

Introdução à Estatística usando o R com Aplicação em Análises Laboratoriais

Profa Carolina & Prof Gilberto

Instituto de Matemática e Estatística
Universidade Federal da Bahia

23 de novembro de 2019

Análise de Regressão

Análise de Regressão: conjunto de técnicas estatísticas utilizada quando há interesse em **investigar o comportamento de uma variável com respeito à um conjunto de outras variáveis**.

Problema de regressão: consiste em **estabelecer e determinar uma função que descreva a relação** entre uma variável, chamada de **variável resposta** ou dependente e **denotada por Y** , e um conjunto de variáveis observáveis, chamadas de variáveis explicativas ou **covariáveis** e **denotadas por x_1, x_2, \dots, x_p** .

Análise de Regressão

Uma vez estabelecida e determinada a relação funcional entre a variável resposta e a(s) variável(is) explicativas, podemos explorar esta relação para obter informações sobre a variável resposta a partir do conhecimento das covariáveis.

Importante: as relações estatísticas não necessariamente implicam em relações causais, mas a presença de qualquer relação estatística fornece um ponto inicial para outras pesquisas.

Modelos de regressão: podem ser usados para predição, estimação, testes de hipótese e para modelar relações casuais.

Modelo de regressão linear simples

Considere o seguinte problema:

Uma engenheira industrial de uma empresa que engarrafa bebidas deve analisar as operações de entrega de produtos e serviços para máquinas de venda automática.

Ela suspeita que o tempo exigido para que um entregador carregue e conserte uma máquina esteja relacionado com o número de caixas de produtos entregues.

A engenheira selecionados aleatoriamente $n = 25$ estabelecimentos de varejo com máquinas de venda automática. Ela então visita os 25 estabelecimentos selecionados e observa o tempo de entrega na entrada (em minutos) e o volume de produto entregue (em caixas) para cada um.

Modelo de regressão linear simples

Seja:

- Y : variável que representa o tempo de entrega;
- x : variável que representa o volume entregue.

Então, para $i = 1, \dots, n$, podemos escrever:

$$Y_i = \beta_0 + \beta_1 x_i. \quad (1)$$

Os parâmetros β_0 e β_1 são chamados de coeficientes da regressão.

Estes coeficientes têm uma interpretação simples e útil:

- o parâmetro β_1 é a mudança na média da distribuição de Y_i por cada aumento unitário de x_i ;
- se a amplitude dos dados dos x_i inclui o zero, β_0 fornece a média da distribuição de Y_i quando $x_i = 0$. Quando a amplitude dos dados dos x_i não inclui o zero, β_0 não tem uma interpretação prática.

Modelo de regressão linear simples

Os dados observados são os pares ordenados (x_i, y_i) , $i = 1, \dots, n$.

```
dados <- read_xlsx("dados_regressao.xlsx",  
                  sheet = "Tempo_Volume_Distancia")  
  
glimpse(dados)
```

Observations: 25
Variables: 3
\$ Volume <dbl> 7, 3, 3, 4, 6, 7, 2, 7, 30, 5, 16
\$ Distancia <dbl> 560, 220, 340, 80, 150, 330, 110,
\$ Tempo <dbl> 16.68, 11.50, 12.03, 14.88, 13.75

Modelo de regressão linear simples

Gráfico de dispersão

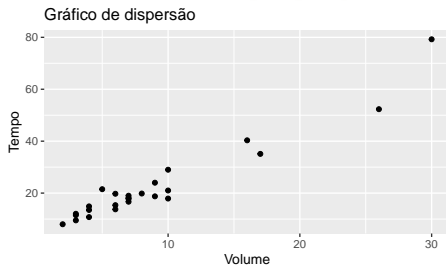
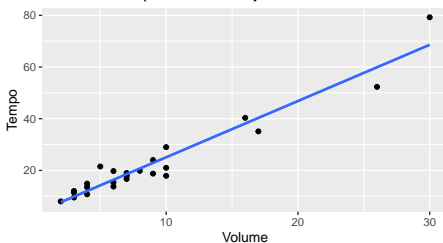


Gráfico de dispersão e reta ajustada



Federal da Bahia

Modelo de regressão linear simples

Na Figura, os dados observados não estão exatamente sobre a reta. Logo, a equação (1) deve ser modificada para acomodar este fenômeno.

Seja:

- ϵ uma variável aleatória (v.a.) não observável representando o erro entre a reta (1) e os valores observados de Y . É um erro estatístico, i.e., uma v.a. que explica a falha do modelo em ajustar os dados com exatidão.

Assim, um modelo estatístico mais plausível para o problema é dado por

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad i = 1, \dots, n, \quad (2)$$

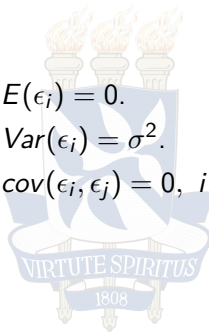
onde Y_i é a v.a. resposta, x_i é a variável preditora e ϵ_i é o erro aleatório.

Suposições do modelo de regressão linear simples

Suposição 1: $E(\epsilon_i) = 0$.

Suposição 2: $Var(\epsilon_i) = \sigma^2$.

Suposição 3: $cov(\epsilon_i, \epsilon_j) = 0, i \neq j, j = 1, 2, \dots, n$.



UFBA
Universidade
Federal da Bahia

Estimação

O método de mínimos quadrados (MMQ) é mais utilizado do que qualquer outro procedimento de estimação em modelos de regressão.

O MMQ fornece os estimadores de β_0 e β_1 tal que a soma de quadrados das diferenças entre as observações y_i 's e a linha reta ajustada seja mínima.

Assim, de todos os possíveis valores de β_0 e β_1 , os estimadores de mínimos quadrados (EMQ) serão aqueles que minimizam a soma de quadrados dos erros.

Estimação

Os estimadores de mínimos quadrados (EMQ) de β_0 e β_1 , denotados $\hat{\beta}_0$ e $\hat{\beta}_1$, são dados por

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

e

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i x_i - n \bar{Y} \bar{x}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2},$$

onde $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ e $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ são as médias dos Y_i 's e x_i 's, respectivamente.

Portanto, $\hat{\beta}_0$ e $\hat{\beta}_1$ são os EMQ para o intercepto e inclinação, respectivamente.

Estimação

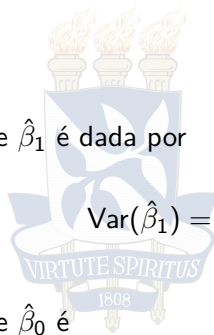
Os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ são não-viciados para os parâmetros do modelo β_0 e β_1 . Isto é,

$$E(\hat{\beta}_1) = \beta_1$$

e

$$E(\hat{\beta}_0) = \beta_0.$$

A variância de $\hat{\beta}_1$ é dada por


$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}.$$

A variância de $\hat{\beta}_0$ é

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).$$

Estimação

Reta de regressão estimada:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

A reta ajustada fornece a estimativa pontual da média de Y_i para um particular x_i .

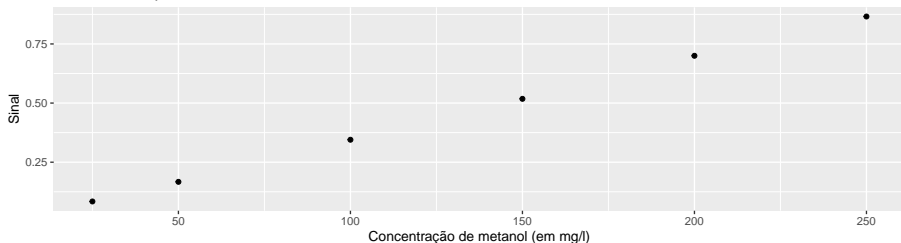
O valor \hat{y}_i é dito ser o valor predito de y_i .

Exemplo

```
dados <- read_xlsx("dados.xlsx", sheet = "Metanol")
```

```
ggplot(data = dados) +  
  geom_point(aes(x = x, y = y)) +  
  labs(x = 'Concentração de metanol (em mg/l)',  
        y = 'Sinal',  
        title = 'Gráfico de dispersão')
```

Gráfico de dispersão



Exemplo (continuação)

```
with(dados,  
      cor.test(x, y,  
                alternative = "two.sided",  
                conf.level = 0.95))  
  
##  
## Pearson's product-moment correlation  
##  
## data: x and y  
## t = 183.93, df = 4, p-value = 5.242e-09  
## alternative hypothesis: true correlation is not equ  
## 95 percent confidence interval:  
## 0.9994318 0.9999939  
## sample estimates:  
## cor  
## 0.9999409
```

Exemplo (continuação)

```
dados <- read_xlsx("dados.xlsx", sheet = "Metanol")
```

```
modelo <- lm(y ~ x, data = dados)
```

```
modelo
```

```
##  
## Call:  
## lm(formula = y ~ x, data = dados)  
##  
## Coefficients: (Intercept) x  
## -0.005112 0.003498
```



UFBA
Universidade
Federal da Bahia

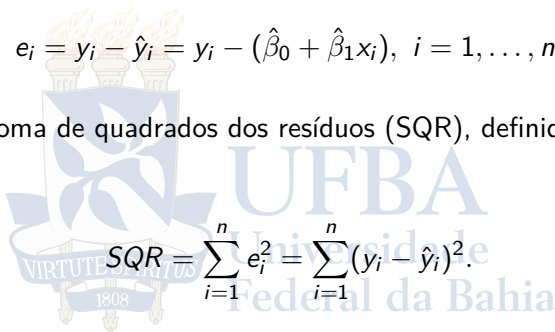
Estimação de σ^2

Resíduos:

A diferença entre o valor observado y_i e o valor predito \hat{y}_i é um resíduo. Matematicamente, o i -ésimo resíduo é

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, \dots, n.$$

Considere a soma de quadrados dos resíduos (SQR), definida por


$$SQR = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Um estimador não viciado de σ^2 é

$$\hat{\sigma}^2 = \frac{SQR}{n-2} = QMR.$$

Teste de hipóteses

Suposição adicional: os erros ϵ_i do modelo são normalmente distribuídos.

Assim, as suposições completas do modelo de regressão linear simples (2) são: os erros são independentes e normalmente distribuídos com média zero e variância σ^2 .

Teste de significância do modelo: um caso especial e muito importante em teste de hipóteses é testar

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_1 : \beta_1 \neq 0.$$

Podemos utilizar a análise de variância (ANOVA) para testar a significância da regressão.

A ANOVA é baseada no particionamento da variabilidade total da variável resposta Y .

Partição da soma de quadrados do modelo

Soma de quadrados total (SQT):

$$SQT = \sum_{i=1}^n (y_i - \bar{y})^2.$$

Soma de quadrados dos resíduos (SQR):

$$SQR = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Soma de quadrados do modelo ou da regressão (SQM):

$$SQM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

Observação: $SQT = SQR + SQM$.

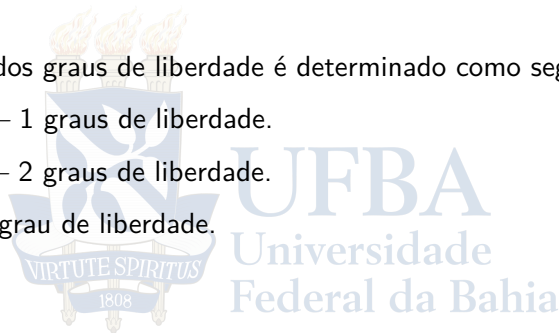
Partição da soma de quadrados do modelo

A separação dos graus de liberdade é determinado como segue.

SQT: tem $n - 1$ graus de liberdade.

SQR: tem $n - 2$ graus de liberdade.

SQM: tem 1 grau de liberdade.



Teste F (ANOVA)

Podemos utilizar o teste F da ANOVA para testar a hipótese $H_0 : \beta_1 = 0$.

Temos que:

$$F_0 = \frac{\frac{SQM}{gl_M}}{\frac{SQR}{gl_R}} = \frac{\frac{SQM}{1}}{\frac{SQR}{(n-2)}} = \frac{QMM}{QMR} \sim F_{1,n-2}.$$

Portanto, para testar a hipótese $H_0 : \beta_1 = 0$, calcula-se a estatística F_0 e rejeita-se H_0 se

$$F_0 > F_{\alpha,1,n-2}.$$

Analogamente, rejeitamos H_0 se o p-valor do teste for menor do que um nível de significância $\alpha \in (0, 1)$ pré-fixado.

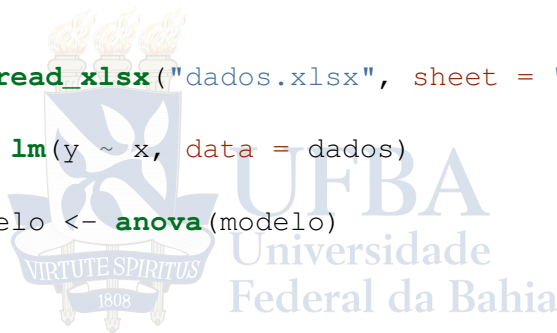
Teste F (ANOVA)

Em resumo, temos a tabela ANOVA para testar a significância da regressão:

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F_0
Modelo	SQM	1	QMM	$\frac{QMM}{QMR}$
Resíduo	SQR	$n - 2$	QMR	
Total	SQT	$n - 1$		

Exemplo

```
dados <- read_xlsx("dados.xlsx", sheet = "Metanol")  
  
modelo <- lm(y ~ x, data = dados)  
  
anova_modelo <- anova(modelo)
```



Exemplo (continuação)

```
anova_modelo
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: y
```

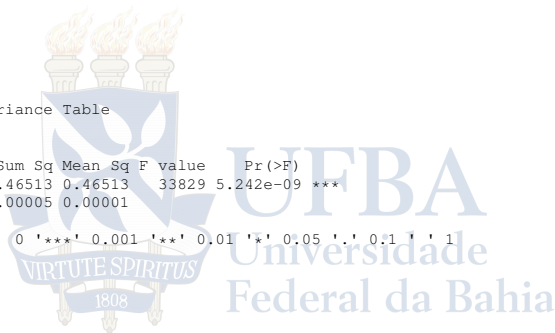
```
##      Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## x      1  0.46513   0.46513    33829 5.242e-09 ***
```

```
## Residuals  4  0.00005   0.00001
```

```
## ---
```

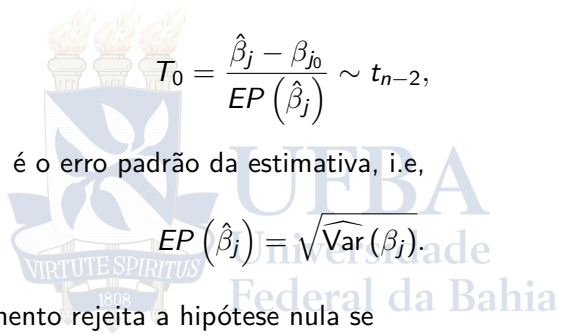
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Teste t para os parâmetros

Considere o teste

$$H_0 : \beta_j = \beta_{j_0} \quad \text{versus} \quad H_1 : \beta_j \neq \beta_{j_0}.$$


$$T_0 = \frac{\hat{\beta}_j - \beta_{j_0}}{EP(\hat{\beta}_j)} \sim t_{n-2},$$

onde $EP(\hat{\beta}_j)$ é o erro padrão da estimativa, i.e.,

$$EP(\hat{\beta}_j) = \sqrt{\widehat{\text{Var}}(\beta_j)}.$$

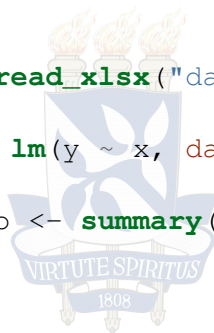
Esse procedimento rejeita a hipótese nula se

$$|T_0| > t_{\alpha/2, n-2}.$$

Analogamente, rejeitamos H_0 se o p-valor do teste for menor do que um nível de significância $\alpha \in (0, 1)$ pré-fixado.

Exemplo

```
dados <- read_excel("dados.xlsx", sheet = "Metanol")  
  
modelo <- lm(y ~ x, data = dados)  
  
sum_modelo <- summary(modelo)
```



UFBA
Universidade
Federal da Bahia

Exemplo (continuação)

```
sum_modelo
```

```
##
## Call:
## lm(formula = y ~ x, data = dados)
##
## Residuals:
##      1      2      3      4      5      6
## 0.0016712 -0.0027699  0.0003479 -0.0015342  0.0055836 -0.0032986
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.112e-03  2.885e-03  -1.772    0.151
## x            3.498e-03  1.902e-05 183.927 5.24e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003708 on 4 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 3.383e+04 on 1 and 4 DF,  p-value: 5.242e-09
```

Coeficiente de determinação

O coeficiente de determinação é definido por

$$R^2 = \frac{SQM}{SQT} = 1 - \frac{SQR}{SQT},$$

- SQT : é uma medida de variabilidade em y sem considerar o efeito da variável regressora x ;
- SQR é uma medida da variabilidade remanescente em y após x ter sido considerada;
- então: R^2 é frequentemente dito ser a proporção de variação explicada pelo regressor x .

Como $0 \leq SQR \leq SQT$, segue que $0 \leq R^2 \leq 1$.

Um valor de R^2 perto de 1 significa que a maior parte da variação em y é explicada pelo modelo de regressão.

Intervalo de confiança para os parâmetros

```
dados <- read_xlsx("dados.xlsx", sheet = "Metanol")
```

```
modelo <- lm(y ~ x, data = dados)
```

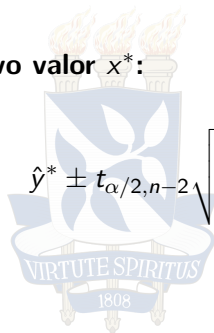
```
ic_parametros <- confint(modelo, level = 0.95)
```

```
ic_parametros
```

```
##                2.5 %          97.5 %  
## (Intercept) -0.013123192 0.002898535  
## x           0.003444846 0.003550442
```

Intervalos de predição

Para um novo valor x^* :


$$\hat{y}^* \pm t_{\alpha/2, n-2} \sqrt{QMR \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]}$$

Universidade Federal da Bahia

Exemplo

```
dados <- read_xlsx("dados.xlsx", sheet = "Metanol")

modelo <- lm(y ~ x, data = dados)

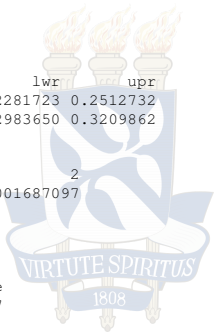
x_novo <- data.frame(x = c(70, 90))

ic_novos_preditos <- stats::predict(modelo,
                                     newdata = x_novo,
                                     se.fit = TRUE,
                                     interval = "prediction",
                                     level = 0.95)
```

Exemplo (continuação)

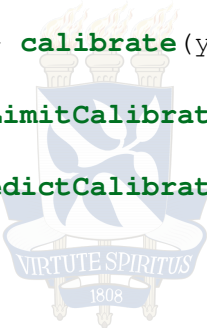
```
ic_novos_preditos
```

```
## $fit
##      fit      lwr      upr
## 1 0.2397227 0.2281723 0.2512732
## 2 0.3096756 0.2983650 0.3209862
##
## $se.fit
##      1      2
## 0.001886132 0.001687097
##
## $df
## [1] 4
##
## $residual.scale
## [1] 0.003708007
```



UFBA
Universidade
Federal da Bahia

Calibração e predição inversa



```
mod_cal <- calibrate(y ~ x, data = dados)

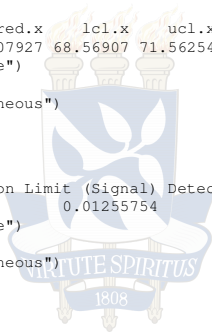
detectionLimitCalibrate(mod_cal, coverage = 0.95)

inversePredictCalibrate(mod_cal, obs.y = 0.01255754,
                           intervals = TRUE,
                           coverage = 0.95)
```

Calibração e predição inversa

```
##      obs.y   pred.x      lcl.x      ucl.x
## [1,]  0.24 70.07927 68.56907 71.56254
## attr("coverage")
## [1] 0.95
## attr("simultaneous")
## [1] FALSE
```

```
##      Decision Limit (Signal) Detection Limit (Concentration)
##      0.01255754      9.89486859
## attr("coverage")
## [1] 0.95
## attr("simultaneous")
## [1] TRUE
```



UFBA
Universidade
Federal da Bahia

Análise de resíduos

As principais suposições feitas até agora no estudo de análise de regressão linear simples foram as seguintes:

- ① Linearidade: a relação entre a resposta Y e as variáveis regressoras é linear, pelo menos aproximadamente.
- ② O termo de erro ϵ tem média zero.
- ③ Homoscedasticidade: o termo de erro ϵ tem variância constante σ^2 .
- ④ Independência: os erros são não correlacionados.
- ⑤ Normalidade: os erros são normalmente distribuídos.

Exemplo

```
tab <- augment(modelo)

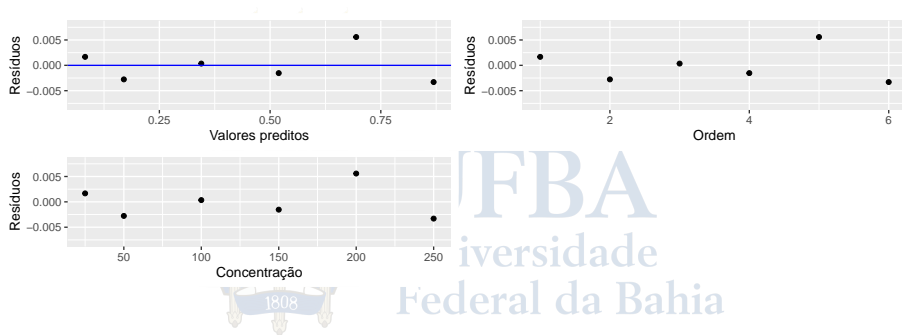
g1 <- ggplot(tab, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 0, color = 'blue') +
  ylim(c(-0.008, 0.008)) +
  labs(x = "Valores preditos", y = "Resíduos")

g2 <- ggplot(tab, aes(x = seq_along(.resid), y = .resid)) +
  geom_point() +
  ylim(c(-0.008, 0.008)) +
  labs(x = "Ordem", y = "Resíduos")

g3 <- ggplot(tab, aes(x = x, y = .resid)) +
  geom_point() +
  ylim(c(-0.008, 0.008)) +
  labs(x = "Concentração", y = "Resíduos")

cowplot::plot_grid(g1, g2, g3)
```

Exemplo



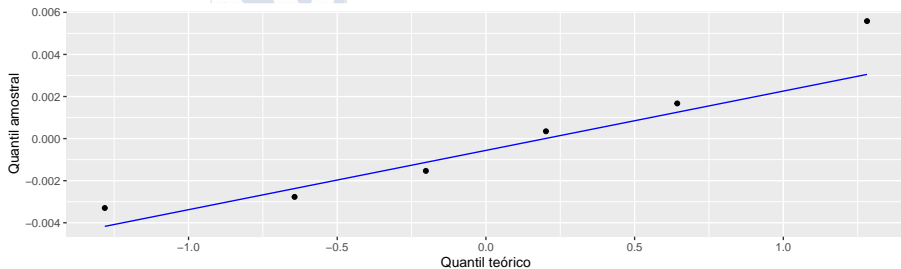
Exemplo

```
shapiro.test(modelo$residuals)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  modelo$residuals  
## W = 0.92274, p-value = 0.5253
```

Exemplo (continuação)

```
ggplot(tab, aes(sample = .resid)) +  
  geom_qq() +  
  geom_qq_line(color = "blue") +  
  labs(x = "Quantil teórico", y = "Quantil amostral")
```



Exemplo (continuação)

```
bptest (modelo)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data: modelo  
## BP = 1.7089, df = 1, p-value = 0.1911
```


Exemplo (continuação)

```
dwtest (modelo)
```

```
##  
## Durbin-Watson test  
##  
## data: modelo  
## DW = 2.9555, p-value = 0.79  
## alternative hypothesis: true autocorrelation is gre
```