Programa Educacional de Vigilância em Saúde nas Fronteiras

Estatística aplicada à estatística

Confundimento e Interação

Leo Bastos

(aula reaproveitada, ministrada online para o programa vigifronteiras)

Sumário

- Confundimento
- Interação
- Estimando OR na presença de interação

Confundimento x Interação

O termo **confundimento** é usado por epidemiologistas para descrever a covariável que está associada com ambos, desfecho e exposição.

Epidemiologistas usam o termo **modificador de efeito** para descrever a variável que interage com a exposição.

Exemplo 1: Índice de Massa Corporal

- Suponha que estamos interessados em estudar a associação entre IMC(desfecho) e hipertensão arterial (exposição).
- Sabemos que o IMC está relacionado com uma série de fatores, um deles a escolaridade(covariável).
- Se a escolaridade for igualmente distribuída entre hipertensos e não hipertensos, então uma comparação de médias seria suficiente para comparar o IMC.
- Se a escolaridade não for igualmente distribuída entre os grupos, então uma comparação do IMC entre os dois grupos perde o sentido.
- Vamos supor que todas os outros fatores estão igualmente distribuídos entre os dois grupos, exceto a escolaridade.

Exemplo 1: Índice de Massa Corporal

Distribuição do IMC e escolaridade(anos de estudo) para hipertensos e não hipertensos

Variável	Hipertensos (Grupo 1)		Não hipertensos (Grupo 2)	
	Média	Desvio padrão	Média	Desvio padrão
IMC	26,9	4,7	25,2	4,6
Anos de estudo	10,4	4,7	12,0	4,3

- A diferença entre o IMC entre hipertensos e não hipertensos pode está sendo afetada pela diferença da escolaridade entre os grupos.

- Como podemos comparar o IMC entre os dois grupos?

Exemplo 1: Índice de Massa Corporal

- Devemos controlar pela variável escolaridade.
- Suponha o seguinte modelo de regressão múltipla

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

onde Y é o IMC, X_1 é hipertensão arterial (X_1 = 1 se hipertenso e X_1 = 0 se não hipertenso) e X_2 é a escolaridade.

Se β_1 for estatisticamente diferente de zero, então a distribuição do IMC entre hipertensos e não hipertensos é estatisticamente diferente.

Exemplo 1: Índice de Massa Corporal

Modelo 1: IMC ~ HA

```
Call:
lm(formula = imc_i ~ hart, data = vigitelSL)

Residuals:
    Min      1Q  Median      3Q     Max
-11.691  -3.026  -0.455   2.577  40.772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.201      0.123 204.956  < 2e-16 ***
hart           1.701      0.220   7.732 1.65e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.633 on 2063 degrees of freedom
Multiple R-squared:  0.02816,   Adjusted R-squared:  0.02769
F-statistic: 59.78 on 1 and 2063 DF,  p-value: 1.647e-14
```

Exemplo 1: Índice de Massa Corporal

Modelo 2: IMC ~ HA + ESCOLARIDADE

```
Call:
lm(formula = imc_i ~ hart + anos_estudo, data = vigitelSL)

Residuals:
     Min       1Q   Median       3Q      Max
-11.670  -3.030  -0.435   2.649  40.364

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.60835    0.30286  84.556  < 2e-16 ***
hart         1.64449    0.22326   7.366 2.53e-13 ***
anos_estudo -0.13382    0.02296  -1.473    0.041 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.632 on 2062 degrees of freedom
Multiple R-squared:  0.02918,    Adjusted R-squared:  0.02824
F-statistic: 30.99 on 2 and 2062 DF,  p-value: 5.484e-14
```

Confundimento

- Então a variável escolaridade, seria uma variável de confundimento.
- **Solução:** Ajustar um modelo controlando pela variável de
- confundimento.
- Se o desfecho fosse binário (ou de qualquer outro tipo) o conceito de variável de confundimento continua válido, e a solução é a mesma.

$$g(\mu_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

onde X_2 é uma variável de confundimento.

Exemplo 2: Índice de Massa Corporal

- Suponha que estamos novamente interessados em estudar a associação entre IMC (desfecho) e hipertensão arterial (exposição).
- Vamos considerar uma terceira variável que pode afetar a relação IMC x hipertensão, a idade(covariável).
- Vamos supor que todas os outros fatores estão igualmente
- distribuídos entre hipertensos e não hipertensos, exceto a idade.
- Vamos ajustar um modelo controlando por idade.

Exemplo 2: Índice de Massa Corporal

Modelo 1: IMC ~ HA + IDADE

```
Call:
lm(formula = imc_i ~ hart + idade, data = vigitelSL)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.899  -3.092  -0.467   2.657  40.562 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.573080   0.306813  80.091  < 2e-16 ***
hart         1.449224   0.247046   5.866 5.18e-09 ***
idade        0.013949   0.006249   2.232   0.0257 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.629 on 2062 degrees of freedom
Multiple R-squared:  0.0305,	Adjusted R-squared:  0.02956 
F-statistic: 32.44 on 2 and 2062 DF,  p-value: 1.348e-14
```

Exemplo 2: Índice de Massa Corporal

Modelo 1: IMC ~ HA + IDADE

```
Call:
lm(formula = imc_i ~ hart + idade, data = vigitelSL)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.899  -3.092  -0.467   2.657  40.562 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.573080   0.306813  80.091  < 2e-16 ***
hart         1.449224   0.247046   5.866 5.18e-09 ***
idade        0.013949   0.006249   2.232   0.0257 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.629 on 2062 degrees of freedom
Multiple R-squared:  0.0305,	Adjusted R-squared:  0.02956 
F-statistic: 32.44 on 2 and 2062 DF,  p-value: 1.348e-14
```

Será que idade é uma variável modificadora de efeito?

Exemplo 2: Índice de Massa Corporal

O modelo com **interação multiplicativa** entre idade e hipertensão é:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e$$

onde

Y é o IMC, X_1 é hipertensão arterial (X_1 = 1 se hipertenso, X_1 = 0 se não)

e X_2 é a idade.

Logo,

$$X_1 = 0 \quad \Rightarrow \quad Y = \beta_0 + \beta_2 X_2 + e$$
$$X_1 = 1 \quad \Rightarrow \quad Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + e$$

Exemplo 2: Índice de Massa Corporal

Modelo 2: IMC ~ HA + IDADE + HA*IDADE

```
Call:
lm(formula = imc_i ~ hart * idade, data = vigitelSL)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.423  -3.096  -0.427   2.608  40.360 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.967085   0.338898  70.721  < 2e-16 ***
hart         5.160369   0.930017   5.549 3.25e-08 ***
idade        0.027419   0.007025   3.903 9.80e-05 ***
hart:idade  -0.062729   0.015160  -4.138 3.65e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.611 on 2061 degrees of freedom
Multiple R-squared:  0.03849,	Adjusted R-squared:  0.03709 
F-statistic:  27.5 on 3 and 2061 DF,  p-value: < 2.2e-16
```

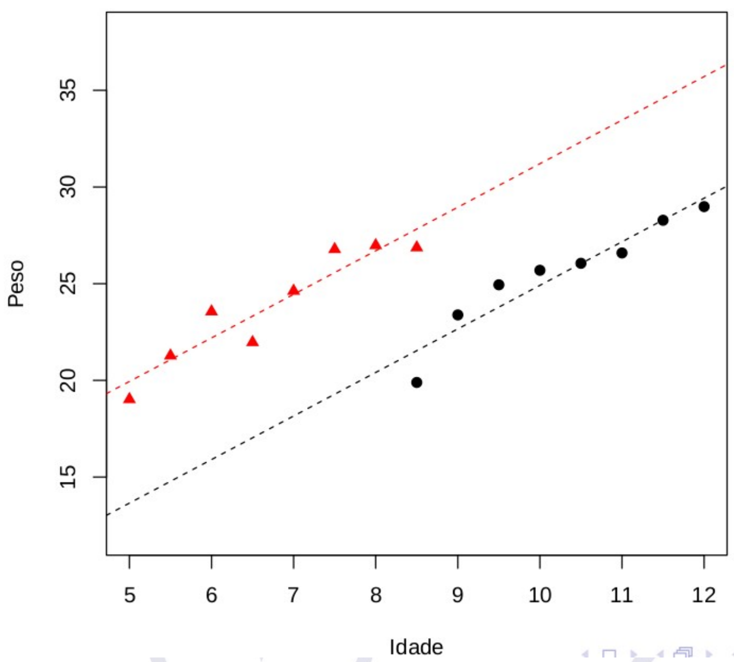
Exemplo 2: Índice de Massa Corporal

Comparando os Modelos 1 e 2

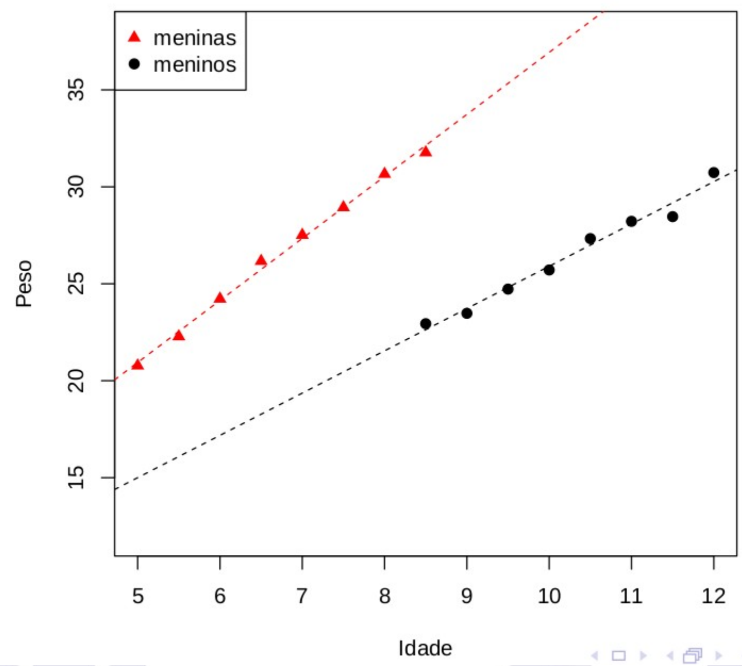
```
Analysis of Variance Table

Model 1: imc_i ~ hart + idade
Model 2: imc_i ~ hart * idade
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   2062 44181
2   2061 43817  1    364.02 17.122 3.646e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Confundimento x interação

Considere o peso de crianças como desfecho, a idade como exposição e o sexo como covariável. Podemos observar graficamente a diferença entre o sexo como variável de confundimento e como modificadora de efeito



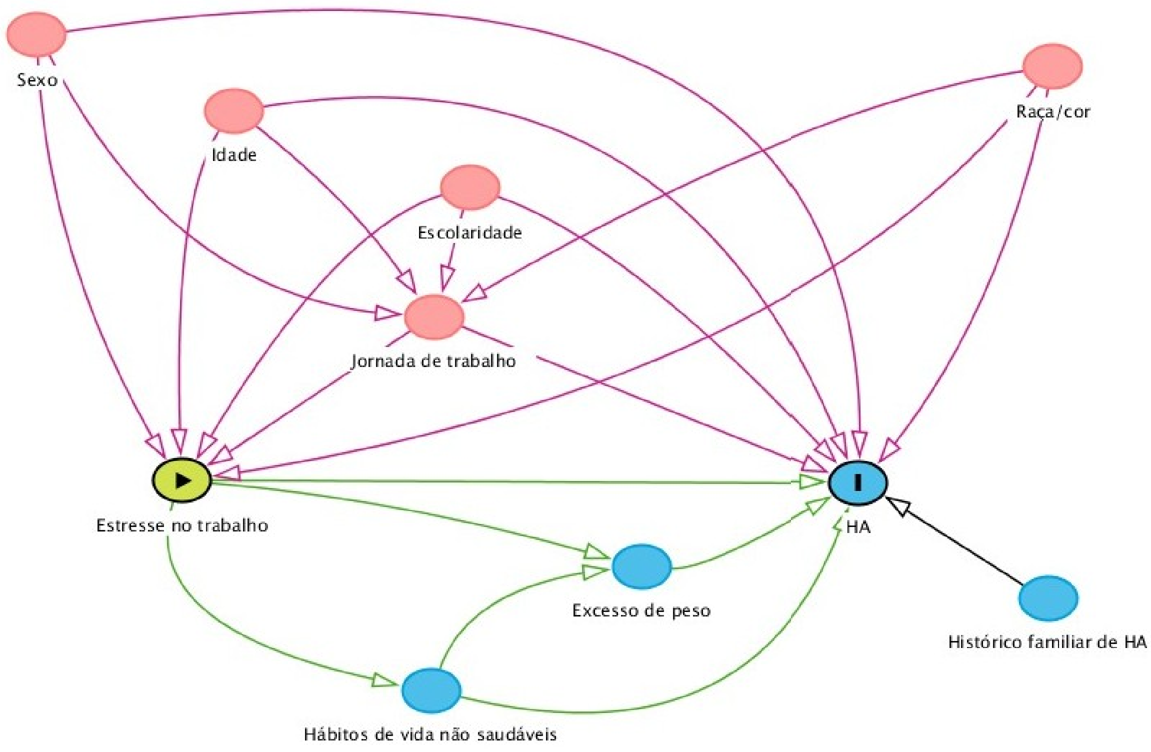
Confundimento



Interação

Exemplo 3: Estresse no trabalho x Hipertensão Arterial

Gráfico Acíclico Direcional para estudar a associação entre estresse no trabalho e hipertensão arterial



Quais as variáveis de confundimento?

Exemplo 4: Doença coronariana

A chance de doença coronariana é diferente entre homens e mulheres? (Hosmer e Lesmeshow, capítulo 3).

CHD: Indicador de presença de doença coronariana: 1 se sim, 0 c.c. (desfecho)

SEX: Sexo do paciente, 1 se masculino, 0 se feminino (exposição).

AGE: Idade do paciente em anos (covariável).

Programa Educacional de Vigilância em Saúde nas Fronteiras

Exemplo 4: Doença coronariana

Vamos ajustar os seguintes modelos:

$$CHD \sim Ber(p)$$

1. $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 SEX$
2. $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 SEX + \beta_2 AGE$
3. $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 SEX + \beta_2 AGE + \beta_3 SEX : AGE$

Exemplo 4: Doença coronariana

Estimativas dos coeficientes e análise da deviance

Modelo	Interc.	SEX	AGE	SEX:AGE	Deviance	Pr(>Chi)
1	0.060	1.981			419.82	
2	-3.374	1.356	0.082		407.78	0.0005
3	-4.216	4.239	0.103	-0.062	406.36	0.2387

A variável idade deve ser considerada, nota-se uma mudança considerável no efeito do sexo.

Idade é uma variável de **confundimento.**

Não existe evidências de **interação** multiplicativa entre sexo e idade.

Estimando OR na presença de interação

Suponha o seguinte modelo com duas variáveis explicativas e a interação entre elas:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Queremos calcular a OR da variável X_1 .Se X_1 = a, então a chance é igual a $\exp\{\beta_0 + \beta_1 a + \beta_2 x_2 + \beta_3 a x_2\}$

Se X_1 = b, então a chance é igual a $\exp\{\beta_0 + \beta_1 b + \beta_2 x_2 + \beta_3 b x_2\}$

Dessa forma a razão de chances de interesse é

$$OR = \frac{\exp\{\beta_0 + \beta_1 a + \beta_2 x_2 + \beta_3 a x_2\}}{\exp\{\beta_0 + \beta_1 b + \beta_2 x_2 + \beta_3 b x_2\}}$$

Estimando OR na presença de interação

Logo
,

$$OR = \exp\{(a-b)(\beta_1 + \beta_3 x_2)\}$$

Se a=1 e
b=0,

$$OR = \exp\{\beta_1 + \beta_3 x_2\}$$

O intervalo de confiança de (1 − α) × 100% para a razão de chances é

$$\exp\left[(a-b)(\hat{\beta}_1 + \hat{\beta}_3 x_2) \pm z_{\alpha/2}\sqrt{\mathbb{V}[(a-b)(\hat{\beta}_1 + \hat{\beta}_3 x_2)]}\right]$$

onde

$$\mathbb{V}[(a-b)(\hat{\beta}_1 + \hat{\beta}_3 x_2)] = (a-b)^2\left(\mathbb{V}[\hat{\beta}_1] + x_2^2\mathbb{V}[\hat{\beta}_3] + 2x_2\mathbb{C}ov[\hat{\beta}_1, \hat{\beta}_3]\right)$$

Programa Educacional de Vigilância em Saúde nas Fronteiras

Exemplo 5: Baixo peso ao nascer

Suponha que estamos interessados em estudar a associação entre o baixo peso ao nascer e o peso da mãe na última menstruação.

low: Baixo peso ao nascer: 1 se peso < 2,5Kg, 0 c.c.(desfecho)

lwd: Peso da mãe na ultima menstruação: 1 se peso < 50Kg, 0 c.c. (exposição)

age: Idade da mãe em anos (covariável).

Exemplo 5: Baixo peso ao nascer

Vamos ajustar uma regressão logística

low ~ Ber(p)

Modelo 1: lwd

Modelo 2: lwd + age

Modelo 3: lwd + age + lwd*age

Programa Educacional de Vigilância em Saúde nas Fronteiras

Exemplo 5: Baixo peso ao nascer

Estimativas dos coeficientes e análise da

Modelo	Intercept	lwd	age	lwd:age	Deviance	Pr(>Chi)
1	-1.05	1.05			226.24	
2	-0.03	1.01	-0.04		224.29	0.1621
3	0.77	-1.94	-0.08	0.13	221.14	0.0761

- A redução na deviance com a inclusão da variável age foi muito pequena (Modelo 2).
- A inclusão do termo de interação fornece uma redução significativa na
- deviance (Modelo 3).

Exemplo 5: Baixo peso ao nascer

- A equação da OR para lwd (mães com peso <50Kg na última menstruação comparado com mães com peso >=50kg) é dada por:
-

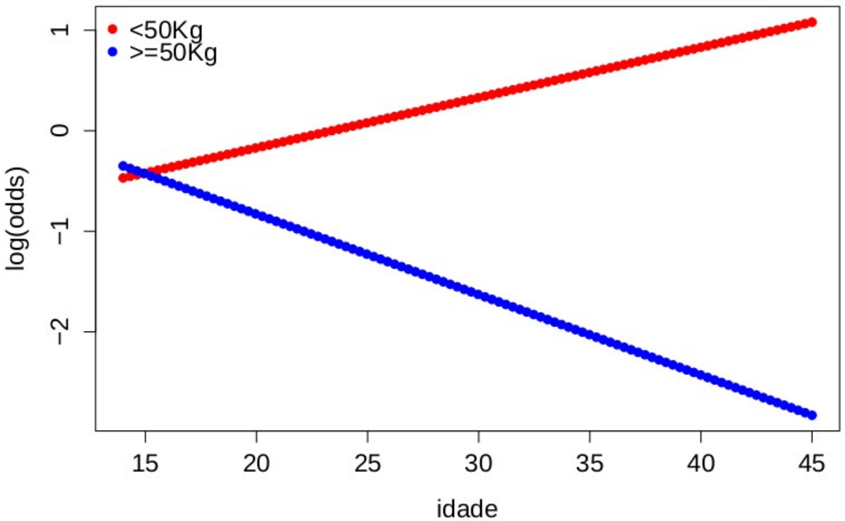
$$OR = \exp\{-1.94 + 0.13 age\}$$

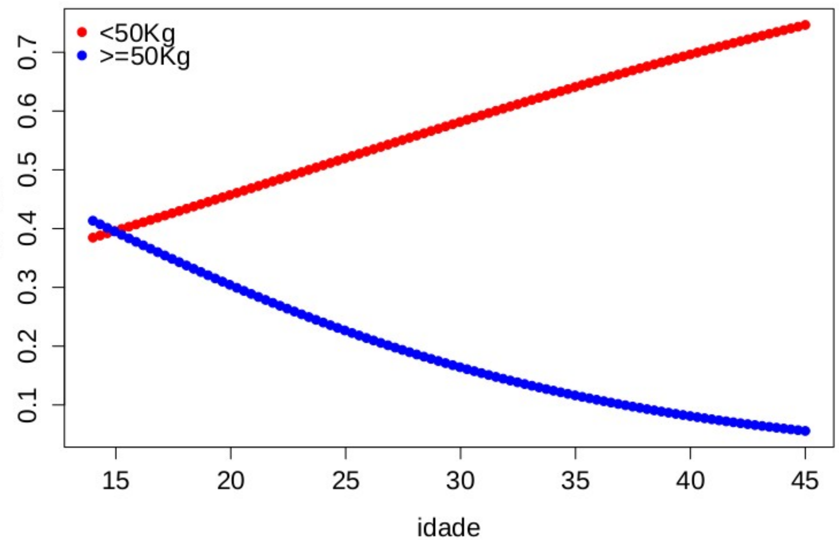
-
- O IC de 95% para OR de lwd é

$$\exp\left\{-1.94 + 0.13 age \pm 1.96\sqrt{2.97 + 0.005 age^2 - 2 \times 0.13 age}\right\}$$

Exemplo 5: Baixo peso ao nascer

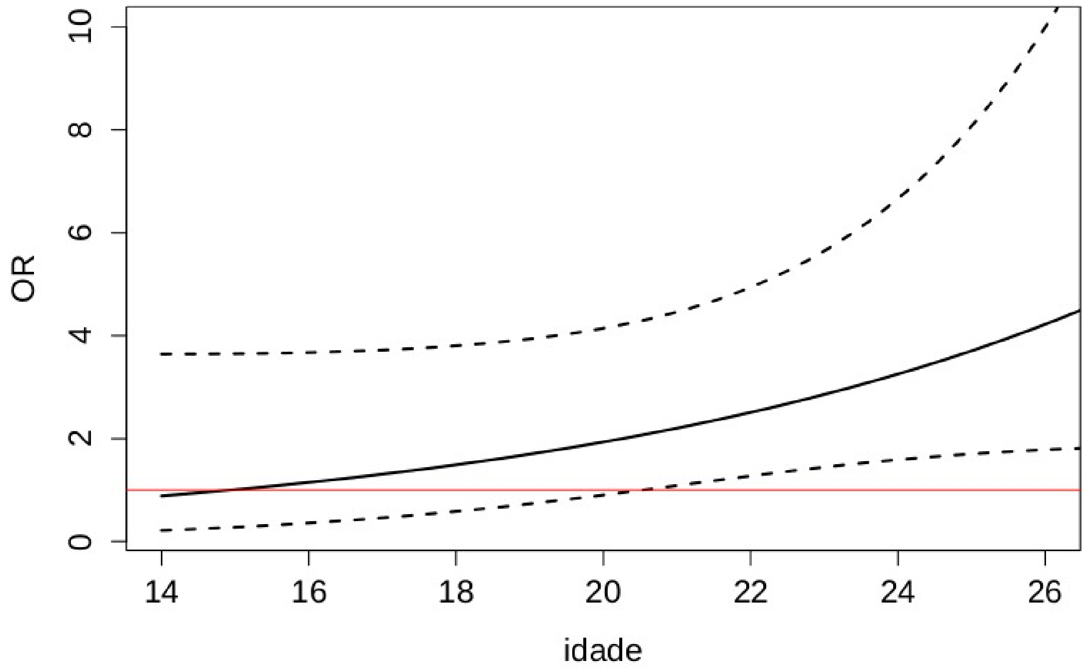
Logaritmo das chances e probabilidade de baixo peso ao nascer por idade da mãe e peso da mãe na última menstruação





Exemplo 5: Baixo peso ao nascer

Razão de chances do peso da mãe na última menstruação
por idade da mãe



Aula Prática

Utilizaremos o banco de dados do VIGITEL "Vigitel2019_SaoLuis.csv"

Suponha a variável auto avaliação de saúde (saruim) como desfecho e as demais variáveis como independentes

Dicionário de dados:

- saruim - avaliação de saúde ruim (0=Não ou 1=Sim).
- q6 - idade.
- q7 - sexo (1=Masculino, 2=Feminino).
- inativo - inatividade física (0=Não ou 1=Sim).
- fesc - faixas de escolaridade (1="0 a 8 anos",2="9 a 11 anos", 3="12 anos ou mais").