ANÁLISE DE DADOS MULTIVARIADOS I - REGRESSÃO

(AULA 09)

Novembro e dezembro de 2018

Reinaldo Soares de Camargo

Normalidade dos Resíduos

- **Normalidade dos resíduos**. A normalidade dos resíduos não é uma condição fundamental para a inferência sobre os coeficientes estimados

- Quando os resíduos não são normais, para amostras suficientemente grandes, as estimativas tornam-se normais

- No entanto, testes de normalidade podem ajudar a indicar outros problemas na regressão, como não linearidades não captadas pelas variáveis preditoras

- Para testar a normalidade dos resíduos no R, vamos utilizar o pacote “olsrr”.

- Esse pacote possui uma gama de testes comumente conhecidos na literatura. Para todos esses testes, a hipótese nula é de que os erros são normais. Caso rejeitemos a hipótese nula, temos indicação de que os resíduos não são normais.

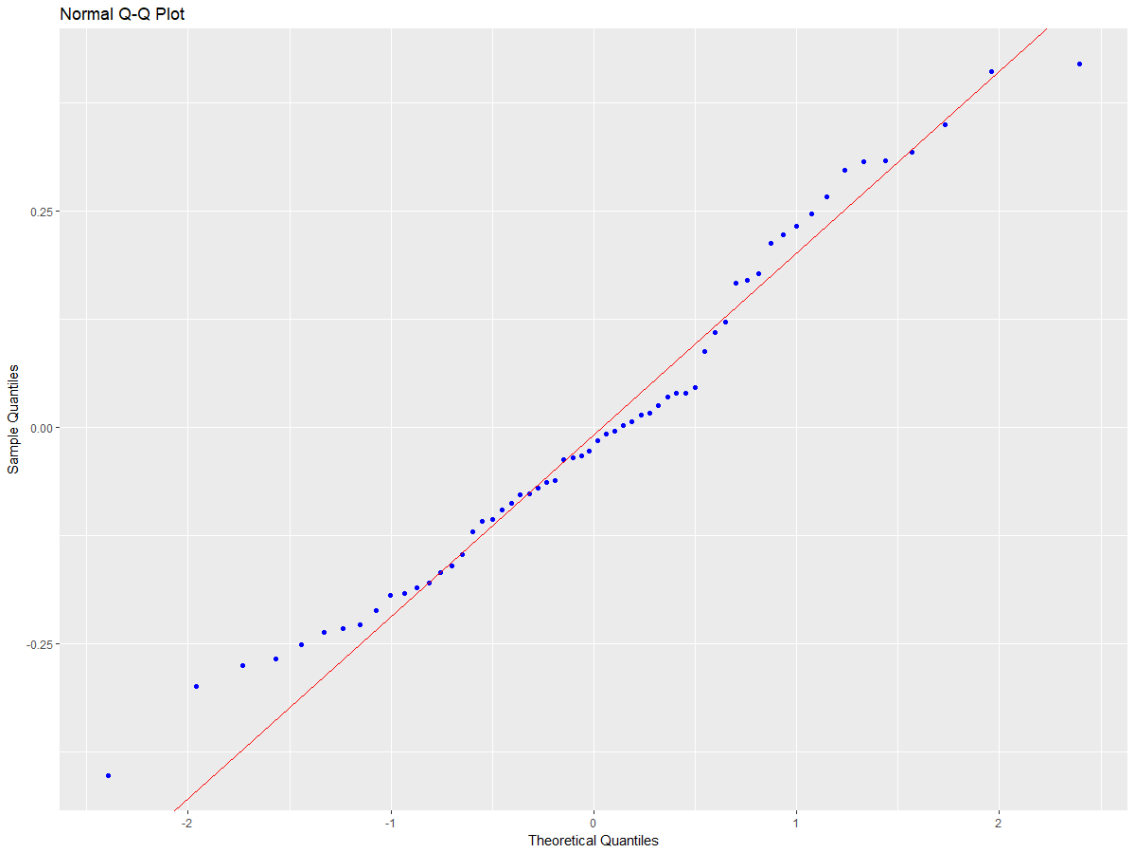
```
-----------------------------------------------
       Test              Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk               0.9711         0.1659
Kolmogorov-Smirnov         0.1062         0.4751
Anderson-Darling           0.6008         0.1137
```

Normalidade dos Resíduos

- Na prática, mesmo quando os erros não são normais, não necessariamente não mais podemos utilizar os resultados da regressão. A não normalidade dos erros tem mais um caráter de *warning* em muitas aplicações

- Há também um conjunto de técnicas para checar a normalidade dos resíduos graficamente. A ferramenta mais comum é o QQ plot (quantile-quantile plot).

- Se os erros forem normais, espera-se que os pontos em azul se posicionem ao longo da reta em vermelho.



Normalidade dos Resíduos

- Utilizando o programa “Ajuste_8.R”, para a regressão a seguir, utilizando o pacote “olsrr”:

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

2. Teste a normalidade dos resíduos utilizando os testes: Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling

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

Multicolinearidade

- **Multicolinearidade**. Além dos problemas apontados acima, um dos problemas comumente discutidos na literatura é o de multicolinearidade entre as variáveis regressoras da equação.

- Um caso particular mais simples de multicolinearidade acontece quando duas das variáveis preditoras são muito correlacionadas.

- No entanto, há tipos de multicolinearidade que não são identificados simplesmente pela correlação simples entre pares de variáveis na regressão.

    - Por exemplo, se uma das variáveis preditoras for gerada por uma equação linear das demais variáveis preditoras, há uma multicoliaridade forte, que não pode ser identificada necessariamente por uma correção simples entre pares de variáveis independentes

- Um dos problemas da presença de multicolinearidade é que os coeficientes de algumas das variáveis preditoras podem variar sensivelmente quando se adicionam ou se retiram variáveis da regressão. Alguns coeficientes podem até mudar de sinal de forma abrupta, quando retiramos alguma variável ou adicionamos.

- A presença de multicolinearidade dificulta a interpretação dos resultados da regressão.

Multicolinearidade

Exemplo: Regressão da taxa de internações por condições sensíveis sobre variáveis representativas da evolução da atenção básica por região

Devido à presença de multicolinearidade, os autores rodaram regressões separadas, com as variáveis da política pública

Internações por condições sensíveis	Região Norte	Região Nordeste	Região Sudeste	Região Sul	Região Centro-Oeste
Variáveis explicativas	Coeficientes e erros-padrão robustos				
Cobertura das ESFs	0,0484 (0,0682)	0,0143 (0,0331)	-0,1103*** (0,0308)	-0,0837** (0,0406)	-0,0677 (0,0644)
Cobertura dos ACS	0,1749** (0,0768)	0,1321*** (0,0426)	-0,1011*** (0,0355)	-0,0153 (0,0406)	-0,0111 (0,0606)
Cobertura dos cadastramentos	0,0039 (0,121)	-0,0305 (0,0442)	-0,1596*** (0,0447)	-0,0015 (0,0501)	-0,1792 (0,1121)
Observações	3592	14322	13239	9444	3671
Grupos (municípios)	449	1792	1664	1187	466

Multicolinearidade

- Uma das medidas para lidar com multicolinearidade é excluir algumas das variáveis do lado direito da equação. Há uma literatura conhecida de seleção de modelos, que permite determinar quais variáveis excluir.

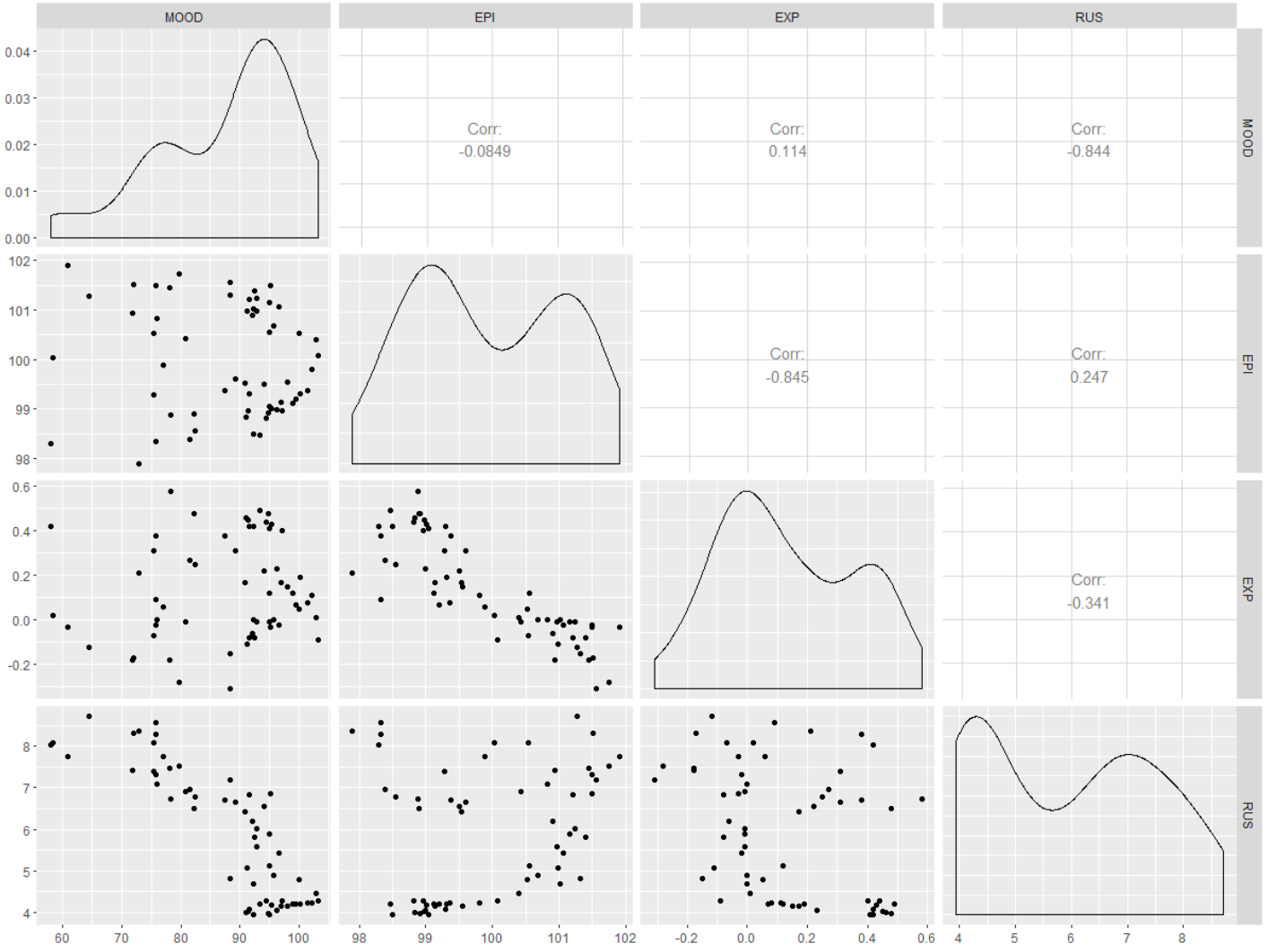
- Há uma gama de técnicas para identificação de colinearidade, havendo também procedimentos heurísticos consolidados.

- No R, podemos usar os pacotes “mctest” e “GGally”.

```
#----- testes de multicolinearidade

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

X <- model.matrix(mod1)
head(X)

Xnoint <- X[, -1]
head(Xnoint)

ggpairs(data.frame(Xnoint))

omcdiag(Xnoint, bondyield$RAARUS)
```

Multicolinearidade

```
ggpairs(data.frame(Xnoint))
```



Multicolinearidade

```
> omcdiag(Xnoint, bondyield$RAARUS)

Call:
omcdiag(x = Xnoint, y = bondyield$RAARUS)


Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.0634         0
Farrar Chi-Square:       153.7544         1
Red Indicator:             0.5200         1
Sum of Lambda Inverse:    15.9992         0
Theil's Method:            1.1768         1
Condition Number:        497.0396         1

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

Multicolinearidade

- **Exercício - continuação**. Utilizando o programa “ajuste_9.R”, para a regressão a seguir, utilizando os pacotes “mctest” e “GGally”:

```
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. 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)?

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

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

Obrigado!