Estatística aplicada à epidemiologia II
Modelos para desfecho binário

Leo Bastos – leonardo.bastos@fiocruz.br

PROCC – Fundação Oswaldo Cruz

https://github.com/lsbastos/eae2

FIOCRUZ

O modelo de regressão logística simples

- O modelo de regressão logistica simples tem a seguinte forma:

$$Y_i \sim Bernoulli(\theta_i), \qquad i = 1, 2, \ldots, n$$

onde $\theta_i = P(Y_i = 1)$, e $logit(\theta_i) = \alpha + \beta X_i$.

- Interpretações de β depende de X:
  - Para X binário, β é a diferença do $\log(odds)$ entre as duas categorias
  - Para X categórico, cada β mostra a diferença do $\log(odds)$ entre a categoria específica e a categoria de base.
  - Para X numérico, β representa a mudança em $\log(odds)$ quando se aumenta X em 1 unidade.

FIOCRUZ

Regressão logística simples

- Seja Y um desfecho binário, e.g. tipo de parto: {cesário, vaginal}.
- Vamos supor que estamos interessados em avaliar a associação entre o desfecho e uma exposição X, e.g. raça da mãe {Branca, Não branca}.
- Se a exposição for binária, podemos montar uma tabela 2x2 e calcular algumas medidas de associação de interesse, e avaliar se a associção é significativa via algum teste estatístico.
- Lembrando que o teste só tem validade, se assumirmos a suposição de independência entre observações.

FIOCRUZ

Regressão logística simples: SINASC Rio 2016

- n = 37279 nascidos vivos no município do Rio de Janeiro, notificados no SINASC (Sistema de Informação sobre Nascidos Vivos).
- Microdado baixado via ftp do datasus. (Formato dbc, posteriormente tratado)
- Selecionamos os nascimentos que:
  - Seja a primeira gestação
  - Tenha a informação do tipo de parto
  - Tenha informação completa para idade, escolaridade e raça/cor da mãe.

FIOCRUZ

Regressão logística simples: Variável binária

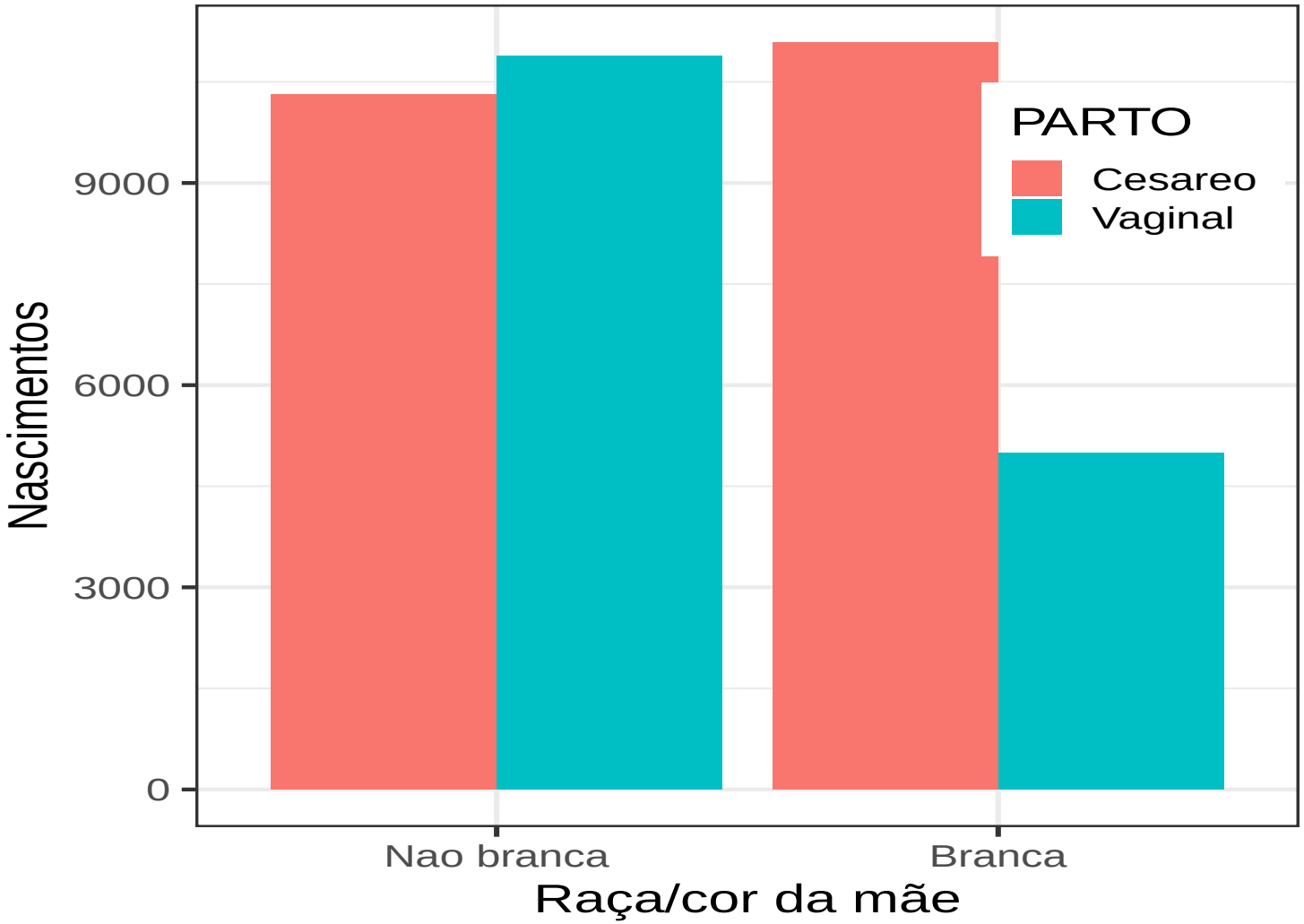
- Tipo de parto versus raça/cor da mãe:

	Nao branca	Branca
Cesareo	10312	11089
Vaginal	10888	4990

  $n = 37279$ nascimentos, e $p.value < 0.0001$. (Qual foi o teste?)
- $OR_{\text{Branca}} = 2.35$.
- Interpretação?

FIOCRUZ

Visualizando



FIOCRUZ

Regressão logística

- Saída do modelo

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.0544	0.0137	-3.96	0.0001
RACACORMAEBranca	0.8529	0.0219	38.95	0.0000

- OR da raça branca com IC de 95%

	OR	2.5 %	97.5 %
1	2.35	2.25	2.45

FIOCRUZ

Regressão logística simples: Variável categórica

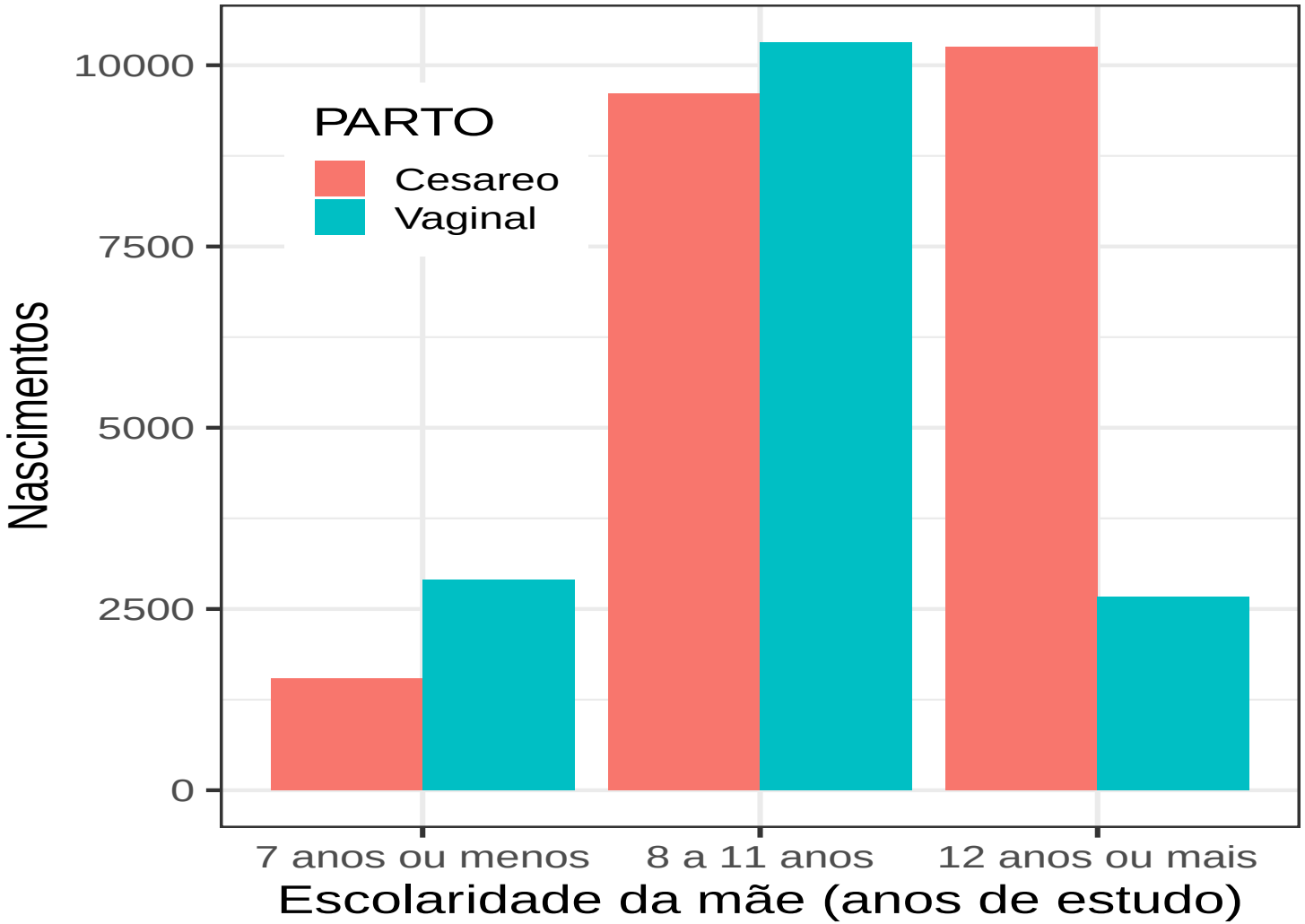
- Tipo de parto versus escolaridade da mãe:

	7 anos ou menos	8 a 11 anos	12 anos ou mais
Cesareo	1535	9614	10252
Vaginal	2896	10317	2665

$n = 37279$ nascimentos, e $p.value < 0.0001$. (Qual foi o teste?)

- $OR_{\text{8 a 11 anos}} = 1.75$ e $OR_{\text{12 anos ou mais}} = 7.26$.

Visualizando



FIOCRUZ

Regressão logística

- Saída do modelo

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.6348	0.0316	-20.11	0.0000
ESCMAE8 a 11 anos	0.5642	0.0346	16.30	0.0000
ESCMAE12 anos ou mais	1.9821	0.0383	51.70	0.0000

- OR com IC de 95%

	OR	2.5 %	97.5 %
ESCMAE8 a 11 anos	1.76	1.64	1.88
ESCMAE12 anos ou mais	7.26	6.73	7.83

- Interpretação?

FIOCRUZ

Regressão logística simples: Numérica

- Tipo de parto versus idade da mãe:

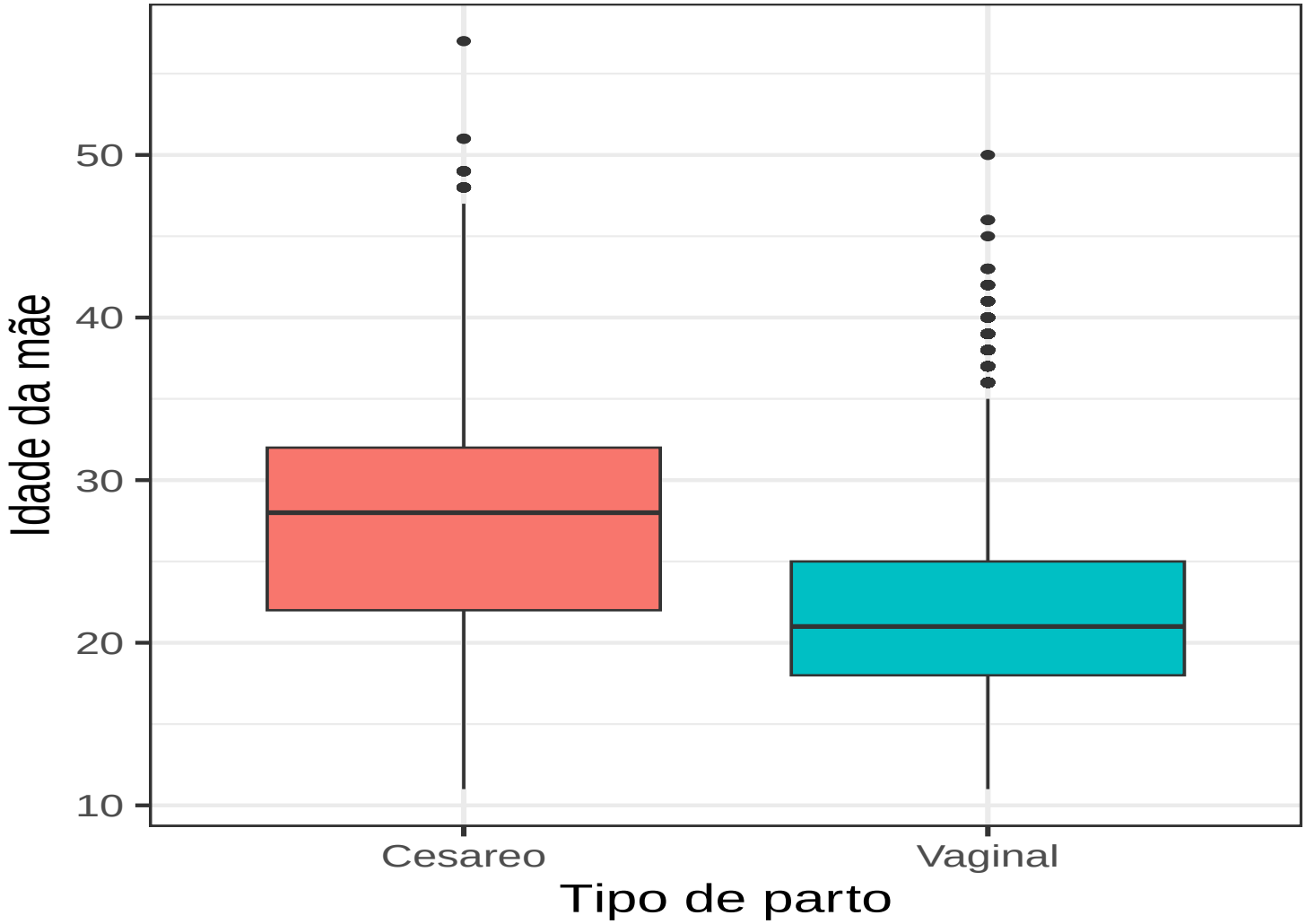
	PARTO	n	Media	sd	LI	LS
1	Cesareo	21401	27.4	6.6	27.3	27.5
2	Vaginal	15878	22.1	5.8	22.0	22.2

$n = 37279$ nascimentos, e $p.value < 0.0001$. (Qual foi o teste?)

FIOCRUZ

Regressão logística simples: Variável numérica

- Tipo de parto versus idade da mãe:



FIOCRUZ

Visualizando

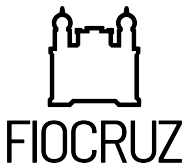
- Saída do modelo

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9550	0.0467	-63.33	0.0000
IDADEMAE	0.1324	0.0019	70.22	0.0000

- OR com IC de 95%

	OR	2.5 %	97.5 %
1	1.14	1.14	1.15

- Interpretação?



Regressão logística múltipla

- Suponha agora que queremos incluir todas as variáveis no modelo
- Isso é feito via modelo múltiplo (~~multivariado~~)
- Seja Y um desfecho binário, e $X_1, X_2, \ldots, X_p$, p variáveis explicativas.
- O modelo de regressão logística é dado por

$$Y_i \sim Bernoulli(\theta_i)$$

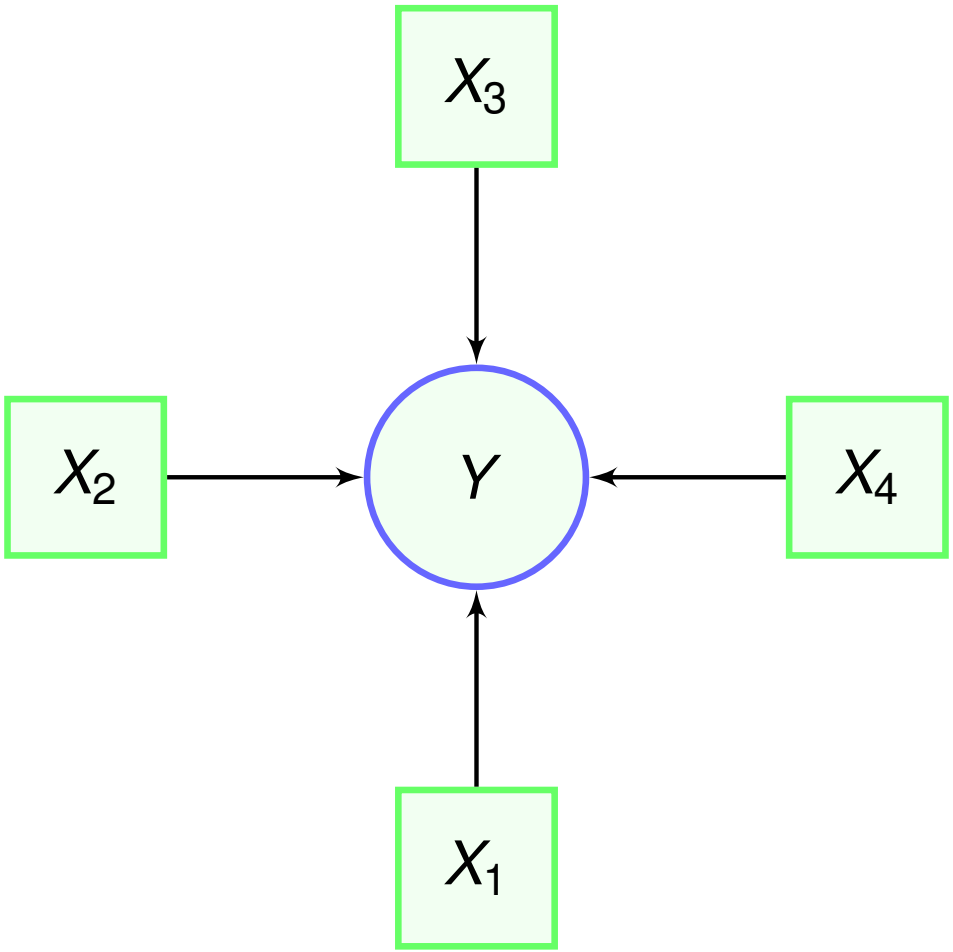
onde

$$logit(\theta_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

- A interpretação dos coeficientes continua a mesma.

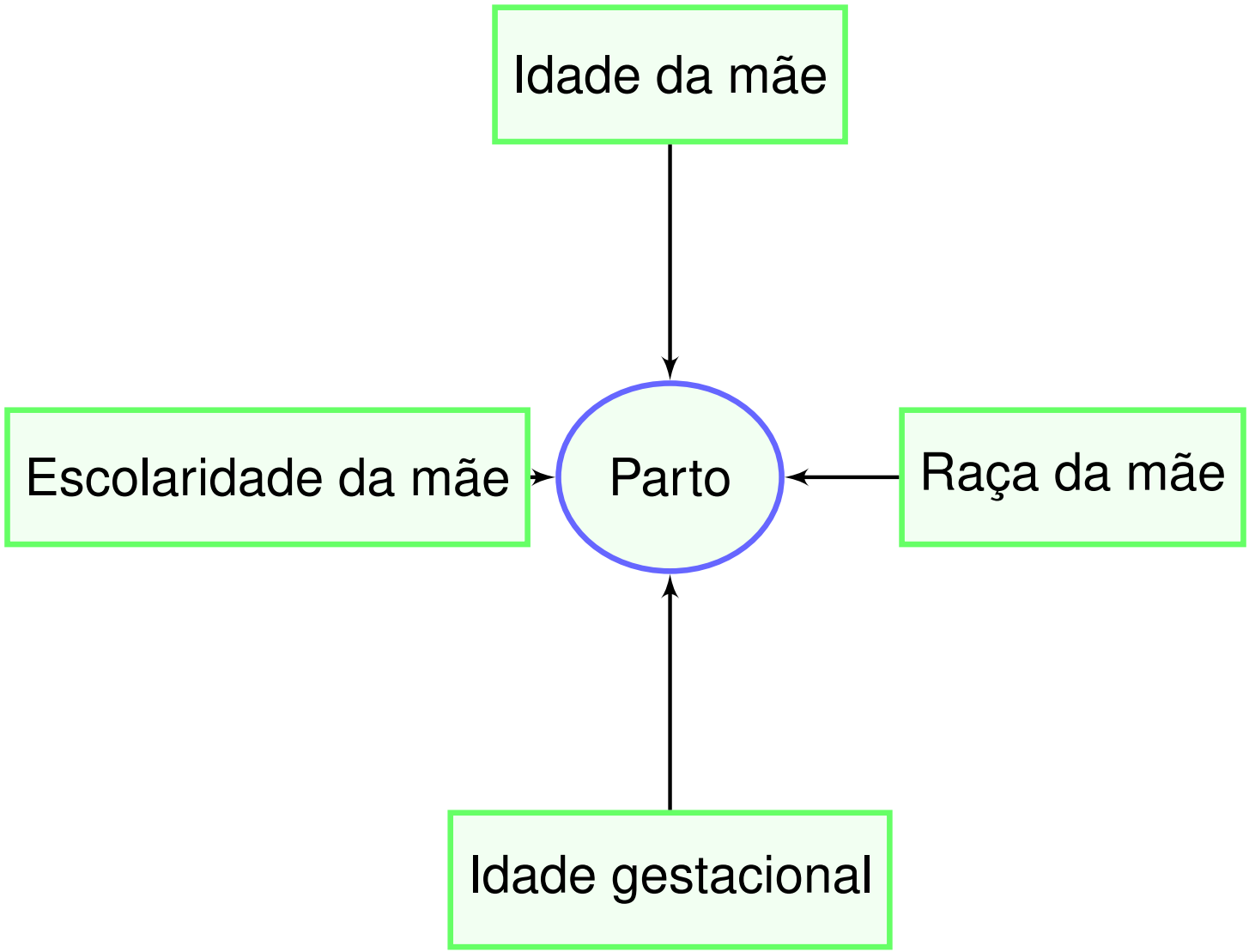
FIOCRUZ

DAG de um modelo de regressão logística



FIOCRUZ

DAG do exemplo



FIOCRUZ

Modelo múltiplo para cesariana

- Ajustando um modelo de regressão logística múltiplo

```
> names(dados2)[4] <- "IDGEST"
> outputB <- glm(I(PARTO == "Cesareo") ~ ESCMAE +
+                  RACACORMAE + IDGEST + IDADEMAE,
+                data = dados2,
+                family = binomial())
```

- A função $I(\cdot)$ é uma função indicadora, no R ela retorna {T,F}, o que é equivalente a {0,1}.

FIOCRUZ

Modelo múltiplo para cesariana

- Os coeficientes estimados

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.7001	0.0558	-48.42	0.0000
ESCMAE8 a 11 anos	0.2152	0.0367	5.87	0.0000
ESCMAE12 anos ou mais	0.8349	0.0464	17.99	0.0000
RACACORMAEBranca	0.2544	0.0252	10.08	0.0000
IDGESTAbaixo de 32 semanas	0.2610	0.0728	3.58	0.0003
IDGESTDe 32 a 37	0.3662	0.0299	12.24	0.0000
IDGESTAcima de 40	0.1942	0.0383	5.07	0.0000
IDADEMAE	0.0983	0.0023	43.35	0.0000

- O que podemos dizer?

FIOCRUZ

Olhando para o efeito da escolaridade

- ORs “brutas” (Parto versus Escolaridade da mãe)

	OR	2.5 %	97.5 %
ESCMAE8 a 11 anos	1.76	1.64	1.88
ESCMAE12 anos ou mais	7.26	6.73	7.83

- OR da escolaridade da mãe controlada por idade gestacional, idade e raça/cor da mãe

	OR	2.5 %	97.5 %
ESCMAE8 a 11 anos	1.24	1.15	1.33
ESCMAE12 anos ou mais	2.30	2.10	2.52

- O que está acontecendo?

FIOCRUZ

Correlação com a exposição

- Raça/Cor versus Escolaridade

	7 ou -	8 a 11	12 ou +
Nao branca	3547	13855	3798
Branca	884	6076	9119

$P.value < 0.0001$ (Que teste foi esse mesmo?)

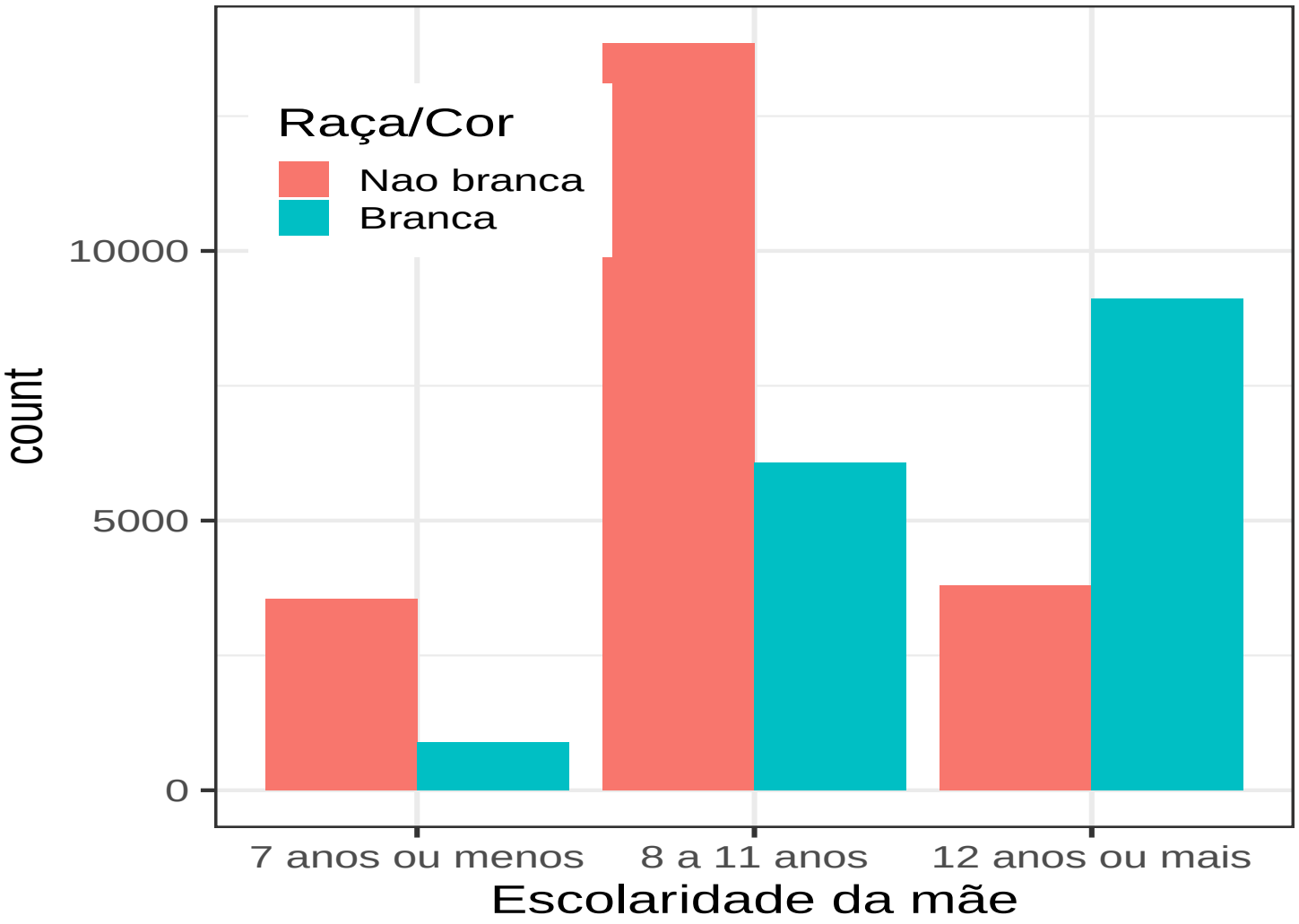
- Idade gestacional versus Escolaridade

	7 ou -	8 a 11	12 ou +
Abaixo de 32 semanas	170	544	252
De 32 a 37	872	3752	2934
De 37 a 40	2877	13279	8914
Acima de 40	512	2356	817

$P.value < 0.0001$ (Que teste foi esse mesmo?)

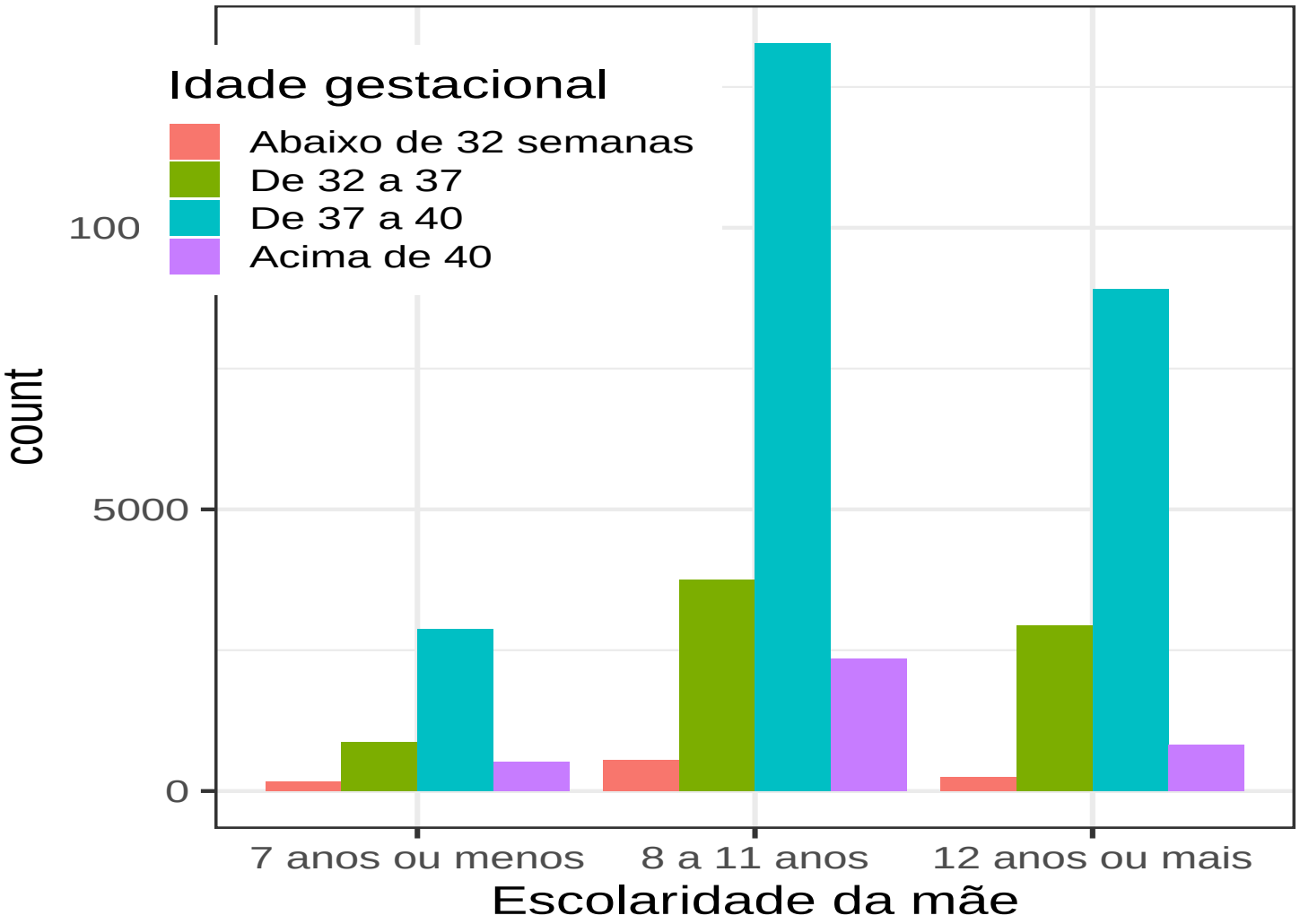
FIOCRUZ

Escolaridade versus Raça/Cor



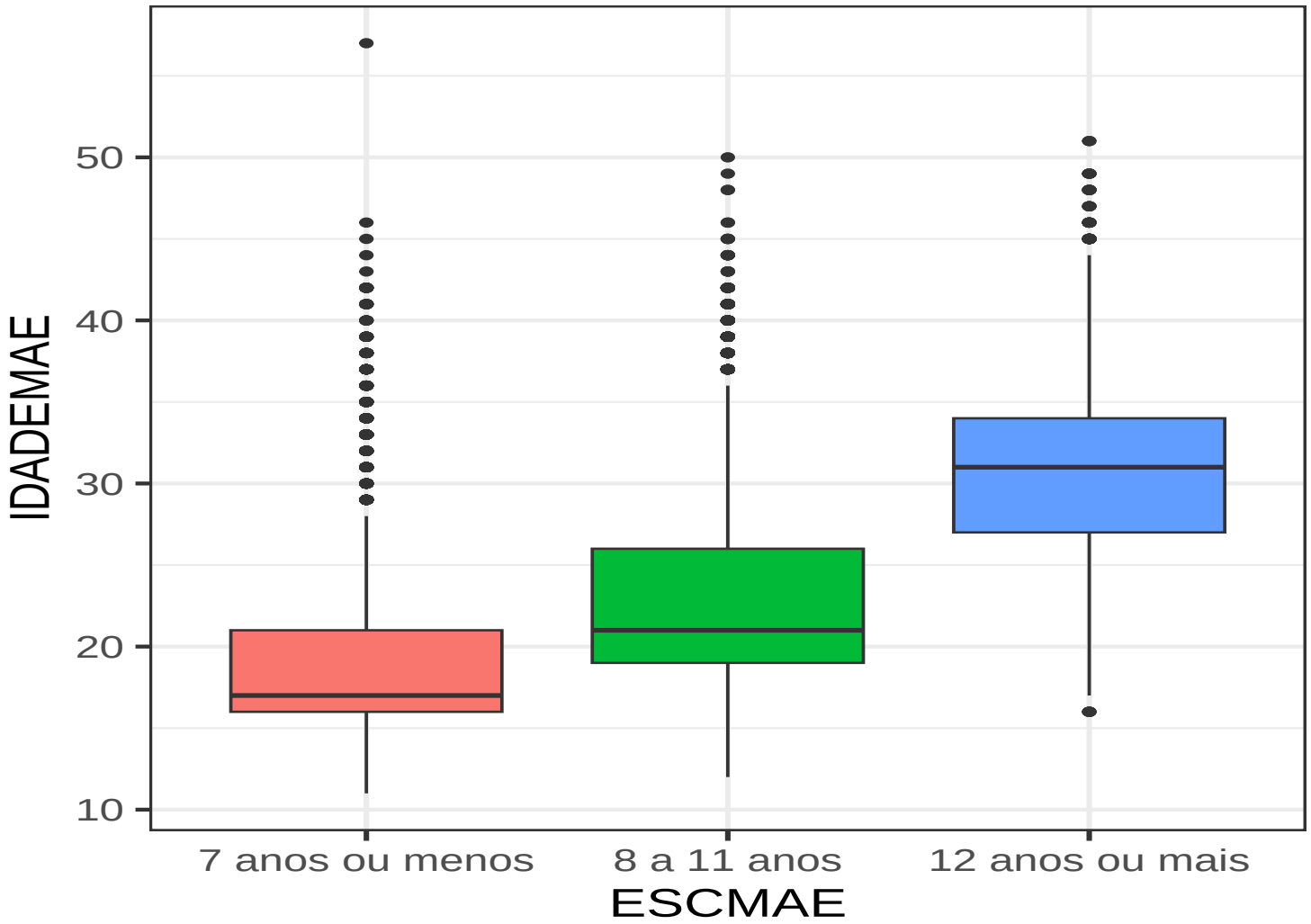
FIOCRUZ

Escolaridade versus Idade gestacional



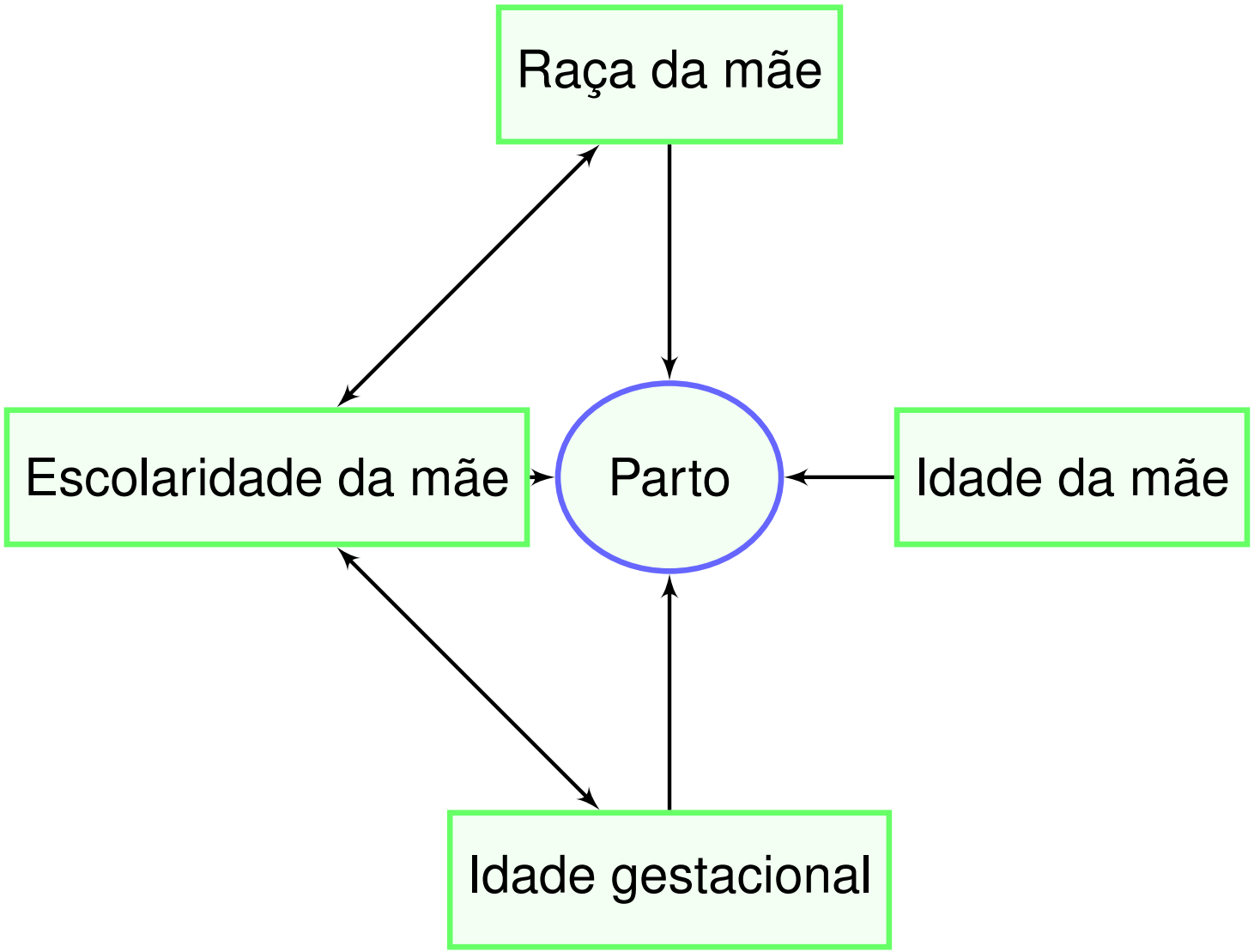
FIOCRUZ

Escolaridade versus Idade



FIOCRUZ

DAG do exemplo



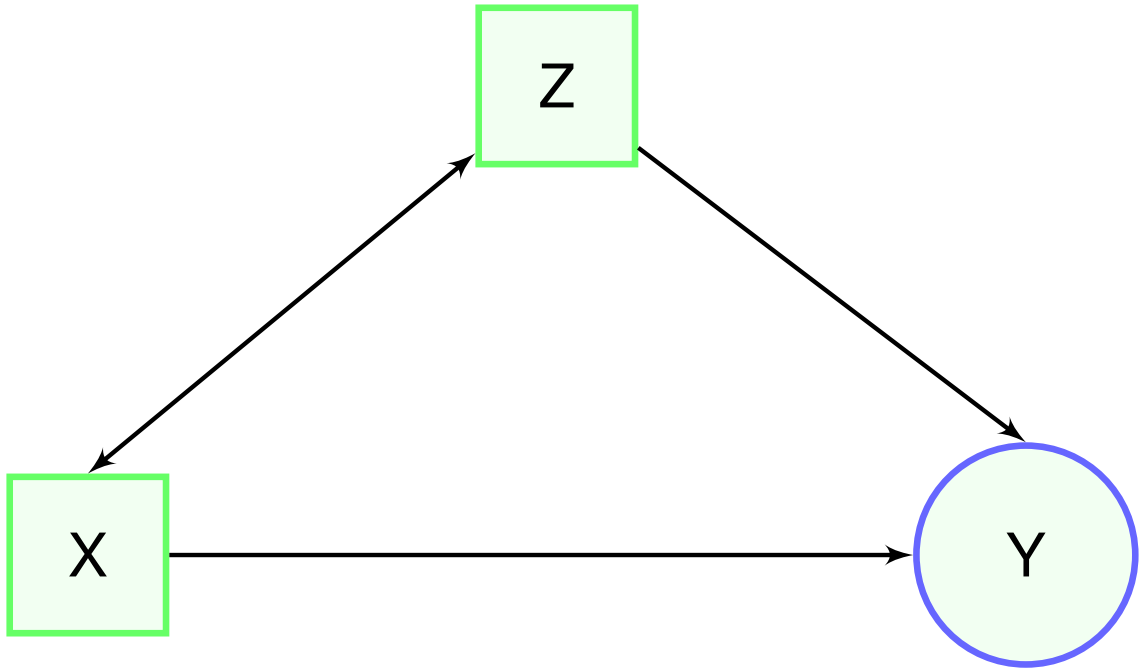
FIOCRUZ

Confundimento

- Seja Y um possível desfecho
- Seja X a exposição de interesse
- Seja Z uma outra variável explicativa (ou um conjunto delas)
- Dizemos que Z é uma variável **confundidora** se ela induz, elimina, reduz ou reforça a associação entre a exposição e desfecho.

FIOCRUZ

Confundimento



FIOCRUZ

Exemplo: Szklo & Javier Neto (Cap. 5)

- Seja um estudo caso-controle hipotético com 300 participantes, 150 com malária e 150 controles.

Sex	Cases	Controls
Females	62	82
Males	88	68

  $p.value = 0.02811$ (Que teste foi esse?)

- OR = 1.71 (O que isso significa?)

FIOCRUZ

Exemplo: Szklo & Javier Neto (Cap. 5)

- Se ajustarmos um modelo de regressão logística temos que:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2796	0.1683	-1.66	0.0967
SexMales	0.5374	0.2332	2.30	0.0212

OR=1.71 com IC 95% (1.09-2.71)

- O exemplo também tem uma terceira variável, que está relacionada com o tipo de ocupação se *indoor* ou *outdoor*.

FIOCRUZ

Exemplo: Szklo & Javier Neto (Cap. 5)

- Essa variável está associada com o desfecho:

Ocupacao	Cases	Controls
Outdoor	63	18
Indoor	87	132

com OR=5.31 com IC 95% (3-9.8)

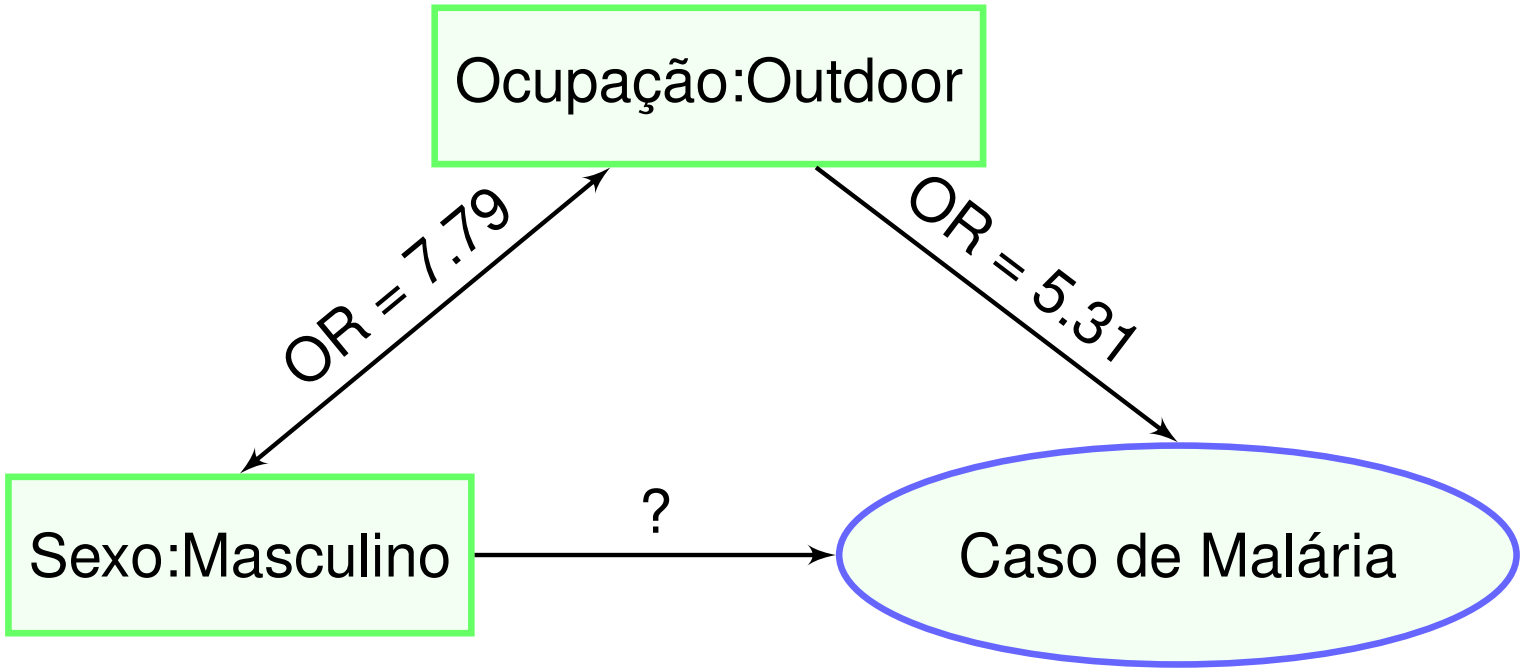
- E também está associada com a exposição:

Sex	Outdoor	Indoor
Males	68	88
Females	13	131

Com OR=7.79 com IC 95% (4.18-15.53)

FIOCRUZ

Confundimento



FIOCRUZ

Controlando pela ocupação

- Nesse exemplo simples, podemos fazer a extratificação e avaliar a OR do sexo para os dois tipos de ocupação.
- Ocupação outdoor:

Sex	Cases	Controls
Males	53	15
Females	10	3

OR=1.06 com IC 95% (0.22-4.01)

- Ocupação indoor:

Sex	Cases	Controls
Males	35	53
Females	52	79

OR=1 com IC 95% (0.58-1.74)

FIOCRUZ

Via modelo

- O modelo de regressão logística com sexo apenas:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2796	0.1683	-1.66	0.0967
SexMales	0.5374	0.2332	2.30	0.0212

- Controlando por ocupação:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.4211	0.1739	-2.42	0.0154
SexMales	0.0105	0.2626	0.04	0.9681
OcupacaoOutdoor	1.6651	0.3219	5.17	0.0000

- Podemos perceber a mudança de efeito da variável sexo ao incluirmos a ocupação.

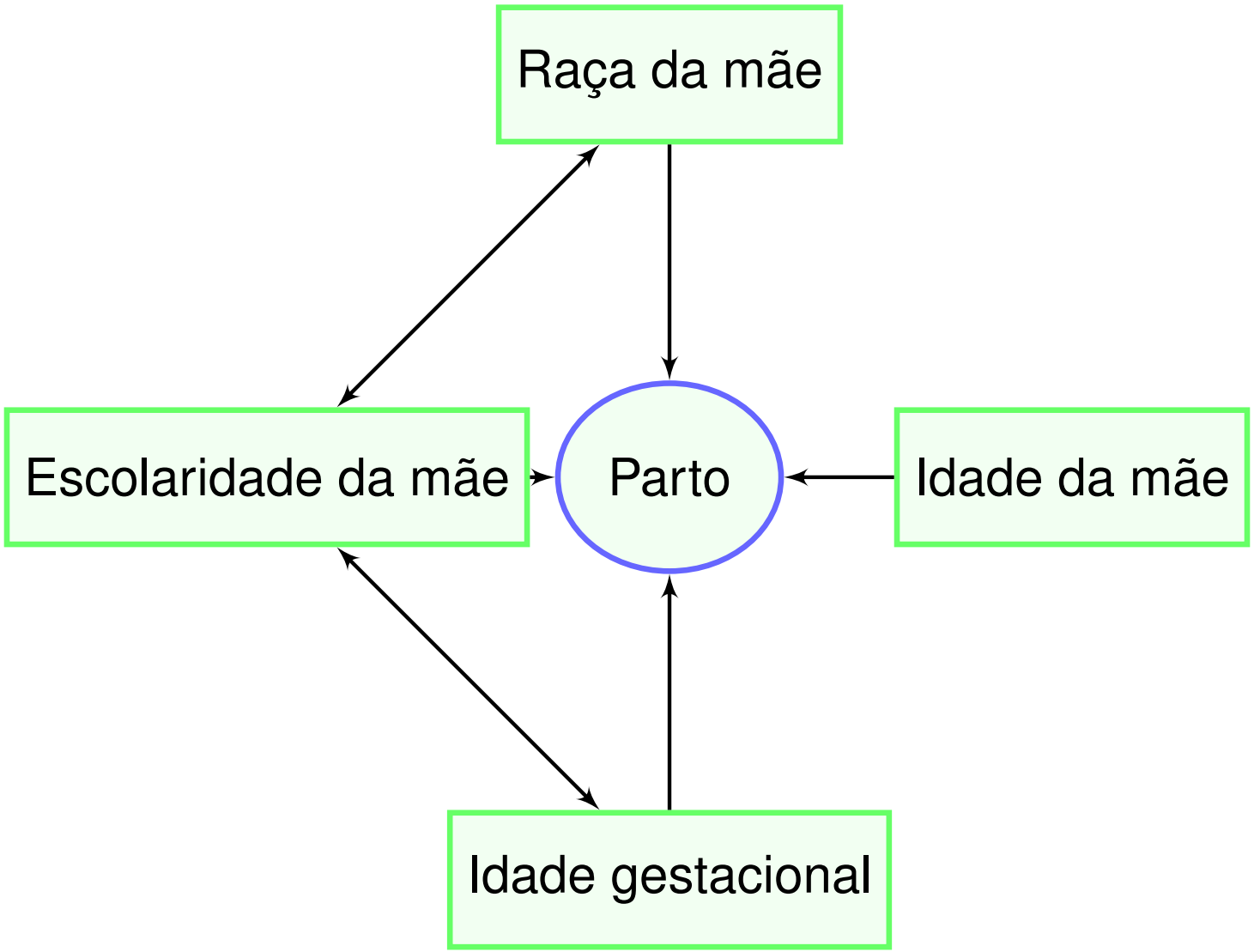
FIOCRUZ

Confundimento sob o modelo múltiplo

- Vamos continuar a pensar em confundimento em um contexto que continuamos interessados na associação entre desfecho Y e exposição X
- Suponha agora que temos outras variáveis que já sabemos que afetam o defecho.
- Podemos usar a mesma lógica anterior de avaliar o efeito com e sem a(s) suposta(s) variável(is) confundidoras.

FIOCRUZ

DAG do exemplo



FIOCRUZ

Olhando para o efeito da escolaridade

- ORs “brutas” (Parto versus Escolaridade da mãe)

	OR	2.5 %	97.5 %
ESCMAE8 a 11 anos	1.76	1.64	1.88
ESCMAE12 anos ou mais	7.26	6.73	7.83

- OR da escolaridade da mãe controlada por idade gestacional, idade e raça/cor da mãe

	OR	2.5 %	97.5 %
ESCMAE8 a 11 anos	1.24	1.15	1.33
ESCMAE12 anos ou mais	2.30	2.10	2.52

- O que está acontecendo?

FIOCRUZ

Confundimento: Regra geral

- A variável confundidora está associada de forma causal com o desfecho
- A variável confundidora está associada com a exposição, ou ela causa a exposição ou ela está associada a algo que causa a exposição.
- A variável confundidora NÃO está no caminho causal entre a exposição e o desfecho.

FIOCRUZ