Modelos Estatísticos I
Modelos para desfecho contínuo

Leo Bastos – leonardo.bastos@fiocruz.br

PROCC – Fundação Oswaldo Cruz

https://github.com/lsbastos/eae2

Outline

O modelo linear múltiplo

- Vamos assumir que temos duas ou mais variáveis explicativas.
- Seja Y um desfecho contínuo avaliado em n observações independentes, e de forma geral p variáveis explicativas X, ou seja,

$$\begin{aligned} Y_i \mid \mathrm{X}_i &\sim N(\mu_i, \sigma^2), \qquad i = 1, 2, \ldots, n, \\ \mu_i &= \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}. \end{aligned}$$

- Isso é o mesmo que escrever

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i} + \epsilon_i$$

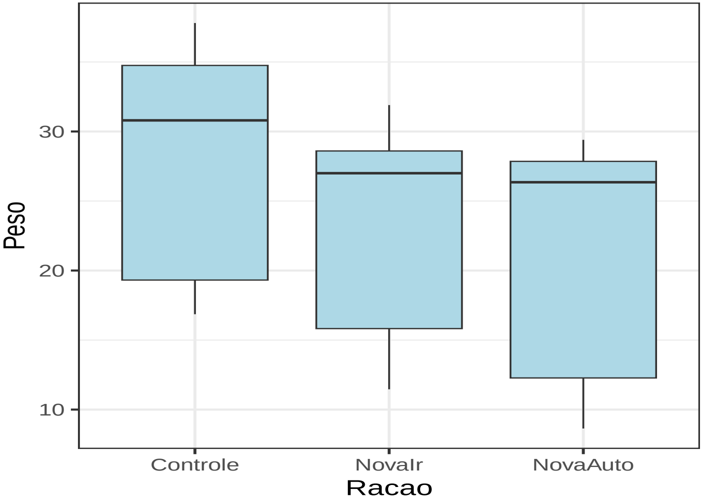
onde $\epsilon_i \sim N(0, \sigma^2)$, e $\epsilon_i \perp \epsilon_j, \forall i \neq j$.

FIOCRUZ

Exemplo do peso dos ratinhos

- Considere um estudo realizado para avaliar a eficiência de ração no ganho de peso de ratos.
- Seja Y peso (em g) de 45 ratos com 20 dias de idade alimentados a partir do desmame por duas semanas com uma das três marcas diferentes de ração.
- Os ratos também pertencem a 3 linhagens distintas.

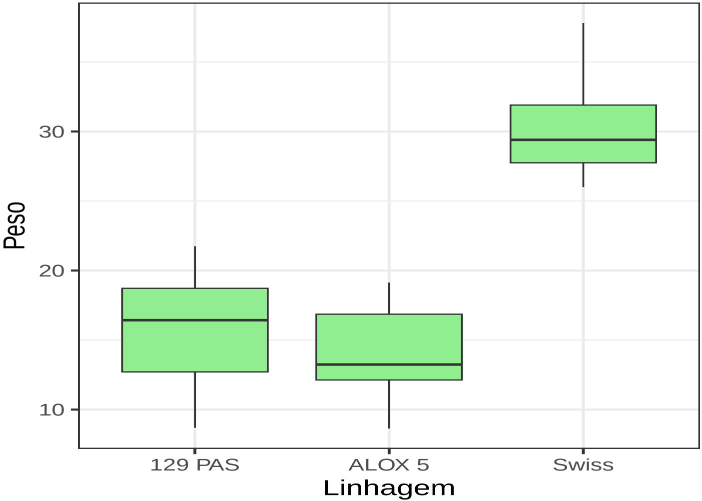
Visualisando os dados



ANOVA

```
> summary(aov(Peso ~ Racao, data = racao))
            Df Sum Sq Mean Sq F value Pr(>F)
Racao        2    394  196.99   3.066 0.0574 .
Residuals   41   2634   64.25
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1 observation deleted due to missingness
```

Visualisando os dados

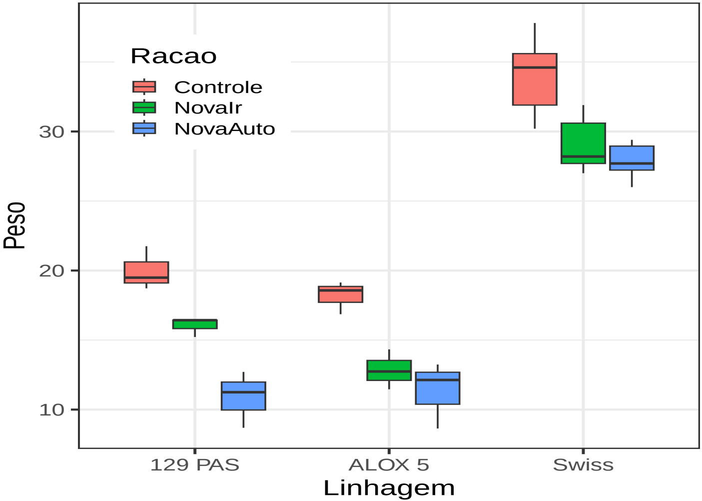


ANOVA

```
> summary(aov(Peso ~ Linhagem, data = racao))

            Df Sum Sq Mean Sq F value Pr(>F)
Linhagem     2 2531.1  1265.5   104.3 <2e-16 ***
Residuals   41  497.4    12.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
1 observation deleted due to missingness
```

Visualisando os dados



```
> summary(aov(Peso ~ Racao + Linhagem, data = racao))
            Df Sum Sq Mean Sq F value   Pr(>F)
Racao        2  394.0   197.0   54.19 5.51e-12 ***
Linhagem     2 2492.7  1246.3  342.87  < 2e-16 ***
Residuals   39  141.8     3.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
1 observation deleted due to missingness
```

O modelo

- O modelo que incorpora a exposição e controla pela linhagem é dado por

$$
\begin{aligned}
Y_i &\sim N(\mu_i, \sigma^2), \qquad 1 = 2, \ldots, 45, \\
\mu_i &= \alpha + \beta_2 X_{1,i}^{(NovaAu)} + \beta_3 X_{1,i}^{(NovaIr)} + \beta_3 X_{2,i}^{(ALOX5)} + \beta_2 X_{2,i}^{(Swiss)}
\end{aligned}
$$

  onde X_1 é o efeito da ração, e X_2 é o efeito da linhagem.

- A notação no R seria: $Peso \sim Racao + Linhagem$

O modelo no R

```
> modelo <- lm(Peso ~ Racao + Linhagem, data = racao)
> summary(modelo)

Call:
lm(formula = Peso ~ Racao + Linhagem, data = racao)
Residuals:
    Min      1Q  Median      3Q     Max
-4.0292 -1.3412  0.1531  1.4275  3.7783

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     19.4941     0.7533  25.879  < 2e-16 ***
RacaoNovaIr     -4.8073     0.6962  -6.905 2.89e-08 ***
RacaoNovaAuto   -6.7749     0.7087  -9.560 9.02e-12 ***
LinhagemALOX 5  -1.5100     0.8988  -1.680    0.101
LinhagemSwiss   14.5277     0.7375  19.698  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.907 on 39 degrees of freedom
```

A partir da saída

Em média

- os ratos da linha 129 PAS tratados com a ração controle pesam 19.5g.
- as duas rações, Nova Irradiada e Nova autoclavada, reduzem o peso médio em 4.8 e 6.8g respectivamente quando comparadas ao controle.
- a linhagem ALOX 5 não difere estatísticamente da 129 PAS. No entanto, a linhagem Swiss em média pesa 14.5g a mais que a 129 PAS.

Análise de resíduos

- Quais forma as suposições assumidas para o modelo anterior?
  - Normalidade
  - Independência
  - Variância constante
- Será que elas são válidas?
- Como podemos checar?

Análise de resíduos

- Vamos definir os resíduos do modelo como

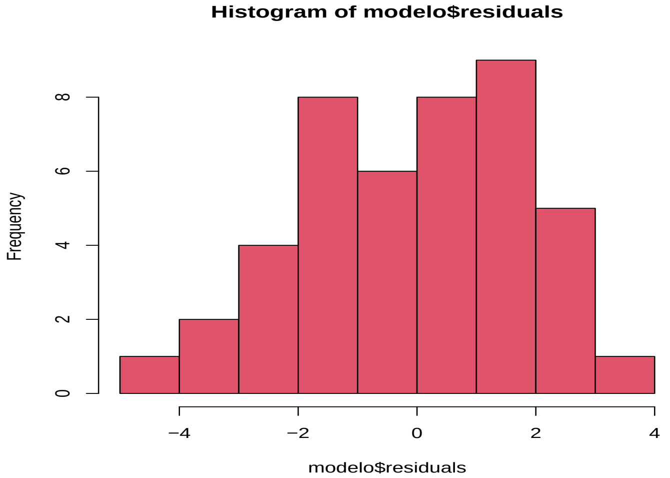
$$e_i = Y_i - \hat{Y}_i$$

- Por construção, a distribuição dos resíduos é

$$e_i \sim N(0, \sigma^2), \qquad \forall i \quad e_i \perp e_j \forall i \neq j$$

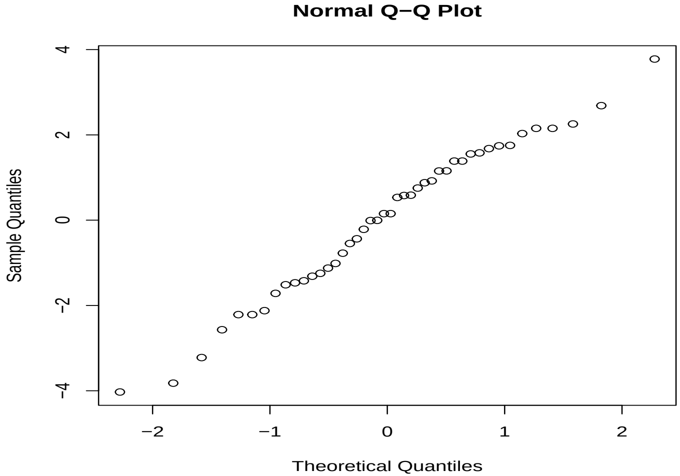
- Podemos então olhar estatísticas descritivas dos resíduos

```
> hist(modelo$residuals, col = 2)
```



Suposição de normalidade

> qqnorm(modelo$residuals)



Suposição de normalidade

Podemos testar a hipótese de normalidade usando o teste de Shapiro-Wilks

$$\begin{cases} H_0: & \text{as observações seguem uma distribuição normal}(\mu, \sigma^2) \\ H_1: & \text{as observações seguem outra distribuição.} \end{cases}$$

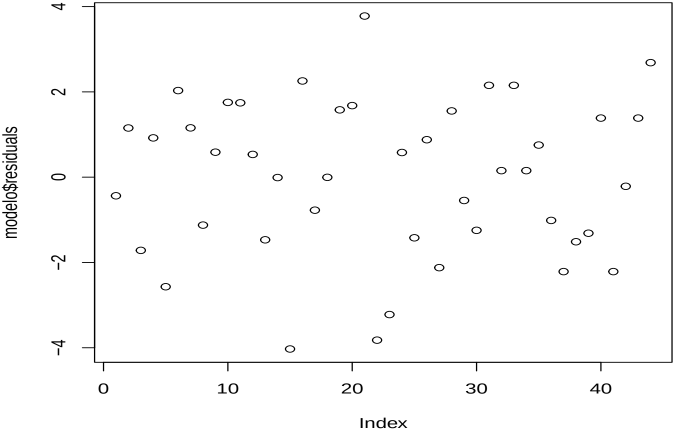
```
> shapiro.test(modelo$residuals)

        Shapiro-Wilk normality test
data:  modelo$residuals
W = 0.9773, p-value = 0.5292
```

Variância constante (e padrões não-lineares)

Faz-se um scatter plot dos resíduos versus a ordem dos dados, e espera-se que a variância não tenha um padrão bem definido

```
> plot(modelo$residuals)
```



Análise de resíduos (resumo)

- A análise de resíduos consite em um conjunto de estatísticas descritivas associadas aos resíduos de um modelo ajustado.
- A analise de resíduos deve ser realizada sempre que um modelo linear for ajustado e os coeficientes testados (teste de Wald $H_0: \quad \beta_k = 0$ versus $H_1: \quad \beta_k \neq 0$).
- As principais suposições de um modelo linear são
  - Normalidade (que pode ser relaxada se a amostra for grande)
  - Variância constante (Homocedasticidade)
  - Independência
- A análise de resíduos pode se estender ad infinitum, pois podemos testar a influencia de outras variáveis, formas não lineares, presença de outlier, etc. (Ler capítulos 7 e 8 do open statistics 3rd ed.)

FIOCRUZ

Comparação de modelos

Suponha agora que gostaríamos de ajustas difentes modelos:

```
> modeloR <- lm(Peso ~ Racao, data = racao)
> modeloL <- lm(Peso ~ Linhagem, data = racao)
> modelo <- lm(Peso ~ Racao + Linhagem, data = racao)
> modeloRL <- lm(Peso ~ Racao * Linhagem, data = racao)
```

Comparação de modelos

- Como podemos comparar diferentes modelos?
- Se os modelos forem aninhados* podemos usar a ANOVA no modelo mais completo
- Outra forma de comparar modelos é usando critérios de informação.

FIOCRUZ

Testando a interação

No modelo mais completo temos o termo de interação *Racao* : *Linhagem*, podemos testar se algum dos termos da interação é significativo

```
> summary(modeloRL)
Call:
lm(formula = Peso ~ Racao * Linhagem, data = racao)
Residuals:
    Min      1Q  Median      3Q     Max 
-3.5889 -1.2083  0.1174  1.4931  4.0111 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   19.9867     1.0700  18.679  < 2e-16 ***
RacaoNovaIr                   -3.9567     1.5132  -2.615   0.0131 *  
RacaoNovaAuto                 -9.1033     1.5132  -6.016 7.36e-07 ***
LinhagemALOX 5                -1.7967     1.5132  -1.187   0.2431    
LinhagemSwiss                 13.8022     1.2355  11.171 4.33e-13 ***
RacaoNovaIr:LinhagemALOX 5    -1.3900     2.1400  -0.650   0.5202    
RacaoNovaAuto:LinhagemALOX 5   2.2500     2.1400   1.051   0.3003    
RacaoNovaIr:LinhagemSwiss     -0.9544     1.7473  -0.546   0.5884    
RacaoNovaAuto:LinhagemSwiss    3.2019     1.7609   1.818   0.0776 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.853 on 35 degrees of freedom
```

Testando a interação

Testando usando a ANOVA de modelos, partindo do modelo nulo até o modelo completo

```
> anova(modeloRL)
```

```
Analysis of Variance Table
Response: Peso
              Df  Sum Sq Mean Sq  F value    Pr(>F)
Racao          2  393.98  196.99  57.3526 9.006e-12 ***
Linhagem       2 2492.67 1246.33 362.8666 < 2.2e-16 ***
Racao:Linhagem  4   21.55    5.39   1.5688    0.2043
Residuals     35  120.21    3.43
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Cada linha testa a igualdade do modelo anterior (o primeiro é o modelo nulo, i.e. $Y \sim 1$) e o modelo com a adição da variável daquela linha.

Testando a interação

Note que a ordem de entrada das covariáveis do modelo é importante

```
> anova(lm(Peso ~ Linhagem*Racao, data = racao))
```

```
Analysis of Variance Table
Response: Peso
               Df  Sum Sq Mean Sq  F value    Pr(>F)
Linhagem        2 2531.06 1265.53 368.4555 < 2.2e-16 ***
Racao           2  355.58  177.79  51.7637 3.502e-11 ***
Linhagem:Racao  4   21.55    5.39   1.5688    0.2043
Residuals      35  120.21    3.43
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dica: Se for usar esse recurso, inclua a sua exposição de interesse primeiro.

Podemos comparar dois modelos

```
> anova(modeloR, modeloRL)
Analysis of Variance Table
Model 1: Peso ~ Racao
Model 2: Peso ~ Racao * Linhagem
  Res.Df     RSS Df Sum of Sq   F    Pr(>F)
1     41 2634.43
2     35  120.21  6    2514.2 122 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Critério de informação

- O método da ANOVA para comparar modelos funciona bem para comparar modelos lineares alinhados
- Uma abordagem mais geral, é baseada em **critérios de informação**
- Os critérios de informação são construídos a partir da função de desvio (*deviance*)
- Essa função é dada por $-2\log_e \hat{L}$, $\hat{L}$ é a verossimilhança ajustada
- Para o modelo linear a função desvio é dada por:

$$-2\log_e \hat{L} = n\log(2\pi\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2}\sum_i (y_i - \hat{y}_i)^2$$

FIOCRUZ

Critério de informação

- Se compararmos a função desvio de dois modelos com o mesmo tamanho de amostra, temos que aquele com o menor valor é o melhor ajustado.
- No entanto, modelos com muitos parâmetros acabam sendo beneficiados. (*overfitting*)
- O critério de informação mais famoso foi proposto por Akaike (1974)

$$AIC = 2p - 2\log_e \hat{L}$$

  onde p é o número de parâmetros
- Outro critério também popular é o BIC, proporto por Schwarz (1978)

$$BIC = p\log_e n - 2\log_e \hat{L}$$

- Ambos critérios penalizam modelos com mais parâmetros

FIOCRUZ

Critérios de informação no R

- Os critérios de informação apresentados estão implementados no R
- As funções associadas se aplicam ao objeto do modelo ajustado
- No exemplo:

```
> AIC(modelo)
[1] 188.3463
> BIC(modelo)
[1] 199.0515
```

Critérios de informação no R

	df	AIC	BIC
modeloR	4.00	312.92	320.06
modeloL	4.00	239.57	246.71
modelo	6.00	188.35	199.05
modeloRL	10.00	189.09	206.93

- Seja uma amostra de 150 participantes do SHHS (Scotish Heart Health Study), e desejamos ver como o IMC depende do histórico de tabagismo e do sexo.

- O desfecho é será o IMC (kg/m^2), e as explicativas são sexo (duas categorias) e tabagismo (3 categorias)

- Teremos quatro modelos candidatos:
  1. IMC (Y) versus sexo ($X_1 = \{M, F\}$)
  2. IMC versus tabagismo ($X_2 = \{\text{current}, \text{ex}, \text{never}\}$)
  3. IMC versus sexo e tabagismo
  4. IMC versus sexo, tabagismo, e sua interação

FIOCRUZ

Modelo 1

$$
\begin{aligned}
Y_i \mid \mathrm{X}_i &\sim N(\mu_i, \sigma^2), \qquad i = 1, 2, \ldots, 150, \quad Y_i \perp Y_j, \forall i \neq j, \\
\mu_i &= \alpha + \beta_1 X_{1,i}^{(2)}.
\end{aligned}
$$

onde $X_1^{(1)}$ é a categorias de referência de sexo.

```
> (modelo1.shhs <- lm(BMI ~ Sex, data = SHHS))
Call:
lm(formula = BMI ~ Sex, data = SHHS)
Coefficients:
(Intercept)         SexF
     26.332       -1.112
```

Modelo 2

$$\begin{aligned} Y_i \mid X_i &\sim N(\mu_i, \sigma^2), \qquad i = 1, 2, \ldots, 150, \quad Y_i \perp Y_j, \forall i \neq j, \\ \mu_i &= \alpha + \beta_2 X_{2,i}^{(2)} + \beta_3 X_{2,i}^{(3)}. \end{aligned}$$

onde $X_2^{(1)}$ é a categoria de referência de tabagismo.

```
> (modelo2.shhs <- lm(BMI ~ Smoking, data = SHHS))
Call:
lm(formula = BMI ~ Smoking, data = SHHS)
Coefficients:
 (Intercept)     Smokingex  Smokingnever
      24.527         1.691         2.279
```

Modelo 3

$$\begin{aligned} Y_i \mid \mathrm{X}_i &\sim N(\mu_i, \sigma^2), \qquad i = 1, 2, \ldots, 150, \quad Y_i \perp Y_j, \forall i \neq j, \\ \mu_i &= \alpha + \beta_1 X_{1,i}^{(2)} + \beta_2 X_{2,i}^{(2)} + \beta_3 X_{2,i}^{(3)}. \end{aligned}$$

```
> (modelo3.shhs <- lm(BMI ~ Sex + Smoking, data = SHHS))
Call:
lm(formula = BMI ~ Sex + Smoking, data = SHHS)
Coefficients:
 (Intercept)           SexF     Smokingex  Smokingnever
      25.112         -1.340         1.655         2.483
```

FIOCRUZ

Modelo 4

$$
\begin{aligned}
Y_i \mid X_i &\sim N(\mu_i, \sigma^2), \qquad i = 1, 2, \ldots, 150, \quad Y_i \perp Y_j, \forall i \neq j, \\
\mu_i &= \alpha + \beta_1 X_{1,i}^{(2)} + \beta_2 X_{2,i}^{(2)} + \beta_3 X_{2,i}^{(3)} + \beta_4 X_{1,i}^{(2)} X_{2,i}^{(2)} + \beta_5 X_{1,i}^{(2)} X_{2,i}^{(3)}.
\end{aligned}
$$

```
> (modelo4.shhs <- lm(BMI ~ Sex * Smoking, data = SHHS))
Call:
lm(formula = BMI ~ Sex * Smoking, data = SHHS)
Coefficients:
      (Intercept)                 SexF           Smokingex       Smokingnever
          25.5287              -2.2954              1.3009             1.3722
   SexF:Smokingex  SexF:Smokingnever
           0.8008               2.1348
```

Escolhendo modelo via ANOVA

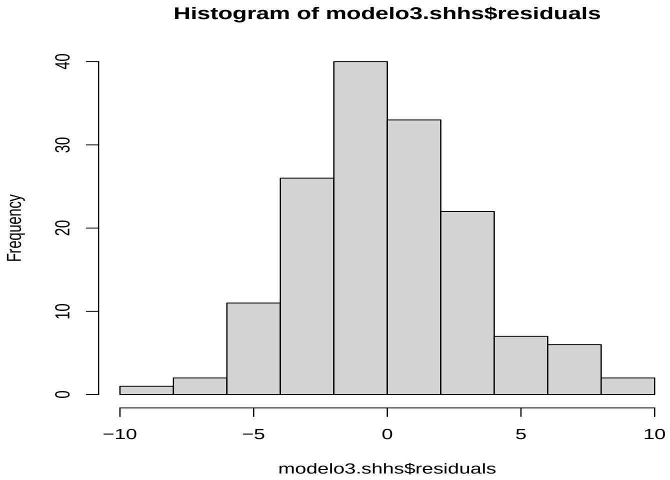
```
> anova(modelo4.shhs)
Analysis of Variance Table
Response: BMI
             Df  Sum Sq Mean Sq F value    Pr(>F)
Sex           1   46.28  46.279  4.2317 0.0414790 *
Smoking       2  167.76  83.880  7.6700 0.0006835 ***
Sex:Smoking   2   29.78  14.892  1.3617 0.2595062
Residuals   144 1574.81  10.936
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Escolhendo modelo via critérios de informação

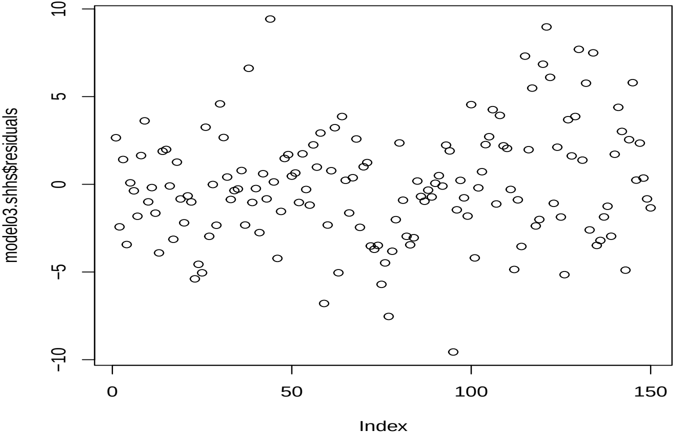
	df	AIC	BIC
modelo1.shhs	3.00	802.10	811.13
modelo2.shhs	4.00	795.19	807.23
modelo3.shhs	5.00	791.18	806.23
modelo4.shhs	7.00	792.37	813.44

Análise de resíduos

```
> hist(modelo3.shhs$residuals)
```



```
> plot(modelo3.shhs$residuals)
```



Modelo final

```
> summary(modelo3.shhs)

Call:
lm(formula = BMI ~ Sex + Smoking, data = SHHS)
Residuals:
    Min      1Q  Median      3Q     Max
-9.5618 -2.1509 -0.2218  1.9846  9.4236

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.1118     0.5071  49.522  < 2e-16 ***
SexF          -1.3400     0.5486  -2.443 0.015779 *
Smokingex      1.6545     0.6707   2.467 0.014785 *
Smokingnever   2.4829     0.6498   3.821 0.000196 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.315 on 146 degrees of freedom
Multiple R-squared:  0.1177,        Adjusted R-squared:  0.09956
F-statistic: 6.492 on 3 and 146 DF,  p-value: 0.0003737
```

Exemplo: DMFT

- DMFT versus consumo de açucar e tipo de país

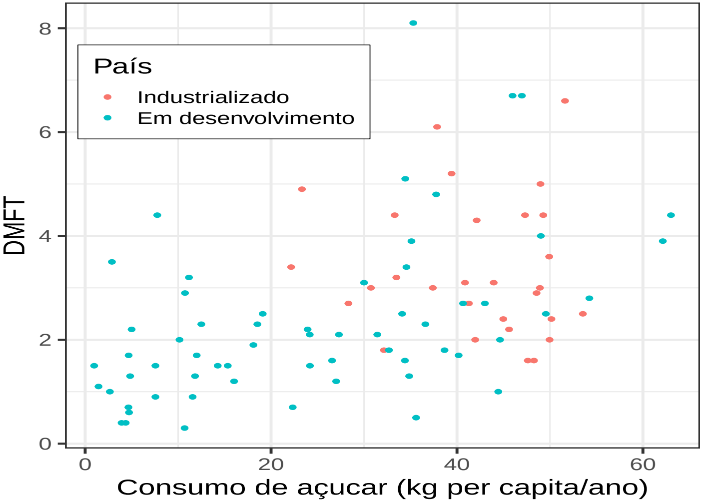
$$
\begin{aligned}
Y_i \mid X_i &\sim N(\mu_i, \sigma^2), \qquad i = 1, 2, \ldots, n, \quad Y_i \perp Y_j, \forall i \neq j, \\
\mu_i &= \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i}.
\end{aligned}
$$

- $n = 90$ países
- X_1 é o consumo de açucar (kg per capita / ano)
- X_2 é a classificação do país

$$
X_2 = \begin{cases} 1 & \text{se país industrializado,} \\ 0 & \text{se país em desenvolvimento.} \end{cases}
$$

FIOCRUZ

Exemplo: DMFT versus consumo de açucar



Quatro modelos distintos

- Modelo 1:

$$\mathbb{E}[DFMT] = \alpha + \beta_1 \text{Consumo}$$

- Modelo 2:

$$\mathbb{E}[DFMT] = \alpha + \beta_1 \text{Consumo} + \beta_2 \text{Pais}$$

- Modelo 3:

$$\mathbb{E}[DFMT] = \alpha + \beta_1 \text{Consumo} + \beta_3 \text{Consumo:Pais}$$

- Modelo 4:

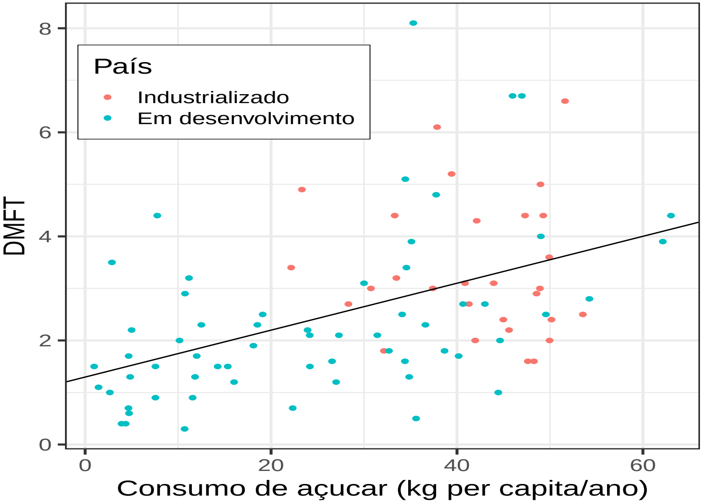
$$\mathbb{E}[DFMT] = \alpha + \beta_1 \text{Consumo} + \beta_2 \text{Pais} + \beta_3 \text{Consumo:Pais}$$

Ajuste dos modelos no R

```
> # Modelo ignorando o tio de país
> modelo1 <- lm(DMFT ~ Consumo, data = dmft)
> #
> # Modelo variando intercepto
> modelo2 <- lm(DMFT ~ Consumo + Pais, data = dmft)
> #
> # Modelo variando slope
> modelo3 <- lm(DMFT ~ Consumo + Consumo:Pais,
+               data = dmft)
> #
> # Modelo variando intercepto e slope
> modelo4 <- lm(DMFT ~ Consumo + Pais + Consumo:Pais,
+               data = dmft)
```

FIOCRUZ

Modelo 1

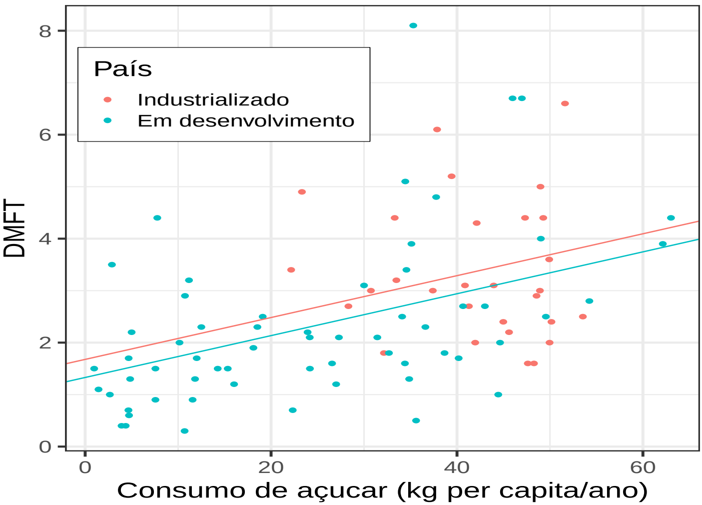

País
Industrializado
Em desenvolvimento
DMFT
8
6
4
2
0
0
20
40
60
Consumo de açucar (kg per capita/ano)

Modelo 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2966	0.3062	4.23	0.0001
Consumo	0.0451	0.0089	5.06	0.0000

```
> AIC(modelo1)
[1] 319.3474
> BIC(modelo1)
[1] 326.8468
```

Modelo 2



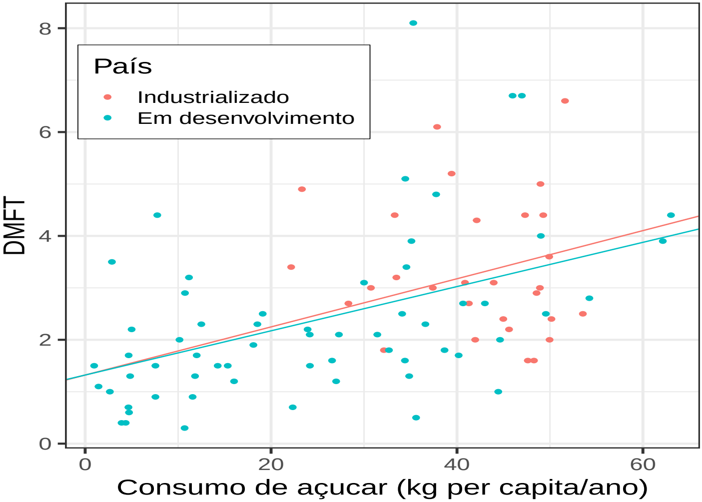
País
Industrializado
Em desenvolvimento
DMFT
0
2
4
6
8
0
20
40
60
Consumo de açucar (kg per capita/ano)

Modelo 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6774	0.4998	3.36	0.0012
Consumo	0.0403	0.0102	3.94	0.0002
PaisEm desenvolvimento	-0.3479	0.3608	-0.96	0.3375

```
> AIC(modelo2)
[1] 320.3903
> BIC(modelo2)
[1] 330.3895
```

Modelo 3

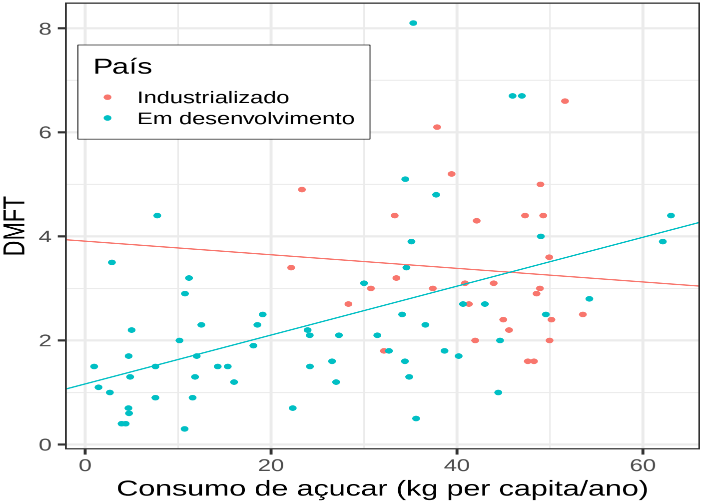

País
Industrializado
Em desenvolvimento
DMFT
0
2
4
6
8
0
20
40
60
Consumo de açucar (kg per capita/ano)

Modelo 3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3211	0.3129	4.22	0.0001
Consumo	0.0464	0.0094	4.92	0.0000
Consumo:PaisEm desenvolvimento	-0.0038	0.0087	-0.43	0.6680

```
> AIC(modelo3)
[1] 321.156
> BIC(modelo3)
[1] 331.1552
```

Modelo 4


País
Industrializado
Em desenvolvimento
DMFT
0
2
4
6
8
0
20
40
60
Consumo de açucar (kg per capita/ano)

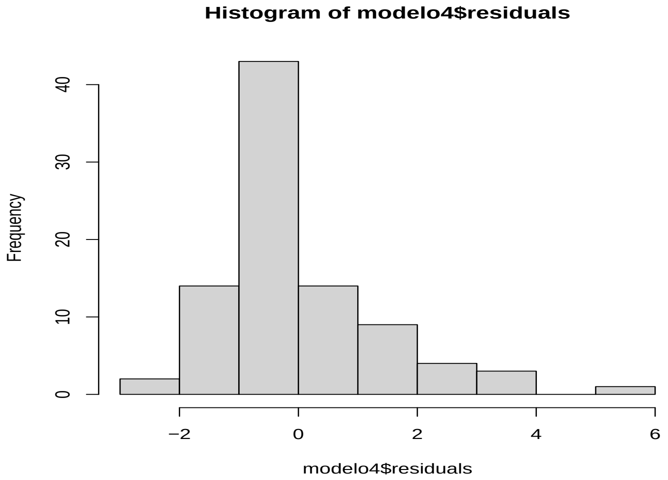
Modelo 4

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.9086	1.2865	3.04	0.0032
Consumo	-0.0131	0.0301	-0.43	0.6658
PaisEm desenvolvimento	-2.7439	1.3248	-2.07	0.0413
Consumo:PaisEm desenvolvimento	0.0600	0.0320	1.88	0.0638

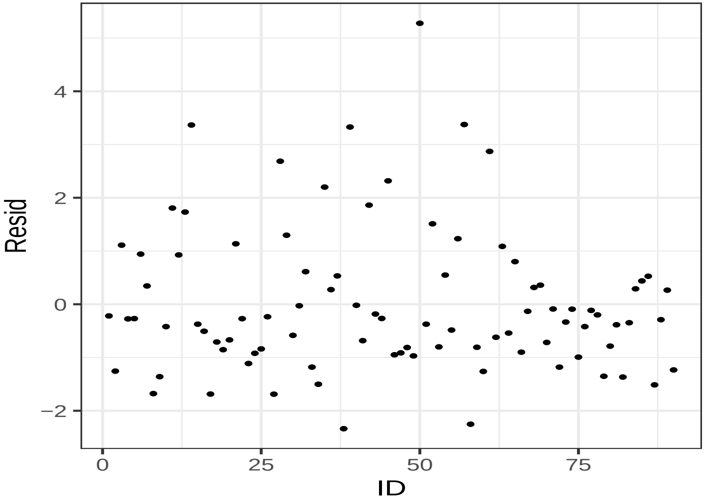
```
> AIC(modelo4)
[1] 318.7751
> BIC(modelo4)
[1] 331.2742
```

FIOCRUZ

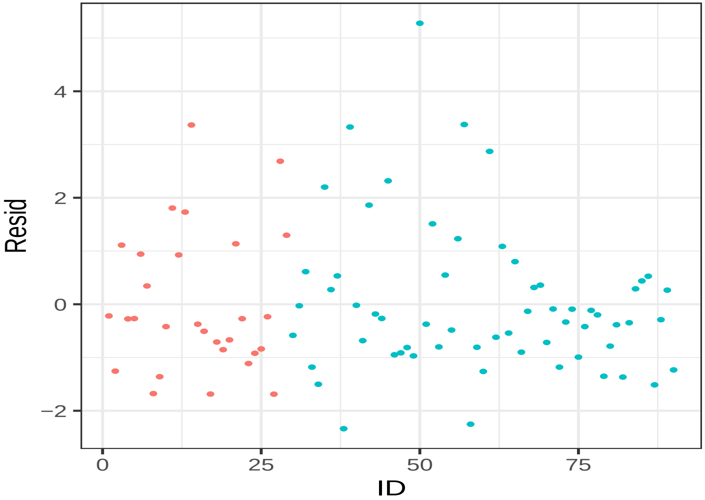
Análise de resíduos do Modelo 4



Análise de resíduos



Análise de resíduos



Quatro modelos distintos na escala do log

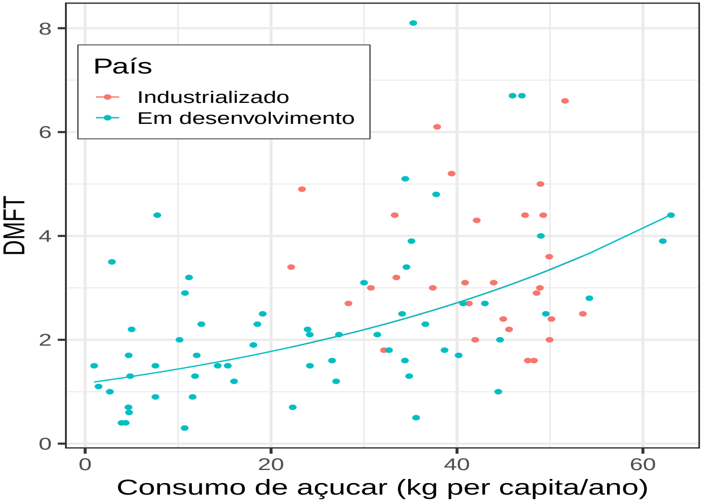
- Modelo 1:
$$\log(\mathbb{E}[DFMT]) = \alpha + \beta_1 \text{Consumo}$$
- Modelo 2:
$$\log(\mathbb{E}[DFMT]) = \alpha + \beta_1 \text{Consumo} + \beta_2 \text{Pais}$$
- Modelo 3:
$$\log(\mathbb{E}[DFMT]) = \alpha + \beta_1 \text{Consumo} + \beta_3 \text{Consumo:Pais}$$
- Modelo 4:
$$\log(\mathbb{E}[DFMT]) = \alpha + \beta_1 \text{Consumo} + \beta_2 \text{Pais} + \beta_3 \text{Consumo:Pais}$$

Ajuste dos modelos no R

```
> # Modelo ignorando o tio de país
> modelo1 <- lm(log(DMFT) ~ Consumo, data = dmft)
> #
> # Modelo variando intercepto
> modelo2 <- lm(log(DMFT) ~ Consumo + Pais, data = dmft)
> #
> # Modelo variando slope
> modelo3 <- lm(log(DMFT) ~ Consumo + Consumo:Pais,
+               data = dmft)
> #
> # Modelo variando intercepto e slope
> modelo4 <- lm(log(DMFT) ~ Consumo + Pais + Consumo:Pais,
+               data = dmft)
```

FIOCRUZ

Modelo 1



País
Industrializado
Em desenvolvimento
DMFT
0
2
4
6
8
0
20
40
60
Consumo de açucar (kg per capita/ano)

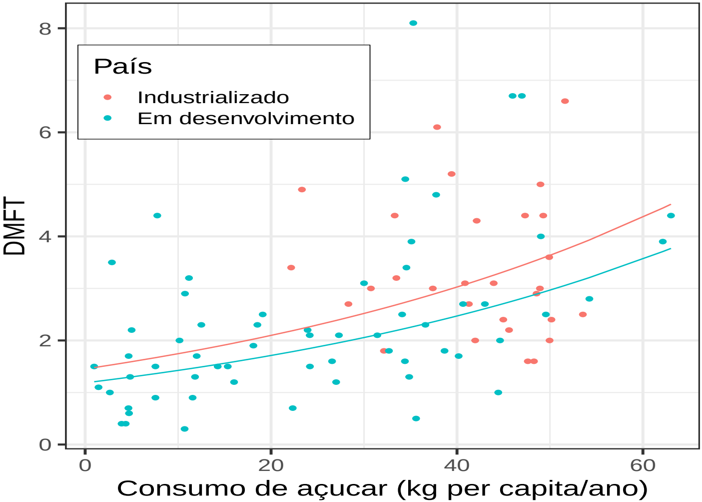
Modelo 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1511	0.1229	1.23	0.2221
Consumo	0.0212	0.0036	5.92	0.0000

```
> AIC(modelo1)
[1] 155.0517
> BIC(modelo1)
[1] 162.5512
```

FIOCRUZ

Modelo 2

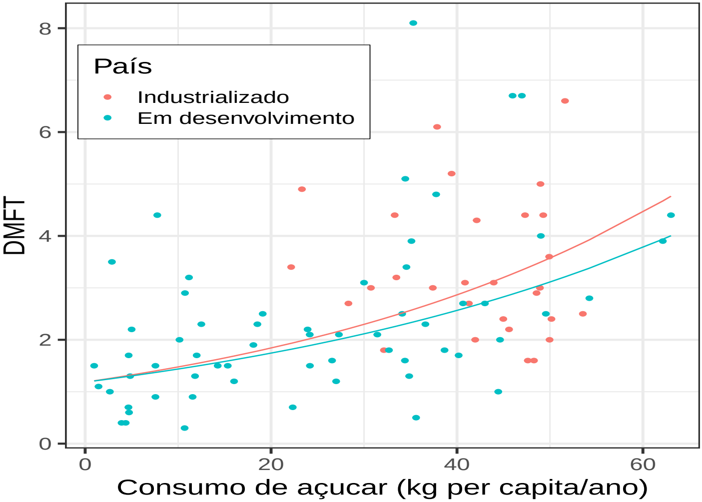

País
Industrializado
Em desenvolvimento
DMFT
0
2
4
6
8
0
20
40
60
Consumo de açucar (kg per capita/ano)

Modelo 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3736	0.1994	1.87	0.0644
Consumo	0.0184	0.0041	4.50	0.0000
PaisEm desenvolvimento	-0.2032	0.1439	-1.41	0.1616

```
> AIC(modelo2)
[1] 155.0132
> BIC(modelo2)
[1] 165.0125
```

Modelo 3



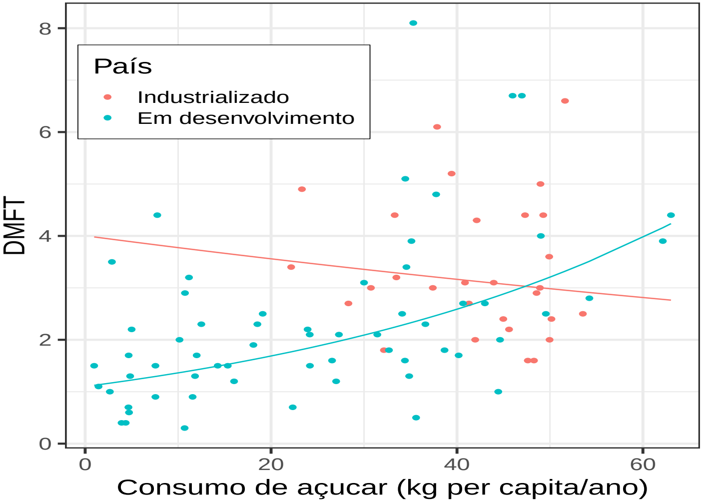
País
Industrializado
Em desenvolvimento
DMFT
0
2
4
6
8
0
20
40
60
Consumo de açucar (kg per capita/ano)

Modelo 3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1691	0.1253	1.35	0.1806
Consumo	0.0221	0.0038	5.85	0.0000
Consumo:PaisEm desenvolvimento	-0.0028	0.0035	-0.79	0.4327

```
> AIC(modelo3)
[1] 156.4114
> BIC(modelo3)
[1] 166.4106
```

Modelo 4



País
Industrializado
Em desenvolvimento
DMFT
Consumo de açucar (kg per capita/ano)
0
2
4
6
8
0
20
40
60

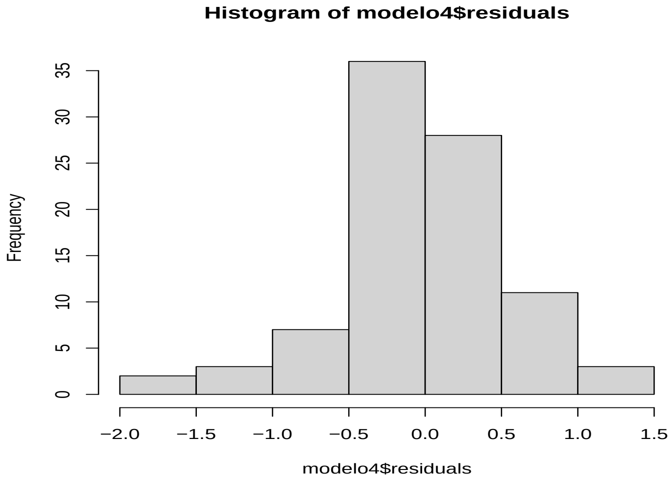
Modelo 4

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3871	0.5102	2.72	0.0079
Consumo	-0.0059	0.0120	-0.49	0.6241
PaisEm desenvolvimento	-1.2916	0.5254	-2.46	0.0160
Consumo:PaisEm desenvolvimento	0.0273	0.0127	2.15	0.0343

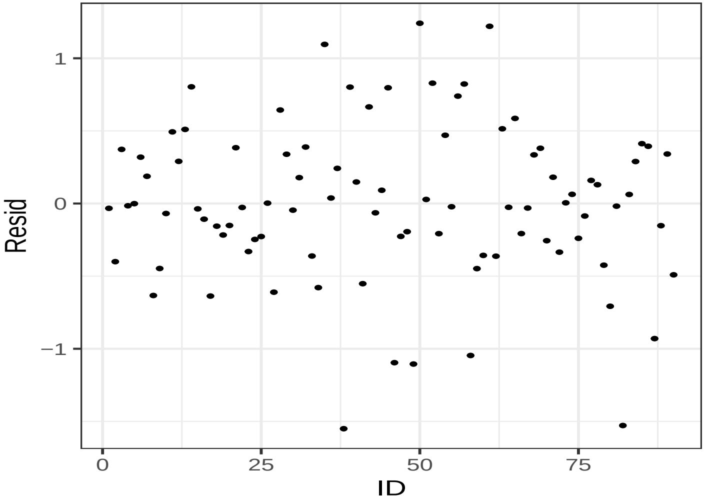
```
> AIC(modelo4)
[1] 152.2993
> BIC(modelo4)
[1] 164.7983
```

FIOCRUZ

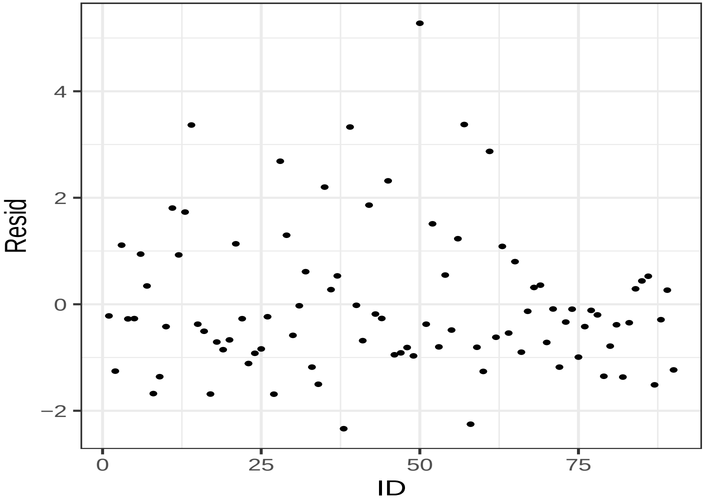
Análise de resíduos do Modelo 1



Análise de resíduos



Melhorou?



Resumo

- Análise de resíduos do modelo linear
  - Normalidade (histograma ou qqplot dos resíduos)
  - Independência (resíduos versus ordem dos dados)
  - Variancia constante (resíduos versus ordem dos dados)
- Comparação de modelos
  - ANOVA
  - Critérios de informação
- Aula prática
  - Replicar os exemplos da aula e outros
  - Veremos métodos de seleção de variáveis

FIOCRUZ