Modelos Estatísticos I
Modelos para desfecho contínuo

Leo Bastos – leonardo.bastos@fiocruz.br

PROCC – Fundação Oswaldo Cruz

Outline

Modelagem estatística, de uma forma mais geral

- Defina seu desfecho, Y, entenda bem sua natureza.
- Represente a aleatoriedade associada ao desfecho usando uma distribuição de probabilidades. (Componente aleatório / Verossimilhança)
- Defina uma variável de exposição, X, e sua relação com a média do desfecho (Componente sistemático)

$$\mathbb{E}[Y] = f(X)$$

- Estime os parâmetros associados (Inferência)

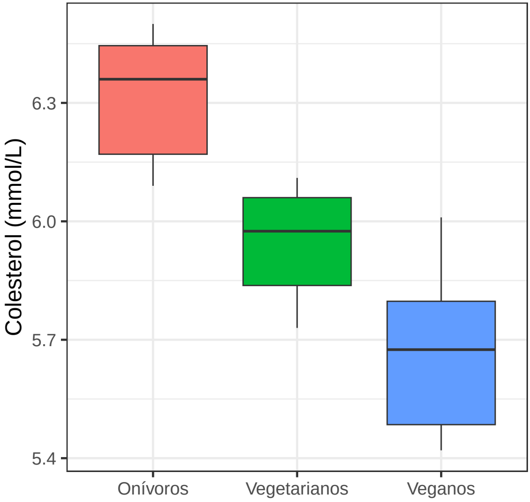
Desfecho contínuo exposição categórica

- Vamos supor que temos um desfecho contínuo, Y, e uma variável categórica com I categorias
- Se fossem duas categorias, $I = 2$, teríamos um teste de comparação de médias. (Qual teste?)
- Para $I > 2$ categorias, podemos usar uma técnica chamada **ANOVA** (**AN**alysis **O**f **VA**riance)
- ANOVA é uma técnica estatística proposta por Sir Ronald A. Fisher (1921)
- E consiste é um método de testar a hipótese de igualdade de médias em diferentes grupos

- Suponha que um estudo de efeitos de dieta restrita em carne tenha 6 veganos, 6 lacto-vegetarianos, e 6 onívoros (sem restrição de dieta). E o nível de coleterol (mmol/l) foi avaliado.
- Os dados estão na tabela abaixo:

	Onívoros	Vegetarianos	Veganos
1	6.35	5.92	6.01
2	6.47	6.03	5.42
3	6.09	5.81	5.44
4	6.37	6.07	5.82
5	6.11	5.73	5.73
6	6.50	6.11	5.62

Exemplo


Colesterol (mmol/L)
6.3
6.0
5.7
5.4
Onívoros
Vegetarianos
Veganos

ANOVA

- De forma geral, na ANOVA queremos testar a seguinte hipótese:

$$
\begin{aligned}
H_0 &: \mu_1 = \mu_2 = \cdots = \mu_I \\
H_1 &: \mu_i \neq \mu_j \text{ para algum } i \text{ e } j.
\end{aligned}
$$

- Construindo a ANOVA
  - Seja y_{ij} o valor do desfecho para a j-ésima observação do grupo i.
  - Mantra da ANOVA

Soma de quadrados total	=	Soma de quadrados entre grupos	+	Soma de quadrados dentro dos grupos

  - Em matematiquês

$$
\begin{aligned}
\sum_i \sum_j (y_{ij} - \bar{y})^2 &= \sum_i n_i(\bar{y}_i - \bar{y})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \\
SQ_t &= SQ_e + SQ_d
\end{aligned}
$$

FIOCRUZ

- A famosa tabela da ANOVA

Fonte de variação	Soma de Quad.	g.l.	MS	F
Entre	SQ_e	$I-1$	$s_e^2 = \dfrac{SQ_e}{I-1}$	$F = \dfrac{s_e^2}{s_d^2}$
Dentro	SQ_d	$n-I$	$s_d^2 = \dfrac{SQ_d}{n-I}$	
Total	SQ_t	$n-1$		

- Sob H_0 (médias iguais), $F \sim F_{(I-1,n-I)}$.
- Esse resultado é válido se:
  - as variâncias dos grupos pode ser considerada a mesma
  - o desfecho segue uma distribuição normal, OU se a amostra é grande o suficiente para usar o TCL.

Teorema central do limite

Esse teorema (e suas variações) tem papel fundamental na inferência estatística, e estabelece que:

Theorem (TCL)

Seja $Y_1, Y_2, Y_3, \ldots$ uma sequência de variáveis aleatórias independentes identicamente distribuídas com média $\mathbb{E}[Y_i] = \mu$ e variância $\mathbb{V}[Y_i] < \infty$. Então quando n se aproxima do infinito,

$$\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

FIOCRUZ

ANOVA no exemplo

- Descritivas

	Dieta	n	Media	Variancia
1	Onívoros	6	6.31	0.03
2	Vegetarianos	6	5.95	0.02
3	Veganos	6	5.67	0.05

- No R

```
> aov(Colesterol ~ Dieta, dados)
```

FIOCRUZ

- Saída editada

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Dieta	2	1.24	0.62	17.62	0.0001
Residuals	15	0.53	0.04		

- Qual a conclusão?

Comparações múltiplas

- Como o teste F da ANOVA foi significativo, então temos que encontrar qual(is) grupo(s) tem a média diferente dos outros.
- Isso é feito fazendo comparações dois a dois, lembrando que a variância é a mesma (por hipótese).

Comparações múltiplas

```
> pairwise.t.test(x = dados$Colesterol,
+                 g = dados$Dieta)
        Pairwise comparisons using t tests with pooled SD
data:  dados$Colesterol and dados$Dieta

             Onívoros Vegetarianos
Vegetarianos 0.0078   -
Veganos      8.5e-05  0.0243

P value adjustment method: holm
```

Alternativa não-paramétrica

- A suposição de normalidade ou que o tamanho da amostra seja grande o suficiente podem ser inapropriadas.
- Uma alternativa é o uso de testes não-paramétricos (teste que não assumem uma distribuição de probablidad epara o desfecho).
- A versão não-paramétrica do test-t é o teste de Wilcoxon.
- A ANOVA tem sua versão não-parametrica é chamada de teste Kruskal-Wallis. (seção 9.12, Woodward)

FIOCRUZ

- Esse teste tem como hipóteses

$$
\begin{aligned}
H_0 &: \text{As distribuições de probabilidade dos diferentes grupos é a mesma;} \\
H_1 &: \text{a distribuição de algum grupo é diferente do demais.}
\end{aligned}
$$

- Esse teste é baseado nos postos (rank) dos dados.

Teste Kruskal-Wallis (intuição)

Dados observados:

	Onívoros	Vegetarianos	Veganos
1	6.35	5.92	6.01
2	6.47	6.03	5.42
3	6.09	5.81	5.44
4	6.37	6.07	5.82
5	6.11	5.73	5.73
6	6.50	6.11	5.62

Teste Kruskal-Wallis (intuição)

Posto dos dados:

	Onívoros	Vegetarianos	Veganos
1	15.00	8.00	9.00
2	17.00	10.00	1.00
3	12.00	6.00	2.00
4	16.00	11.00	7.00
5	13.50	4.50	4.50
6	18.00	13.50	3.00

FIOCRUZ

Teste Kruskal-Wallis

```
> kruskal.test(Colesterol ~ Dieta, data = dados)

        Kruskal-Wallis rank sum test
data:  Colesterol by Dieta
Kruskal-Wallis chi-squared = 12.52, df = 2, p-value = 0.001911
```

Comparações múltiplas

```
> pairwise.wilcox.test(x = dados$Colesterol,
+                      g = dados$Dieta)
        Pairwise comparisons using Wilcoxon rank sum test with
data:  dados$Colesterol and dados$Dieta

             Onívoros Vegetarianos
Vegetarianos 0.0205   -
Veganos      0.0065   0.0542

P value adjustment method: holm
```

Variável dummy

- Variáveis categóricas podem ser representadas por variáveis indicadoras ou dummy.
- Uma variável categórica com I categorias pode ser representada com I variáveis dummy, uma por categoria.
- No exemplo, portanto teremos portanto 3 variáveis dummy:

$$
\begin{aligned}
x^{(1)} &= \begin{cases} 1 & \text{para onívoros} \\ 0 & \text{caso contrário} \end{cases}, \\
x^{(2)} &= \begin{cases} 1 & \text{para vegetarianos} \\ 0 & \text{caso contrário} \end{cases}, \\
x^{(3)} &= \begin{cases} 1 & \text{para veganos} \\ 0 & \text{caso contrário} \end{cases}.
\end{aligned}
$$

FIOCRUZ

O modelo da ANOVA

- Para evitar problemas numéricos (identificabilidade), uma variável com I categorias deve ter $I - 1$ variáveis dummy.
- Uma das categorias deve ser escolhida como base.
(Na epidemiologia, usa-se a categoria de menor ou maior risco)
- No exemplo, suponha que escolhemos os onívoros como categoria de base, então temos agora as seguintes combinações:

Dieta	$X^{(2)}$	$X^{(3)}$
Onívoro	0	0
Vegetariano	0	1
Vegano	1	0

FIOCRUZ

- De forma geral, o modelo estatístico que gere resultados equivalentes a ANOVA é

$$y = \alpha + x^{(2)}\beta_2 + x^{(3)}\beta_3 + \cdots + x^{(I)}\beta_I + \epsilon$$

onde $\epsilon \sim N(0, \sigma^2)$.

- Notem que não tem a variável dummy da primeira categoria (que é escolhida arbitrariamente).
- No exemplo, temos agora as médias

Dieta	$X^{(2)}$	$X^{(3)}$	Média
Onívoro	0	0	α
Vegetariano	0	1	$\alpha + \beta_2$
Vegano	1	0	$\alpha + \beta_3$

- Usualmente não a necessidade de criar variáveis dummy, elas são criadas automaticamente nos softwares.
- Diferentes softwares usam diferentes formas para escolha da categoria de base. No R, isso é feito automaticamente ao definir uma variável como fator.

```
> # O primeiro valor do vetor levels é a referencia
> # default: ordem alfabetica
> factor(x, levels)
> #
> # Redefinindo a categoria de referencia
> relevel(x, ref)
```

ANOVA: Resumindo

- A análise de variância foi uma técnica de alta relevância no milênio passado.
- Houve muito desenvolvimento em torno dessa técnica, como a Two way ANOVA, Three Way ANOVA, MANOVA, etc.
- Por outro lado, a ANOVA e sua versão não paramétrica (Kruskall-Wallis) são bastante usadas na epidemiologia como medida descritiva.
- Útil por exemplo, na "tabela 1" (ou "tabela 2") quando há cruzamento entre uma variável categórica versus uma contínua.
- Como a ANOVA pode ser escrita como um modelo linear, vamos olhar para essa modelagem a partir de agora.

FIOCRUZ

Modelagem estatística, de uma forma mais geral

- Defina seu desfecho, Y, entenda bem sua natureza.
- Represente a aleatoriedade associada ao desfecho usando uma distribuição de probabilidades. (Componente aleatório / Verossimilhança)
- Defina uma variável de exposição, X, e sua relação com a média do desfecho (Componente sistemático)

$$\mathbb{E}[Y] = f(X; \theta)$$

- Estime os parâmetros θ (Inferência)

FIOCRUZ

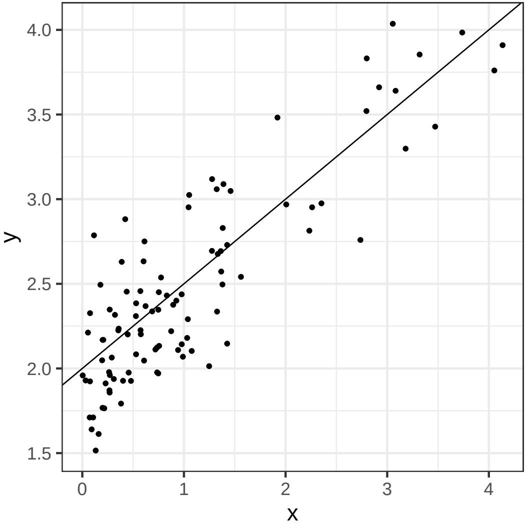
Desfecho contínuo exposição contínua

- Modelo linear simples
$$Y = \alpha + \beta X + \epsilon,$$
onde $\epsilon \sim N(0, \sigma^2)$.
- O modelo linear é um dos modelos mais úteis da estatística, e é aplicado em praticamente todas as áreas do conhecimento.
- Também conhecido por regressão linear, o termo regressão foi proposto por Francis Galton em 1886 em seu estudo de eugenia. (Vejam um pouco mais sobre isso em `http://chance.amstat.org/2013/09/1-pagano/`)
- É possivel construir uma tabela ANOVA para desfecho e exposição contínuas.

FIOCRUZ

- Modelo linear simples
$$Y = \alpha + \beta X + \epsilon,$$
onde $\epsilon \sim N(0, \sigma^2)$.
- α é usualmente chamado de intercepto, pois é o valor de y quando $x = 0$.
- β é chamado de coeficiente de inclinação (slope), ele representa a contribuição em y quando x aumenta em um unidade.
- σ^2 representa a variabilidade associada aos erros quando uma reta é ajustada.

Grafico de dispersão

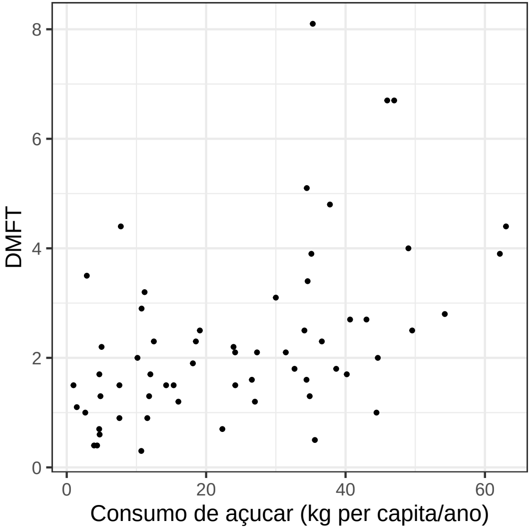


Exemplo: DMFT versus consumo de açucar

- Esse exemplo considera a relação entre o consumo de açucar e cáries em 61 países em desenvolvimento. Dados explorados em Woodward and Walker (1994).
- DMFT (Decayed, missing, or filled teeth) é um valor médio de DMFT em crianças de 12 anos por país, extraído do *WHO Oral Disease Data Bank*.
- Consumo de açucar (Kg per capita/ano) foi derivado de fontes dos governos e indústrias desses países.

FIOCRUZ

Exemplo: DMFT versus consumo de açucar



FIOCRUZ

- Modelo linear simples
$$Y = \alpha + \beta X + \epsilon,$$
onde $\epsilon \sim N(0, \sigma^2)$.
- O método mais antigo para estimar (α, β) é o método de mínimos quadrados (OLS) que tem a seguinte formulação matemática:
$$(\hat{\alpha}, \hat{\beta})_{OLS} : arg\, \min \sum_i (y_i - \alpha - \beta x_i)^2$$
- O estimador de máxima verossimilhança é para $(\alpha, \beta, \sigma^2)$ é dado por
$$(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)_{MLE} : arg\, \max (\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (y_i - \alpha - \beta x_i)^2 \right\}$$

FIOCRUZ

- Vantagem do MLE, seja $\hat{\theta}$ um MLE para θ, então:

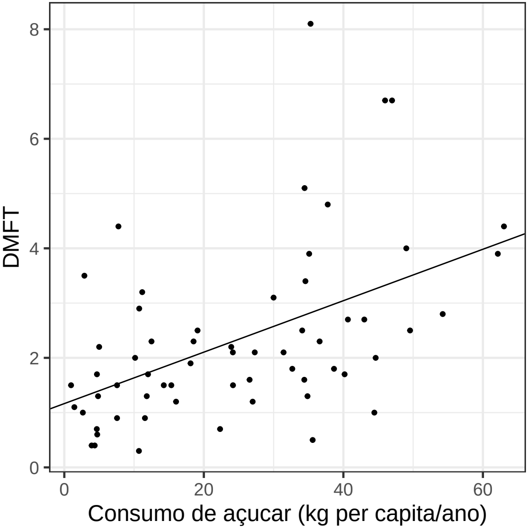
$$\hat{\theta} \xrightarrow{d} N(\theta, \mathbb{V}[\hat{\theta}])$$

- Suposição de independência, e normalidade de Y ou grandes amostras.
- Isso permite a construção de ICs e a realização de testes de hipóteses.
- Contas já feitas e implementadas para modelos lineares em qualquer software estatístico com alguma dignidade.

Voltando ao exemplo

```
> output <- lm(DMFT ~ Consumo, data = dmft2)
> output
Call:
lm(formula = DMFT ~ Consumo, data = dmft2)
Coefficients:
(Intercept)      Consumo
    1.16468      0.04698
```

Exemplo: DMFT versus consumo de açucar



FIOCRUZ

Saida completa

```
> summary(output)
Call:
lm(formula = DMFT ~ Consumo, data = dmft2)
Residuals:
    Min      1Q  Median      3Q     Max
-2.3370 -0.8124 -0.2896  0.4381  5.2771

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.16468    0.32044   3.635 0.000585 ***
Consumo      0.04698    0.01082   4.340 5.66e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.394 on 59 degrees of freedom
Multiple R-squared:  0.242,	Adjusted R-squared:  0.2292
F-statistic: 18.84 on 1 and 59 DF,  p-value: 5.662e-05
```

Saida completa

- Podemos prever o valor médio de Y para qualquer valor de X usando a equação
$$\mathbb{E}[Y \mid X] = \alpha + \beta X$$
- Segundo esse modelo qual o valor de DMFT esperado para o Brasil cujo consumo foi de 46.98 kg per capita/ano?
- Aplicando direto na fórmula:

$$\mathbb{E}[Y \mid X = 46.98] = 1.165 + 0.047 \times 46.98 = 3.37$$

  No entanto o valor observado foi de 6.7.
- Se consumo de açucar do Brasil fosse reduzido em 10 unidades qual seria o DMFT esperado?

$$\mathbb{E}[Y \mid X = 36.98] = 1.165 + 0.047 \times 36.98 = 2.90$$

- E a incerteza associada?

FIOCRUZ

Incerteza para valores preditos

- Graças a uma propriedade esperta dos MLEs (invariância), temos que o estimador do valor esperado de Y também é um MLE.

$$\widehat{\mathbb{E}[Y]} = \hat{\alpha} + \hat{\beta}X$$

- E usando outro resultado, sabemos que a distribuição de $\widehat{\mathbb{E}[Y]}$ é assintoticamente normal com média $\mathbb{E}[Y]$ e variância conhecida.
- Na verdade a variância de $\widehat{\mathbb{E}[Y]}$ tem a seguinte forma:

$$\begin{aligned}\mathbb{V}\left[\widehat{\mathbb{E}[Y]}\right] &= \mathbb{V}\left[\hat{\alpha} + \hat{\beta}X\right] \\ &= \mathbb{V}\left[\hat{\alpha}\right] + X^2\mathbb{V}\left[\hat{\beta}\right] + 2X\mathbb{C}ov\left[\hat{\alpha}, \hat{\beta}\right]\end{aligned}$$

- A matriz de covariância de um modelo linear é dada no R pela função: *vcov(modelo)*

Variância do valor esperado

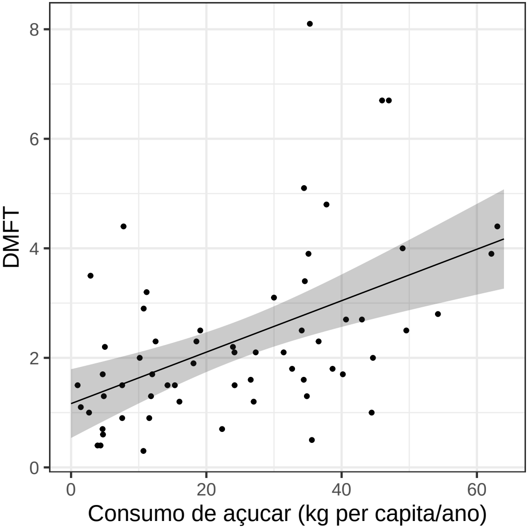
```
> (COV <- vcov(output))
            (Intercept)       Consumo
(Intercept)  0.102681069 -0.0028804239
Consumo     -0.002880424  0.0001171536
> #
> # A variancia da previsao para o valor esperado
> COV[1,1] + 46.98^2*COV[2,2] + 2*46.98*COV[1,2]
[1] 0.09060863
> #
> # Variancia para o novo valor
> COV[1,1] + 36.98^2*COV[2,2] + 2*36.98*COV[1,2]
[1] 0.04985491
```

Variância do valor esperado

```
> # Usando a funcao predict
> previsao <- predict(output, se.fit = T,
+                     newdata = data.frame(
+                       Consumo = c(46.98, 36.98)
+                       )
+                     )
> previsao$fit
       1        2
3.371624 2.901862
> previsao$se.fit^2
         1          2
0.09060863 0.04985491
```

FIOCRUZ

Exemplo: DMFT versus consumo de açucar



Funções "não lineares"

- Em algumas situações a relação linear entre desfecho Y e exposição X pode não ser razoável.
- Podemos escrever alguns modelos não-lineares e ainda assim usar métodos de estimação de modelos lineares.
- São exemplos:
    - Modelo de exponencial:

$$Y = A\exp\{BX\}\epsilon \quad \equiv \quad \log(Y) = \log(A) + BX + log(\epsilon)$$

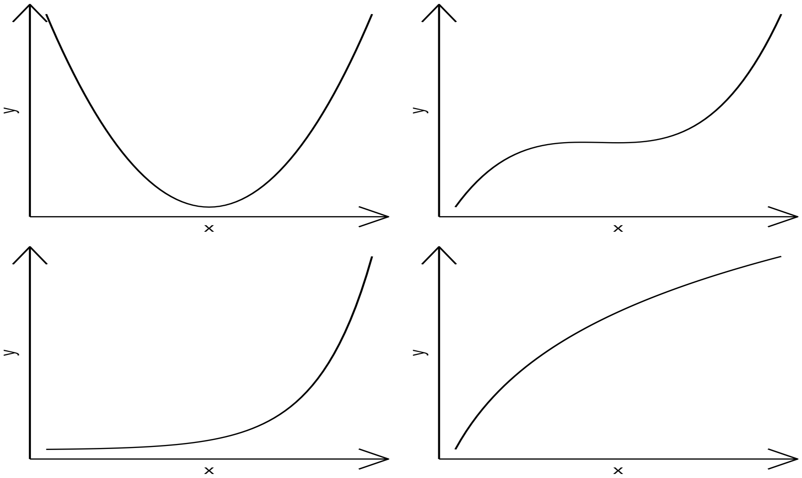
    - Modelo polinomial:

$$Y = \alpha + X\beta_1 + X^2\beta_2 + \cdots + X^p\beta_p$$

    - Modelo logaritmo

$$Y = \alpha + \beta\log(X) + \epsilon$$

FIOCRUZ


y
x
y
x
y
x
y
x

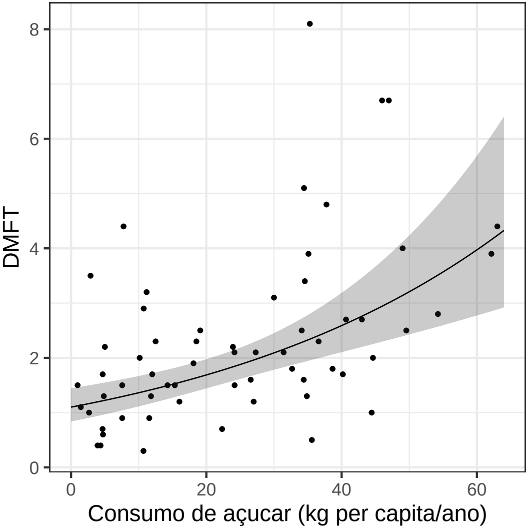
Ajustando um modelo exponencial ao exemplo

```
> output2 <- lm(log(DMFT) ~ Consumo,  data = dmft2)
> summary(output2)
Call:
lm(formula = log(DMFT) ~ Consumo, data = dmft2)
Residuals:
     Min       1Q   Median       3Q      Max
-1.55029 -0.33483  0.00487  0.38049  1.24113

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.095507   0.138912   0.688    0.494
Consumo     0.021394   0.004692   4.560 2.64e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6043 on 59 degrees of freedom
Multiple R-squared:  0.2606,        Adjusted R-squared:  0.248
F-statistic: 20.79 on 1 and 59 DF,  p-value: 2.637e-05
```

Ajustando um modelo exponencial ao exemplo



O modelo linear múltiplo

- Vamos assumir que temos duas ou mais variáveis explicativas.
- Note que podemos continuar interessados na relação desfecho-1-exposição, mas vamos considerar que outras variáveis também explicam o desfecho.
- Continuamos com um desfecho contínuo, Y, 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

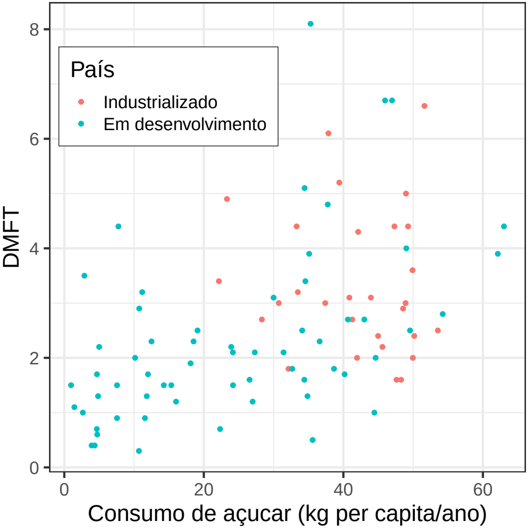
- Vamos considerar o caso em que $p = 2$:

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

- Vamos incluir 29 países industrializados ao nosso exemplo de DMFT versus consumo de açucar.
- Agora termos 90 observações, com uma segunda variável explicativa

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

Exemplo: DMFT versus consumo de açucar

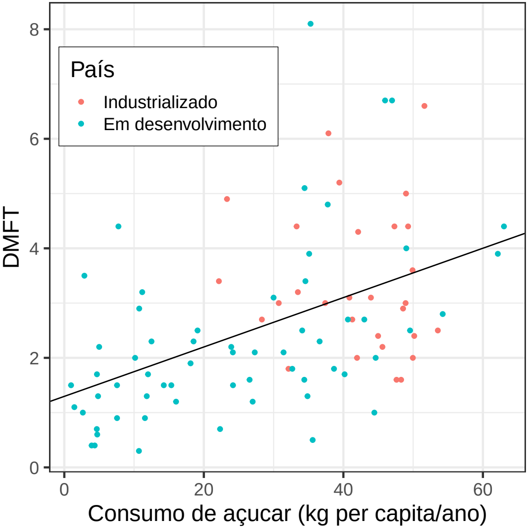


Quatro modelos distintos

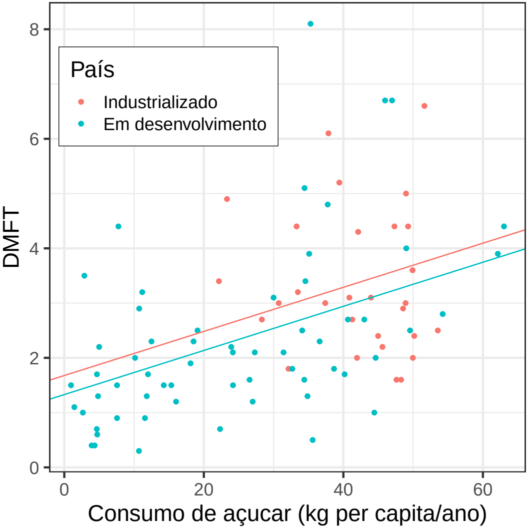
- Modelo 1:
$$\mathbb{E}[\mathit{DFMT}] = \alpha + \beta_1 \text{Consumo}$$
- Modelo 2:
$$\mathbb{E}[\mathit{DFMT}] = \alpha + \beta_1 \text{Consumo} + \beta_2 \text{Pais}$$
- Modelo 3:
$$\mathbb{E}[\mathit{DFMT}] = \alpha + \beta_1 \text{Consumo} + \beta_3 \text{Consumo:Pais}$$
- Modelo 4:
$$\mathbb{E}[\mathit{DFMT}]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

```
> # 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)
```

Modelo 1

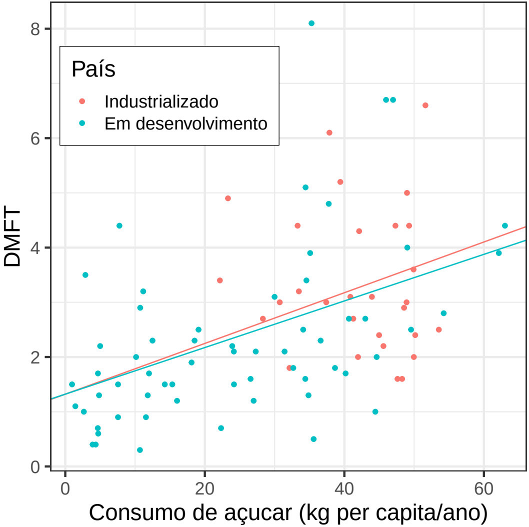


País
Industrializado
Em desenvolvimento
DMFT
Consumo de açucar (kg per capita/ano)
0
2
4
6
8
0
20
40
60


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

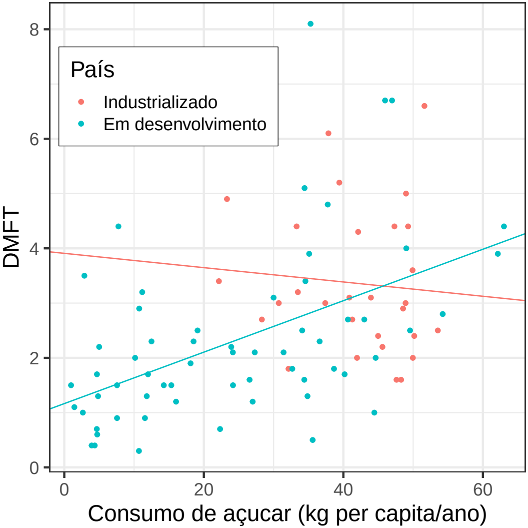
FIOCRUZ

Modelo 3



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

Modelo 4



FIOCRUZ

Duas explicativas discretas: Exemplo SHHS

- 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

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

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
```

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

```
> (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
```

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
```