

ANÁLISE DE DADOS MULTIVARIADOS I - REGRESSÃO

(3ª Lista de Exercícios)

Novembro e dezembro de 2018

Reinaldo Soares de Camargo

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$perc_crianças_extrem_pobres  
              + dados3$perc_crianças_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 mantêm.

teste de Breusch-Pagan

- Hipótese nula: $H_0: \delta_1 = \delta_2 = \dots = \delta_k = 0$
- Se a hipótese nula for verdadeira, tem-se uma indicação de que não há heteroscedasticidade nos resíduos da regressão (erros homocedasticos)

```
> bptest(dados3$mort_infantil ~ dados3$renda_per_capita
+       + dados3$indice_gini
+       + dados3$salario_medio_mensal
+       + dados3$perc_crianças_extrem_pobres
+       + dados3$perc_crianças_pobres
+       + dados3$perc_pessoas_dom_agua_estogo_inadequados
+       + dados3$perc_pessoas_dom_paredes_inadequadas
+       + dados3$perc_pop_dom_com_coleta_lixo)

studentized Breusch-Pagan test

data: dados3$mort_infantil ~ dados3$renda_per_capita + dados3$indice_gini +      dados3$salario_medio_mensal + dados3$perc_cr
ianças_extrem_pobres +      dados3$perc_crianças_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +      dados3$perc_pe
ssoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo
BP = 726.01, df = 8, p-value < 2.2e-16
```

Como $p\text{-value} < 0,05$ rejeita-se H_0 , portanto , o modelo possui erros heterocedasticos.

Modelos de Regressão

Modelo original

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.936e+01  8.196e-01  23.627 < 2e-16 ***
dados3$renda_per_capita -1.278e-03  5.784e-04  -2.209 0.02721 *
dados3$indice_gini      -1.430e+01  1.247e+00 -11.470 < 2e-16 ***
dados3$salario_medio_mensal -1.775e-01  9.515e-02  -1.866 0.06212 .
dados3$perc_crianças_extrem_pobres 3.854e-02  1.216e-02   3.169 0.00154 **
dados3$perc_crianças_pobres      2.159e-01  1.148e-02  18.812 < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados 5.055e-02  6.021e-03   8.397 < 2e-16 ***
dados3$perc_pessoas_dom_paredes_inadequadas 4.297e-02  7.924e-03   5.423 6.12e-08 ***
dados3$perc_pop_dom_com_coleta_lixo -7.045e-03  6.520e-03  -1.080 0.27999
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.038 on 5555 degrees of freedom
Multiple R-squared:  0.6804,    Adjusted R-squared:  0.6799
F-statistic: 1478 on 8 and 5555 DF, p-value: < 2.2e-16
```

Estimadores robustos para erros heterocedasticos

```
> coeftest(mod1.ex, vcov = vcovHC(mod1.ex, "HC3"))
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.9365e+01  1.0962e+00  17.6660 < 2.2e-16 ***
dados3$renda_per_capita -1.2777e-03  5.0962e-04  -2.5071 0.0122003 *
dados3$indice_gini      -1.4298e+01  1.2633e+00 -11.3173 < 2.2e-16 ***
dados3$salario_medio_mensal -1.7752e-01  1.1386e-01  -1.5592 0.1190152
dados3$perc_crianças_extrem_pobres 3.8540e-02  1.5262e-02   2.5252 0.0115914 *
dados3$perc_crianças_pobres      2.1590e-01  1.3541e-02  15.9438 < 2.2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados 5.0554e-02  8.5008e-03   5.9470 2.898e-09 ***
dados3$perc_pessoas_dom_paredes_inadequadas 4.2969e-02  1.1962e-02   3.5921 0.0003309 ***
dados3$perc_pop_dom_com_coleta_lixo -7.0450e-03  9.8643e-03  -0.7142 0.4751344
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Resposta: Os coeficientes significantes a 5% se mantiveram, embora o erro padrão, o t-value e p-value tenha se modificado com a estimação robusta para erros heterocedasticos.

Modelos de Regressão

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

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.936e+01  8.196e-01  23.627 < 2e-16 ***
dados3$renda_per_capita -1.278e-03  5.784e-04  -2.209 0.02721 *
dados3$indice_gini    -1.430e+01  1.247e+00 -11.470 < 2e-16 ***
dados3$salario_medio_mensal -1.775e-01  9.515e-02  -1.866 0.06212 .
dados3$perc_crianças_extrem_pobres  3.854e-02  1.216e-02   3.169 0.00154 **
dados3$perc_crianças_pobres    2.159e-01  1.148e-02  18.812 < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  5.055e-02  6.021e-03   8.397 < 2e-16 ***
dados3$perc_pessoas_dom_paredes_inadequadas    4.297e-02  7.924e-03   5.423 6.12e-08 ***
dados3$perc_pop_dom_com_coleta_lixo   -7.045e-03  6.520e-03  -1.080 0.27999
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.038 on 5555 degrees of freedom
Multiple R-squared:  0.6804,    Adjusted R-squared:  0.6799
F-statistic: 1478 on 8 and 5555 DF,  p-value: < 2.2e-16
```

```
> coeftest(mod1, vcov = vcovHAC(mod1))
```

t test of coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.9365e+01  1.1948e+00 16.2077 < 2.2e-16 ***
dados3$renda_per_capita -1.2777e-03  5.7838e-04  -2.2090 0.02721 *
dados3$indice_gini    -1.4298e+01  1.6030e+00 -8.9193 < 2.2e-16 ***
dados3$salario_medio_mensal -1.7752e-01  1.1288e-01  -1.5727 0.11584
dados3$perc_crianças_extrem_pobres  3.8540e-02  1.7761e-02   2.1699 0.03006 *
dados3$perc_crianças_pobres    2.1590e-01  1.7189e-02  12.5601 < 2.2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  5.0554e-02  1.1224e-02   4.5042 6.799e-06 ***
dados3$perc_pessoas_dom_paredes_inadequadas    4.2969e-02  1.3885e-02   3.0947 0.00198 **
dados3$perc_pop_dom_com_coleta_lixo   -7.0450e-03  1.0729e-02  -0.6567 0.51143
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

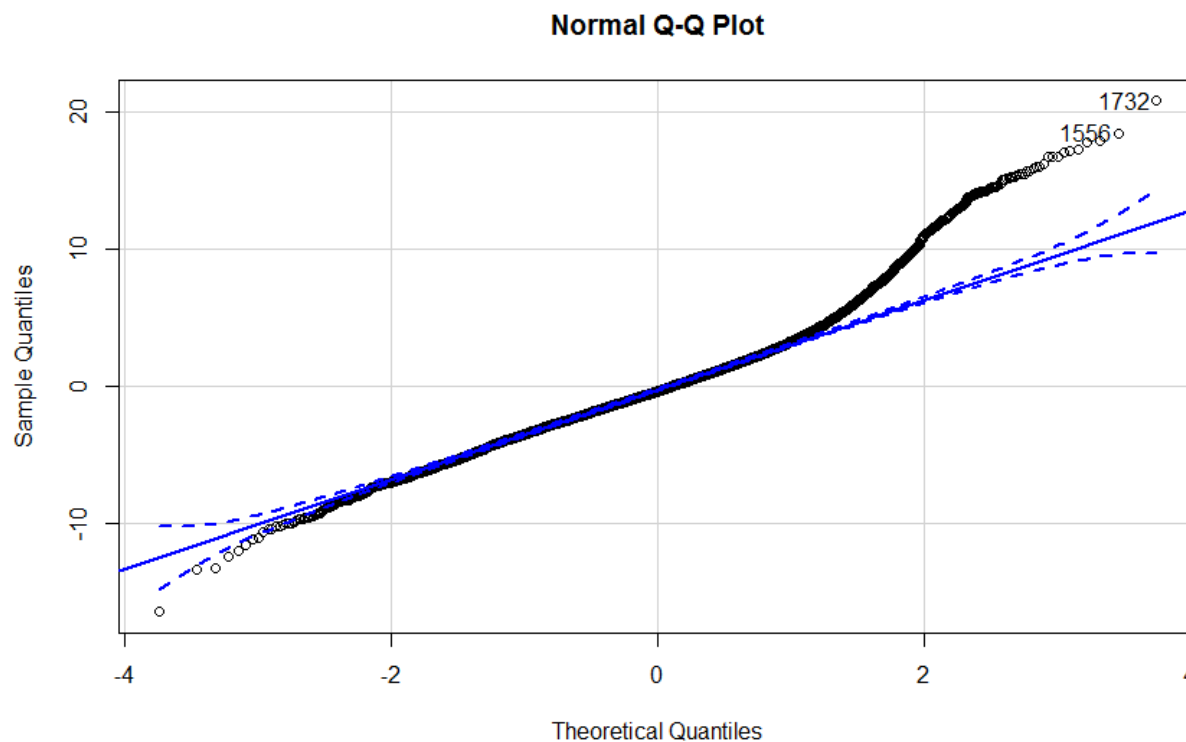
Modelo
original

Resposta: Os coeficientes significantes a 5% se mantiveram, embora o erro padrão, o t-value e p-value tenha se modificado com a estimação robusta para heteroscedasticidade e autocorrelação serial.

Modelos de Regressão

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

```
> qqPlot(residuos, main = "Normal Q-Q Plot",
```



Resposta: Pela análise da QQ_plot suspeita-se da não normalidade dos resíduos dados que os dados do quartil da amostra se afastam dos quantis da distribuição teórica nas caldas.

Modelos de Regressão

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

```
> shapiro.test(residuos)
Error in shapiro.test(residuos) : sample size must be between 3 and 5000
> ks.test(residuos,dados3$mort_infantil )

      Two-sample Kolmogorov-Smirnov test

data:  residuos and dados3$mort_infantil
D = 0.96352, p-value < 2.2e-16
alternative hypothesis: two-sided

warning message:
In ks.test(residuos, dados3$mort_infantil) :
  p-value will be approximate in the presence of ties
> ad.test(residuos)

      Anderson-Darling normality test

data:  residuos
A = 55.043, p-value < 2.2e-16
```

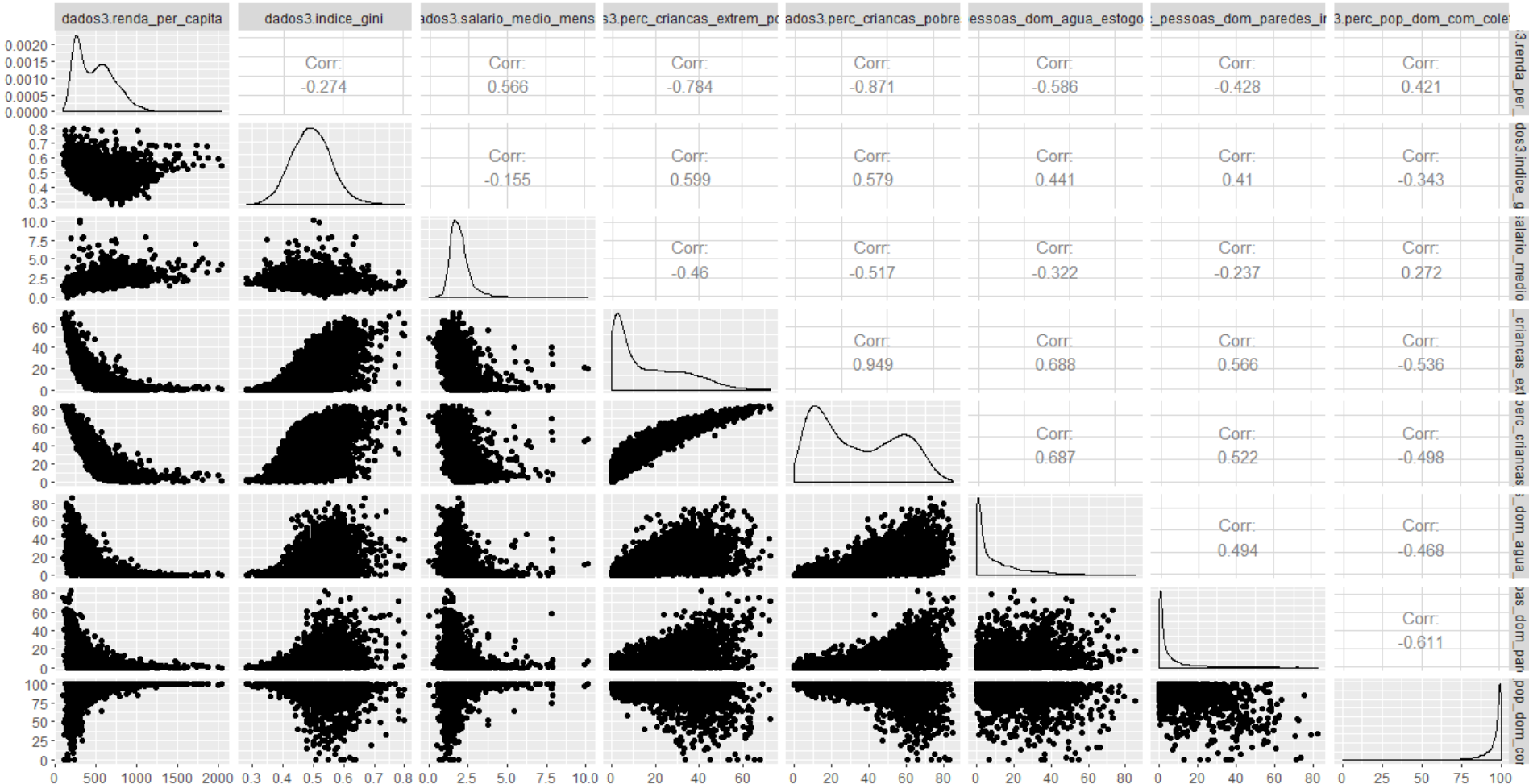
Resposta: os testes KS e AD cuja hipostese nula é de normalidade dos resíduos, têm p-valor menor do que 5%, portanto rejeita-se a hipótese nula de normalidade dos resíduos.

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

Resposta: Os testes estatísticos de KS e AD confirmam a suspeita da análise gráfica qq-plot, assim sendo os resíduos da regressão não são normais.

Modelos de Regressão

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



Resposta: A correlação entre Perc_Crianças_Pobres X Renda Percapita = -0,871; e entre Perc_Crianças_Pobres X Perc_Crianças _Extramente_Pobres = 0,949.

Modelos de Regressão

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

```
> omcdiag(Xnoint, dados3$mort_infantil)

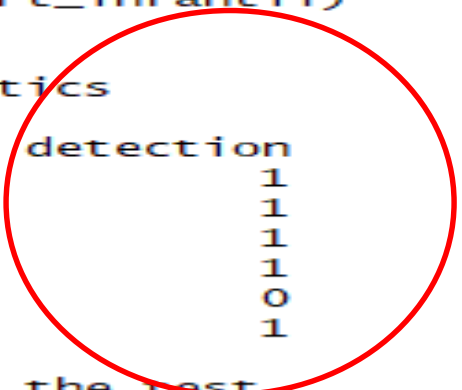
Call:
omcdiag(x = Xnoint, y = dados3$mort_infantil)

Overall Multicollinearity Diagnostics


```

	MC Results	detection
Determinant $ X'X $:	0.0012	1
Farrar Chi-Square:	37235.4553	1
Red Indicator:	0.5410	1
Sum of Lambda Inverse:	49.8862	1
Theil's Method:	0.2715	0
Condition Number:	47.5672	1

```
1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```



Resposta: Como 5 dos seis 6 testes de multicolinearidade detectaram a presença de colinearidade, podemos concluir que existe multicolinearidade nas variáveis explicativas do modelo.

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

Resposta: Tanto o teste gráfico do teste estatístico confirma a presença de multicolinearidade entres as variáveis explicativas.

Obrigado!