# Violando as hipóteses básicas

Celso J. Costa Junior

INTRODUÇÃO À ECONOMETRIA
19 de setembro de 2019

*cjcostaj@yahoo.com.br*

# Sumário


1 Introdução

2 Violando a hipótese VI: a Multicolinearidade

3 Violando a hipótese V: a autocorrelação

4 Violando a hipótese IV: a heteroscedasticidade

5 Exemplos no R

# Introdução

Hipóteses básicas sobre o modelo de regressão linear:

I $E(\epsilon_t) = 0$

II erros são normalmente distribuídos

III $X_t$ são fixos (não estocásticos)

IV $var(\epsilon_t) = \sigma^2$

V $E(\epsilon_t \epsilon_i) = 0, t \neq i$

VI cada variável independente $X$ não pode ser combinação linear das demais.

# Violando a hipótese VI: a Multicolinearidade

Multicolinearidade é a (alta) correlação entre duas (ou mais) variáveis em um modelo de regressão múltipla.

- Exemplos de Multicolinearidade:

$$X_1 = 2X_2$$

$$X_1 = X_2 + 3$$

$$X_1 = 4X_2 - 5$$

ou,

$$X_1 = 2X_2 + X_3 + 4$$

- Problema da Multicolinearidade: Exemplo, se $X_1 = 2X_2$, qualquer variação da segunda implicará em variação proporcionalmente idêntica da primeira. É impossível distinguir qual é a influência de uma ou de outra para a variável dependente Y.

# Violando a hipótese VI: a Multicolinearidade (cont.)

Exemplo 1.1: Suponha que o consumo é função da renda e da taxa real de juros. Se assumirmos ainda que esta relação é linear, teremos então que a especificação do modelo econométrico a ser estimado será dada por:

$$C_t = \beta_0 + \beta_1 Y_t + \beta_2 r_t + \mu_t$$

# Violando a hipótese VI: a Multicolinearidade (cont.)

| ano/trimestre | consumo (US$ bilhões) | renda (US$ bilhões) | taxa de juros real (% a.a.) |
|---|---|---|---|
| 1990/1 | 72,2 | 105,6 | 12,00 |
| 1990/2 | 75,6 | 97,4 | 12,50 |
| 1990/3 | 89,6 | 112,0 | 11,00 |
| 1990/4 | 93,7 | 128,0 | 10,00 |
| 1991/1 | 92,2 | 120,2 | 10,50 |
| 1991/2 | 84,6 | 115,3 | 10,75 |
| 1991/3 | 90,8 | 105,4 | 11,25 |
| 1991/4 | 82,9 | 103,6 | 12,00 |
| 1992/1 | 65,8 | 102,7 | 12,25 |
| 1992/2 | 70,9 | 93,2 | 13,00 |
| 1992/3 | 63,1 | 98,3 | 12,50 |
| 1992/4 | 86,3 | 108,1 | 11,75 |
| 1993/1 | 87,2 | 115,8 | 11,50 |
| 1993/2 | 79,3 | 99,8 | 11,00 |
| 1993/3 | 87,4 | 110,5 | 10,50 |
| 1993/4 | 100,6 | 127,8 | 10,25 |

# Violando a hipótese VI: a Multicolinearidade (cont.)

| | *coeficiente* | *desvio-padrão* | *estatística t* |
|---|---|---|---|
| constante | 111,487 | 66,840 | 1,667 |
| renda | 0,374 | 0,288 | 1,298 |
| taxa de juros real | -6,097 | 3,314 | 1,840 |

estatística F = 17,645

Repare que o valor tabelado da estatística t considerando-se 10% de significância e 13 graus de liberdade é 1,771, ou seja, apenas o coeficiente da taxa de juros é significante; se considerarmos 5% (2,160 como valor tabelado), todos os coeficientes não são significantes.

# Violando a hipótese VI: a Multicolinearidade (cont.)

**Consequências da multicolinearidade**

- os testes t podem resultar insignificantes, ainda que as variáveis sejam relevantes. Isto ocorre porque a variância dos coeficientes das variáveis explicativas aumenta quando ocorre multicolinearidade e daí o motivo dos testes t apresentarem baixa significância (ou mesmo não serem significantes).
- isto não significa que os testes t sejam inválidos.
- mesmo na presença de multicolinearidade, são mantidas as propriedades usuais do estimador de mínimos quadrados, isto é, continuam não viesados, eficientes e consistentes.

# Violando a hipótese VI: a Multicolinearidade (cont.)

**Como identificar a multicolinearidade?**

- um teste F bastante significante acompanhado de estatísticas t para os coeficientes pouco significantes, ou até mesmo não significantes.
- o cálculo direto da correlação entre as variáveis também é uma forma de identificar a presença de multicolinearidade.
- O cálculo da correlação, no entanto, pode não funcionar muito bem quando temos mais do que duas variáveis no modelo.

# Violando a hipótese VI: a Multicolinearidade (cont.)

**O que fazer quando há multicolinearidade?**

- A providência óbvia é retirar variáveis correlacionadas do modelo.
- Alterando o exemplo 1.1, $C_t = \beta_0 + \beta_1 Y_t + \mu_t$

| | *coeficiente* | *desvio-padrão* | *estatística t* |
|---|---|---|---|
| constante | -7,859 | 17,405 | 0,452 |
| renda | 0,830 | 0,159 | 5,221 |

estatística F = 27,264

Neste caso, evidentemente, a multicolinearidade necessariamente foi eliminada pois sobrou apenas uma variável explicativa.

# Violando a hipótese VI: a Multicolinearidade (cont.)

**O que fazer quando há multicolinearidade?**

- Muitas vezes é possível reduzir os efeitos da multicolinearidade através do aumento da amostra.
- Em alguns casos, seria possível reespecificar o modelo.
- Há ainda a alternativa de não se fazer nada.

# Violando a hipótese V: a autocorrelação

- Autocorrelação significa a correlação de uma variável com valores defasados (com diferenças no tempo) dela mesmo.
- A hipótese V faz menção a autocorrelação dos erros.
- O erro não é uma variável especificamente, mas um conjunto de diversas influências que, pela sua própria natureza, são difíceis de serem medidas, mas não exercem influência uma sobre a outra.
- Mas, e se exercerem? E por que exerceriam? A omissão desta variável "joga" sua influência, sistemática, para o termo de erro.
- Outro tipo de erro que poderia levar a autocorrelação seria a má especificação funcional.
- Mas a autocorrelação pode ocorrer pela própria natureza do processo.

# Violando a hipótese V: a autocorrelação (cont.)

- Um modelo de regressão em que a autocorrelação esteja presente:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \epsilon_t$$

onde:

$$\epsilon_t = \rho \epsilon_{t-1} + \mu_t$$

# Violando a hipótese V: a autocorrelação (cont.)

**Consequências da autocorrelação**

- A hipótese de não existência de autocorrelação nos erros é um pré-requisito para a demonstração do Teorema de Gauss-Markov, como o qual se mostra que o estimador de mínimos quadrados de uma regressão linear é um MELNV.
- Entretanto, que a hipótese necessária para que o estimador seja não viesado e consistente (que é a de que os regressores, os "X", não sejam correlacionados com o erro) não é violada e, portanto, ainda que não tenha a menor variância, o estimador continua, em geral, não viesado e consistente.
- As exceções são os modelos que incluem, entre as variáveis dependentes (regressores), defasagens da variável independente, $Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \epsilon_t$ e $\epsilon_t = \rho\epsilon_{t-1} + \mu_t$.

# Violando a hipótese V: a autocorrelação (cont.)

**Como identificar a autocorrelação?**

- A maneira mais comum de identificar a existência de autocorrelação é através do teste de Durbin-Watson:

$$DW = \frac{\sum_{t=2}^{n}(\widehat{\epsilon}_t - \widehat{\epsilon}_{t-1})^2}{\sum_{t=2}^{n}\widehat{\epsilon}_t^2} \approx 2(1 - \widehat{\rho})$$

- $\rho \approx 0$ e $DW \approx 2$, não há autocorrelação.
- $\rho \approx 1$ e $DW \approx 0$, há autocorrelação positiva.
- $\rho \approx -1$ e $DW \approx 4$, há autocorrelação negativa.
- Limitações do teste Durbin-Watson:
  - existe uma região em que o teste é inconclusivo.
  - a regressão não incluir o intercepto (termo constante).
  - a regressão incluir, como variáveis explicativas, defasagens da variável dependente.

# Violando a hipótese V: a autocorrelação (cont.)

- Mas quão distante de 2 deve estar o valor da estatística DW para que possamos concluir que existe, de fato, autocorrelação?
- Usando a tabela Durbin-Watson.
- Se, por exemplo, estivermos testando um modelo com duas variáveis explicativas, com 20 observações, para um nível de significância de 5%, encontramos os valores $d_i = 1,10$ e $d_s = 1,54$.
  - Se o valor de DW for abaixo de 1,10, rejeitamos a hipótese nula de não autocorrelação, isto é, concluímos que existe autocorrelação.
  - Se DW estiver entre 1,54 e 2, concluímos que não há autocorrelação (aceitamos a hipótese nula).
  - Se, entretanto, o valor de DW cair entre 1,10 e 1,54, o teste é inconclusivo, não dá para dizer se há ou não autocorrelação.

# Violando a hipótese V: a autocorrelação (cont.)

Exemplo 2.1: Na tabela abaixo encontramos dados de consumo e renda trimestrais de um país durante 5 anos. Estime a função consumo (consumo como função da renda) e teste a existência de autocorrelação, com 5% de significância.

# Violando a hipótese V: a autocorrelação (cont.)

| ano/trimestre | consumo (US$ bilhões) | renda (US$ bilhões) |
|---|---|---|
| **1994/3** | 757,6 | 970,0 |
| **1994/4** | 745,2 | 988,5 |
| **1995/1** | 673,4 | 866,5 |
| **1995/2** | 652,2 | 812,4 |
| **1995/3** | 676,2 | 845,3 |
| **1995/4** | 709,1 | 891,9 |
| **1996/1** | 704,7 | 899,3 |
| **1996/2** | 691,8 | 911,2 |
| **1996/3** | 696,6 | 903,2 |
| **1996/4** | 667,6 | 904,5 |
| **1997/1** | 667,2 | 906,7 |
| **1997/2** | 671,0 | 920,2 |
| **1997/3** | 716,9 | 958,4 |
| **1997/4** | 698,4 | 934,1 |
| **1998/1** | 676,7 | 944,4 |
| **1998/2** | 661,4 | 956,3 |
| **1998/3** | 686,8 | 971,7 |
| **1998/4** | 685,2 | 958,9 |
| **1999/1** | 684,9 | 961,9 |
| **1999/2** | 675,1 | 966,4 |
| **1999/3** | 663,1 | 977,5 |
| **1999/4** | 672,8 | 988,5 |
| **2000/1** | 675,2 | 1001,2 |
| **2000/2** | 693,1 | 996,7 |
| **2000/3** | 721,6 | 1005,6 |
| **2000/4** | 747,5 | 1011,2 |
| **2001/1** | 742,4 | 1004,2 |
| **2001/2** | 740,5 | 997,4 |
| **2001/3** | 741,5 | 1000,4 |
| **2001/4** | 722,6 | 1006,6 |

# Violando a hipótese V: a autocorrelação (cont.)

| | *coeficiente* | *desvio-padrão* | *estatística t* |
|---|---|---|---|
| constante | 402,672 | 87,676 | 4,59 |
| renda | 0,311 | 0,092 | 3,37 |

estatística F = 11,32

- O coeficiente da renda foi significante (a 1%) e a regressão foi válida ("aprovada" pelo teste F, a 1%). Agora, convém testar a existência de autocorrelação.
- $DW = 0,4454$.
- Como o limite inferior da tabela de DW é, para 5% de significância, 30 observações e uma variável explicativa, $d_i = 1,35$, ou, para 1% de significância, 1,20 (em ambos os casos, maior do que 0,4454), concluímos que existe autocorrelação.

# Violando a hipótese V: a autocorrelação (cont.)

**O que fazer quando há autocorrelação?**

- Primeiro há a questão de qual é a causa da autocorrelação. Se o problema é de especificação, ele pode ser corrigido com a inclusão de mais variáveis ou com a alteração da forma funcional.
- Se não é este o caso, ou seja, a autocorrelação é uma "parte integrante" do modelo estimado:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t \tag{1}$$

$$\epsilon_t = \rho \epsilon_{t-1} + \mu_t$$

# Violando a hipótese V: a autocorrelação (cont.)

Suponhamos ainda que o coeficiente $\rho$ seja conhecido, pois $\widehat{\rho} \approx 1 - \frac{DW}{2}$,

$$\rho Y_{t-1} = \rho\beta_0 + \rho\beta_1 X_{t-1} + \rho\epsilon_{t-1} \tag{2}$$

Subtraindo (2) de (1):

$$Y_t - \rho Y_{t-1} = \beta_0 - \rho\beta_0 + \beta_1 X_t - \rho\beta_1 X_{t-1} + \epsilon_t - \rho\epsilon_{t-1} \tag{3}$$

Sabendo que $\mu_t = \epsilon_t - \rho\epsilon_{t-1}$, e denominando, $Y_t^* = Y_t - \rho Y_{t-1}$, $\beta_0^* = \beta_0 - \rho\beta_0$ e $X_t^* = X_t - \rho X_{t-1}$. A equação (3) fica:

$$Y_t^* = \beta_0^* + \beta_1 X_t^* + \mu_t \tag{4}$$

# Violando a hipótese V: a autocorrelação (cont.)

Refazendo o exemplo 2.1, corrigindo o problema da autocorrelação, $\widehat{\rho} \approx 0,777$,

$$Y_t^* = Y_t - 0,777 Y_{t-1}$$

$$X_t^* = X_t - 0,777 X_{t-1}$$

| | *coeficiente* | *desvio-padrão* | *estatística t* |
|---|---|---|---|
| constante | 33,401 | 17,210 | 1,94 |
| $X^*$ | 0,566 | 0,081 | 6,97 |

estatística F = 48,52
DW = 1,3716

# Violando a hipótese V: a autocorrelação (cont.)

Se compararmos o valor encontrado (1,3716) com a tabela para 29 observações, veremos que, para 5% de significância, $d_i = 1,34$ e $d_s = 1,48$, o teste é inconclusivo. A 1% de significância, entretanto, os valores tabelados são $d_i = 1,12$ e $d_s = 1,25$, portanto aceitamos a hipótese de não existência de autocorrelação com esta significância.

# Violando a hipótese IV: a heteroscedasticidade

A hipótese IV estabelece que a variância dos erros deve ser constante (o que é conhecido como homoscedasticidade).

**Consequências da heteroscedasticidade:**

- A hipótese IV (assim como a hipótese V) é uma hipótese necessária para a demonstração do Teorema de Gauss-Markov. Desta forma, as conseqüências são basicamente as mesmas da presença da autocorrelação: os estimadores de mínimos quadrados ordinários continuam não viesados, mas já não são aqueles de menor variância. As variâncias dos estimadores são viesadas, invalidando assim os testes de hipóteses.

# Violando a hipótese IV: a heteroscedasticidade (cont.)

**Como identificar a heteroscedasticidade?**

- O teste de Goldfeld e Quandt consiste em separar a regressão em duas, uma com valores menores de X e outra com valores maiores e aí fazer um teste para comparar a variância em cada regressão (um teste comum de comparação de variâncias, isto é, um teste F).

# Violando a hipótese IV: a heteroscedasticidade (cont.)

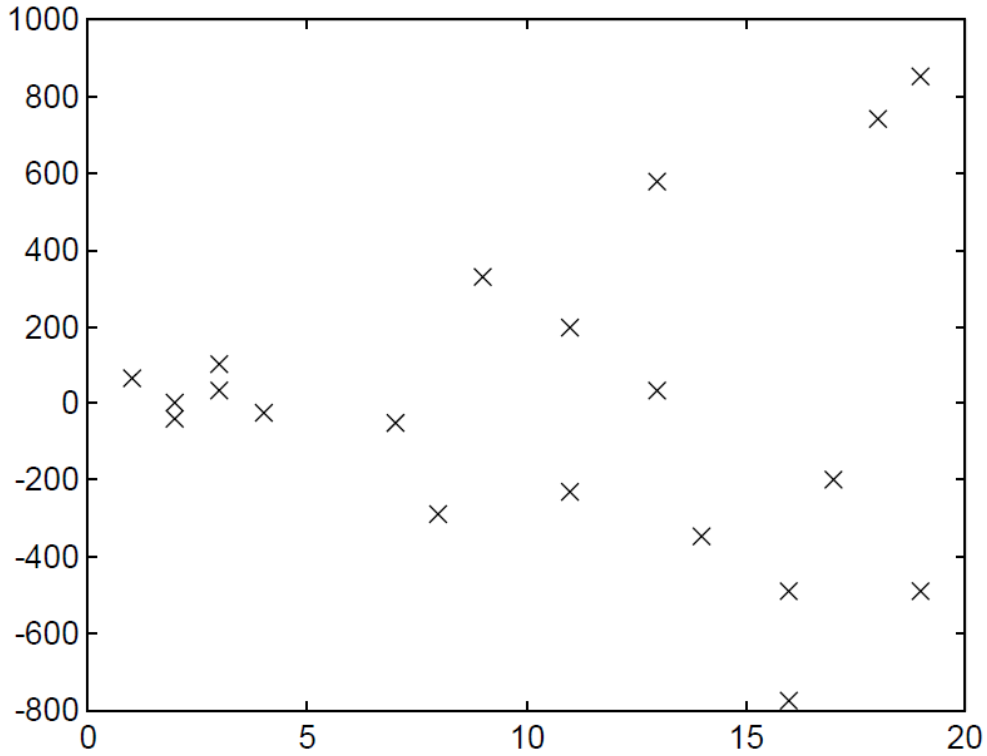
Exemplo 3.1: São dados na tabela abaixo os dados dos salários de 20 trabalhadores e os anos de estudo de cada um. Faça uma regressão dos salários em função dos anos de estudo e teste para a existência de heteroscedasticidade utilizando o teste de Goldfeld e Quandt.

| anos de estudo | salários (R$) |
|---|---|
| 1 | 410,00 |
| 2 | 508,90 |
| 3 | 857,70 |
| 2 | 551,30 |
| 3 | 789,20 |
| 4 | 935,50 |
| 7 | 1529,30 |
| 8 | 1497,50 |
| 9 | 2317,70 |
| 11 | 2169,50 |
| 11 | 2596,80 |
| 13 | 2844,60 |
| 13 | 3391,00 |
| 14 | 2671,20 |
| 16 | 2653,80 |
| 16 | 2939,10 |
| 17 | 3437,00 |
| 18 | 4583,30 |
| 19 | 3559,30 |
| 19 | 4896,70 |

# Violando a hipótese IV: a heteroscedasticidade (cont.)

| | *coeficiente* | *desvio-padrão* | *estatística t* |
|---|---|---|---|
| constante | 139,074 | 184,155 | 0,755 |
| anos de estudo | 205,621 | 15,400 | 13,35 |

F = 178,28



# Violando a hipótese IV: a heteroscedasticidade (cont.)

Para testarmos a heteroscedasticidade, dividiremos os dados em dois grupos como manda o teste de Goldfeld e Quandt. Esta divisão é arbitrária, mas o teste tende a ser mais eficiente se dividir os dados ao "meio".

$$\frac{var_{II}}{var_{I}} = \frac{345890,73}{3673,60} = 94,16$$

Como o valor limite na tabela F, com 5% de significância, para 5 graus de liberdade no numerador e 4 graus de liberdade no denominador é 6,26, rejeitamos a hipótese de que as variâncias sejam iguais (vale a hipótese de que a variância da segunda regressão é maior) e, portanto, rejeitamos a hipótese nula de homoscedasticidade.

# Violando a hipótese IV: a heteroscedasticidade (cont.)

**Como identificar a heteroscedasticidade?**

- Outro teste que pode ser usado para detecção do problema de heteroscedasticidade é o teste de White que consiste em, a partir de um modelo de regressão qualquer:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \epsilon_t$$

É feita uma regressão auxiliar onde a variável dependente é o resíduo ao quadrado e os regressores são os próprios regressores da regressão original, seus quadrados e os produtos cruzados, desta forma:

$$\widehat{\epsilon}_t^2 = \gamma_0 + \gamma_1 X_{1,t} + \gamma_2 X_{2,t} + \gamma_3 X_{1,t}^2 + \gamma_4 X_{2,t}^2 + \gamma_5 X_{1,t} X_{2,t} + \mu_t$$

Um $R^2$ elevado nesta regressão auxiliar é um indício de que há heteroscedasticidade. Pode-se demonstrar que $nR^2$ segue uma distribuição de $\chi^2$ com o número de graus de liberdade igual ao número de regressores da regressão auxiliar (menos o intercepto).

# Violando a hipótese IV: a heteroscedasticidade (cont.)

**O que fazer quando há heteroscedasticidade?**
Havendo heteroscedasticidade, o procedimento de "correção" é mais simples se soubermos (ou pelo menos, suspeitarmos) qual é o padrão da heteroscedasticidade.
Tomemos um modelo de regressão abaixo e suponhamos que exista heteroscedasticidade.

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \epsilon_t$$

Digamos que seja conhecido que a variância dos erros é dada por: $var(\epsilon_t) = \sigma_t^2 = Z_t \sigma^2$.

# Violando a hipótese IV: a heteroscedasticidade (cont.)

Dividindo a equação anterior por $\sqrt{Z_t}$:

$$\frac{Y_t}{\sqrt{Z_t}} = \beta_0 \frac{1}{\sqrt{Z_t}} + \beta_1 \frac{X_{1,t}}{\sqrt{Z_t}} + \beta_2 \frac{X_{2,t}}{\sqrt{Z_t}} + \frac{\epsilon_t}{\sqrt{Z_t}}$$

Assim,

$$var\left(\frac{\epsilon_t}{\sqrt{Z_t}}\right) = \frac{1}{Z_t} var(\epsilon_t) = \frac{1}{Z_t} Z_t \sigma^2 = \sigma^2 = var(\mu_t)$$

Então,

$$\frac{Y_t}{\sqrt{Z_t}} = \beta_0 \frac{1}{\sqrt{Z_t}} + \beta_1 \frac{X_{1,t}}{\sqrt{Z_t}} + \beta_2 \frac{X_{2,t}}{\sqrt{Z_t}} + \mu_t$$

Obs: Muitas vezes funciona assumir $Z_t = dp(\epsilon_t)$.

# Violando a hipótese IV: a heteroscedasticidade (cont.)

Corrigindo o problema da heteroscedasticidade do exemplo 3.1:

$$Y_t = \beta_1 + \beta_2 X_t + \epsilon_t$$

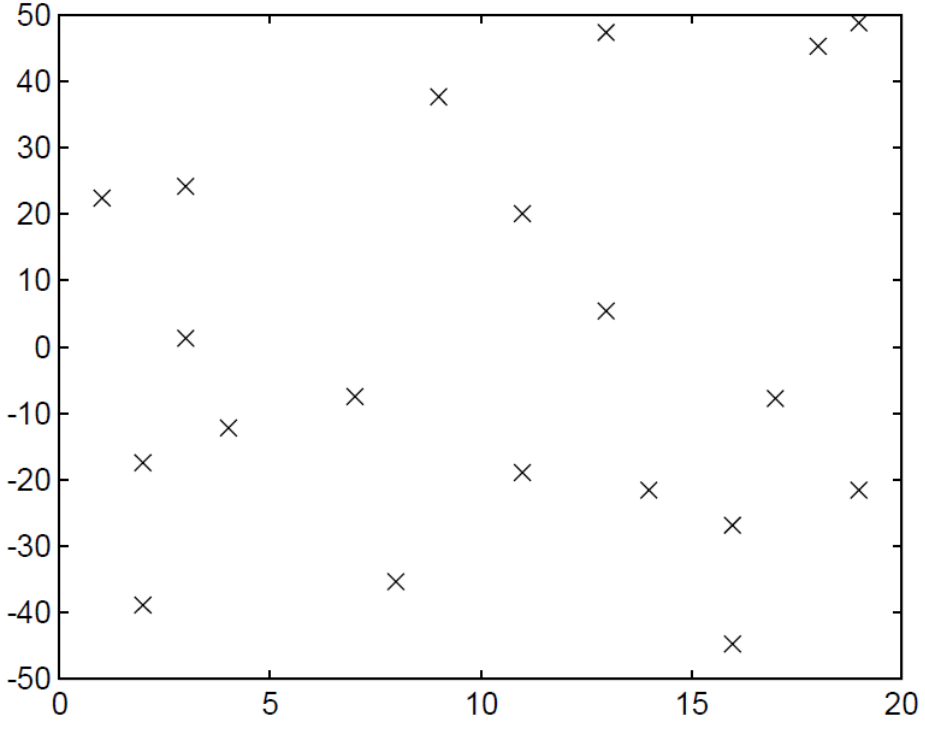
$$\frac{Y_t}{X_t} = \beta_1 \frac{1}{X_t} + \beta_2 \frac{X_t}{X_t} + \frac{\epsilon_t}{X_t}$$

$$\frac{Y_t}{X_t} = \beta_1 \frac{1}{X_t} + \beta_2 + \mu_t$$

# Violando a hipótese IV: a heteroscedasticidade (cont.)

| | *coeficiente* | *desvio-padrão* | *estatística t* |
|---|---|---|---|
| $\hat{\beta}_2$ | 198,869 | 9,126 | 21,79 |
| $\hat{\beta}_1$ | 188,745 | 29,716 | 6,35 |

$F = 40{,}34$



# Exemplos no R

Exemplo 1.1:

```
base1 <- ts(read.csv2('Exemplo1_1.csv', header=T, sep=';',
  dec=',')[,2:4], start=c(1990,01), freq=4)


reg1<- lm(base1[,1]~1+base1[,2]+base1[,3])

print(summary(reg1))

reg2<- lm(base1[,1]~1+base1[,2])

print(summary(reg2))
```

# Exemplos no R (cont.)

Exemplo 2.1:

```
library(lmtest)
library(xts)

base2 <- ts(read.csv2('Exemplo2_1.csv', header=T, sep=';',
  dec=',')[,2:3], start=c(1994,03), freq=4)


reg3<- lm(base2[,1]~1+base2[,2])
print(summary(reg3))

DW<-dwtest(reg3)

rho<-1-0.4454081/2
print(rho)

base2_star  <- (base2 - rho*lag.xts(base2, k = 1))

reg4<- lm(base2_star[,1]~1+base2_star[,2])
print(summary(reg4))
dwtest(reg4)
```

# Exemplos no R (cont.)

Exemplo 3.1:

```
library(lmtest)

base3 <- read.csv2('Exemplo3_1.csv', header=T, sep=';',dec=',')

reg5<- lm(base3[,2]~1+base3[,1])

print(summary(reg5))

bptest(reg5)
bptest(reg5,~ fitted(reg5) + I(fitted(reg5)^2))

Y_X<-base3[,2]/base3[,1]
X_1<-1/base3[,1]


reg6<- lm(Y_X~1+X_1)
print(summary(reg6))

bptest(reg6)
bptest(reg6,~ fitted(reg6) + I(fitted(reg6)^2))
```

END