# ANÁLISE DE DADOS MULTIVARIADOS I - REGRESSÃO

(AULA 07)

**Novembro e dezembro de 2018**

Reinaldo Soares de Camargo

# Violação dos pressupostos do modelo de regressão linear

# Modelos de Regressão - Pressupostos

$$Y = \beta_1 + \beta_2 X_1 + \ldots + \beta_k X_k + u_i \quad \text{ou} \quad \mathrm{Y} = \mathrm{X}\beta + \mathrm{u}$$

1. O termo de erro possui média zero: E ($u_i$) = 0
2. Os erros são normalmente distribuídos.
3. A variável *x* não é aleatória e deve assumir pelo menos dois valores diferentes.
4. **Variância do erro é constante: var($u_i$) = $\sigma^2$**
5. **Os erros não são autocorrelacionados : E($u_i u_j$) = 0, I ≠ j**
6. Cada variável independente $X_i$ não pode ser combinação linear das demais

Os erros não são autocorrelacionados : E($u_i u_j$) = 0, I ≠ j

# Modelos de Regressão

Teorema de Gaus-Markov

Dados os pressupostos do modelo de regressão linear, os estimadores por mínimos quadrados ordinários (MQO) na classe dos estimadores lineares não-enviesados, têm variância mínima, ou seja, são os Melhores Estimadores Lineares Não Viesados (MELNV), ou BLUE do inglês (Best Linear Unbiased Estimator)

# Modelos de Regressão

- Testes comumente empregados para as hipóteses de regressão:
    - Testes de heteroscedasticidade
    - Testes de autocorrelação serial
    - Testes de normalidade dos resíduos
    - Testes de multicolinearidade
- Os testes podem indiretamente indicar problemas na forma funcional do modelo. Nesse caso, talvez seja necessário incluir outras variáveis, ou aplicar transformações nas varáveis (log, segunda potência, incluir mais defasagens etc.)
- A maioria dos procedimentos descritos a seguir aplicam-se a regressão com dados *cross-section*. No entanto, podemos aplicar técnicas análogas para dados de painel.
- Há várias funcionalidades no R para tratar de check dos resíduos em modelos de regressão.
- Nesta aula, vamos utilizar o programa em R, “ajuste_7.R”
- O pacote utilizado para testes dos resíduos nesse caso é o pacote “lmtest”
- O pacote utilizado para correção das estimativas é o pacote “sandwich”

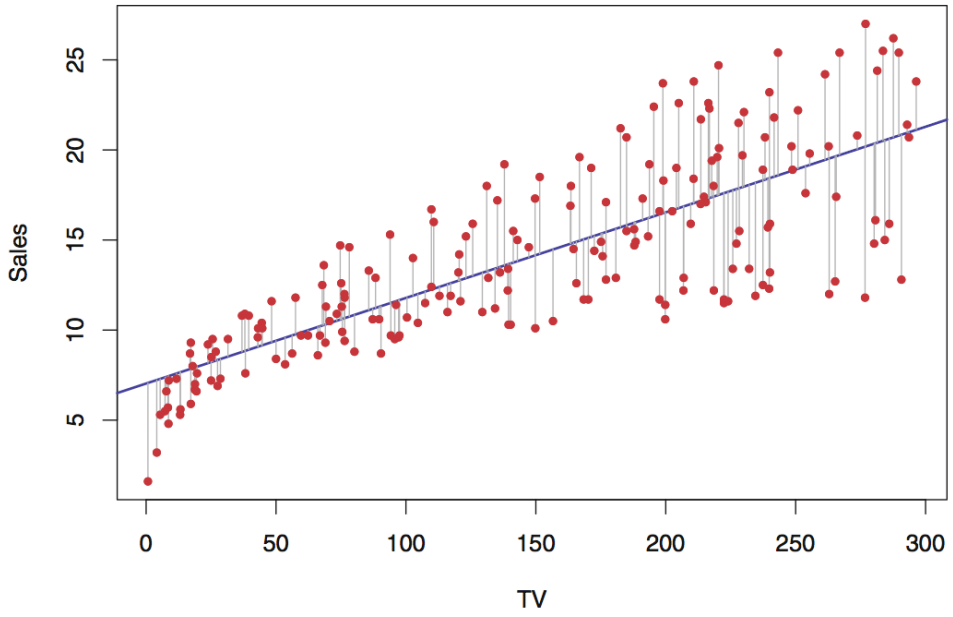
# Heteroscedasticidade

- **Testes de heteroscedasticidade**. Como a grande maioria de testes de checks dos resíduos, esses testes são feitos em dois estágios

- Em um primeiro estágio, é feita uma regressão linear tradicional, com equação

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \epsilon_i$$

- Calculam-se os resíduos da regressão, $\hat{\epsilon}_i = y_i - \hat{y}_i$



# Heteroscedasticidade

- Em um segundo estágio, roda-se uma regressão com os resíduos da regressão ao quadrado como variável dependente e as variáveis preditoras da regressão original como variáveis independentes

$$\hat{\epsilon}_i^2 = \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \ \ldots + \delta_k x_{ki} + \nu_i \qquad \text{(A)}$$

- O teste de heteroscedasticidade consiste simplesmente em um teste F, no qual se testam a significância conjunta dos parâmetros $\delta_1, \ldots, \delta_k$ no segundo estágio

- Hipótese nula: $H_0$: $\delta_1 = \delta_2 = \ \ldots = \delta_k = 0$

- Se a hipótese nula for verdadeira, tem-se uma indicação de que não há heteroscedasticidade nos resíduos da regressão (erros homocedasticos)

- Portanto, caso rejeitemos a hipótese nula, há indícios de heteroscedasticidade nos erros do modelo

- Esse teste é conhecido como teste de Breusch-Pagan

- Resumidamente, se não existe heteroscedasticidade, é de se esperar que os resíduos ao quadrado não aumentem ou diminuam com o aumento do valor predito, $\hat{y}$ e assim, a estatística de teste deveria ser insignificante.

# Heteroscedasticidade

**No R:** Utilizando o programa “ajuste_7.R”, rode testes de heteroscedasticidade. Com dados da aula 6: censo_2000.csv.

```
mod1.ex <- lm(dados$mort_infantil ~ dados$renda_per_capita
              + dados$indice_gini
              + dados$perc_criancas_extrem_pobres
              + dados$perc_criancas_pobres
              + dados$perc_pessoas_dom_agua_estogo_inadequados
              + dados$perc_pessoas_dom_paredes_inadequadas
              + dados$perc_pop_dom_com_coleta_lixo)
```

# Heteroscedasticidade (1º estágio)

```
mod1.ex <- lm(dados$mort_infantil ~ dados$renda_per_capita
              + dados$indice_gini
              + dados$perc_criancas_extrem_pobres
              + dados$perc_criancas_pobres
              + dados$perc_pessoas_dom_agua_estogo_inadequados
              + dados$perc_pessoas_dom_paredes_inadequadas
              + dados$perc_pop_dom_com_coleta_lixo)
```

```
Call:
lm(formula = dados$mort_infantil ~ dados$renda_per_capita + dados$indice_gini +
    dados$perc_criancas_extrem_pobres + dados$perc_criancas_pobres +
    dados$perc_pessoas_dom_agua_estogo_inadequados + dados$perc_pessoas_dom_paredes_inadequadas +
    dados$perc_pop_dom_com_coleta_lixo)

Residuals:
    Min      1Q  Median      3Q     Max
-16.520  -2.502  -0.370   1.922  20.815

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                1.920e+01  8.150e-01  23.558  < 2e-16 ***
dados$renda_per_capita                    -1.476e-03  5.687e-04  -2.595  0.00948 **
dados$indice_gini                         -1.443e+01  1.245e+00 -11.593  < 2e-16 ***
dados$perc_criancas_extrem_pobres          3.769e-02  1.216e-02   3.101  0.00194 **
dados$perc_criancas_pobres                 2.178e-01  1.143e-02  19.046  < 2e-16 ***
dados$perc_pessoas_dom_agua_estogo_inadequados  5.009e-02  6.017e-03   8.325  < 2e-16 ***
dados$perc_pessoas_dom_paredes_inadequadas  4.255e-02  7.923e-03   5.370 8.17e-08 ***
dados$perc_pop_dom_com_coleta_lixo        -7.756e-03  6.511e-03  -1.191  0.23362
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.039 on 5556 degrees of freedom
Multiple R-squared:  0.6802,    Adjusted R-squared:  0.6798
F-statistic:  1688 on 7 and 5556 DF,  p-value: < 2.2e-16
```

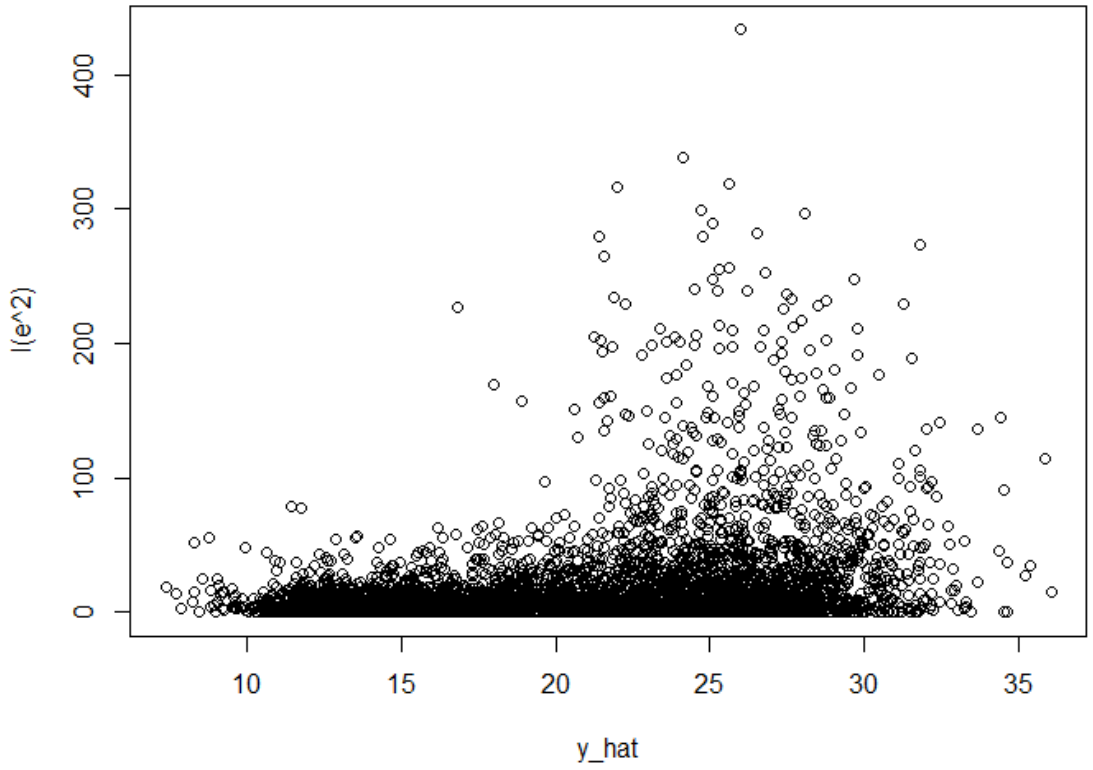
# Heteroscedasticidade (2º estágio)

```
> #---- obtem residuos da regressão -------#
>
> e <- mod.ex$residuals
> summary (e)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-16.520  -2.502  -0.370   0.000   1.922  20.815
```

```
> #----Estima o modelo do erro quadrado em relação as variáveis explicativas -------#
> mod2.ex <- lm(I(e^2) ~ dados$renda_per_capita
+              + dados$indice_gini
+              + dados$perc_criancas_extrem_pobres
+              + dados$perc_criancas_pobres
+              + dados$perc_pessoas_dom_agua_estogo_inadequados
+              + dados$perc_pessoas_dom_paredes_inadequadas
+              + dados$perc_pop_dom_com_coleta_lixo)
> summary (mod2.ex)

Call:
lm(formula = I(e^2) ~ dados$renda_per_capita + dados$indice_gini +
    dados$perc_criancas_extrem_pobres + dados$perc_criancas_pobres +
    dados$perc_pessoas_dom_agua_estogo_inadequados + dados$perc_pessoas_dom_paredes_inadequadas +
    dados$perc_pop_dom_com_coleta_lixo)

Residuals:
   Min     1Q Median     3Q    Max
-49.02 -12.06  -3.78   2.18 404.17

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   -4.670776   6.248909  -0.747  0.45482
dados$renda_per_capita                         0.007655   0.004360   1.756  0.07920 .
dados$indice_gini                            -22.877892   9.544517  -2.397  0.01656 *
dados$perc_criancas_extrem_pobres              0.041831   0.093204   0.449  0.65358
dados$perc_criancas_pobres                     0.559044   0.087676   6.376 1.96e-10 ***
dados$perc_pessoas_dom_agua_estogo_inadequados 0.081220   0.046135   1.760  0.07838 .
dados$perc_pessoas_dom_paredes_inadequadas     0.225316   0.060747   3.709  0.00021 ***
dados$perc_pop_dom_com_coleta_lixo             0.074504   0.049921   1.492  0.13565
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.97 on 5556 degrees of freedom
Multiple R-squared:  0.1306,    Adjusted R-squared:  0.1295
F-statistic: 119.3 on 7 and 5556 DF,  p-value: < 2.2e-16
```

# Heteroscedasticidade

```
> #-------- Gráficos dos residuos ao quadradro em relação ao y_hat ----------------------#
> #---- Caso não exista heterocedasticidade é de se esperar que residuos ao quadrado não---#
> #---- não aumentem ou diminuem com o aumento do valor do Y_hat---------------------------#
>
> plot(I(e^2)~y_hat)
>
```




Alguma suspeita?

# Heteroscedasticidade

```
> #-----Teste de Breusch-Pagan -------#
> # H0: os coeficientes estimados são iguais a zero.
>
> bptest( dados$mort_infantil ~ dados$renda_per_capita
+         + dados$indice_gini
+         + dados$perc_criancas_extrem_pobres
+         + dados$perc_criancas_pobres
+         + dados$perc_pessoas_dom_agua_estogo_inadequados
+         + dados$perc_pessoas_dom_paredes_inadequadas
+         + dados$perc_pop_dom_com_coleta_lixo, data=dados)

        studentized Breusch-Pagan test

data:  dados$mort_infantil ~ dados$renda_per_capita + dados$indice_gini +     dados$perc_criancas_extrem_pobres + dados$perc
_criancas_pobres +     dados$perc_pessoas_dom_agua_estogo_inadequados + dados$perc_pessoas_dom_paredes_inadequadas +     dad
os$perc_pop_dom_com_coleta_lixo
BP = 726.83, df = 7, p-value < 2.2e-16
```

Qual conclusão?

# Heteroscedasticidade

**Exercício no R:** Baixe a base de dados dados_turma.csv e teste a presença de heterocedasticidade no modelo que relaciona altura ao peso e ao gênero.

```
dados2 <- read.table("https://raw.githubusercontent.com/Cayan-Portela/ENAP_regressao/master/Aula%2002/dados_turma.csv",
                     header = TRUE)

mod3.exe <- lm(Altura~Peso + Genero, data = dados2)
summary(mod3.exe)
```

# Heteroscedasticidade

- Uma outra versão para o teste de heteroscedasticidade discutido anteriormente é obtida adicionando-se mais termos à segunda regressão
- Nesse caso, usam-se termos quadráticos das variáveis da regressão original

$$\hat{\epsilon}_i^2 = \delta_0 + \delta_1 x_{1i} + \cdots + \delta_k x_{ki} + \lambda_1 x_{1i}^2 + \cdots + \lambda_k x_{ki}^2 + \nu_i \qquad \text{(B)}$$

- O teste de heteroscedasticidade consiste em um teste F, no qual se testam a significância conjunta dos parâmetros $\delta_1, \dots, \delta_k, \lambda_1, \dots, \lambda_k$, no segundo estágio
- Hipótese nula: $H_0$: $\delta_1 = \cdots = \delta_k = \lambda_1 = \cdots = \lambda_k = 0$
- Alternativamente, podem-se inserir produtos das variáveis da regressão original

$$\hat{\epsilon}_i^2 = \delta_0 + \delta_1 x_{1i} + \cdots + \delta_k x_{ki} + \lambda_1 x_{1i}^2 + \cdots + \lambda_k x_{ki}^2 + \kappa_1 x_{1i} x_{2i} + \kappa_2 x_{1i} x_{3i} + \nu_i \qquad \text{(C)}$$

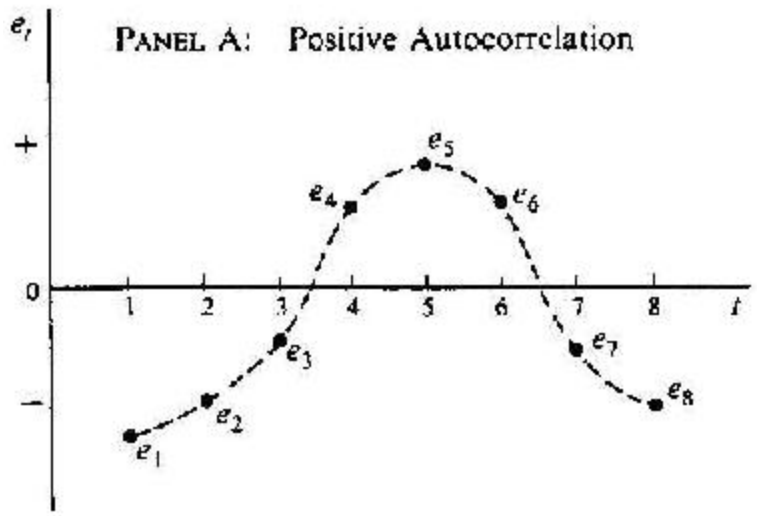
- O teste de heteroscedasticidade consiste em um teste F, no qual se testam a significância conjunta dos parâmetros $\delta_1, \dots, \delta_k, \lambda_1, \dots, \lambda_k, \kappa_1, \kappa_2$, no segundo estágio
- Hipótese nula: $H_0$: $\delta_1 = \cdots = \delta_k = \lambda_1 = \cdots = \lambda_k = \kappa_1 = \kappa_2 = \cdots = 0$
- Esse teste é conhecido como teste de White

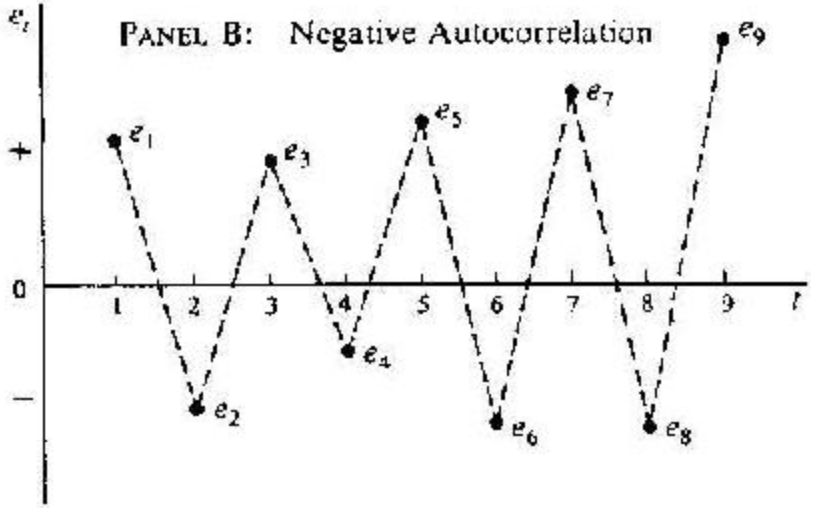
# Autocorrelação Serial

- **Testes de autocorrelação serial**. Os testes de autocorrelação serial também são feitos em dois estágios
- Esses testes fazem mais sentido quando temos observações ao longo do tempo
- Em um primeiro estágio, é feita uma regressão linear tradicional, com equação

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \epsilon_i$$

- Calculam-se os resíduos da regressão, $\hat{\epsilon}_i = y_i - \hat{y}_i$



# Autocorrelação Serial

- Em um segundo estágio, roda-se uma regressão com os resíduos da regressão versus defasagens dos próprios resíduos

$$\hat{\epsilon}_i = \delta_0 + \delta_1 \hat{\epsilon}_{i-1} + \delta_2 \hat{\epsilon}_{i-2} + \ \ldots + \delta_m \hat{\epsilon}_{i-m} + \nu_i \qquad \text{(D)}$$

- O teste de autocorrelação serial consiste simplesmente em um teste F, no qual se testam a significância conjunta dos parâmetros $\delta_1$, ... , $\delta_k$ no segundo estágio

- Hipótese nula: $H_0$: $\delta_1 = \delta_2 = \ \ldots = \delta_k = 0$

- Se a hipótese nula for verdadeira, tem-se uma indicação de que não há autocorrelação serial nos resíduos da regressão

- Portanto, caso rejeitemos a hipótese nula, há indícios de autocorrelação serial nos erros do modelo

- Podemos especificar diferentes valores para o número defasagens $m$ para verificar a robustez dos resultados do teste de autocorrelação serial. Em geral, para dados anuais podemos usar *m* = 1 ou 2. Para dados trimestrais, usamos *m* = 5 ou 9.

- Esse teste é conhecido como teste de Breusch-Godfrey

# Autocorrelação Serial e Heteroscedasticidade

- **No R:**

```
#---- testes de heteroscedasticidade dos resíduos

modelCH <- RAARUS ~ MOOD + EPI + EXP + RUS
mod1 <- lm(modelCH, data=bondyield)
summary(mod1)

bptest(modelCH, data=bondyield)
bptest(modelCH, varformula = RAARUS ~ MOOD + EPI + EXP + RUS, data=bondyield)
bptest(modelCH, varformula = RAARUS ~ MOOD + EPI + EXP + RUS +
                          I(MOOD^2) + I(EPI^2) + I(EXP^2) + I(RUS^2), data=bondyield)
bptest(modelCH, varformula = RAARUS ~ MOOD + EPI + EXP + RUS +
                          I(MOOD^2) + I(EPI^2) + I(EXP^2) + I(RUS^2) +
                          I(MOOD*EPI), data=bondyield)

#---- testes de autocorrelação serial dos resíduos

modelCH <- RAARUS ~ MOOD + EPI + EXP + RUS
mod1 <- lm(modelCH, data=bondyield)
summary(mod1)

bgtest(modelCH, data=bondyield)          #--- uma defasagem (default)
bgtest(modelCH, order = 1, data=bondyield)   #--- uma defasagem
bgtest(modelCH, order = 4, data=bondyield)   #--- quatro defasagens
```

# Autocorrelação Serial e Heteroscedasticidade

- **No R:**

```
> bptest(modelCH, data=bondyield)

        studentized Breusch-Pagan test

data:  modelCH
BP = 2.9784, df = 4, p-value = 0.5614

> bgtest(modelCH, order = 4, data=bondyield)    #--- quatro defasagens

        Breusch-Godfrey test for serial correlation of order up to 4

data:  modelCH
LM test = 22.421, df = 4, p-value = 0.0001652
```

# Modelos de Regressão

- **Quais as implicações de identificarmos a presença de autocorrelação serial e/ou heteroscedasticidade nos resíduos da regressão?**

- Em primeiro lugar, a esses dois problemas não implicam que as **estimativas dos coeficientes sejam viesadas**. Portanto, em média, a regressão continua "acertando" nos valores verdadeiros dos parâmetros da equação.

- Podemos continuar usando as estimativas dos coeficientes que obtemos nos outputs da regressão.

- Dois problemas acontecem quando há autocorrelação serial ou heteroscedasticidade:

    - *Problema 1*. Em primeiro lugar, as **estimativas dos coeficientes via MQO não são eficientes**. Portanto, há estimadores diferentes, também lineares, que provêm coeficientes com menor dispersão em torno dos coeficientes verdadeiros. A imprecisão desses outros coeficientes são mais precisas. Esses são conhecidos como GLS (*generalized least squares*) ou estimadores de mínimos quadrados generalizados.

    - *Problema 2*. Em segundo lugar, **as estimativas de erros padrões que saem no output da regressão (do R, por exemplo) não são mais válidas**. Nesse caso, as estatísticas *t*, os p-valores e demais testes de hipótese, os intervalos de confiança, não são mais válidos.

# Modelos de Regressão

- A abordagem tradicional mais antiga comumente utilizada para o problema 1 era tentar modelar o tipo de autocorrelação ou o tipo de heteroscedasticidade. Dessa forma, buscava-se resolver os dois problemas ao mesmo tempo.

    - Softwares como SAS possuem funções específicas para modelar parametricamente a heteroscedasticidade ou a autocorrelação serial, obtendo-se supostamente estimativas mais precisas para os coeficientes. Esses estimadores têm o nome de FGLS (*feasible generalized least squares*).

    - Obtém-se automaticamente estimativas "corretas" para os erros padrões, e consequentemente para as estatísticas *t*, para os p-valores, e para os intervalos de confiança

- O problema com essas abordagens é que podemos estar resolvendo um problema criando-se outro. Quando tentamos modelar explicitamente a heteroscedasticidade ou a autocorrelação serial, podemos estar errando na forma dessa correção, e criando uma imprecisão ainda maior.

- Atualmente, principalmente em econometria, parte-se para uma alternativa mais simples. Não mais se procura modelar explicitamente a autocorrelação serial ou a heteroscedasticidade.

# Modelos de Regressão

- Abordagem comumente utilizada atualmente:
  - Não nos preocupamos com eficiência do estimador de mínimos quadrados ordinários. Portanto, dado que os coeficientes estimados são não viesados, não se tenta mais obter estimadores mais eficientes (com maior precisão) do que os estimadores de MQO
  - No entanto, ainda há o problema de erros padrões estarem incorretos. Como resolver esse problema?
  - A literatura evoluiu na tentativa de estimar os erros padrões de forma robusta. Portanto, o objetivo agora é: dado que existe autocorrelação serial ou heteroscedasticidade nos resíduos (de forma geral), como podemos estimar corretamente os erros padrões?
  - Uma vez corrigidos os erros padrões, os testes de hipótese, p-valores, estatísticas *t*, e intervalos de confiança também são mais críveis
  - No R, há funcionalidades para tratar essas correções. Um pacote comumente utilizado é o pacote “sandwich”. Vide programa “Analise_de_Regressao_com_Cheque_Residuos.R”

# Modelos de Regressão

- Abordagem comumente utilizada atualmente (continuação):
  - Quando empregamos as diferentes correções para os erros padrões, observa-se que as estimativas dos coeficientes não se alteram
  - Dependendo da correção, os erros padrões, e consequentemente as estatísticas *t* e os p-valores apresentam valores diferentes
  - As correções comumente utilizadas dividem-se em dois grupos: correções apenas para heteroscedasticidade e correções para heteroscedasticidade e autocorrelação serial
  - No primeiro caso, as diferentes correções têm abreviação HC (*heteroscedasticity correction*)
  - No segundo caso, a abreviação utilizada é HAC (*heteroscedasticity and autocorrelation correction*)
  - Na prática, vez determinada qual dos dois grupos acima se deseja utilizar, sugere-se empregar diferentes correções e verificar se as conclusões se mantém (em geral, as conclusões não mudam – dentre as diferentes correções). As conclusões podem mudar quando comparadas ao estimador MQO original.

# Modelos de Regressão

- **No R:**

```
#---- estimadores robustos para erros heteroscedasticos

summary(mod.ex)

coeftest(mod.ex, vcov = sandwich)                  # robust; sandwich
coeftest(mod.ex, vcov = vcovHC(mod1, "HC0"))   # robust; HC0

coeftest(mod.ex, vcov = vcovHC(mod1, "HC1"))   # robust; HC1
coeftest(mod.ex, vcov = vcovHC(mod1, "HC2"))   # robust; HC2
coeftest(mod.ex, vcov = vcovHC(mod1, "HC3"))   # robust; HC3

#---- estimadores robustos para erros heteroscedasticos e autocorrelacionados

summary(mod.ex)

coeftest(mod.ex, vcov = vcovHAC(mod.ex))
```

# Modelos de Regressão

- Utilizando o programa “ajuste_7.R”, para a regressão a seguir, utilizando o pacote “sandwich”:

```
mod1.ex <- lm(dados$mort_infantil ~ dados$renda_per_capita
   + dados$indice_gini
   + dados$perc_criancas_extrem_pobres
   + dados$perc_criancas_pobres
   + dados$perc_pessoas_dom_agua_estogo_inadequados
   + dados$perc_pessoas_dom_paredes_inadequadas
   + dados$perc_pop_dom_com_coleta_lixo)
```

1. Aplique correções para heteroscedasticidade e verifique se as conclusões sobre a significância dos parâmetros se mantém

2. Aplique correções para heteroscedasticidade e autocorrelação serial e verifique se as conclusões sobre a significância dos parâmetros se mantém

# Modelos de Regressão

- **3ª Lista de exercícios para entregar em 26/11/2018.**
  - Os exercícios podem ser entregues em grupos de 2 alunos, e o grupo deve submeter o código em R utilizado para responder ao exercício, juntamente com a discussão dos resultados.
  - Utilize a base de dados do IDH brasil 2010 (IDH_Brasil_2010.csv)
  - Rode a regressão de acordo com o modelo abaixo:

```
mod1.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
                                   + dados3$indice_gini
                                   + dados3$salario_medio_mensal
                                   + dados3$perc_criancas_extrem_pobres
                                   + dados3$perc_criancas_pobres
                                   + dados3$perc_pessoas_dom_agua_estogo_inadequados
                                   + dados3$perc_pessoas_dom_paredes_inadequadas
                                   + dados3$perc_pop_dom_com_coleta_lixo)


summary(mod1.ex)
```

# Modelos de Regressão

1. Aplique correções para heteroscedasticidade e verifique se as conclusões sobre a significância dos parâmetros se mantém

2. Aplique correções para heteroscedasticidade e autocorrelação serial e verifique se as conclusões sobre a significância dos parâmetros se mantém

3. Faço o QQ-plot dos resíduos da regressão. Pelo QQ-plot, há indícios de violação da hipótese de normalidade dos resíduos da regressão?

4. Teste a normalidade dos resíduos utilizando os testes: Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling

5. Qual a sua conclusão geral sobre a normalidade dos resíduos da regressão?

7. Plote o gráfico com a função “ggpairs” para checar a correlação entre pares de variáveis preditoras. Há algum par com correlação alta (maior do que 0.8)?

8. Teste a presença de multicolinearidade no modelo, utilizando a função “omcdiag”.

9. Qual a sua conclusão sobre a presença de multicolinearidade na regressão?

Obrigado!