ln y} = \hat{\beta}_0 + \hat{\beta}_1 \ln x_1 + \hat{\beta}_2 x_2$$

$\hat{\beta}_1$ – uma variação de 1% em x_1 origina, tudo o resto constante, uma variação percentual aproximada de $\hat{y}$ dada por $\hat{\beta}_1$

$$\%\Delta x_1 = 1 \rightarrow \%\Delta\hat{y} \cong \beta_1\%$$

$\hat{\beta}_2$ – uma variação unitária de x_2 origina, tudo o resto constante, uma variação percentual aproximada de $\hat{y}$ dada por $100\ \hat{\beta}_2$

$$\Delta x_2 = 1 \rightarrow \%\Delta\hat{y} \cong 100\ \hat{\beta}_2\%$$

Em qualquer dos casos a aproximação é valida para pequenas variações de y, nomeadamente para valores pequenos de $\hat{\beta}_j$.

O propósito é estabelecer, nestes 2 casos, uma fórmula exata.

Complemento sobre o uso de logaritmos

Caso da semi-elasticidade constante $\hat{\beta}_2$

Modelo inicial → $\widehat{\ln y} = \hat{\beta}_0 + \hat{\beta}_1 \ln x_1 + \hat{\beta}_2 x_2$

Modelo pós-incremento → $\widehat{\ln y^*} = \hat{\beta}_0 + \hat{\beta}_1 \ln x_1 + \hat{\beta}_2 (x_2 + \Delta x_2)$

Como $\widehat{\ln y^*} - \widehat{\ln y} = \widehat{\ln\left(\frac{y^*}{y}\right)} = \widehat{\ln\left(\frac{y+\Delta y}{y}\right)} = \ln\left(1 + \frac{\Delta\hat{y}}{\hat{y}}\right)$,

donde $\widehat{\ln y^*} - \widehat{\ln y} = \hat{\beta}_2 \Delta x_2 \Leftrightarrow \ln\left(1 + \frac{\Delta\hat{y}}{\hat{y}}\right) = \hat{\beta}_2 \Delta x_2$

$$1 + \frac{\Delta\hat{y}}{\hat{y}} = e^{\hat{\beta}_2 \Delta x_2} \Leftrightarrow \frac{\Delta\hat{y}}{\hat{y}} = e^{\hat{\beta}_2 \Delta x_2} - 1$$

Isto é

$$\%\Delta\hat{y} = 100(e^{\hat{\beta}_2 \Delta x_2} - 1)$$

$$\Delta x_2 = 1 \rightarrow \%\Delta\hat{y} = 100(e^{\hat{\beta}_2} - 1)$$

quanto menor fosse $\hat{\beta}_j$.

Vamos agora alargar horizontes e definir esta variação quando $\hat{\beta}_j$ não é tão pequeno assim.

Complemento sobre o uso de logaritmos

Caso da elasticidade constante $\hat{\beta}_1$

Modelo inicial → $\widehat{\ln y} = \hat{\beta}_0 + \hat{\beta}_1 \ln x_1 + \hat{\beta}_2 x_2$

Modelo pós-incremento → $\widehat{\ln y^*} = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_1 + \Delta x_1) + \hat{\beta}_2 x_2$

$$\widehat{\ln y^*} - \widehat{\ln y} = \hat{\beta}_1(\ln(x_1 + \Delta x_1) - \ln x_1)$$

$$\ln\left(1 + \frac{\Delta \hat{y}}{\hat{y}}\right) = \hat{\beta}_1 \ln\left(\frac{x_1 + \Delta x_1}{x_1}\right) = \hat{\beta}_1 \ln\left(1 + \frac{\Delta x_1}{x_1}\right) = \ln\left(1 + \frac{\Delta x_1}{x_1}\right)^{\widehat{\beta_1}}$$

Logo $1 + \frac{\Delta \hat{y}}{\hat{y}} = \left(1 + \frac{\Delta x_1}{x_1}\right)^{\widehat{\beta_1}}$ ou seja $\frac{\Delta \hat{y}}{\hat{y}} = \left(1 + \frac{\Delta x_1}{x_1}\right)^{\widehat{\beta_1}} - 1$

E portanto $\%\Delta \hat{y} = 100\left(\left(1 + \frac{\Delta x_1}{x_1}\right)^{\widehat{\beta_1}} - 1\right)$

$$\%\Delta x_1 = 1 \rightarrow \%\Delta \hat{y} = 100(1.01^{\widehat{\beta_1}}\text{-}1)$$

Complemento sobre o uso de logaritmos

Qualidade da aproximação: Depende de $\hat{\beta}_j$ e de Δx ou $\%\Delta x$

½ Elasticidade Constante Δx=1				Elasticidade Constante $\%\Delta x$=1		
$\hat{\beta}$	$\%\Delta y$ apox $100 \times \hat{\beta}$	$\%\Delta y$		$\hat{\beta}$	$\%\Delta y$ apox $\hat{\beta}$	$\%\Delta y$
-0.05	-5	-4.877		-0.05	-0.05	-0.050
0.05	5	5.127		0.05	0.05	0.050
0.1	10	10.517		0.1	0.1	0.100
0.3	30	34.986		0.3	0.3	0.299
0.5	50	64.872		0.5	0.5	0.499
0.7	70	101.375		0.7	0.7	0.699
1	100	171.828		1	1	1.000

A aproximação é bem mais robusta no modelo log-log

Complemento sobre o uso de logaritmos

Exemplo:

$$\widehat{ln(preço)} = 1.289 + 0.810\ ln(area) + 0.038\ quartos$$

Para uma variação de 5% da área do imóvel a aproximação aponta para uma variação percentual de 0.810×5, isto é, 4.05% enquanto o valor exato origina, em %, $100 \times (1.05^{0.810} - 1)$, isto é 4.03%

Para uma variação de 2 quartos a aproximação indica uma variação 7.6% no preço enquanto o valor percentual exato será $100 \times (e^{0.038 \times 2} - 1)$ ou seja 7.9%.

Como acontece em muitas situações práticas as aproximações funcionam bem.

Logaritmizar ou não as variáveis

O livro refere algumas **regras práticas** para logaritmizar (ou não) as variáveis sublinhando que nenhuma delas é *"written in stone".*

Assim as afirmações que seguem devem ser tomadas como hipóteses a estudar em cada caso.

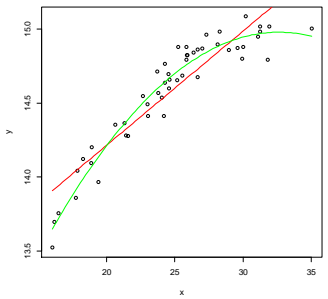
- *Logaritmizar* variáveis expressas em unidades monetárias como salários, preços, vendas, ….
- *Logaritmizar* contagens (valores inteiros elevados) como populações, nº de empregados (grandes empresas) , nº de estudantes numa universidade, ….
- *Não logaritmizar* variáveis medidas em unidades de tempo (anos, semanas,…) como educação, experiência profissional, idade ,….
- Situação menos clara para variáveis medidas em % ou para proporções como taxa de desemprego, % de sucesso num exame, …. (para estas variáveis é importante sublinhar a diferença entre variação percentual e pontos percentuais).

Ter ainda presente que logaritmizar y pode diminuir uma possível assimetria da variável e reduzir o peso de alguns *outliers.*

Termos quadráticos

Ideia → "suavizar" o impacto marginal da variável x_j na variável y pela introdução de um termo quadrático.

Exemplo: Tendo-se observado uma amostra de 50 observações do par (y_i, x_i) , obteve-se o diagrama de dispersão que se segue



Em vermelho traçou-se o MRL ajustado: $\hat{y} = 12.688 + 0.0733\ x$ enquanto a verde se traçou o modelo com termo quadrático em x, $\hat{y} = 9.833 + 0.3157\ x - 0.0048\ x^2$

A ideia por trás do termo quadrático é flexibilizar o modelo e assim melhorar o ajustamento.

Termos quadráticos

Em termos formais

$$y = \beta_0 + \beta_1 x_1 + \beta_2\ x_1^2 + \beta_3\ x_3 + \cdots + \beta_k x_k + u$$

e portanto

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2\ x_1^2 + \hat{\beta}_3\ x_3 + \cdots + \hat{\beta}_k x_k$$

Agora o impacto de x_1 em $\hat{y}$ será dado por

$$\frac{\partial \hat{y}}{\partial x_1} = \hat{\beta}_1 + 2\hat{\beta}_2\ x_1$$

Isto é, deixa de ser constante e pode até trocar de sinal (caso $\hat{\beta}_1$ e $\hat{\beta}_2$ tenham sinais trocados).

O caso mais frequente é $\hat{\beta}_1 > 0$ e $\hat{\beta}_2 < 0$ como na figura anterior mas nada obriga.

	$\hat{\beta}_2 < 0$	$\hat{\beta}_2 > 0$
$\frac{\partial \hat{y}}{\partial x_1} > 0$	$x_1 < \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$	$x_1 > \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$
$\frac{\partial \hat{y}}{\partial x_1} = 0$	$x_1 = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$	$x_1 = \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$
$\frac{\partial \hat{y}}{\partial x_1} < 0$	$x_1 > \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$	$x_1 < \frac{-\hat{\beta}_1}{2\hat{\beta}_2}$

Termos quadráticos - exemplo

Retomemos o exemplo dos salários horários wage função da educação, tenure, género e experiência, introduzindo um termo quadrático na variável experiência (dados no ficheiro WAGE1 e output no slide que se segue)

$$\widehat{wage} = -2.120 + 0.530\ educ + 0.134\ tenure - 1.790\ female + 0.205\ exper - 0.004\ exper^2$$
$$(0.049) \quad (0.021) \quad (0.258) \quad (0.034) \quad (0.001)$$

Como se pode verificar o modelo passa os testes habituais e iremos portanto interpretar o resultado.

Do quadro anterior tiramos que o impacto de $exper$ será positivo no salario enquanto $exper < \frac{0.205}{2\times 0.004} = 24.82$ (conta feita com todas as casas decimais). A experiência profissional beneficia o trabalhador até aos 24.82 anos e começa a prejudica-lo de aí em diante. Neste caso concreto vale a pena considerar o possível efeito da variável idade (que não está no modelo). A sua omissão conjugada com a sua correlação com $exper$ pode aliás levantar problemas de endogeneidade que não iremos discutir aqui.

Termos quadráticos - exemplo

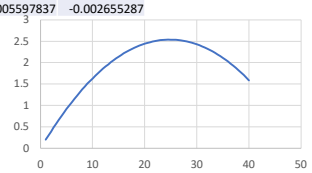
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.631388521
R Square	0.398651465
Adjusted R Square	0.392869268
Standard Error	2.877600715
Observations	526

ANOVA

	df	SS	MS	F	Significance F
Regression	5	2854.509655	570.901931	68.94463019	3.01595E-55
Residual	520	4305.904656	8.280585876		
Total	525	7160.41431			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-2.120092201	0.712046238	-2.977464224	0.003041951	-3.518933027	-0.721251375
educ.years	0.53009073	0.048588114	10.90988476	4.27791E-25	0.434637605	0.625543854
tenure.years	0.133669401	0.020632481	6.478590626	2.14967E-10	0.093136138	0.174202663
female	-1.790225875	0.257691666	-6.94716249	1.11859E-11	-2.296470559	-1.283981192
exper.years	0.204841793	0.034457597	5.944749778	5.0807E-09	0.137148586	0.272535001
exper^2	-0.004126562	0.000748917	-5.510039898	5.65241E-08	-0.005597837	-0.002655287



Interacção entre variáveis explicativas

A introdução de um termo de interação é feita quando o efeito parcial de uma variável explicativa depende de outra variável explicativa. Supondo, sem perda de generalidade que se assume que o efeito parcial de x_1 depende de x_2,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_4 + \cdots + \hat{\beta}_k x_k$$

isto é constroi-se uma variável x_3 em que cada observação é o produto das observações de x_1 e de x_2 ou seja $x_{i3} = x_{i1} \times x_{i2}, i = 1,2, \ldots, n$. A estimação e a inferência são feitas nos termos habituais.

Temos agora $\frac{\partial \hat{y}}{\partial x_1} = \hat{\beta}_1 + \hat{\beta}_3 x_2$. O efeito parcial de x_1 depende de x_2 (e inversamente, o efeito parcial de x_2 também depende de x_1)

Interacção entre variáveis explicativas – exemplo

Retome-se o exemplo anterior (*wage* função de *educ*, *tenure*, *female* e *exper*) com termo quadrático e introduza-se uma interação entre educ e tenure. A ideia é que a antiguidade na empresa será mais valorizada para os quadros médios ou superiores e portanto com mais anos de educação).

Define-se uma nova variável (educ*tenure no output) como sendo o produto de $educ_i$ por $tenure_i$ e estima-se a regressão (output no slide que se segue)

$$\widehat{wage} = -0.452 + 0.406\ educ - 0.104\ tenure - 1.833\ female + 0.186\ exper$$
$$-0.004\ exper^2 + 0.020\ (educ \times tenure)$$

Repare-se

- Na alteração dos coeficientes, nomeadamente de *educ* e *tenure*. Este último mudou até de sentido e deixou de ser estatisticamente significante.
- A interação é estatisticamente significante, tal como *educ* mas *tenure* deixou de o ser. Reforça a ideia que o impacto de *tenure* é feito em termos de *educ,* não existindo um impacto autónomo significativo (isto é independente de *educ*) . Já *educ* tem um impacto autónomo muito forte, com seria aliás de esperar.
- Neste como em todas as regressões baseadas neste conjunto de dados a discriminação de género é muito marcada!

Interacção entre variáveis explicativas – exemplo

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.642556997					
R Square	0.412879494					
Adjusted R Square	0.406091974					
Standard Error	2.846092569					
Observations	526					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	6	2956.388241	492.7313735	60.82920957	5.66707E-57	
Residual	519	4204.02607	8.100242909			
Total	525	7160.41431				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.451983047	0.84688127	-0.533702967	0.59377555	-2.115719691	1.211753598
educ.years	0.406363274	0.059384744	6.84289003	2.19453E-11	0.289699252	0.523027296
tenure.years	-0.104344931	0.07014746	-1.48750832	0.137487889	-0.242152796	0.033462935
female	-1.832956317	0.255154726	-7.183705154	2.36962E-12	-2.334219342	-1.331693292
exper.years	0.186334006	0.034477558	5.404501249	9.92236E-08	0.11860128	0.254066732
exper^2	-0.003727954	0.000749196	-4.975940865	8.84564E-07	-0.005199783	-0.002256125
educ*tenure	0.019806316	0.005584847	3.546438498	0.000425829	0.008834631	0.030778001

Teste à forma funcional - RESET

Modelo → $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

Objetivo: O teste RESET procura detetar uma má especificação da forma funcional, nomeadamente a omissão de uma variável relevante que esteja correlacionada com as variáveis explicativas incluídas no modelo ou uma não transformação da variável dependente.

Intuição: Se o modelo estiver bem especificado então nenhuma função das variáveis explicativas acrescenta algo ao modelo. Assim vamos compara o modelo estimado com um modelo auxiliar onde se acrescente às variáveis explicativas $\hat{y}^2, \hat{y}^3$.

Teste à forma funcional - RESET

Procedimento:

1. Estimar (se não estiver já feito) o modelo
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$
e guardar os valores ajustados $\hat{y}_i$
2. Definir o modelo auxiliar
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + \varepsilon$$
3. Testar $H_0: \gamma_1 = \gamma_2 = 0$ contra $H_1: \gamma_1 \neq 0$ ou $\gamma_2 \neq 0$
utilizando um teste F para a nulidade de um sub-conjunto de parâmetros. O teste F pode ser feito com base na SSR ou no R^2 já que a variável dependente é a mesma nas 2 regressões.

Nota: a rejeição de H_0 requer que se procure outra forma funcional alternativa - o modelo alternativo é apenas uma regressão auxiliar.

Teste à forma funcional - RESET

Comentários ao teste RESET:

- Porquê as potências 2 e 3 para $\hat{y}$ na regressão auxiliar? Simplicidade e as não linearidades são geralmente bem apanhadas;
- Como se dá o benefício da dúvida a H_0 **a não rejeição de H_0 não garante que o nosso modelo é o mais adequado**. Apenas nos diz que não é rejeitado. Para quase todos os fenómenos é habitual encontrar vários modelos que passam o teste RESET!
- Ao invés, **se rejeitarmos H_0 o modelo tem de ser reformulado**
- O teste à nulidade conjunta também pode ser baseado no teste LM (multiplicadores de Lagrange)

Forma funcional – RESET – exemplo

Exemplo: testar a forma funcional de

$$preço = \beta_0 + \beta_1 area + \beta_2 quartos + u$$

```
 regress preco quartos area

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   72.95
       Model |  579971.198     2  289985.599            Prob > F      =  0.0000
    Residual |  337883.308    85  3975.09774            R-squared     =  0.6319
-------------+------------------------------           Adj R-squared =  0.6232
       Total |  917854.506    87  10550.0518            Root MSE      =  63.048
------------------------------------------------------------------------------
       preco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quartos |   15.12134   9.488598     1.59   0.115    -3.744538    33.98721
        area |   1.383606   .1489435     9.29   0.000     1.087467    1.679746
       _cons |  -19.2855   31.04753    -0.62   0.536    -81.0163     42.4453
------------------------------------------------------------------------------


. predict precohat
(option xb assumed; fitted values)

. generate precohat2=precohat^2

. generate precohat3=precohat^3
```


Forma funcional – RESET – exemplo

Exemplo (continuação):

```
. regress preco quartos area precohat2 precohat3

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  4,    83) =   41.17
       Model |  610249.039     4   152562.26            Prob > F      =  0.0000
    Residual |  307605.467    83  3706.08996            R-squared     =  0.6649
-------------+------------------------------           Adj R-squared =  0.6487
       Total |  917854.506    87  10550.0518            Root MSE      =  60.878

------------------------------------------------------------------------------
       preco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quartos |  -58.37904   38.71904    -1.51   0.135    -135.3897    18.63158
        area |  -5.680895   3.613211    -1.57   0.120    -12.86743    1.505637
   precohat2 |   .0133394   .0076821     1.74   0.086    -.0019399    .0286187
   precohat3 |  -.0000109   7.20e-06    -1.52   0.133    -.0000252    3.40e-06
       _cons |   675.0476   328.2222     2.06   0.043     22.22683    1327.868
------------------------------------------------------------------------------
```

H_0: $\gamma_1 = \gamma_2 = 0 \rightarrow$ FF correcta

$$F_{obs} = \frac{(0.6649 - 0.6319)/2}{(1 - 0.6649)/(88 - 5)} = 4.08$$

A 5% de significância, $F(2,83) \simeq 3.15$

Rejeita-se H_0: a FF não é válida

Forma funcional – RESET – exemplo

Exemplo (continuação): procura de outra FF – adiciona-se o quadrado de area

```
. generate area2=area^2
. regress preco quartos area area2
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =   52.11
       Model |  597041.642     3  199013.881            Prob > F      =  0.0000
    Residual |  320812.864    84  3819.20076            R-squared     =  0.6505
-------------+------------------------------           Adj R-squared =  0.6380
       Total |  917854.506    87  10550.0518            Root MSE      =    61.8
------------------------------------------------------------------------------
       preco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quartos |   14.48683   9.305514     1.56   0.123    -4.018205    32.99187
        area |  -.2476931   .7852994    -0.32   0.753    -1.809347    1.313961
       area2 |   .0036535   .0017281     2.11   0.037      .000217    .0070901
       _cons |   149.9207   85.62566     1.75   0.084    -20.35527    320.1968
------------------------------------------------------------------------------

. predict precohatA
(option xb assumed; fitted values)
. generate precohatA2=precohatA^2
. generate precohatA3=precohatA^3
```


Forma funcional – RESET – exemplo

Exemplo (continuação):

```
. regress preco quartos area area2 precohatA2 precohatA3
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  5,    82) =   33.15
       Model |  614077.42      5  122815.484            Prob > F      =  0.0000
    Residual |  303777.086    82  3704.59861            R-squared     =  0.6690
-------------+------------------------------           Adj R-squared =  0.6489
       Total |  917854.506    87  10550.0518            Root MSE      =  60.865
------------------------------------------------------------------------------
       preco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quartos |  -138.0947   77.28978    -1.79   0.078    -291.8487    15.65929
        area |   6.140899   4.174249     1.47   0.145    -2.163011    14.44481
       area2 |  -.0453309   .0264509    -1.71   0.090    -.0979502    .0072884
  precohatA2 |   .0277233   .0134951     2.05   0.043     .0008772    .0545695
  precohatA3 |  -.0000214   .0000101    -2.12   0.037    -.0000415   -1.33e-06
       _cons |  -529.7486   404.1627    -1.31   0.194    -1333.757    274.2597
------------------------------------------------------------------------------
```

$H_0: \gamma_1 = \gamma_2 = 0 \rightarrow$ FF correcta

$$F_{obs} = \frac{(0.6690 - 0.6505)/2}{(1 - 0.6690)/(88 - 6)} = 2.30$$

A 5% de significância, $F(2,82) \simeq 3.15$

Não se rejeita H_0: a FF que inclui o quadrado de área **não é rejeitada (passa o RESET)**

Forma funcional – RESET

Recorda-se que para um mesmo problema existem geralmente vários modelos que passam o teste RESET.

Voltando ao exemplo e considerando agora o dataset completo (ver ficheiro "hprice1 com nomes.xlsx") onde apenas se transformou a unidade de medida das áreas de pés-quadrados para m2, pode verificar-se que o modelo

$$\widehat{\ln preço} = 0.766 + 0.168 \ln arealote + 0.700 \ln areacasa + 0.037\ quartos$$
$$(0.0383) \qquad (0.0929) \qquad (0.0275)$$

Também passa o teste RESET

$$F_{obs} = \frac{(0.66399 - 0.64300)/2}{(1 - 0.66399)/(88 - 6)} = 2.565$$

A 5% de significância, $F(2{,}82) \simeq 3.15$ ou $p - value = 0.083$

Em suma: O teste RESET serve essencialmente para rejeitar formalizações incorretas e não tanto para escolher **a** formalização adequada

Previsão

Modelo → $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

$$E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Modelo estimado → $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$

Como já se viu anteriormente o modelo estimado, depois de devidamente testado, pode ser utilizado para fazer previsões.

Duas questões surgem:

1. Que prever: valor de y, conhecidos os valores de $x_1, x_2, \ldots, x_k$, ou o valor esperado condicionado de y?
2. Como prever: previsão pontual vs previsão por intervalo?

Em qualquer dos casos assume-se que os valores das variáveis explicativas para os quais se quer fazer previsão são conhecidos, $\boldsymbol{x} = \mathbf{c}$ ou seja $x_1 = c_1, x_2 = c_2, \ldots., x_k = c_k$.

Previsão em média $E(y|\boldsymbol{x}=\boldsymbol{c})$

Modelo → $E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$

Uma previsão pontual para $E(y|\boldsymbol{x}=\boldsymbol{c}) = \beta_0 + \beta_1 c_1 + \cdots + \beta_k c_k$ é obtida substituindo os β_j pelos seus estimadores/estimativas $\hat{\beta}_j$ já que os valores das variáveis explicativas para os quais se quer uma previsão são conhecidos. Neste quadro $E(y|\boldsymbol{x}=\boldsymbol{c})$ pode ser considerada uma combinação linear de β_j e pode-se definir $E(y|\boldsymbol{x}=\boldsymbol{c}) = \theta_0$ com $\theta_0 = \beta_0 + \beta_1 c_1 + \cdots + \beta_k c_k$.

A previsão pontual será então $\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \cdots + \hat{\beta}_k c_k$

Para uma previsão por intervalos (ou simplesmente para aferir a variabilidade do previsor) é necessário obter o erro padrão de $\hat{\theta}_0$.

Previsão em média $E(y|\boldsymbol{x}=\boldsymbol{c})$

Prever $\theta_0 = E(y|\boldsymbol{x}=\boldsymbol{c}) = \beta_0 + \beta_1 c_1 + \cdots + \beta_k c_k$

Previsor $\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \cdots + \hat{\beta}_k c_k$

Como obter $\hat{\sigma}_{\hat{\theta}_0}$?

O problema é em tudo semelhante ao que se abordou na inferência para uma combinação linear de coeficientes. Existem 2 métodos:

1. Recorrer à matriz estimada das variâncias/covariâncias dos $\hat{\beta}_j$
2. Utilizar a regressão auxiliar:
   - Fazer $\beta_0 = \theta_0 - \beta_1 c_1 - \cdots - \beta_k c_k$
   - Construir a regressão auxiliar substituindo β_0 pela expressão anterior
     $y = \theta_0 - \beta_1 c_1 - \cdots - \beta_k c_k + \beta_1 x_1 + \cdots + \beta_k x_k + u$
     $y = \theta_0 + \beta_1 (x_1 - c_1) + \cdots + \beta_k (x_k - c_k) + u$
   - Ao estimar a regressão obtém-se diretamente $\hat{\theta}_0$ e $\hat{\sigma}_{\hat{\theta}_0}$.

Obtidos $\hat{\theta}_0$ e $\hat{\sigma}_{\hat{\theta}_0}$ a construção do intervalo de previsão não levanta problema.

Previsão em média - exemplo

Voltemos ao exemplo do preço de um imóvel com função da área, da área ao quadrado e do número de quartos. O propósito agora é obter um intervalo de precisão a 95% para o valor esperado de um imóvel (preço de mercado, preço esperado) com 4 quartos e 220 m2 de área.

Modelo estimado:

$$\widehat{preço} = 149.92 - 0.2477\ area + 0.0037\ area^2 + 14.487\ quartos$$
$$(85.626)\quad (0.7853)\qquad (0.0017)\qquad (9.3055)$$

Modelo auxiliar estimado:

$$\widehat{preço} = 330.206 - 0.4277\ (area - 220) + 0.0037\ (area^2 - 220^2) + 14.487\ (quartos - 4)$$
$$(10.726)\quad (0.7853)\qquad (0.0017)\qquad (9.3055)$$

No Excel ou num software que gere um IC a 95% para os parâmetros, é só ler. Caso o output não gere este intervalo, calcular $(\hat{\theta}_0 - t_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\theta}_0}; \hat{\theta}_0 + t_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\theta}_0})$

No exemplo obtém-se $330.206 \pm 1.989 \times 10.726$, isto é, (308.87; 351.54). A amplitude do intervalo depende da variabilidade do modelo **e dos valores escolhidos para as variáveis explicativas**.

Previsão em média - exemplo

Output computador

preço	area	area^2	quartos	area-220	area^2-220^2	quartos-4
300	226	51076	4	6	2676	0
370	193	37249	3	-27	-11151	-1
191	128	16384	3	-92	-32016	-1
195	135	18225	3	-85	-30175	-1
373	234	54756	4	14	6356	0
466.275	256	65536	5	36	17136	1
332.5	192	36864	3	-28	-11536	-1

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.80652046					
R Square	0.650475252					
Adjusted R Square	0.637992226					
Standard Error	61.79968268					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	597041.6428	199013.8809	52.1087768	4.01148E-19	
Residual	84	320812.8654	3819.200779			
Total	87	917854.5083				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	330.2059972	10.72571732	30.78637888	1.60398E-47	308.8767321	351.5352624
area-220	-0.247693118	0.785299393	-0.315412339	0.753230872	-1.809347083	1.313960846
area^2-220^2	0.003653521	0.001728126	2.114151745	0.037467195	0.000216953	0.00709009
quartos-4	14.48683014	9.305514014	1.556800636	0.123277661	-4.018204989	32.99186526

Previsão para um caso particular $y|\boldsymbol{x} = \boldsymbol{c}$

Modelo → $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

Problema → Prever $y^0 = \beta_0 + \beta_1 c_1 + \cdots + \beta_k c_k + u^0$ em que u^0 é uma realização da variável aleatória u

- *Previsor pontual* → $\hat{y}^0 = \hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \cdots + \hat{\beta}_k c_k$ já que $E(u^0) = 0$
- *Erro de previsão* → $\hat{e}^0 = y^0 - \hat{y}^0 = \theta_0 + u^0 - \hat{\theta}_0 = \theta_0 - \hat{\theta}_0 + u^0$
  - $E(\hat{e}^0) = E(\theta_0 - \hat{\theta}_0 + u^0) = E(\theta_0 - \hat{\theta}_0) + E(u^0) = 0$
  - $var(\hat{e}^0) = var(\theta_0 - \hat{\theta}_0 + u^0) = var(-\hat{\theta}_0 + u^0) = var\hat{\theta}_0 + var\, u^0$
    $= var\hat{\theta}_0 + \sigma^2$

    já que $\hat{\theta}_0$ é uma combinação linear de $\hat{\beta}_j$ e u^0 *é não correlacionado com os* u_i da amostra.

Previsão para um caso particular $y|\boldsymbol{x} = \boldsymbol{c}$

Modelo → $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$

Problema → Prever $y^0 = \beta_0 + \beta_1 c_1 + \cdots + \beta_k c_k + u^0$ em que u^0 é uma realização da variável aleatória u

- *Previsor pontual* → $\hat{y}^0 = \hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \cdots + \hat{\beta}_k c_k$
- *Erro de previsão* → $\hat{e}^0 = y^0 - \hat{y}^0$; $E(\hat{e}^0) = 0$; $var(\hat{e}^0) = var\hat{\theta}_0 + \sigma^2$

No quadro da hipótese MRL 6, $\hat{e}^0$ terá distribuição normal (quer y^0 quer $\hat{y}^0$ têm distribuição normal) e portanto $\hat{e}^0 \sim N(0; var\hat{\theta}_0 + \sigma^2)$.

Pela razão habitual (σ^2 desconhecido) recorre-se à *t-Student* $\frac{y^0 - \hat{y}^0}{\sqrt{\hat{\sigma}^2_{\hat{\theta}_0} + \hat{\sigma}^2}} \sim t(n-k-1)$

Previsão para um caso particular $y|\boldsymbol{x} = \boldsymbol{c}$

- Utilizando a técnica da regressão auxiliar que se viu no caso anterior obtém-se sem dificuldade $\hat{y}^0$ e $\sqrt{\hat{\sigma}^2_{\hat{\theta}_0} + \hat{\sigma}^2}$. Neste último caso tem de se fazer uma pequena conta auxiliar.
- O intervalo de previsão vem então $\left(\hat{y}^0 - t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2_{\hat{\theta}_0} + \hat{\sigma}^2}; \hat{y}^0 + t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2_{\hat{\theta}_0} + \hat{\sigma}^2}\right)$

  já que $\hat{\theta}_0$ é uma combinação linear de $\hat{\beta}_j$ e u^0 *é não correlacionado com os* u_i da amostra.

Previsão para um caso particular - exemplo

Voltemos ao exemplo do preço de um imóvel com função da área, da área ao quadrado e do número de quartos. O propósito agora é obter um intervalo de precisão a 95% para o valor de um imóvel específico com 4 quartos e 220 m2 de área.

Modelo auxiliar estimado (o mesmo do caso anterior)

$$\widehat{preço} = \underset{(10.726)}{330.206} - \underset{(0.7853)}{0.4277}\,(area - 220) + \underset{(0.0017)}{0.0037}\,(area^2 - 220^2) + \underset{(9.3055)}{14.487}\,(quartos - 4)$$

com $\hat{\sigma}^2 = 3819.20$

Assim $\hat{y}^0$=330.206 e $\sqrt{\hat{\sigma}^2_{\hat{\theta}_0} + \hat{\sigma}^2}$=62.7235 . Como é habitual, $\hat{\sigma}^2$ é muito superior a $\hat{\sigma}^2_{\hat{\theta}_0}$ o que tende a gerar intervalos com uma amplitude excessiva que os torna de fraca utilidade

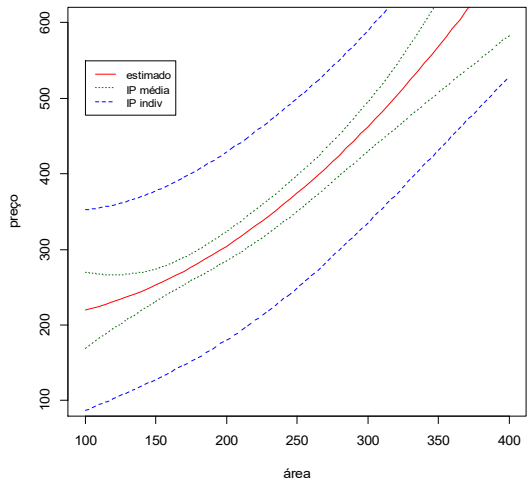
O intervalo de previsão vem então $(205.47; 454.94)$ o que, dada a sua amplitude, o torna de utilidade quase "nula" .

Previsão para um caso particular - exemplo

Como referido anteriormente, a amplitude do intervalo de previsão vai depender de 3 fatores:

1. A previsão para um caso particular vai originar intervalos com (muito) maior amplitude face à previsão em média.
2. O nível de confiança
3. Os valores das variáveis explicativas. Quanto mais próximos das médias amostrais, menor a amplitude.

O gráfico ilustra a situação para o exemplo fazendo variar a área e mantendo sempre $quartos = 4$



Previsão: ln(y) é a variável dependente

As previsões obtidas até agora dizem respeito a y. Quando y resulta de uma transformação as previsões em termos da variável original (aquelas que nos vão interessar) deixam de ser diretas.

Das várias transformações possíveis, apenas se vai ver o caso em que a variável de interesse foi logaritmizada, já que é a situação mais frequente.

A primeira ideia que surge é aplicar a transformação inversa (isto é a exponencial), ideia que apenas funciona nos intervalos de previsão para um caso particular já que

$$e^{E(\ln(y|\boldsymbol{x}))} < E\left(e^{\ln(y|\boldsymbol{x})}\right) = E(y|\boldsymbol{x})$$

Consequentemente a simples aplicação da transformação inversa leva a sub-prever o valor esperado condicionado de y.

Como corrigir este enviesamento?

Previsão: ln(y) é a variável dependente

Previsão pontual (quer para a média quer para um caso particular)

Se a hipótese de normalidade da distribuição de u não levantar problemas, então a melhor solução (que tira partido da distribuição lognormal) para estimar o valor esperado condicionado consiste em utilizar

$$\hat{y} = \widehat{E(y|\boldsymbol{x})} = \exp\left(\frac{\hat{\sigma}^2}{2}\right) \exp\left(\widehat{\ln y}\right)$$

Este previsor é enviesado mas é consistente.

Previsão: ln(y) é a variável dependente

Caso se procure uma solução que não se encontre tão vinculada à hipótese de normalidade (grandes amostras) a solução passa por estimar a constante de proporcionalidade entre $\hat{y}$ e $\exp\left(\widehat{\ln y}\right)$ em vez de utilizar o valor $\exp\left(\frac{\hat{\sigma}^2}{2}\right)$. Assim

1. Obter $m_i = \exp\left(\widehat{\ln y_i}\right)$ para os n valores da amostra.
2. Estimar o parâmetro α_o da regressão simples sem termo independente $y_i = \alpha_0\, m_i$, obtendo-se $\hat{\alpha}_0$.
3. As previsões pontuais para y ou $E(y|\boldsymbol{x})$ serão dadas por
$$\hat{y} = \hat{\alpha}_o \exp\left(\widehat{lny}\right)$$

Este previsor, tal como o anterior, é enviesado mas consistente.

Previsão: ln(y) é a variável dependente

Na previsão por intervalos teremos de considerar 2 casos:

- **Previsão de um caso particular** – Dada a invariância dos quantis não será necessário fazer nenhuma correção, isto é, se o intervalo de previsão para $\ln y$ for dado por $(a; b)$, o intervalo de previsão para y será dado por $\left(e^{a};\ e^{b}\right)$.
- **Previsão para a média** – Como o estimador (1º caso) é dado por $\exp\left(\frac{\hat{\sigma}^2}{2}\right)\ \exp\left(\widehat{\ln y}\right)$ será necessário ter em conta não só a correção anterior à localização como também a variabilidade de cada uma destas componentes.

  O IC é então dado por $exp\left(\left(\widehat{\ln y} + \hat{\sigma}^2/2\right) \pm t_{\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\theta}_0} + \frac{\hat{\sigma}^4}{2(n-k-1)}}\right)$.

  Como, em geral, $\frac{\hat{\sigma}^4}{2(n-k-1)}$ é muito inferior a $\widehat{var}\left(\widehat{\ln y}\right)$ uma solução aproximada passa por não considerar esta parcela, ou seja utilizar $exp\left(\left(\widehat{\ln y} + \hat{\sigma}^2/2\right) \pm z_{\alpha/2}\ \hat{\sigma}_{\hat{\theta}_0}\right)$ ou seja "exponenciar" o IC dado na regressão auxiliar e corrigi-lo multiplicando as extremidades por $exp(\hat{\sigma}^2/2)$.

Exemplo

Retomemos o exemplo anterior, considerando agora que se tinha definido como variável dependente o logaritmo do preço e não o preço. Com base neste **novo modelo** vai-se obter uma previsão pontual e um intervalo de precisão a 95% para:

1. O valor de um imóvel particular com 4 quartos e 220 m2 de área.
2. O valor esperado de um imóvel (preço de mercado, preço esperado) com as características anteriores (4 quartos e 220 m2 de área).

Em qualquer dos casos utiliza-se:

Modelo estimado: $\widehat{\ln preço} = 1.2893 + 0.8101 \ln area + 0.0376\ quartos$

$$(0.0988) \qquad (0.0303)$$

Modelo auxiliar estimado (ver slide seguinte):

$$\widehat{\ln preço} = 5.8090 + 0.8101\ (\ln area - \ln 220) + 0.0376\ (quartos - 4)$$

$$(0.0275)\ \ (0.0988) \qquad\qquad (0.0303)$$

Fazendo a conta (ou output) vem o IP para o VE de $\ln preço$: (5.7544; 5.8637)

Aplicando a exponencial a cada extremidade obtém-se (315.588; 352.016).

Este intervalo pode comparar-se com aquele que se obteve com o modelo alternativo.

Algo de semelhante pode ser feito para a previsão para um valor particular.

Exemplo

Retomemos o exemplo anterior, considerando agora que se tinha definido como variável dependente o logaritmo do preço e não o preço. Com base neste **novo modelo** vai-se obter uma previsão pontual e um intervalo de precisão a 95% para:

1. O valor de um imóvel particular com 4 quartos e 220 m2 de área.
2. O valor esperado de um imóvel (preço de mercado, preço esperado) com as características anteriores (4 quartos e 220 m2 de área).

Em qualquer dos casos utilizam-se as regressões:

Modelo estimado: $\widehat{\ln preço} = 1.2893 + 0.8101 \ln area + 0.0376\ quartos$

$$(0.0988) \qquad (0.0303)$$

$$\hat{\sigma}^2 = 0.04134$$

Modelo auxiliar estimado (ver slide seguinte):

$$\widehat{\ln preço} = 5.8090 + 0.8101\ (\ln area - \ln 220) + 0.0376\ (quartos - 4)$$

$$(0.0275)\ \ (0.0988) \qquad\qquad (0.0303)$$

Exemplo – output 1

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.749479308					
R Square	0.561719234					
Adjusted R Square	0.551406745					
Standard Error	0.203324195					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	4.503642728	2.251821364	54.46980398	5.94946E-16	
Residual	85	3.513961901	0.041340728			
Total	87	8.017604629				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	**5.809057324**	**0.027471396**	211.4583978	1.8092E-117	**5.754436829**	**5.863677818**
ln(area)-ln(220)	0.810063655	0.098761107	8.202253716	2.2161E-12	0.613700116	1.006427194
quartos-4	0.037646372	0.030344583	1.240629064	0.21815627	-0.022686789	0.097979533

Exemplo

Previsões pontuais (idênticas, quer para um caso particular, quer para o valor esperado)

Solução 1 – (baseada na normal)

$$\widehat{preço}^0 = \exp\left(\frac{0.20332^2}{2}\right)\exp(5.80906) = 340.266$$

Solução 2 – (mais robusta) – ver slide seguinte

- Obter $m_i = \exp\left(\widehat{\ln y_i}\right)$
- $\hat{y} = \hat{\alpha}_o \exp\left(\widehat{lny}\right) = 1.0286 \exp\left(\widehat{lny}\right)$
- $\widehat{preço}^0 = 1.0286 \ \exp(5.80906) = 342.842$

Qualquer destas previsões pode ser utilizada para prever o valor de mercado, isto é, o valor esperado do preço.

Exemplo – output 2

Regressão auxiliar

preço	ln(preço)	Predicted	exp(lny)
300	5.703782475	5.830854073	340.6494945
370	5.913503006	5.665343357	288.6870869
191	5.252273428	5.332682677	206.9925242
195	5.272999559	5.375814123	216.1157456
373	5.92157842	5.859033042	350.3851723

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.979692285					
R Square	0.959796974					
Adjusted R Square	0.948302721					
Standard Error	62.6755834					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	8158994.73	8158994.73	2077.016192	5.22273E-62	
Residual	87	341755.9016	3928.228754			
Total	88	8500750.632				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
exp(lny)	**1.028612766**	0.022570021	45.57429311	1.67839E-62	0.983752405	1.073473128

Exemplo

Obtenção dos intervalos de previsão:

- **Imóvel particular**

  IP para $\ln preço$: (5.4011; 6.2170)

  $$5.80906 \pm 1.988\sqrt{0.02747^2 + 0.04134}$$

  Aplicando a exponencial a cada extremidade obtém-se **(221.65; 501.19)**.

- **Valor de mercado Imóvel particular (solução aproximada)**

  IP para o VE de $\ln preço$: (5.7544; 5.8637)

  Aplicar a exponencial a cada extremidade (315.59; 352.02)

  Aplicar a correção às extremidades por forma a obter uma solução aproximada

  - Normal: $\exp\left(\frac{0.20332^2}{2}\right) = 1.0209$ **(322.18; 359.37)**
  - Robusta: $\hat{\alpha} = 1.0286$ **(324.61; 362.08)**

Exemplo

- **Valor de mercado Imóvel particular (solução mais exata baseada na Normal)**

  Fórmula a aplicar: $exp\left(\left(\widehat{\ln y} + \hat{\sigma}^2/2\right) \pm t_{\alpha/2}\sqrt{\hat{\sigma}^2_{\hat{\theta}_0} + \frac{\hat{\sigma}^4}{2(n-k-1)}}\right)$

  vindo (ver output) $exp\left(5.80906 + \frac{0.04134}{2} \pm 1.988\sqrt{0.02747^2 + \frac{0.04134^4}{2\times 85}}\right)$

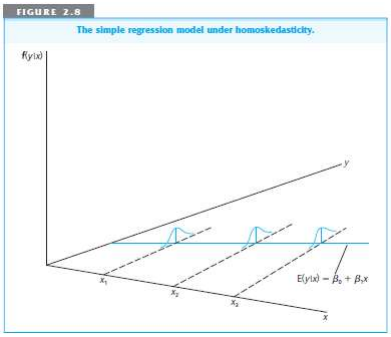
  que origina **(322.06; 359.50)**

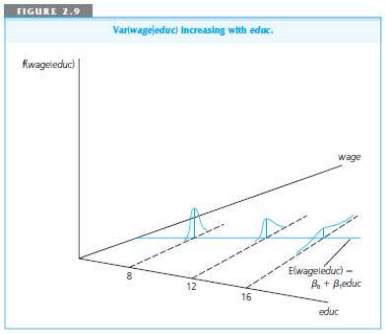
  A comparação com os resultados da fórmula aproximada **(322.18; 359.37)** mostra bem porque se pode utilizar a fórmula aproximada, bem menos trabalhosa

Heterocedasticidade: O que é?

Violação da hipótese MRL 5 – Homocedasticidade – $var(u|\boldsymbol{x}) = \sigma^2$, isto é **a variância da variável residual u deixa de ser constante.**

Por vezes a variância pode modelar-se como função (total ou parcialmente conhecida) das variáveis explicativas mas pode também depender de variáveis que não estão no modelo, isto é, as variáveis que explicam a variância podem não ser (total ou parcialmente) as mesmas que explicam o valor esperado ou pode ainda ter um padrão desconhecido.





Heterocedasticidade: porque acontece?

A heterocedasticidade pode acontecer por:

- Má especificação do modelo nomeadamente:
  - ausência de algumas variáveis relevantes
  - Não transformação de variáveis (logaritmização)
- Natureza do problema

  É natural que um modelo onde a poupança de um agregado familiar é explicada pelo rendimento disponível do agregado, pela sua dimensão e pelo número de elementos com menos de 22 anos sofra de heterocedasticidade e que esta seja de alguma forma "proporcional" ao rendimento disponível: Agregados com fraco rendimento disponível não apresentação grande variabilidade nas poupanças por razões óbvias enquanto em agregados com maior rendimento disponível se observará uma maior variabilidade nas poupanças.

Heterocedasticidade: consequências

Ao cair a hipótese MRL 5 o estimador OLS

- continua a ser estimador centrado e consistente de β (estas propriedades apenas necessitam das hipóteses MRL 1 a MRL 4)
- **deixa de ser o mais eficiente**
- a estimação da variância do estimador, tal como foi feita, deixa de ser válida
- as estatísticas t e F deixam de ter distribuição t e F respetivamente, apenas se mantendo válidas as estatísticas R^2 e $\bar{R}^2$.

Está-se a assumir que se mantêm as restantes hipóteses do modelo, nomeadamente MRL 6 (distribuição normal) e MRL 2 que tem por consequência a ausência de correlação entre os u_i e portanto que a matriz de variância/covariância dos u_i é matriz diagonal.

Heterocedasticidade: soluções

A solução está dependente do conhecimento que se tenha (ou não) do que motiva a heterocedasticidade.

Mais especificamente, a questão passa por saber se se está em condições de assumir que $var(u|\boldsymbol{x}) = \sigma^2\, h(\boldsymbol{x})$ em que $h(\boldsymbol{x})$ (designada por "skedastic function") é uma função conhecida das variáveis explicativas.

Neste caso, substitui-se MRL 5 por uma hipótese mais fraca – MRL 5' $var(u|\boldsymbol{x}) = \sigma^2\, h(\boldsymbol{x})$ – mas os resultados que se vierem a obter irão depender da validade desta nova hipótese. Note-se que o caso de homocedasticidade corresponde a assumir $h(\boldsymbol{x}) = 1$.

Assim temos 2 alternativas:

- Estimação robusta – Assumir apenas que $var(u_i|\boldsymbol{x}) = \sigma_i^2$
- Estimação por GLS (*skedastic function* conhecida) – Assumir $var(u|\boldsymbol{x}) = \sigma^2\, h(\boldsymbol{x})$

Heteroscedasticidade: estimação robusta

Dado que os estimadores OLS continuam a ser **centrados** e **consistentes**, corrige-se apenas a sua variância por forma a que o teste *t* seja válido.

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$$

Como estimar $var(\hat{\beta})$?

Para o teste F a situação é mais complicada e requer software adequado (embora com heterocedasticidade moderada se possa continuar a utilizar o teste). Pode no entanto utilizar-se o teste LM (multiplicadores de Lagrange) adaptado para testar a nulidade de vários coeficientes. Neste caso convém que a amostra tenha dimensão adequada dado que o teste é assintótico.

Sabe-se que:

- $var(u_i|\boldsymbol{X}) = \sigma_i^2$
- $cov(u_i, u_r|\boldsymbol{X}) = 0$, para $i \neq r$ (a hipótese MRL 2 continua válida)

Heteroscedasticidade: estimação robusta de $var(\hat{\beta})$

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$$

$$var(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T var(U|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Enquanto com MLR 5 se utilizava $var(U|\boldsymbol{X}) = \sigma^2\mathbf{I}$, tem-se agora de assumir $var(U|\boldsymbol{X}) = \boldsymbol{\Sigma}$, sendo $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$.

$$var(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{\Sigma}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Como estimar Σ de forma robusta ?

Estimador de White:

$\hat{\sigma}_i^2 = \hat{u}_i^2$ e portanto $\widehat{\boldsymbol{\Sigma}} = diag(\widehat{\boldsymbol{\sigma}}_1^2, \widehat{\boldsymbol{\sigma}}_2^2, \dots, \widehat{\boldsymbol{\sigma}}_n^2)$

Intuição: $var(u_i|\boldsymbol{X}) = E(u_i^2|\boldsymbol{X}) - \big(E(u_i|\boldsymbol{X})\big)^2 = E(u_i^2|\boldsymbol{X})$ MLR4

Heteroscedasticidade: estimação robusta de $var(\hat{\beta})$

A matriz $(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{X})$ vem assim

$$(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{X}) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ \dots & \dots & & \dots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix} \times \begin{bmatrix} \hat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2^2 & \dots & 0 \\ \dots & \dots & & \dots \\ 0 & 0 & \dots & \hat{\sigma}_n^2 \end{bmatrix} \times \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & & \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_2^2 & \dots & \hat{\sigma}_n^2 \\ \hat{\sigma}_1^2 x_{11} & \hat{\sigma}_2^2 x_{21} & \dots & \hat{\sigma}_n^2 x_{n1} \\ \dots & \dots & & \dots \\ \hat{\sigma}_1^2 x_{1k} & \hat{\sigma}_2^2 x_{2k} & \dots & \hat{\sigma}_n^2 x_{nk} \end{bmatrix} \times \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & & \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} \sum \hat{\sigma}_i^2 & \sum \hat{\sigma}_i^2 x_{i1} & \dots & \sum \hat{\sigma}_i^2 x_{ik} \\ \sum \hat{\sigma}_i^2 x_{i1} & \sum \hat{\sigma}_i^2 x_{i1}^2 & \dots & \sum \hat{\sigma}_i^2 x_{i1} x_{ik} \\ \dots & \dots & & \dots \\ \sum \hat{\sigma}_i^2 x_{ik} & \sum \hat{\sigma}_i^2 x_{ik} x_{i1} & \dots & \sum \hat{\sigma}_i^2 x_{ik}^2 \end{bmatrix}$$

Heteroscedasticidade: estimação robusta de $var(\hat{\beta})$

Assim,

$$\widehat{var}(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

Mostra-se que este **estimador é consistente** mas trata-se de uma demonstração complicada.

Wooldridge apresenta uma forma alternativa de obter os elementos da diagonal desta matriz (os estimadores das variâncias) mais fácil de implementar.

$$\widehat{var}(\hat{\beta}_j) = \frac{\sum_{i=1}^n \tilde{r}_{ij}^2 \, \hat{u}_i^2}{SSR_j^2}$$

onde

$\tilde{r}_{ij}$ resíduos da regressão de x_j nas restantes variáveis explicativas

SSR_j soma do quadrado dos resíduos desta regressão $SSR_j = \sum_{i=1}^n \tilde{r}_{ij}^2$

Heteroscedasticidade: estimação robusta de $var(\hat{\beta})$

Exemplo: Considere-se o exemplo habitual dos imóveis (na sua versão mais simples) em que o modelo estimado era

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$

e calcule-se os erros-padrão de forma robusta.

Sol 1.

Com base na regressão original, obtiveram-se os $\hat{u}_i$ e construi-se $\widehat{\boldsymbol{\Sigma}} = \text{diag}\ (\hat{u}_i^2)$

Recorrendo ao software R obteve-se

$$\widehat{var}(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} 1666.8 & -5.9835 & 179.80 \\ -5.9835 & 0.04307 & -0.4474 \\ -179.80 & -0.4474 & 77.6484 \end{bmatrix}$$

Isto é os erros-padrão robustos para $\hat{\beta}_1$ e $\hat{\beta}_2$ são respetivamente 0.2075 e 8.8118.

Sol 2.

Regressões auxiliares (próximo slide)

Heteroscedasticidade: estimação robusta de $var(\hat{\beta})$

Sol 2.

Com base na regressão original, obtiveram-se os $\hat{u}_i$

Temos 2 declives logo 2 regressões auxiliares

- $\widehat{area} = -19.286 + 15.121\ quartos \rightarrow$ obter $\tilde{r}_{i1}$ e calcular

$$\widehat{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n \tilde{r}_{i1}^2\ \hat{u}_i^2}{SSR_1^2} = 0.04307$$ note-se que $SSR_1^2 = \left(\sum_{i=1}^n \tilde{r}_{i1}^2\right)^2$

- $\widehat{quartos} = -19.286 + 1.384\ area \rightarrow$ obter $\tilde{r}_{i2}$ e calcular

$$\widehat{var}(\hat{\beta}_2) = \frac{\sum_{i=1}^n \tilde{r}_{i2}^2\ \hat{u}_i^2}{SSR_2^2} = 77.6484$$

Heteroscedasticidade: estimação robusta de $var(\hat{\beta})$

A solução habitual passa por **recorrer a um software** que produza os erros-padrão robustos à heterocedasticidade.

Estes estimadores são conhecidos como White, Huber ou Eicker (ou combinações com 2 ou 3 destes nomes).

Aparecem geralmente na forma "corrigida" que consiste em multiplicar as variâncias referidas anteriormente por $\frac{n}{n-k-1}$ para melhorar a compatibilidade com o caso homocedástico.

Como é evidente o fator $\frac{n}{n-k-1}$ não altera a validade assintótica do estimador, (já que $\frac{n}{n-k-1} \to 1$ quando $n \to \infty$)

Corrigidos os erros-padrão a inferência é feita nos termos habituais.

Heteroscedasticidade: estimação robusta de $\hat{\sigma}_{\hat{\beta}_j}$

Exemplo anterior feito em **STATA** – erros-padrão robustos (os $\hat{\beta}_j$ são os mesmos do OLS)

```
. regress preco quartos area, robust
Linear regression                               Number of obs =      88
                                                F(  2,    85) =   27.22
                                                Prob > F      =  0.0000
                                                R-squared     =  0.6319
                                                Root MSE      =  63.048
------------------------------------------------------------------------------
             |               Robust
       preco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quartos |   15.12134    8.96599     1.69   0.095    -2.705452    32.94813
        area |   1.383606   .2111629     6.55   0.000     .9637578    1.803455
       _cons |  -19.2855   41.54017    -0.46   0.644    -101.8785    63.30749
------------------------------------------------------------------------------
```

Apresentação habitual

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$
$$(0.149) \qquad (9.489)$$
$$[0.211] \qquad [8.966]$$

Heteroscedasticidade: estimação robusta de $\hat{\sigma}_{\hat{\beta}_j}$

Exemplo anterior feito em **Eviews**

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$

(0.149) (9.489)

[0.211] [8.966]

Dependent Variable: PRECO
Method: Least Squares
Date: 04/27/20 Time: 17:33
Sample: 1 88
Included observations: 88
Huber-White-Hinkley (HC1) heteroskedasticity consistent standard errors and covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-19.28550	41.54017	-0.464261	0.6436
AREA_M2_	1.383606	0.211163	6.552317	0.0000
QUARTOS	15.12134	8.965990	1.686522	0.0954

R-squared	0.631877	Mean dependent var	293.5460
Adjusted R-squared	0.623215	S.D. dependent var	102.7134
S.E. of regression	63.04838	Akaike info criterion	11.15918
Sum squared resid	337883.3	Schwarz criterion	11.24363
Log likelihood	-488.0038	Hannan-Quinn criter.	11.19320
F-statistic	72.95056	Durbin-Watson stat	1.857617
Prob(F-statistic)	0.000000	Wald F-statistic	27.22081
Prob(Wald F-statistic)	0.000000		

Observação: Como é evidente os erros-padrão robustos são iguais àqueles que se obtiveram no slide 9 depois de multiplicados pelo fator (88/85).

Heterocedasticidade: estimação WLS e GLS

Quando se substitui MRL 5 por $var(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$ existem duas alternativas:

1. $h(\boldsymbol{x})$ é uma função conhecida **sem** parâmetros desconhecidos
2. $h(\boldsymbol{x})$ é uma função conhecida **com** parâmetros desconhecidos

O caso 1 onde se assume que a heterocedasticidade é conhecida a menos de uma constante (estimação WLS) tem naturalmente uma solução mais simples do que o caso 2 (estimação GLS).

Os slides apenas abordam o caso 1, sendo que o caso 2 pode ser visto no Wooldridge (sub-secção *Feasible GLS* no quadro da secção *Weighted Least Squares Estimation*)

Heterocedasticidade: estimação WLS

$var(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x})$ sendo $h(\boldsymbol{x})$ uma função conhecida **sem** parâmetros desconhecidos

O método **WLS** produz estimadores centrados, consistentes e eficientes (independentemente da dimensão da amostra) cuja validade está obviamente ligada à hipótese que se assumiu.

A ideia base é partir do modelo de interesse

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \text{ com } var(u|\boldsymbol{x}) = \sigma^2\, h(\boldsymbol{x})$$

e obter um modelo transformado que verifique MRL 5.

Ao dividir os 2 lados da igualdade por $\sqrt{h(\boldsymbol{x})}$ teremos no termo de erro $u^* = \frac{u}{\sqrt{h(\boldsymbol{x})}}$ e portanto $var(u^*|\boldsymbol{x}) = var\left(\frac{u}{\sqrt{h(\boldsymbol{x})}}|\boldsymbol{x}\right) = \frac{var(u|\boldsymbol{x})}{h(\boldsymbol{x})} = \sigma^2$.

Heterocedasticidade: estimação WLS

Modelo de interesse (escrito em termos da amostra)

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$ com $var(u_i|\boldsymbol{x}_i) = \sigma^2 h_i$ e $h_i = \mathrm{h}(\boldsymbol{x}_i)$

Modelo transformado:

$$\frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_{i1}}{\sqrt{h_i}} + \cdots + \beta_k \frac{x_{ik}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}}$$

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + +\beta_2 x_{i2}^* + \cdots + \beta_1 x_{ik}^* + u_i^*$$

com $y_i^* = \frac{y_i}{\sqrt{h_i}}$, $x_{i0}^* = \frac{1}{\sqrt{h_i}}$, $x_{ij}^* = \frac{x_{ij}}{\sqrt{h_i}}$ $(j = 1,2,\ldots,k)$, $u_i^* = \frac{u_i}{\sqrt{h_i}}$

Este modelo é estimado pelo OLS

Notas:

- O modelo transformado não tem constante
- A interpretação dos β's é feita em função do modelo original.

Heterocedasticidade: estimação WLS

Modelo de interesse: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$

Modelo transformado: $y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + +\beta_2 x_{i2}^* + \cdots + \beta_1 x_{ik}^* + u_i^*$

Observações:

- A inferência estatística (R^2, $\bar{R}^2$, testes t, teste F, …) é feita com base no modelo transformado.
- **A interpretação dos $\boldsymbol{\beta}$'s é feita em função do modelo original**.
- O estimador GLS (*Generalized Least Squares*) para esta situação é conhecido como WLS (*Weighted Least Squares*) uma vez que minimiza a soma dos quadrados dos resíduos ponderada em que cada termo é ponderado por $\frac{1}{h_i}$,

$$\sum \hat{u}_i^{*2} = \sum (y_i^* - \hat{y}_i^*)^2 = \sum \left(\frac{y_i}{\sqrt{h_i}} - \frac{\hat{y}_i}{\sqrt{h_i}} \right)^2 = \sum \frac{(y_i - \hat{y}_i)^2}{h_i}$$

sendo $\hat{y}_i$ obtido com os estimadores do modelo transformado.

Heterocedasticidade: estimação WLS - exemplo

Retome-se o exemplo habitual do preço de um imóvel como função da área e do número de quartos.

A estimação do modelo feita anteriormente originava

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$
$$\begin{matrix} & (0.149) & (9.489) \\ & [0.211] & [8.966] \end{matrix}$$

Modelo transformado assumindo $var(u|\boldsymbol{x}) = \sigma^2 area$

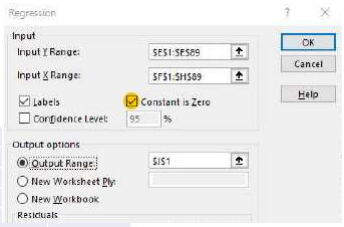
$$\frac{preço_i}{\sqrt{area_i}} = \beta_0 \frac{1}{\sqrt{area_i}} + \beta_1 \frac{area_i}{\sqrt{area_i}} + \beta_2 \frac{quartos_i}{\sqrt{area_i}} + u_i^*$$

Estimação OLS (ver slides seguintes)

$$\left(\frac{\widehat{preço}_i}{\sqrt{area_i}} \right) = 7.8960 \frac{1}{\sqrt{area_i}} + 1.3080 \frac{area_i}{\sqrt{area_i}} + 11.4685 \frac{quartos_i}{\sqrt{area_i}}$$
$$\begin{matrix} & (0.1555) & (8.9630) \end{matrix}$$

Heterocedasticidade: estimação WLS EXCEL

preço	area	quartos	preço*	x*_0	area*	quartos*
300	226	4	19.956	0.067	15.033	0.266
370	193	3	26.633	0.072	13.892	0.216
191	128	3	16.882	0.088	11.314	0.265
195	135	3	16.783	0.086	11.619	0.258



SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.979954982					
R Square	0.960311767					
Adjusted R Square	0.947613221					
Standard Error	4.442557176					
Observations	88					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	40591.53428	13530.51143	685.564247	6.25012E-59	
Residual	85	1677.586712	19.73631426			
Total	88	42269.12099				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
x*_0	7.895976487	30.70298548	0.257172922	0.79766681	-53.1497842	68.94173718
area*	1.307992309	0.155539391	8.409395838	8.46173E-13	0.998738329	1.617246289
quartos*	11.46850858	8.963038244	1.279533599	0.204190848	-6.352412713	29.28942988

Heterocedasticidade: estimação WLS EVIEWS

Definindo $\mathrm{Y} = \frac{preço}{\sqrt{area}}, \mathrm{X0} = \frac{1}{\sqrt{area}}, \ X1 = \frac{area}{\sqrt{area}} = \sqrt{area}, \ X2 = \frac{quartos}{\sqrt{area}}$

Dependent Variable: Y				
Method: Least Squares				
Date: 04/26/20 Time: 14:55				
Sample: 1 88				
Included observations: 88				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
X0	7.895976	30.70299	0.257173	0.7977
X1	1.307992	0.155539	8.409396	0.0000
X2	11.46851	8.963038	1.279534	0.2042
R-squared	0.244431	Mean dependent var		21.33308
Adjusted R-squared	0.226653	S.D. dependent var		5.051797
S.E. of regression	4.442557	Akaike info criterion		5.853834
Sum squared resid	1677.587	Schwarz criterion		5.938288
Log likelihood	-254.5687	Hannan-Quinn criter.		5.887858
Durbin-Watson stat	1.852266			

Heterocedasticidade: estimação WLS STATA

Modelo transformado:

$$\frac{preço_i}{\sqrt{area_i}} = \beta_0 \frac{1}{\sqrt{area_i}} + \beta_1 \frac{area_i}{\sqrt{area_i}} + \beta_2 \frac{quartos_i}{\sqrt{area_i}} + u_i^*$$

Ouput STATA

```
gen precoh= preco/sqrt(area)
gen X0h=1/sqrt(area)
gen areah=area/sqrt(area)
gen quartosh= quartos/sqrt(area)
regress  precoh X0h areah quartosh, noconst
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    85) =  685.56
       Model |   40591.534     3  13530.5113            Prob > F      =  0.0000
    Residual |  1677.58665    85  19.7363135            R-squared     =  0.9603
-------------+------------------------------           Adj R-squared =  0.9589
       Total |  42269.1207    88  480.330917            Root MSE      =  4.4426
------------------------------------------------------------------------------
      precoh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         X0h |   7.895981   30.70298     0.26   0.798    -53.14978    68.94174
       areah |   1.307992   .1555394     8.41   0.000     .9987383    1.617246
    quartosh |   11.46851   8.963038     1.28   0.204    -6.352413    29.28943
------------------------------------------------------------------------------
```


Heterocedasticidade: testes para detecção

Três testes, todos baseados em regressões auxiliares com variável dependente $\hat{u}^2$, onde a significância global dos parâmetros é testada através de um teste F / LM.

Modelo base:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

H_0: Homocedasticidade $\rightarrow var(u|\boldsymbol{x}) = \sigma^2$ (usar OLS)

H_1: Heterocedasticidade (usar GLS ou OLS com erros-padrão robustos)

Tal como os testes estão formulados, só se rejeita a homocedasticidade quando os dados aponta claramente para a heterocedasticidade.

Heterocedasticidade: testes para detecção

1. **Procedimento comum**: estimar o modelo base por OLS e obter $\hat{u}^2$

2. Estimar a regressão auxiliar (depende do teste escolhido)

- Breusch-Pagan: $\hat{u}^2 = \gamma_0 + \gamma_1 x_1 + \cdots + \gamma_k x_k + e$

  estamos a assumir que a heterocedasticidade é função das variáveis explicativas. Na prática, pode restringir-se a um subconjunto das variáveis explicativas se pensarmos que o padrão de heterocedasticidade está relacionado apenas com.este subconjunto

- White : Incluir todas as variáveis mais os seus quadrados e produtos cruzados – Exemplo para $k = 2$

$$\hat{u}^2 = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_1^2 + \gamma_4 x_2^2 + \gamma_5 x_1 x_2 + e$$

- White simplificado: $\hat{u}^2 = \gamma_0 + \gamma_1 \hat{y} + \gamma_2 \hat{y}^2 + e$

em qualquer das alternativas, obter $R^2_{\hat{u}^2}$

Heterocedasticidade: testes para detecção

3. Estatística de teste e distribuição:

$$F = \frac{R^2_{\hat{u}^2}/m}{\left(1 - R^2_{\hat{u}^2}\right)/(n-m-1)} \sim F(m, n-m-1)$$

ou

$$LM = nR^2_{\hat{u}^2} \sim \chi^2_m$$

onde m corresponde ao número de declives na regressão auxiliar

BP→ $m = k$ (na versão habitual); White simplificado → $m = 2$

White → $m = k + k + \frac{k(k-1)}{2} = \frac{k(k+3)}{2}$.

Em termos de filosofia, o teste BP procura um padrão para a heterocedasticidade enquanto o teste de White tem uma filosofia mais "robusta"

Heterocedasticidade: testes para detecção – exemplo 1

Exemplo: Considere-se o modelo habitual

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$

BP

Output Eviews

Heteroskedasticity Test: Breusch-Pagan-Godfrey
Null hypothesis: Homoskedasticity

F-statistic	5.873654	Prob. F(2,85)	0.0041
Obs*R-squared	10.68519	Prob. Chi-Square(2)	0.0048
Scaled explained SS	23.45408	Prob. Chi-Square(2)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/26/20 Time: 14:40
Sample: 1 88
Included observations: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-8300.608	3911.614	-2.122042	0.0367
AREA_M2_	42.12340	18.76508	2.244776	0.0274
QUARTOS	1193.551	1195.449	0.998413	0.3209

R-squared	0.121423	Mean dependent var	3839.583
Adjusted R-squared	0.100750	S.D. dependent var	8376.501
S.E. of regression	7943.334	Akaike info criterion	20.83155
Sum squared resid	5.36E+09	Schwarz criterion	20.91600
Log likelihood	-913.5882	Hannan-Quinn criter.	20.86557
F-statistic	5.873654	Durbin-Watson stat	2.090187
Prob(F-statistic)	0.004080		

Rejeita-se a **homocedasticidade**

Heterocedasticidade: testes para detecção – exemplo 1

Exemplo: Considere-se o modelo habitual

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$

White

Output Eviews

Heteroskedasticity Test: White
Null hypothesis: Homoskedasticity

F-statistic	4.007899	Prob. F(5,82)	0.0027
Obs*R-squared	17.28228	Prob. Chi-Square(5)	0.0040
Scaled explained SS	37.93476	Prob. Chi-Square(5)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/26/20 Time: 14:42
Sample: 1 88
Included observations: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10706.38	13167.06	0.813118	0.4185
AREA_M2_^2	0.465846	0.255698	1.821857	0.0721
AREA_M2_*QUARTOS	21.37083	19.60682	1.089970	0.2789
AREA_M2_	-251.8630	108.9744	-2.311212	0.0233
QUARTOS^2	-1281.901	840.9803	-1.524294	0.1313
QUARTOS	7025.944	5680.170	1.236925	0.2196

R-squared	0.196390	Mean dependent var	3839.583
Adjusted R-squared	0.147389	S.D. dependent var	8376.501
S.E. of regression	7734.605	Akaike info criterion	20.81054
Sum squared resid	4.91E+09	Schwarz criterion	20.97945
Log likelihood	-909.6639	Hannan-Quinn criter.	20.87859
F-statistic	4.007899	Durbin-Watson stat	2.075294
Prob(F-statistic)	0.002651		

Rejeita-se a **homocedasticidade**

Heterocedasticidade: testes para detecção – exemplo 1

Exemplo: Considere-se o modelo habitual

$$\widehat{preço} = -19.286 + 1.384\ area + 15.121\ quartos$$

White simplificado RES2, YC, YC2

Dependent Variable: RES2
Method: Least Squares
Date: 04/26/20 Time: 15:15
Sample: 1 88
Included observations: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.160192	0.115627	1.385425	0.1695
YC	-0.000731	0.000697	-1.048718	0.2973
YC2	1.02E-06	9.84E-07	1.032361	0.3048

R-squared	0.012779	Mean dependent var	0.039931
Adjusted R-squared	-0.010450	S.D. dependent var	0.083026
S.E. of regression	0.083458	Akaike info criterion	-2.095442
Sum squared resid	0.592050	Schwarz criterion	-2.010987
Log likelihood	95.19945	Hannan-Quinn criter.	-2.061417
F-statistic	0.550126	Durbin-Watson stat	2.246235
Prob(F-statistic)	0.578917		

$RES2 = (preço - \widehat{preço})^2$

$YC = \widehat{preço}$

$YC2 = \widehat{preço}^2$

$F_{obs} = \frac{0.012779/2}{(1-0.012779)/85} = 0.550$

$p - value = 0.579$

Dado no output da regressão

$Q_{obs} = 88 \times 0.012779 = 1.125$

$p - value = 0.570$

Heterocedasticidade: testes para detecção – exemplo 2

Exemplo: testar heterocedasticidade no âmbito do modelo

$$\widehat{ln(preço)} = 1.289 + 0.8101 \ln(\text{area}) + 0.0376 \text{ quartos}$$

```
gen larea=log(area)
gen lpreço=log(preco)
regress  lpreço larea quartos
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =   54.47
       Model |  4.50364223     2  2.25182112           Prob > F      =  0.0000
    Residual |  3.51396129    85  .041340721           R-squared     =  0.5617
-------------+------------------------------           Adj R-squared =  0.5514
       Total |  8.01760352    87  .092156362           Root MSE      =  .20332
------------------------------------------------------------------------------
      lpreço |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |   .8100637   .0987611     8.20   0.000     .6137002    1.006427
     quartos |   .0376464   .0303446     1.24   0.218    -.0226868    .0979795
       _cons |    1.28929   .4666125     2.76   0.007     .3615395    2.217041
------------------------------------------------------------------------------
predict uhat, resid
gen uhat2=uhat^2
predict yhat
gen yhat2=yhat^2
```

Heterocedasticidade: testes para detecção – exemplo 2

Exemplo (continuação):

Teste BP:

```
regress uhat2 larea quartos
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =    1.93
       Model |  .026033827     2  .013016913           Prob > F      =  0.1516
    Residual |  .573679783    85  .006749174           R-squared     =  0.0434
-------------+------------------------------           Adj R-squared =  0.0209
       Total |   .59971361    87   .00689326           Root MSE      =  .08215
------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       larea |  -.0607216   .0399045    -1.52   0.132    -.1400625    .0186193
     quartos |   .0227115   .0122608     1.85   0.067    -.0016662    .0470892
       _cons |   .2744371   .1885353     1.46   0.149    -.1004216    .6492957
------------------------------------------------------------------------------
* Teste LM
display 88*0.0434
3.8192
display chi2tail(2,3.8192)
.14813963
```

Não se rejeita H_0: (homocedasticidade) tanto pelo teste F como pelo LM

Heterocedasticidade: testes para detecção – exemplo 2

Exemplo (continuação):

Teste White:

```
gen la2=larea^2
gen qu2=quartos^2
gen laqu= larea*quartos
regress uhat2 larea quartos la2 qu2 laqu
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  5,    82) =    1.30
       Model |   .04396734     5  .008793468           Prob > F      =  0.2731
    Residual |   .55574627    82  .006777394           R-squared     =  0.0733
-------------+------------------------------           Adj R-squared =  0.0168
       Total |   .59971361    87   .00689326           Root MSE      =  .08232
------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
…
------------------------------------------------------------------------------
display 88*0.0733
6.4504
display chi2tail(5,6.4504)
.26482504
```

Não se rejeita H_0: (homocedasticidade) tanto pelo teste F como pelo LM

Heterocedasticidade: testes para detecção – exemplo 2

Exemplo (continuação):

Teste White simplificado:

```
regress uhat2 yhat yhat2
      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  2,    85) =    0.60
       Model |  .008374399     2    .0041872           Prob > F      =  0.5501
    Residual |  .591339211    85  .006956932           R-squared     =  0.0140
-------------+------------------------------           Adj R-squared = -0.0092
       Total |   .59971361    87   .00689326           Root MSE      =  .08341
------------------------------------------------------------------------------
       uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yhat |  -1.488234   1.453357    -1.02   0.309    -4.377898     1.40143
       yhat2 |   .1286466   .1269994     1.01   0.314    -.1238623    .3811554
       _cons |   4.334533   4.153408     1.04   0.300    -3.923556    12.59262
------------------------------------------------------------------------------
display 88*0.0140
1.232
display chi2tail(2,1.232)
.54010052
```

Não se rejeita H_0: (homocedasticidade) tanto pelo teste F como pelo LM

Heterocedasticidade: testes para detecção – exemplo 2

Teste BP - Eviews

Heteroskedasticity Test: Breusch-Pagan-Godfrey				
Null hypothesis: Homoskedasticity				
F-statistic	1.928666	Prob. F(2,85)		0.1516
Obs*R-squared	3.820114	Prob. Chi-Square(2)		0.1481
Scaled explained SS	7.616429	Prob. Chi-Square(2)		0.0222
Test Equation:				
Dependent Variable: RESID^2				
Method: Least Squares				
Date: 04/26/20 Time: 14:49				
Sample: 1 88				
Included observations: 88				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.274437	0.188535	1.455625	0.1492
LAREA	-0.060722	0.039905	-1.521670	0.1318
QUARTOS	0.022712	0.012261	1.852374	0.0674
R-squared	0.043410	Mean dependent var		0.039931
Adjusted R-squared	0.020902	S.D. dependent var		0.083026
S.E. of regression	0.082153	Akaike info criterion		-2.126962
Sum squared resid	0.573680	Schwarz criterion		-2.042507
Log likelihood	96.58631	Hannan-Quinn criter.		-2.092937
F-statistic	1.928666	Durbin-Watson stat		2.182135
Prob(F-statistic)	0.151649			

Não se rejeita H0, isto é, **não se rejeita a homocedasticidade**

Heterocedasticidade: testes para detecção – exemplo 2

Teste White – Eviews

Heteroskedasticity Test: White
Null hypothesis: Homoskedasticity

F-statistic	1.297465	Prob. F(5,82)	0.2731
Obs*R-squared	6.451600	Prob. Chi-Square(5)	0.2647
Scaled explained SS	12.86301	Prob. Chi-Square(5)	0.0247

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 04/26/20 Time: 14:50
Sample: 1 88
Included observations: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.673631	3.110976	1.502304	0.1369
LAREA^2	0.167924	0.123342	1.361455	0.1771
LAREA*QUARTOS	-0.007376	0.042017	-0.175542	0.8611
LAREA	-1.805730	1.235645	-1.461366	0.1477
QUARTOS^2	-0.008546	0.008595	-0.994323	0.3230
QUARTOS	0.127937	0.197192	0.648794	0.5183

R-squared	0.073314	Mean dependent var	0.039931
Adjusted R-squared	0.016808	S.D. dependent var	0.083026
S.E. of regression	0.082325	Akaike info criterion	-2.090539
Sum squared resid	0.555746	Schwarz criterion	-1.921630
Log likelihood	97.98372	Hannan-Quinn criter.	-2.022490
F-statistic	1.297465	Durbin-Watson stat	2.163854
Prob(F-statistic)	0.273087		

Não se rejeita H0, isto é, **não se rejeita a homocedasticidade**

Heterocedasticidade: testes para detecção – exemplo 2

Teste White simplificado– Eviews

Dependent Variable: RES2
Method: Least Squares
Date: 04/26/20 Time: 15:08
Sample: 1 88
Included observations: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.334571	4.153411	1.043617	0.2996
YC	-1.488248	1.453358	-1.024006	0.3087
YC2	0.128648	0.126999	1.012978	0.3139

R-squared	0.013964	Mean dependent var	0.039931
Adjusted R-squared	-0.009237	S.D. dependent var	0.083026
S.E. of regression	0.083408	Akaike info criterion	-2.096643
Sum squared resid	0.591339	Schwarz criterion	-2.012189
Log likelihood	95.25231	Hannan-Quinn criter.	-2.062619
F-statistic	0.601882	Durbin-Watson stat	2.247270
Prob(F-statistic)	0.550097		

$$RES2 = \left(\ln preço - \ln \widehat{preço}\right)^2$$
$$YC = \ln \widehat{preço}$$
$$YC2 = \left(\ln \widehat{preço}\right)^2$$

$$F_{obs} = 0.602$$
$$p - value = 0.550$$

Dado no output da regressão

$$Q_{obs} = 88 \times 0.013964 = 1.229$$
$$p - value = 0.541$$

rejeita a homocedasticidade