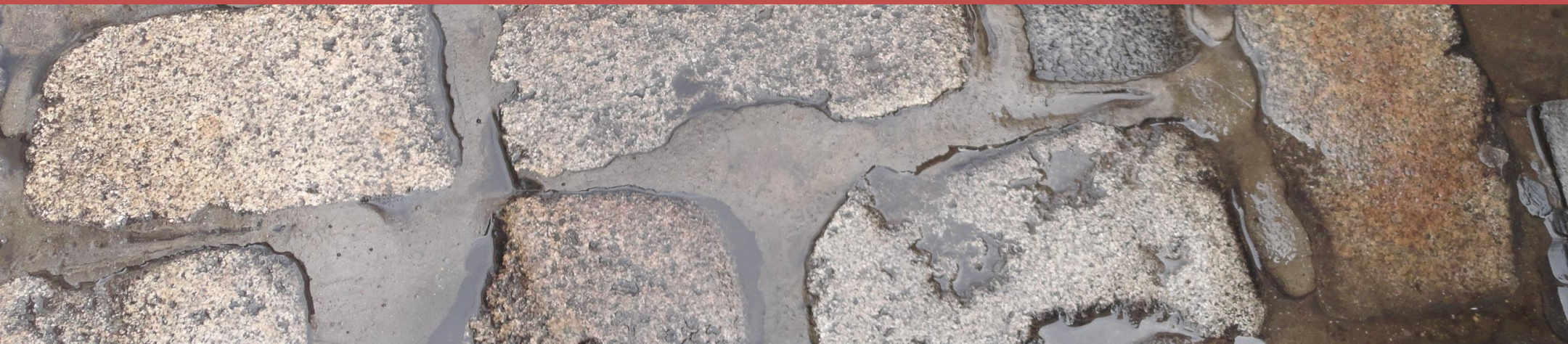
MODELOS DE REGRESSÃO LOGÍSTICA
ANÁLISIS DE REGRESIÓN LOGÍSTICA
LOGISTIC REGRESSION MODELS

2





MODELOS DE REGRESSÃO LOGÍSTICA
ANÁLISIS DE REGRESIÓN LOGÍSTICA
LOGISTIC REGRESSION MODELS

Prof. Edson Zangiacomi Martinez

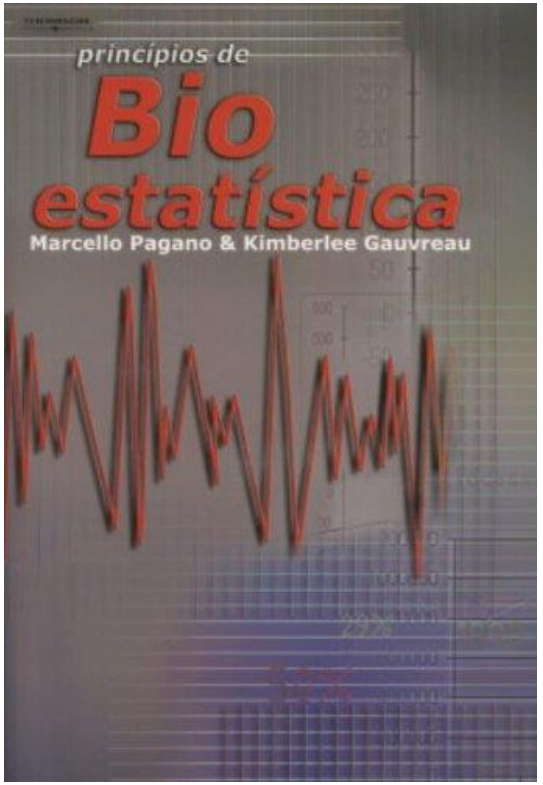
Faculdade de Medicina de Ribeirão Preto
Universidade de São Paulo (USP)
Ribeirão Preto, Brasil
edson@fmrp.usp.br

MODELOS DE REGRESSÃO LOGÍSTICA
ANÁLISIS DE REGRESIÓN LOGÍSTICA
LOGISTIC REGRESSION MODELS

REFERÊNCIA
REFERENCIA
REFERENCE

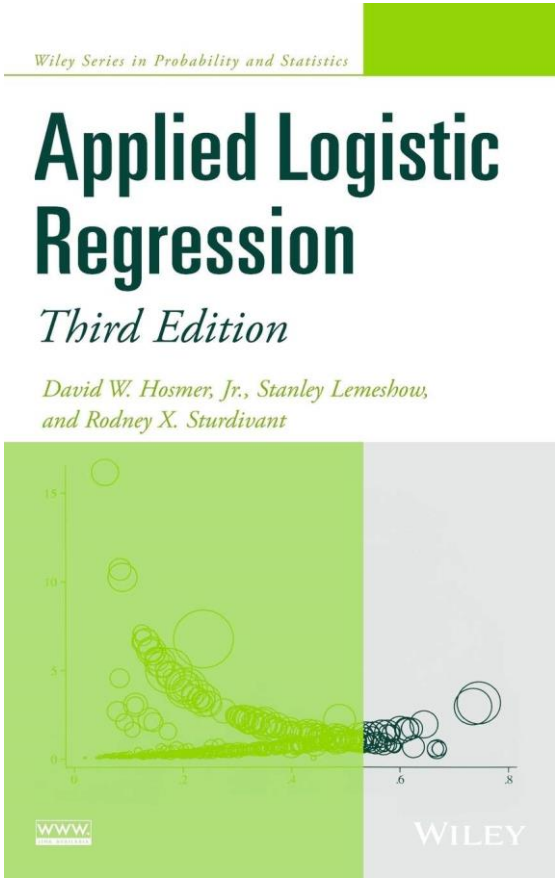
Capítulo 20 – “Regressão Logística” do livro:

Pagano M, Gauvreau K
Princípios de Bioestatística
São Paulo: Pioneira Thomson Learning, 2004.



REFERÊNCIA
REFERENCIA
REFERENCE

Jr. Hosmer DW, Lemeshow S, Sturdivant RX
Applied Logistic Regression
Third Edition
John Wiley & Sons, 2013.





MODELOS DE REGRESSÃO LOGÍSTICA
ANÁLISIS DE REGRESIÓN LOGÍSTICA
LOGISTIC REGRESSION MODELS

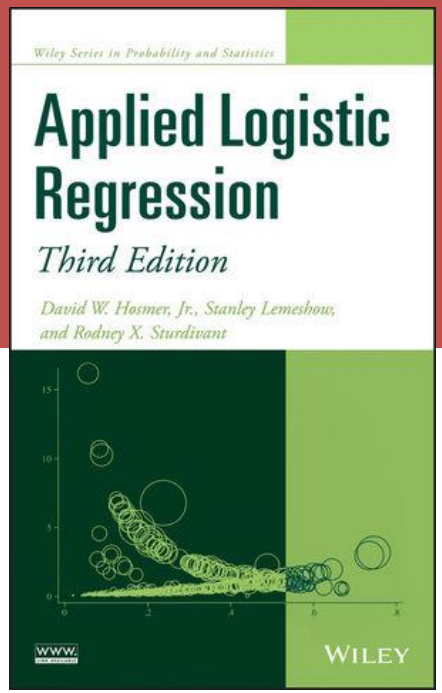
Regressão logística simples

Relaciona uma variável independente a uma variável dependente qualitativa.

Regressão logística múltipla

Relaciona, simultaneamente, mais de uma variável independente a uma variável dependente qualitativa.

EXEMPLO – HOSMER ET AL. (2013)
EJEMPLO – HOSMER ET AL. (2013)
EXAMPLE – HOSMER ET AL. (2013)



NAME: LOW BIRTH WEIGHT DATA (LOWBWT.DAT)

KEYWORDS: Logistic Regression

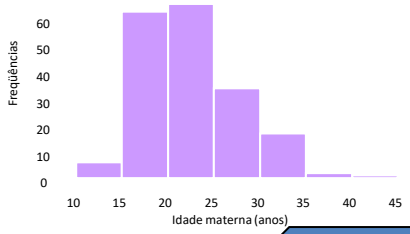
SIZE: 189 observations, 11 variables

SOURCE: Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013)

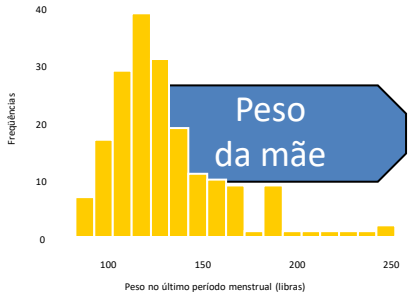
Applied Logistic Regression: Third Edition.

These data are copyrighted by John Wiley & Sons Inc. and must be acknowledged and used accordingly. Data were collected at Baystate Medical Center, Springfield, Massachusetts during 1986.

VARIÁVEIS INDEPENDENTES
VARIABLES INDEPENDIENTES
INDEPENDENT VARIABLES



Idade da mãe



Peso da mãe

Tabagismo

sim	74	(39%)
não	115	(61%)

branca 96 (51%)
negra 26 (14%)
outra 67 (35%)

Cor da pele

0	100	(53%)
1	47	(25%)
2	30	(16%)
3	7	(4%)
4	4	(2%)
6	1	(<1%)

Visitas ao médico

Baixo peso ao nascer

Irritabilidade uterina

sim	28	(15%)
não	161	(85%)

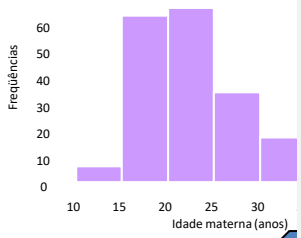
História de hipertensão

sim	12	(6%)
não	177	(94%)

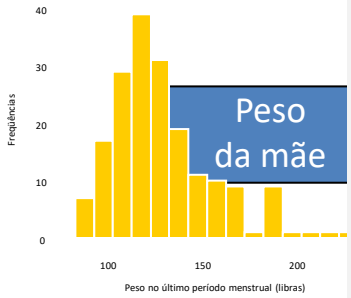
Partos prematuros

0	159	(84%)
1	24	(13%)
2	5	(3%)
3	1	(<1%)

VARIÁVEIS INDEPENDENTES
VARIABLES INDEPENDIENTES
INDEPENDENT VARIABLES



Visitas ao médico



Peso da mãe

Irritabilidade uterina

Partos prematuros

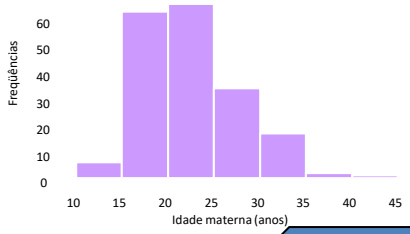
```
> table(w$PTL) # partos prematuros

  0   1   2   3
159  24   5   1
> table(w$FTV) # visitas ao médico

  0   1   2   3   4   6
100  47  30   7   4   1
> w$PTL.c <- w$PTL
> w$PTL.c[w$PTL>1] <- 1
> w$FTV.c <- w$FTV
> w$FTV.c[w$FTV>2] <- 2
> table(w$PTL.c)

  0   1
159  30
> table(w$FTV.c)

  0   1   2
100  47  42
```

0	100	(53%)
1	47	(25%)
2 ou +	42	(22%)

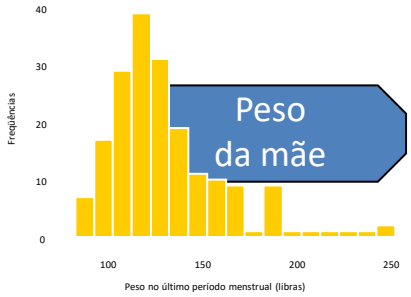
sim	28	(15%)
não	161	(85%)

não	159	(84%)
sim	30	(16%)

VARIÁVEIS INDEPENDENTES
VARIABLES INDEPENDIENTES
INDEPENDENT VARIABLES



Idade da mãe



Peso da mãe

Tabagismo

sim 74 (39%)
não 115 (61%)

branca 96 (51%)
negra 26 (14%)
outra 67 (35%)

Cor da pele

Baixo peso ao nascer

História de hipertensão

sim 12 (6%)
não 177 (94%)

0	100	(53%)
1	47	(25%)
2 ou +	42	(22%)

Visitas ao médico

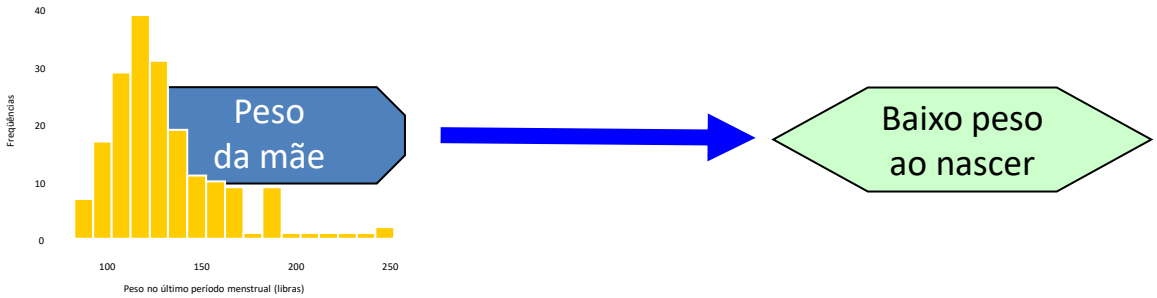
Irritabilidade uterina

sim 28 (15%)
não 161 (85%)

Partos prematuros

não	159	(84%)
sim	30	(16%)

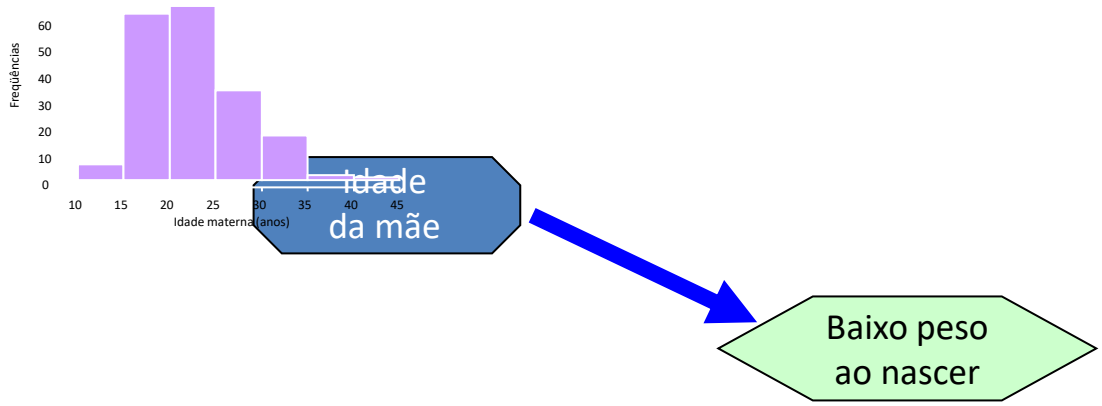
VARIÁVEIS INDEPENDENTES
VARIABLES INDEPENDIENTES
INDEPENDENT VARIABLES



Peso da mãe	*n*	%
Até 50 kg	53	22,2
50 -| 55	43	17,5
55 -| 65	49	33,9
Mais de 65 kg	44	26,5

```
> w$LWTKG.c <- cut(w$LWTKG,breaks=c(30,50,55,65,120))
> table(w$LWTKG.c)

 (30,50]  (50,55]  (55,65] (65,120]
      53       43       49       44
> round(100*prop.table(table(w$LWTKG.c)),1)

 (30,50]  (50,55]  (55,65] (65,120]
    28.0     22.8     25.9     23.3
```

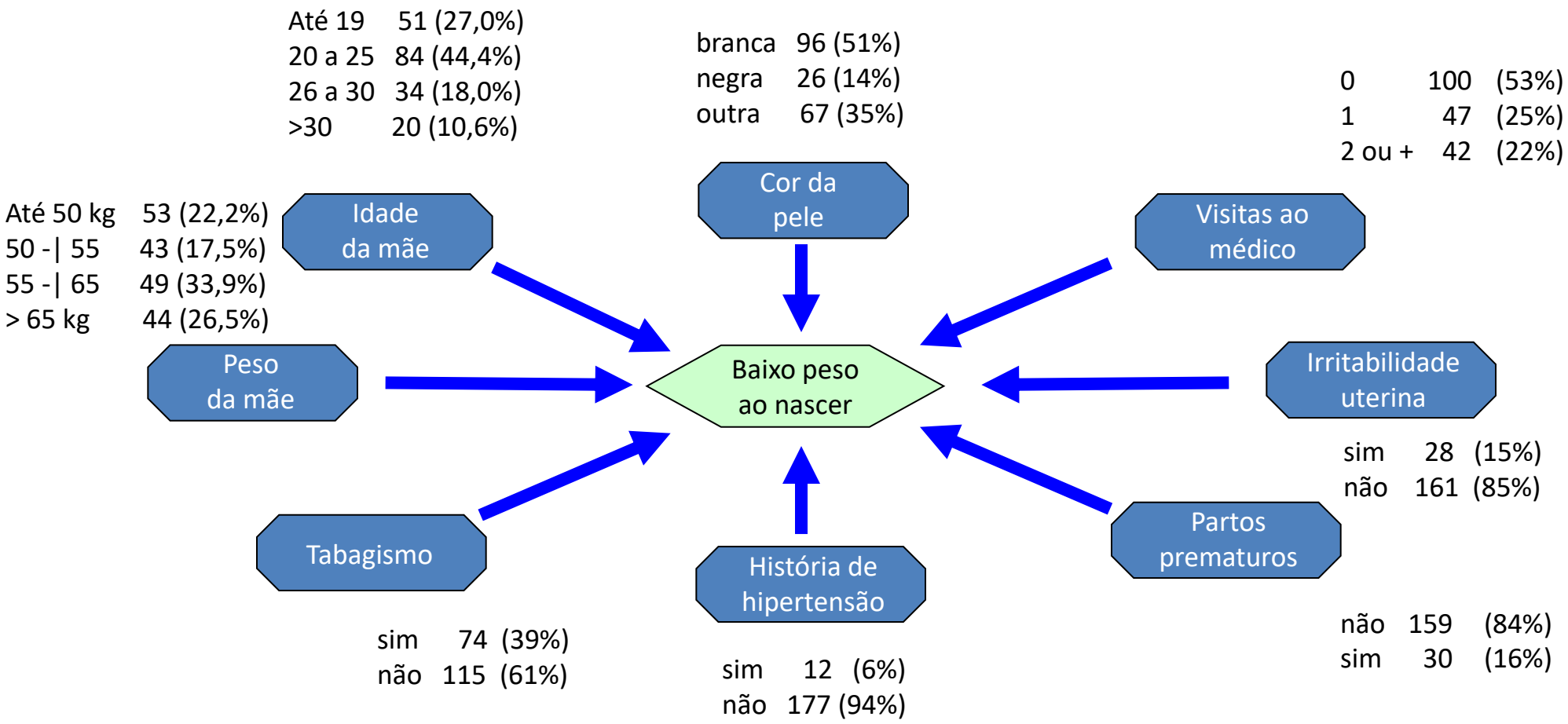
VARIÁVEIS INDEPENDENTES
VARIABLES INDEPENDIENTES
INDEPENDENT VARIABLES



Idade da mãe	*n*	%
Até 19 anos	51	27,0
20 a 25 anos	84	44,4
26 a 30 anos	34	18,0
Mais de 30 anos	20	10,6

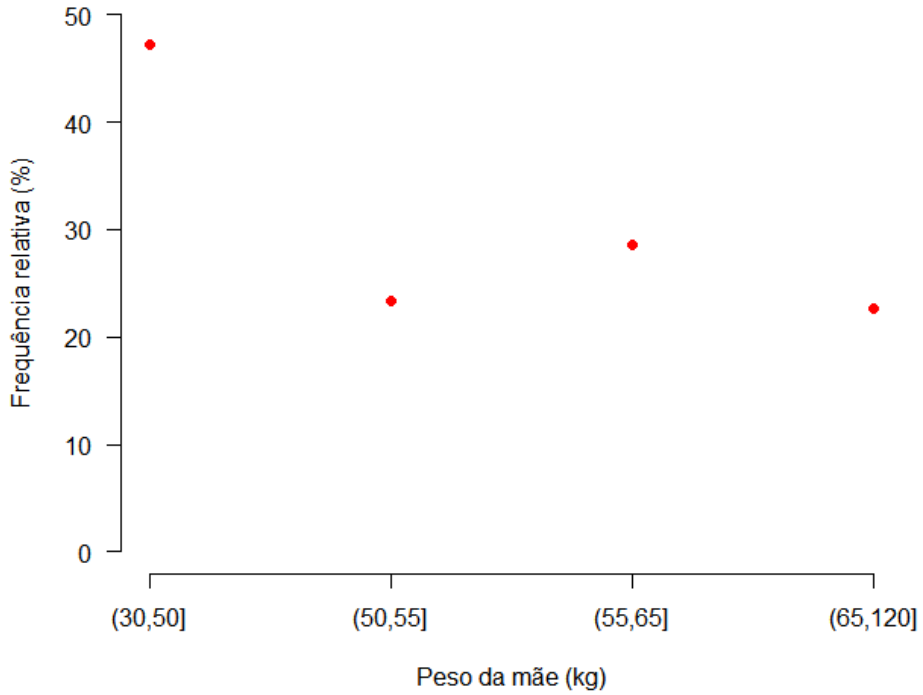
```
> AGE.c   <- cut(w$AGE,  breaks=c(10,19,25,30,50))
> table(w$AGE.c)

(10,19] (19,25] (25,30] (30,50]
     51      84      34      20
> round(100*prop.table(table(w$AGE.c)),1)

(10,19] (19,25] (25,30] (30,50]
   27.0    44.4    18.0    10.6
```

VARIÁVEIS INDEPENDENTES
VARIABLES INDEPENDIENTES
INDEPENDENT VARIABLES



PESO NO ÚLTIMO PERÍODO MENSTRUAL
PESO EN EL ÚLTIMO PERÍODO MENSTRUAL
WEIGHT AT THE LAST MENSTRUAL PERIOD

Peso da mãe	*n*	%	*Baixo peso ao nascer n (%)*
Até 50 kg	53	22,2	25 (47,2%)
50 -| 55	43	17,5	10 (23,3%)
55 -| 65	49	33,9	14 (28,6%)
Mais de 65 kg	44	26,5	10 (22,7%)



PESO NO ÚLTIMO PERÍODO MENSTRUAL
PESO EN EL ÚLTIMO PERÍODO MENSTRUAL
WEIGHT AT THE LAST MENSTRUAL PERIOD

Peso da mãe	*Total*	*Baixo peso ao nascer n (%)*	*OR* (IC 95%)
Até 50 kg	53	25 (47,2%)	
50 -| 55	43	10 (23,3%)	
55 -| 65	49	14 (28,6%)	
Mais de 65 kg	44	10 (22,7%)	Referência

Variáveis indicadoras:

	X_1	X_2	X_3
Mais de 65	0	0	0
55 -| 65	1	0	0
50 -| 55	0	1	0
Até 50	0	0	1

PESO NO ÚLTIMO PERÍODO MENSTRUAL
PESO EN EL ÚLTIMO PERÍODO MENSTRUAL
WEIGHT AT THE LAST MENSTRUAL PERIOD

Peso da mãe	*Total*	*Baixo peso ao nascer n (%)*	*OR* (IC 95%)
Até 50 kg	53	25 (47,2%)	
50 -| 55	43	10 (23,3%)	
55 -| 65	49	14 (28,6%)	
Mais de 65 kg	44	10 (22,7%)	Referência

$$P(Y = 1|X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}$$

PESO NO ÚLTIMO PERÍODO MENSTRUAL
PESO EN EL ÚLTIMO PERÍODO MENSTRUAL
WEIGHT AT THE LAST MENSTRUAL PERIOD

```
> # Regressão logística simples
> # Peso da mãe (categorizado) como variável independente
> model <- glm(y ~ relevel(LWTKG.c, ref = "(65,120]"),family=binomial(link='logit'),data=w)
> summary(model)

Call:
glm(formula = y ~ relevel(LWTKG.c, ref = "(65,120]"), family = binomial(link = "logit"),
    data = w)

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -1.22378    0.35974  -3.402 0.000669 ***
relevel(LWTKG.c, ref = "(65,120]")(30,50]  1.11045    0.45291   2.452 0.014214 *
relevel(LWTKG.c, ref = "(65,120]")(50,55]  0.02985    0.50962   0.059 0.953288
relevel(LWTKG.c, ref = "(65,120]")(55,65]  0.30748    0.47897   0.642 0.520891
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 225.74  on 185  degrees of freedom
AIC: 233.74

Number of Fisher Scoring iterations: 4
```



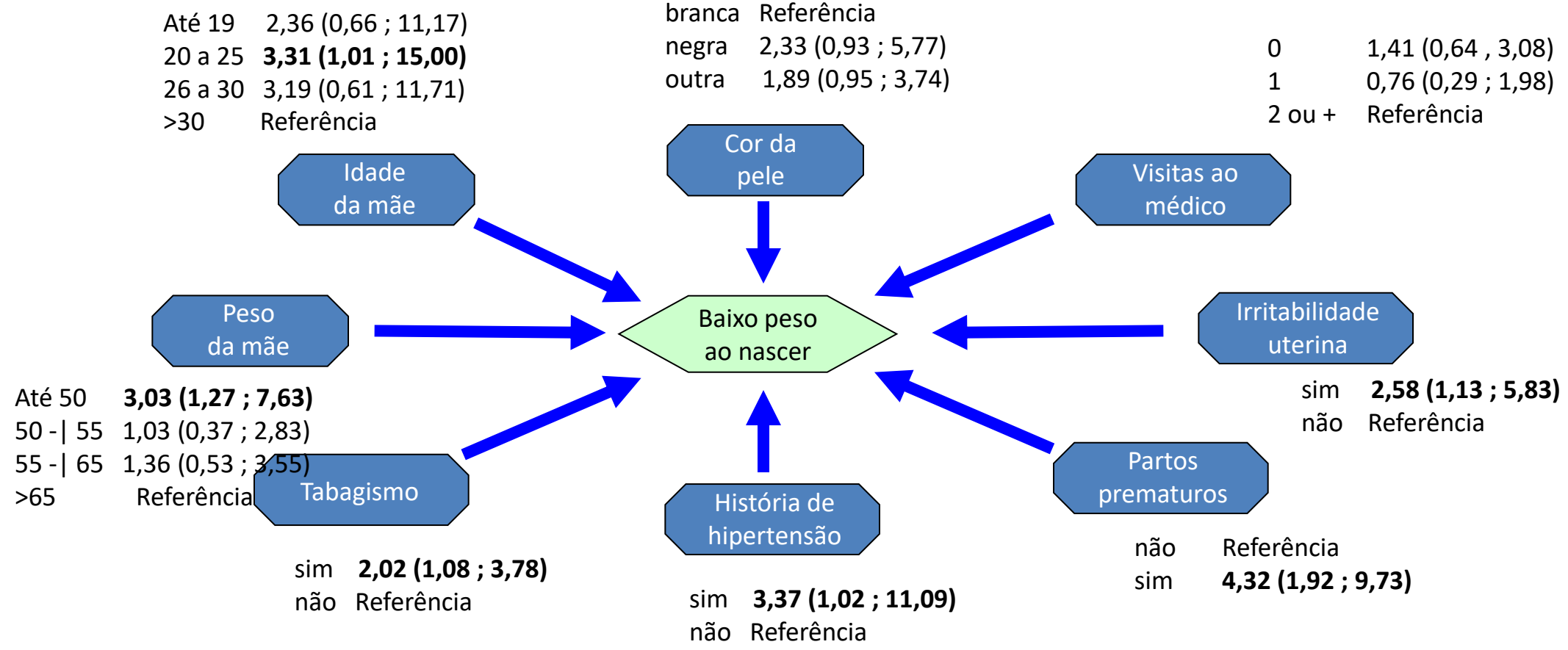
PESO NO ÚLTIMO PERÍODO MENSTRUAL
PESO EN EL ÚLTIMO PERÍODO MENSTRUAL
WEIGHT AT THE LAST MENSTRUAL PERIOD

```
> exp(cbind(OR = coef(model), confint(model)))
Waiting for profiling to be done...
                                                      OR     2.5 %    97.5 %
(Intercept)                                    0.2941176 0.1376950 0.5732701
relevel(LWTKG.c, ref = "(65,120]")(30,50] 3.0357143 1.2773243 7.6308643
relevel(LWTKG.c, ref = "(65,120]")(50,55] 1.0303030 0.3757700 2.8256358
relevel(LWTKG.c, ref = "(65,120]")(55,65] 1.3600000 0.5351722 3.5536114
```

PESO NO ÚLTIMO PERÍODO MENSTRUAL
PESO EN EL ÚLTIMO PERÍODO MENSTRUAL
WEIGHT AT THE LAST MENSTRUAL PERIOD

Peso da mãe	*Total*	*Baixo peso ao nascer n (%)*	*OR* (IC 95%)
Até 50 kg	53	25 (47,2%)	3,03 (1,27; 7,63)
50 -| 55	43	10 (23,3%)	1,03 (0,37; 2,83)
55 -| 65	49	14 (28,6%)	1,36 (0,53; 3,55)
Mais de 65 kg	44	10 (22,7%)	Referência

ODDS RATIO BRUTO
RAZÓN DE MOMIOS CRUDA
CRUDE ODDS RATIO



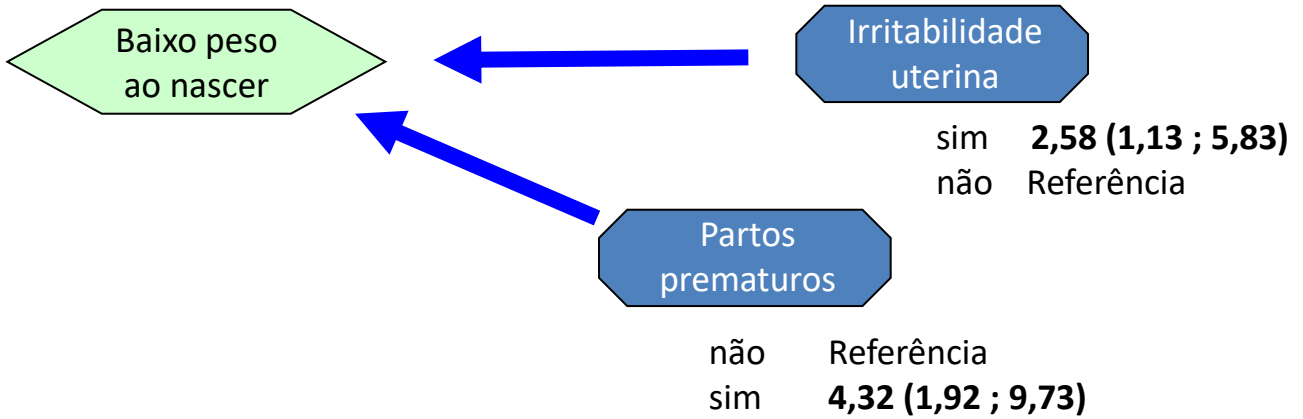
REGRESSÃO LOGÍSTICA MÚLTIPLA
MODELO DE REGRESIÓN LOGÍSTICA MÚLTIPLE
MULTIPLE LOGISTIC REGRESSION MODEL

Sejam k variáveis independentes.

$$P(Y = 1|X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \theta(x_1, x_2, \dots, x_k)$$

$$= \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

ODDS RATIO BRUTO
RAZÓN DE MOMIOS CRUDA
CRUDE ODDS RATIO



REGRESSÃO LOGÍSTICA MÚLTIPLA
MODELO DE REGRESIÓN LOGÍSTICA MÚLTIPLE
MULTIPLE LOGISTIC REGRESSION MODEL

Sejam $k = 2$ variáveis independentes.

$$P(Y = 1|X_1 = x_1, X_2 = x_2) = \theta(x_1, x_2) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

X_1: Irritabilidade uterina

Não: $x_1 = 0$
Sim: $x_1 = 1$

X_2: Partos prematuros

Não: $x_2 = 0$
Sim: $x_2 = 1$

REGRESSÃO LOGÍSTICA MÚLTIPLA
MODELO DE REGRESIÓN LOGÍSTICA MÚLTIPLE
MULTIPLE LOGISTIC REGRESSION MODEL

```
> # Regressão logística múltipla
> # Irritabilidde uterina e história de partos prematuros como variáveis independentes
> model <- glm(y ~ UI + PTL.c, family=binomial(link='logit'), data=w)
> summary(model)

Call:
glm(formula = y ~ UI + PTL.c, family = binomial(link = "logit"), data = w)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1580     0.1942  -5.962  2.5e-09 ***
UI            0.7383     0.4382   1.685  0.09206 .
PTL.c         1.3558     0.4219   3.214  0.00131 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 219.12  on 186  degrees of freedom
AIC: 225.12

Number of Fisher Scoring iterations: 4

> exp(cbind(OR = coef(model), confint(model)))
Waiting for profiling to be done...
                  OR     2.5 %    97.5 %
(Intercept) 0.3141198 0.2117861 0.4545757
UI          2.0923395 0.8757125 4.9486837
PTL.c       3.8798743 1.7097242 9.0514117
```

ODDS RATIO BRUTO E AJUSTADO
RAZÓN DE MOMIOS CRUDA Y AJUSTADA
CRUDE AND ADJUSTED ODDS RATIO

	Total	*Baixo peso ao nascer n (%)*	*OR* bruto (IC 95%)	*OR* ajustado (IC 95%)
Irritabilidade uterina				
não	161	45 (28,0%)	Referência	Referência
sim	28	14 (50,0%)	**2,58** (1,13 ; 5,83)	2,09 (0,87 ; 4,95)
Partos prematuros				
não	159	41 (25,8%)	Referência	Referência
sim	30	18 (60,0%)	**4,32** (1,92 ; 9,73)	**3,88** (1,70 ; 9,05)

HIPÓTESE
HIPÓTESIS
HYPOTHESIS

Am J Obstet Gynecol. 1995 Jan;172(1 Pt 1):138-42.

The irritable uterus: a risk factor for preterm birth?

Roberts WE[1], Perry KG Jr, Naef RW 3rd, Washburne JF, Morrison JC.

Author information

Abstract

OBJECTIVE: Our aim was to determine the incidence and preterm delivery rate along with the indication for delivery in patients with uterine irritability.

STUDY POPULATION: In this retrospective, descriptive study, 17,186 patients with well-defined high-risk factors were compared with 2637 women with uterine irritability.

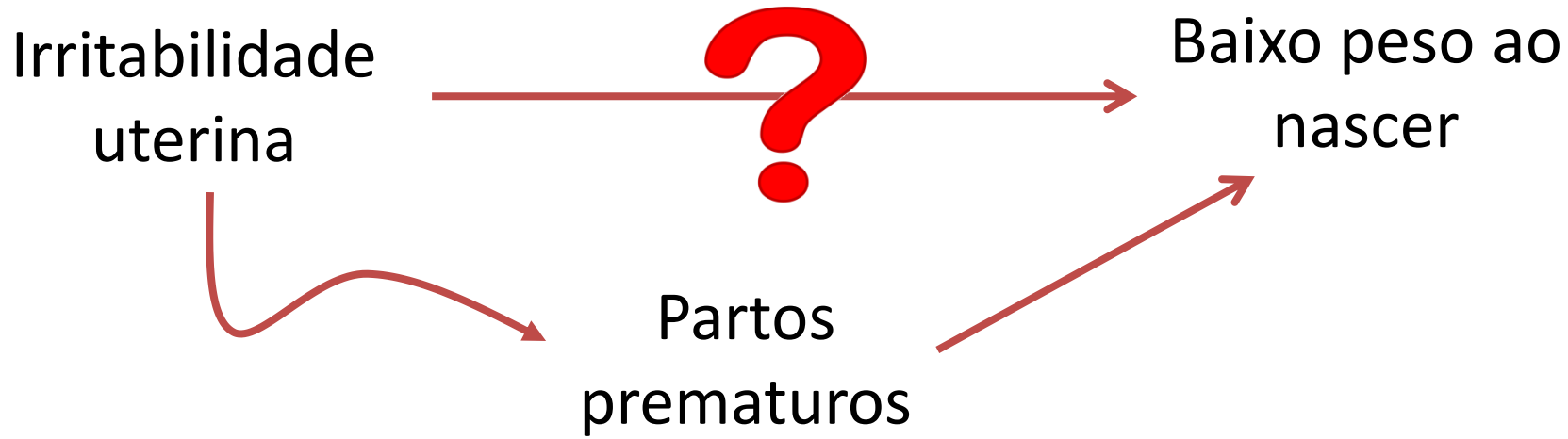
RESULTS: The incidence of preterm labor in patients with uterine irritability was 18.7%, significantly less than in those with other high-risk factors (odds ratio 0.35, 0.31 < odds ratio < 0.38). However, women with uterine irritability who experience preterm labor, compared with other high-risk factors, are much more likely to deliver before 34 weeks' gestation (odds ratio 2.50, 2.07 < odds ratio < 3.03) and more than twice as likely to deliver as a result of advanced preterm labor or membrane rupture (odds ratio 2.20, 1.75 < odds ratio < 2.78).

CONCLUSIONS: The incidence of preterm labor in women with uterine irritability is not as frequent as in patients with other high-risk factors. However, preterm labor does occur in patients with uterine irritability at a rate higher than that in the general obstetric population (18.7% vs 11.0%). Because it appears that women with uterine irritability have more resistance to conventional tocolytic therapy, this condition should prompt the physician to use more aggressive perinatal assessment.

HIPÓTESE
HIPÓTESIS
HYPOTHESIS

		Total	Partos prematuros *n* (%)
Irritabilidade uterina	sim	28	9 (32,1%)
	não	161	21 (13,0%)

HIPÓTESE
HIPÓTESIS
HYPOTHESIS





MODELO COM TODAS AS VARIÁVEIS
MODELO CON TODAS LAS VARIABLES
MODEL WITH ALL VARIABLES

```
> # Modelo com todas as variáveis
> model <- glm(y ~ relevel(AGE.c, ref = "(30,50]") + relevel(LWTKG.c, ref = "(65,120]")
+       + factor(RACE) + SMOKE + UI + PTL.c + factor(FTV.c) +
HT,family=binomial(link='logit'),data=w)
> summary(model)

Call:
glm(formula = y ~ relevel(AGE.c, ref = "(30,50]") + relevel(LWTKG.c,
    ref = "(65,120]") + factor(RACE) + SMOKE + UI + PTL.c + factor(FTV.c) +
    HT, family = binomial(link = "logit"), data = w)

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                -3.3660     0.9472  -3.554  0.00038 ***
relevel(AGE.c, ref = "(30,50]")(10,19]      0.7226     0.8065   0.896  0.37026
relevel(AGE.c, ref = "(30,50]")(19,25]      1.0362     0.7569   1.369  0.17099
relevel(AGE.c, ref = "(30,50]")(25,30]      0.9234     0.8328   1.109  0.26750
relevel(LWTKG.c, ref = "(65,120]")(30,50]   1.1190     0.5568   2.010  0.04446 *
relevel(LWTKG.c, ref = "(65,120]")(50,55]   0.1560     0.5845   0.267  0.78949
relevel(LWTKG.c, ref = "(65,120]")(55,65]   0.4962     0.5533   0.897  0.36983
...
```

MODELO COM TODAS AS VARIÁVEIS
MODELO CON TODAS LAS VARIABLES
MODEL WITH ALL VARIABLES

```
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               -3.3660     0.9472  -3.554  0.00038 ***
relevel(AGE.c, ref = "(30,50]")(10,19]     0.7226     0.8065   0.896  0.37026
relevel(AGE.c, ref = "(30,50]")(19,25]     1.0362     0.7569   1.369  0.17099
relevel(AGE.c, ref = "(30,50]")(25,30]     0.9234     0.8328   1.109  0.26750
relevel(LWTKG.c, ref = "(65,120]")(30,50]  1.1190     0.5568   2.010  0.04446 *
relevel(LWTKG.c, ref = "(65,120]")(50,55]  0.1560     0.5845   0.267  0.78949
relevel(LWTKG.c, ref = "(65,120]")(55,65]  0.4962     0.5533   0.897  0.36983
factor(RACE)2                              1.1971     0.5340   2.242  0.02498 *
factor(RACE)3                              0.7467     0.4677   1.597  0.11034
SMOKE                                      0.8086     0.4256   1.900  0.05747 .
UI                                         0.6791     0.4706   1.443  0.14903
PTL.c                                      1.2178     0.4824   2.525  0.01158 *
factor(FTV.c)1                            -0.3882     0.4852  -0.800  0.42369
factor(FTV.c)2                             0.1741     0.4658   0.374  0.70859
HT                                         1.5751     0.6906   2.281  0.02256 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 194.28  on 174  degrees of freedom
AIC: 224.28
```

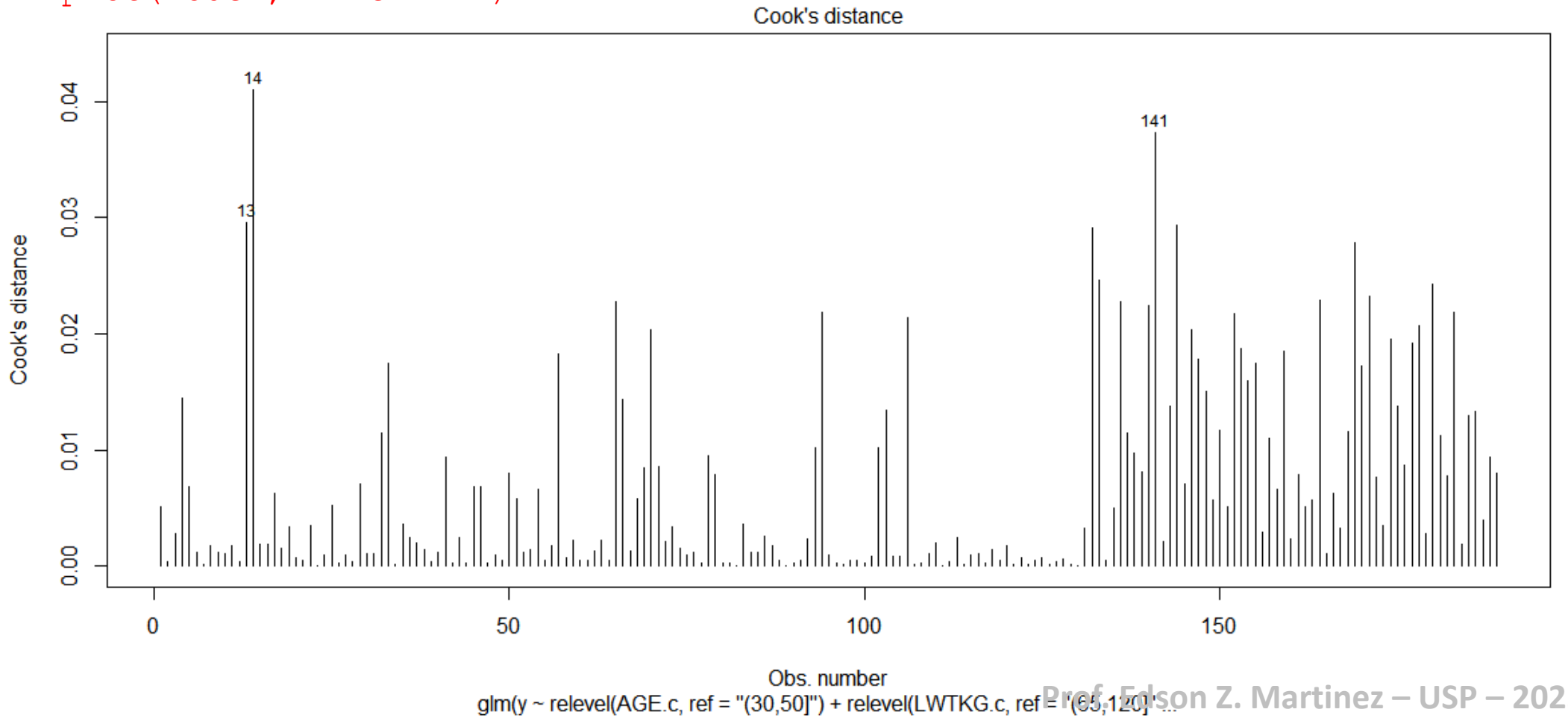


MODELO COM TODAS AS VARIÁVEIS
MODELO CON TODAS LAS VARIABLES
MODEL WITH ALL VARIABLES

```
> round(exp(cbind(OR = coef(model), confint(model))),2)
Waiting for profiling to be done...
                                                    OR 2.5 % 97.5 %
(Intercept)                                       0.03  0.00   0.19
relevel(AGE.c, ref = "(30,50]")(10,19]            2.06  0.46  11.66
relevel(AGE.c, ref = "(30,50]")(19,25]            2.82  0.70  14.76
relevel(AGE.c, ref = "(30,50]")(25,30]            2.52  0.53  14.90
relevel(LWTKG.c, ref = "(65,120]")(30,50] 3.06  1.06   9.53
relevel(LWTKG.c, ref = "(65,120]")(50,55] 1.17  0.37   3.75
relevel(LWTKG.c, ref = "(65,120]")(55,65] 1.64  0.57   5.04
factor(RACE)2                                     3.31  1.16   9.58
factor(RACE)3                                     2.11  0.85   5.38
SMOKE                                             2.24  0.98   5.27
UI                                                1.97  0.78   4.98
PTL.c                                             3.38  1.33   8.96
factor(FTV.c)1                                    0.68  0.25   1.72
factor(FTV.c)2                                    1.19  0.47   2.96
HT                                                4.83  1.27  19.96
```

DISTÂNCIA DE COOK
DISTANCIA DE COOK
COOK'S DISTANCE

```
> plot(model, which = 4)
```



FATOR DE INFLAÇÃO DA VARIÂNCIA GENERALIZADO (GFIV)

FACTOR DE INFLACIÓN DE LA VARIANZA GENERALIZADO (GFIV)

GENERALIZED VARIANCE INFLATION FACTOR (GVIF)

```
> car::vif(model)
                                       GVIF Df GVIF^(1/(2*Df))
relevel(AGE.c, ref = "(30,50]")    1.303540  3        1.045171
relevel(LWTKG.c, ref = "(65,120]") 1.370852  3        1.053979
factor(RACE)                       1.615292  2        1.127360
SMOKE                              1.447308  1        1.203041
UI                                 1.059388  1        1.029266
PTL.c                              1.165686  1        1.079669
factor(FTV.c)                      1.268134  2        1.061186
HT                                 1.135326  1        1.065517
```

Fox, J. and Monette, G. (1992) Generalized collinearity diagnostics. *JASA*, **87**, 178--183.

FATOR DE INFLAÇÃO DA VARIÂNCIA GENERALIZADO (GFIV)
FACTOR DE INFLACIÓN DE LA VARIANZA GENERALIZADO (GFIV)
GENERALIZED VARIANCE INFLATION FACTOR (GVIF)

Adjusted generalized standard error inflation factor (aGSIF)

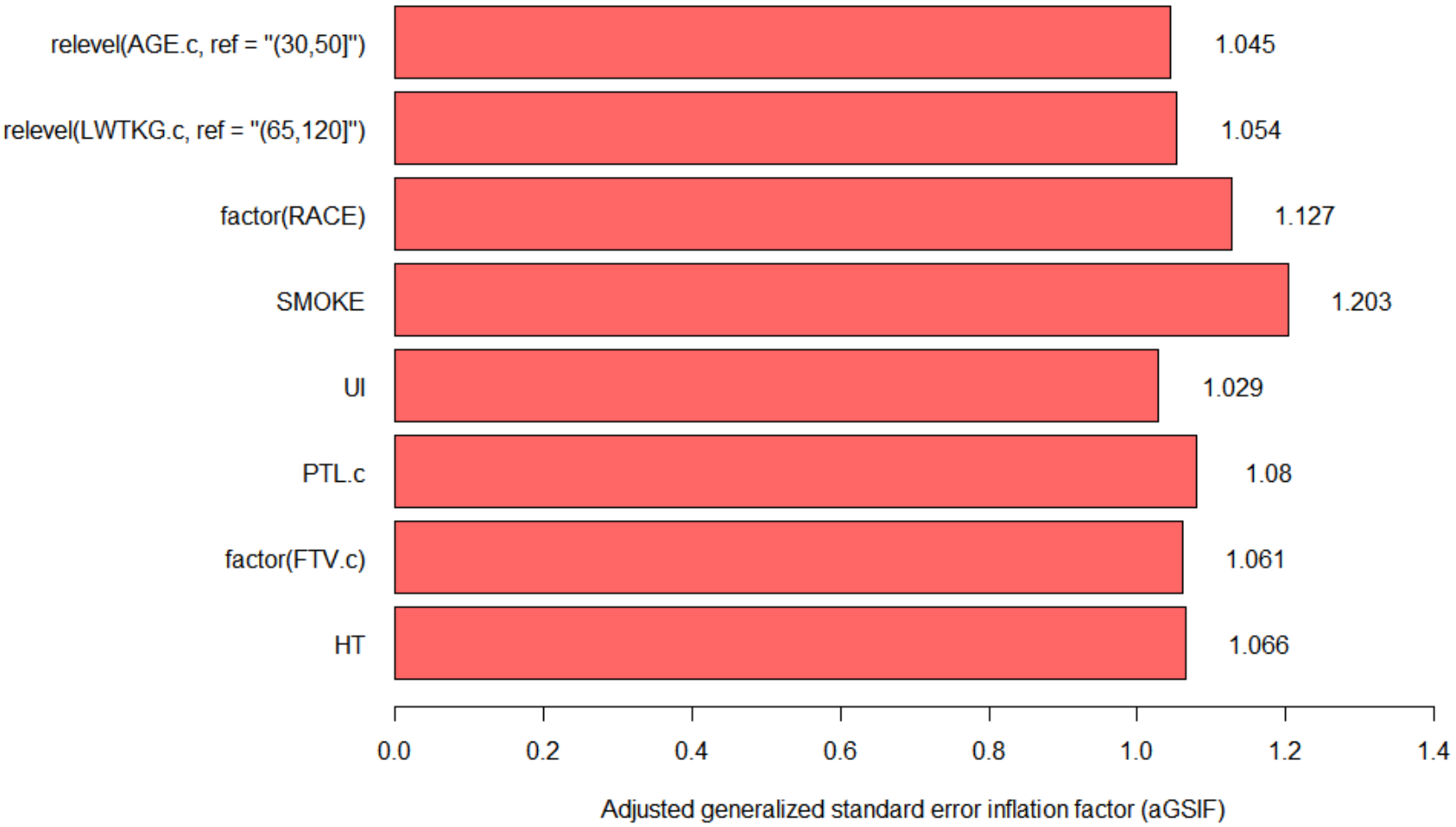
```
> car::vif(model)
                                        GVIF Df GVIF^(1/(2*Df))
relevel(AGE.c, ref = "(30,50]")     1.303540  3        1.045171
relevel(LWTKG.c, ref = "(65,120]") 1.370852  3        1.053979
factor(RACE)                        1.615292  2        1.127360
SMOKE                               1.447308  1        1.203041
UI                                  1.059388  1        1.029266
PTL.c                               1.165686  1        1.079669
factor(FTV.c)                       1.268134  2        1.061186
HT                                  1.135326  1        1.065517
```

Fox, J. and Monette, G. (1992) Generalized collinearity diagnostics. *JASA*, **87**, 178--183.

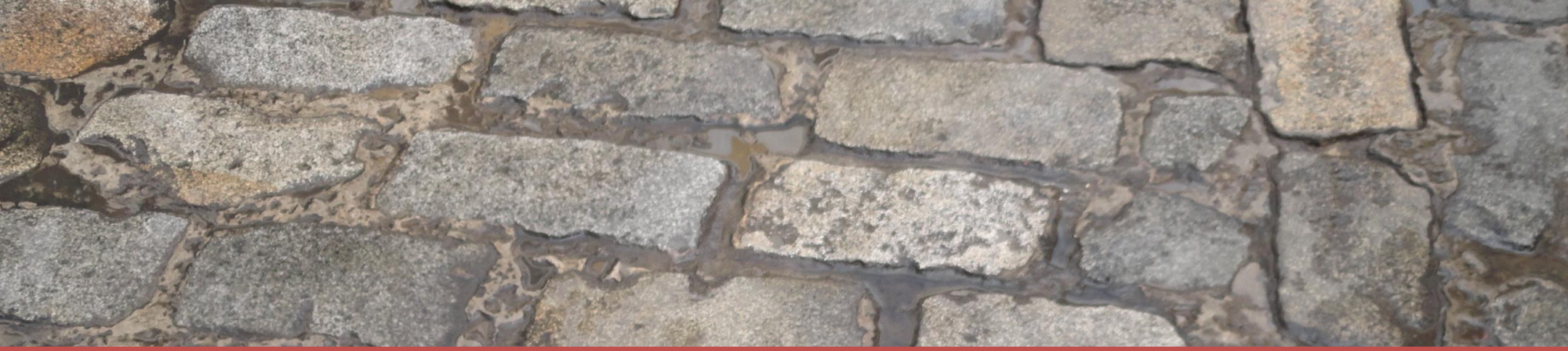
FATOR DE INFLAÇÃO DA VARIÂNCIA GENERALIZADO (GFIV)

FACTOR DE INFLACIÓN DE LA VARIANZA GENERALIZADO (GFIV)

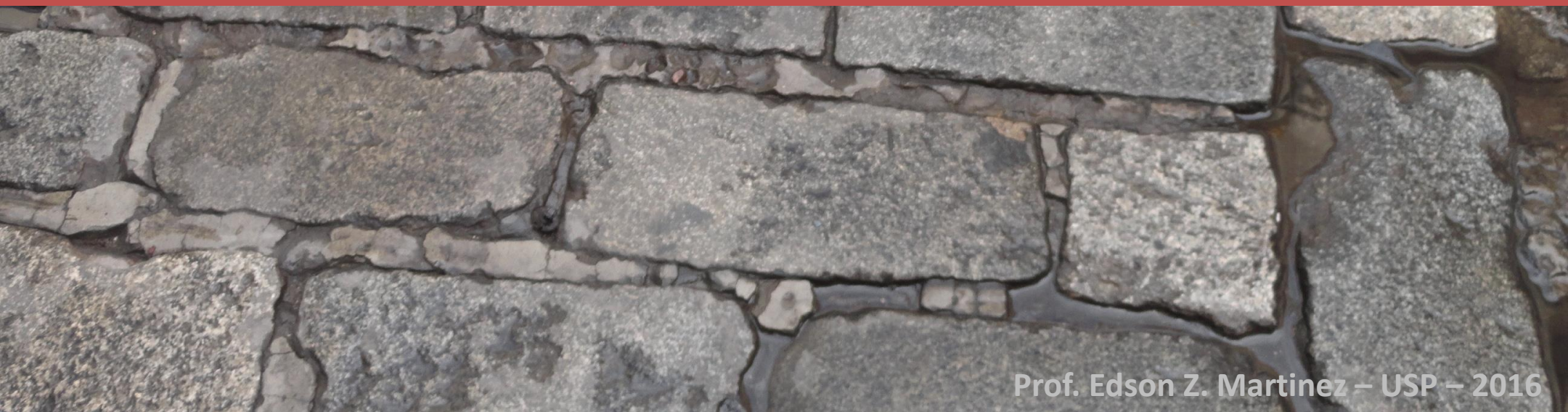
GENERALIZED VARIANCE INFLATION FACTOR (GVIF)



	Total	Baixo peso ao nascer *n* (%)	OR bruto (IC 95%)	OR ajustado (IC 95%)
Idade da mãe				
mais de 30 anos	20	3 (15,0%)	Referência	Referência
26 a 30 anos	34	10 (29,4%)	3,19 (0,61 ; 11,71)	2,52 (0,53 ; 14,90)
20 a 25 anos	84	31 (36,9%)	**3,31** (1,01 ; 15,00)	2,82 (0,70 ; 14,76)
até 19 anos	51	15 (29,4%)	2,36 (0,66 ; 11,17)	2,06 (0,46 ; 11,66)
Peso da mãe				
mais de 65 kg	44	10 (22,7%)	Referência	Referência
55 -| 65	49	14 (28,6%)	1,36 (0,53 ; 3,55)	1,64 (0,57 ; 5,04)
50 -| 55	43	10 (23,3%)	1,03 (0,37 ; 2,83)	1,17 (0,37 ; 3,75)
até 50 kg	53	25 (47,2%)	**3,03** (1,27 ; 7,63)	**3,06** (1,06 ; 9,53)
Cor da pele				
branca	96	23 (24,0%)	Referência	Referência
negra	26	11 (42,3%)	2,33 (0,93 ; 5,77)	**3,31** (1,16 ; 9,58)
outra	67	25 (37,3%)	1,89 (0,95 ; 3,74)	2,11 (0,85 ; 5,38)
Tabagismo				
não	115	29 (25,2%)	Referência	Referência
sim	74	30 (40,5%)	**2,02** (1,08 ; 3,78)	2,24 (0,98 ; 5,27)
Irritabilidade uterina				
não	161	45 (28,0%)	Referência	Referência
sim	28	14 (50,0%)	**2,58** (1,13 ; 5,83)	1,97 (0,78 ; 4,98)
Partos prematuros				
não	159	41 (25,8%)	Referência	Referência
sim	30	18 (60,0%)	**4,32** (1,92 ; 9,73)	**3,38** (1,33 ; 8,96)
Visitas ao médico				
0	100	36 (36,0%)	Referência	Referência
1	47	11 (23,4%)	1,41 (0,64 ; 3,08)	0,68 (0,25 ; 1,72)
2 ou mais	42	12 (28,6%)	0,76 (0,29 ; 1,98)	1,19 (0,47 ; 2,96)
História de hipertensão				
não	177	52 (29,4%)	Referência	Referência
sim	12	7 (58,3%)	**3,36** (1,02 ; 11,09)	**4,83** (1,27 , 19,96)

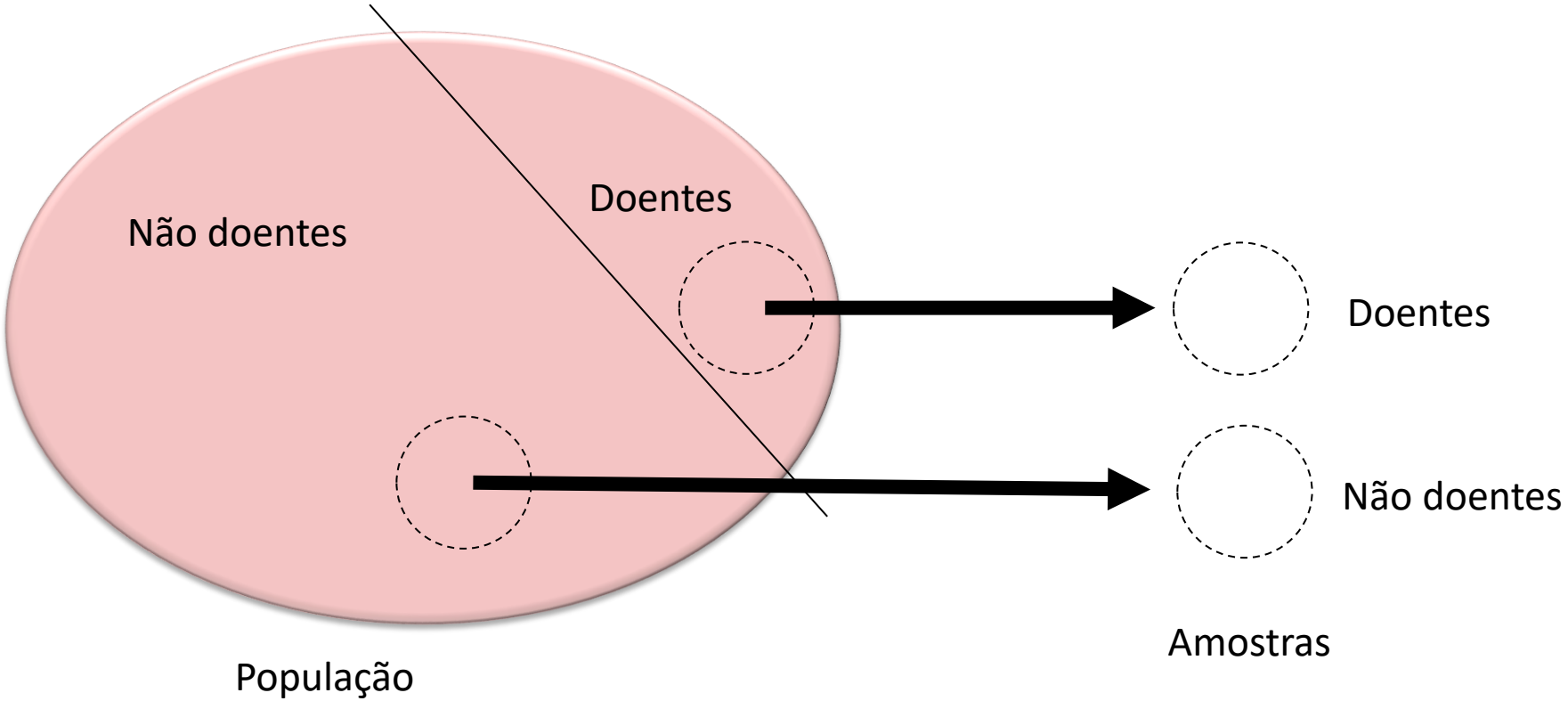


REGRESSÃO LOGÍSTICA CONDICIONAL
REGRESIÓN LOGÍSTICA CONDICIONAL
CONDITIONAL LOGISTIC REGRESSION



Prof. Edson Z. Martinez – USP – 2016

ESTUDOS CASO-CONTROLE
ESTUDIOS DE CASOS Y CONTROLES
CASE-CONTROL STUDIES



ESTUDOS CASO-CONTROLE NÃO PAREADOS
ESTUDIOS DE CASOS Y CONTROLES NO EMPAREJADOS
NON-PAIRED CASE-CONTROL STUDIES

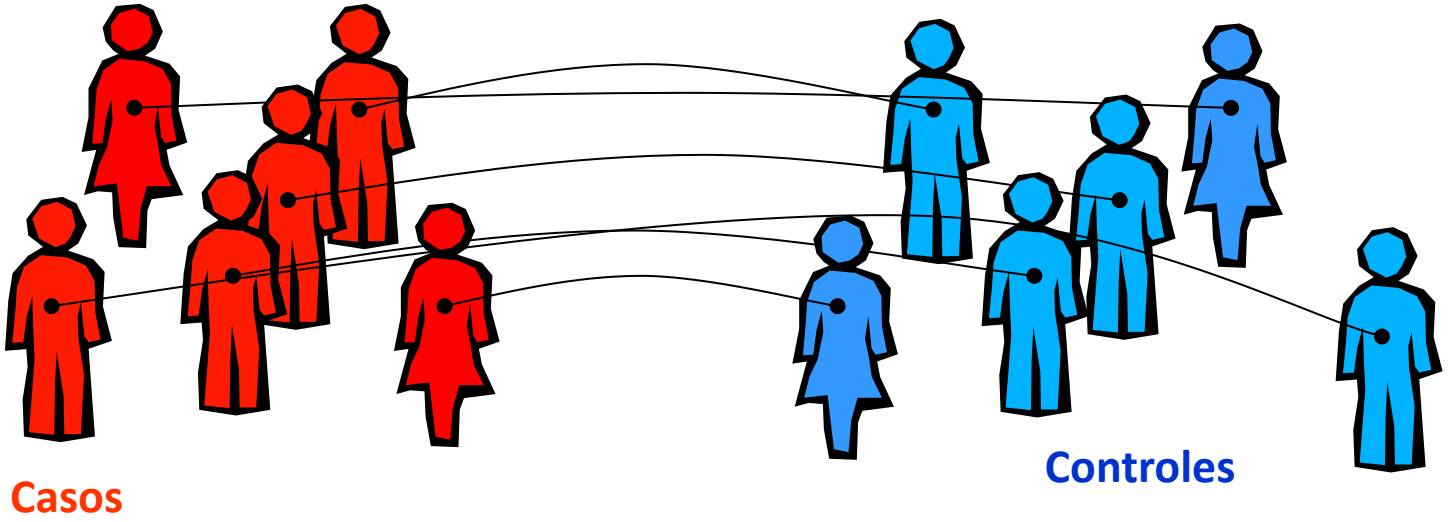
Casos e controles são escolhidos de acordo com os mesmos critérios de inclusão, a não ser a doença em estudo.

Casos

Controles

ESTUDOS CASO-CONTROLE PAREADOS
ESTUDIOS DE CASOS Y CONTROLES EMPAREJADOS
PAIRED CASE-CONTROL STUDIES

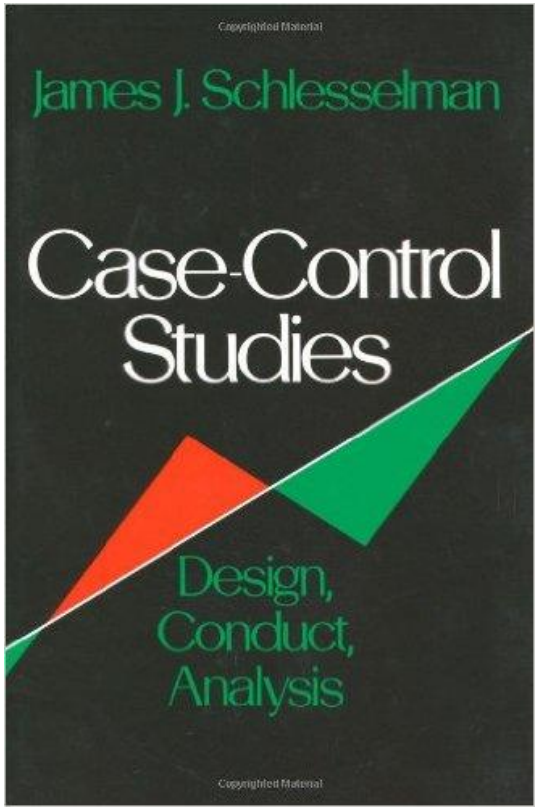
Para cada indivíduo portador da doença, é escolhido um controle com algumas características comuns (p.ex., sexo, idade).



ESTUDOS CASO-CONTROLE PAREADOS
ESTUDIOS DE CASOS Y CONTROLES EMPAREJADOS
PAIRED CASE-CONTROL STUDIES

- *If one matches cases and controls on a variable associated with the study exposure, an analysis that ignores the matching will yield a disease-exposure odds ratio that is biased toward unity. Thus, in some circumstances, unmatched analyses of matched data can spuriously diminish the estimate of an exposure's effect.*

(Schlesselman, 1982, p.137)





MODELOS DE REGRESSÃO LOGÍSTICA
ANÁLISIS DE REGRESIÓN LOGÍSTICA
LOGISTIC REGRESSION MODELS

Estudos caso-controle não pareados

- Regressão logística não condicional

Estudos caso-controle pareados

- Regressão logística condicional

EXEMPLO – LEE & WANG (2003)

EJEMPLO – LEE & WANG (2003)

EXAMPLE – LEE & WANG (2003)

Um estudo teve por objetivo avaliar o efeito da obesidade, do histórico familiar e de atividades físicas no desenvolvimento de diabetes não dependente de insulina.
Vinte e oito indivíduos diabéticos não dependentes de insulina (casos) foram emparelhados com 28 indivíduos não diabéticos (controles) pela idade e pelo sexo.
As variáveis independentes são:

- IMC: índice de massa corporal
- HF: histórico familiar de diabetes (0: ausente, 1: presente)
- ATF: atividade física (0: ausente, 1: presente)

EXEMPLO – LEE (1991)

EJEMPLO – LEE (1991)

EXAMPLE – LEE (1991)

	CASOS			CONTROLES		
Par	IMC	HF	ATF	IMC	HF	ATF
1	22,1	1	1	26,7	0	1
2	31,3	0	0	24,4	0	1
3	33,8	1	0	29,4	0	0
4	33,7	1	1	26,0	0	0
5	23,1	1	1	24,2	1	0
6	26,8	1	0	29,7	0	0
7	32,3	1	0	30,2	0	1
8	31,4	1	0	23,4	0	1
9	37,6	1	0	42,4	0	0
10	32,4	1	0	25,8	0	0
11	29,1	0	1	39,8	0	1
12	28,6	0	1	31,6	0	0
13	35,9	0	0	21,8	1	1
14	30,4	0	0	24,2	0	1

	CASOS			CONTROLES		
Par	IMC	HF	ATF	IMC	HF	ATF
15	39,8	0	0	27,8	1	1
16	43,3	1	0	37,5	1	1
17	32,5	0	0	27,9	1	1
18	30,3	0	0	31,3	0	1
19	32,5	1	0	34,5	1	1
20	32,5	1	0	25,4	0	1
21	21,6	1	1	27,0	1	1
22	24,4	0	1	31,1	0	0
23	46,7	1	0	27,3	0	1
24	28,6	1	1	24,0	0	0
25	29,7	0	0	33,5	0	0
26	29,6	0	1	20,7	0	0
27	34,8	1	0	30,0	0	1
28	37,3	1	0	26,5	0	0

REGRESSÃO LOGÍSTICA CONDICIONAL
REGRESIÓN LOGÍSTICA CONDICIONAL
CONDITIONAL LOGISTIC REGRESSION

- Seja uma única variável independente
- Seja uma amostra de tamanho n
- A função de verossimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp\left[\beta\left(x_i^{(1)} - x_i^{(2)}\right)\right]}{1 + \exp\left[\beta\left(x_i^{(1)} - x_i^{(2)}\right)\right]}$$

em que $x_i^{(1)}$ é uma observação da variável independente para os casos

e $x_i^{(2)}$ é uma observação da variável independente para os controle

REGRESSÃO LOGÍSTICA CONDICIONAL
REGRESIÓN LOGÍSTICA CONDICIONAL
CONDITIONAL LOGISTIC REGRESSION

Par	CASOS $x_i^{(1)}$	CONTROLES $x_i^{(2)}$	$x_i^{(1)} - x_i^{(2)}$
1	1	0	1
2	0	0	0
3	1	0	1
4	1	0	1
5	1	1	0
6	1	0	1
7	1	0	1
8	1	0	1
9	1	0	1
10	1	0	1
11	0	0	0
12	0	0	0
13	0	1	-1
...	...	...	...

Seja a variável independente histórico familiar (HF)

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp\left[\beta\left(x_i^{(1)} - x_i^{(2)}\right)\right]}{1 + \exp\left[\beta\left(x_i^{(1)} - x_i^{(2)}\right)\right]}$$

Notar que o modelo não considera um intercepto.

REGRESSÃO LOGÍSTICA CONDICIONAL
REGRESIÓN LOGÍSTICA CONDICIONAL
CONDITIONAL LOGISTIC REGRESSION

```
> # Pacote survival
> library(survival)
> w$y <- ifelse(w$grupo=="caso",1,0)
>
> # Modelo de regressão logística condicional simples
> summary(clogit(y ~ HF + strata(ID),data=w))
Call:
coxph(formula = Surv(rep(1, 56L), y) ~ HF + strata(ID), data = w,
    method = "exact")

  n= 56, number of events= 28

     coef exp(coef) se(coef)     z Pr(>|z|)
HF 1.4663    4.3333   0.6405 2.289   0.0221 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

   exp(coef) exp(-coef) lower .95 upper .95
HF     4.333     0.2308     1.235     15.21

Concordance= 0.679  (se = 0.089 )
Likelihood ratio test= 6.74  on 1 df,   p=0.009
Wald test            = 5.24  on 1 df,   p=0.02
Score (logrank) test = 6.25  on 1 df,   p=0.01
```

ODDS RATIO, DADOS PAREADOS
RAZÓN DE MOMIOS, DATOS PAREADOS
ODDS RATIO, PAIRED DATA

	Casos HF presente	Casos HF ausente
Controles HF presente	4	3
Controles HF ausente	13	8

$$\widehat{OR} = \frac{13}{3} = 4{,}333$$

REGRESSÃO LOGÍSTICA MÚLTIPLA CONDICIONAL
REGRESIÓN LOGÍSTICA MÚLTIPLE CONDICIONAL
CONDITIONAL MULTIPLE LOGISTIC REGRESSION

- Sejam três variáveis independentes (IMC, HF e ATF)
- Seja uma amostra de tamanho n
- A função de verossimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp\left[\beta_1\left(x_{1i}^{(1)} - x_{1i}^{(2)}\right) + \beta_2\left(x_{2i}^{(1)} - x_{2i}^{(2)}\right) + \beta_3\left(x_{3i}^{(1)} - x_{3i}^{(2)}\right)\right]}{1 + \exp\left[\beta_1\left(x_{1i}^{(1)} - x_{1i}^{(2)}\right) + \beta_2\left(x_{2i}^{(1)} - x_{2i}^{(2)}\right) + \beta_3\left(x_{3i}^{(1)} - x_{3i}^{(2)}\right)\right]}$$

- Em que: β_1 é o efeito do IMC (x_1)
  β_2 é o efeito do HF (x_2)
  β_3 é o efeito da ATF (x_3)

REGRESSÃO LOGÍSTICA MÚLTIPLA CONDICIONAL
REGRESIÓN LOGÍSTICA MÚLTIPLE CONDICIONAL
CONDITIONAL MULTIPLE LOGISTIC REGRESSION

```
> # Modelo múltiplo
> summary(clogit(y ~ IMC + HF + ATF + strata(ID),data=w))
Call:
coxph(formula = Surv(rep(1, 56L), y) ~ IMC + HF + ATF + strata(ID),
    data = w, method = "exact")

  n= 56, number of events= 28

        coef exp(coef) se(coef)      z Pr(>|z|)
IMC  0.10506   1.11078  0.07516  1.398   0.1621
HF   1.69726   5.45900  0.77363  2.194   0.0282 *
ATF -0.70066   0.49626  0.64657 -1.084   0.2785
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    exp(coef) exp(-coef) lower .95 upper .95
IMC    1.1108     0.9003    0.9586     1.287
HF     5.4590     0.1832    1.1984    24.867
ATF    0.4963     2.0151    0.1397     1.762

Concordance= 0.75  (se = 0.116 )
Likelihood ratio test= 12.3  on 3 df,   p=0.006
Wald test            = 6.88  on 3 df,   p=0.08
Score (logrank) test = 10.07  on 3 df,   p=0.02
```

ANÁLISE NÃO PAREADA DE DADOS PAREADOS
ANÁLISIS SIN PAREAMIENTO DE DATOS EMPAREJADOS
UNMATCHED ANALYSES OF MATCHED DATA

```
data exemplo;
input IMC HF ATF y;
cards;
22.1   1   1    1
31.3   0   0    1
33.8   1   0    1
33.7   1   1    1
...
20.7   0   0    0
29.2   1   1    0
30.0   0   1    0
26.5   0   0    0
;;;
proc logistic descending
data=exemplo;
 model y = IMC HF ATF;
run;
```

INADEQUADO!

ANÁLISE NÃO PAREADA DE DADOS PAREADOS
ANÁLISIS SIN PAREAMIENTO DE DATOS EMPAREJADOS
UNMATCHED ANALYSES OF MATCHED DATA

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.0970	1.7596	1.4204	0.2333
IMC	1	0.0669	0.0557	1.4432	0.2296
HF	1	1.1375	0.5787	3.8631	0.0494
ATF	1	-0.8608	0.5928	2.1088	0.1465

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect	Unit	Estimate	95% Confidence Limits	
IMC	1.0000	1.069	0.962	1.200
HF	1.0000	3.119	1.029	10.171
ATF	1.0000	0.423	0.127	1.332

ANÁLISE NÃO PAREADA DE DADOS PAREADOS
ANÁLISIS SIN PAREAMIENTO DE DATOS EMPAREJADOS
UNMATCHED ANALYSES OF MATCHED DATA

Regressão logística condicional

Efeito	*OR*	IC 95%	Valor *p*
β_1 (IMC)	1.111	(0.967 ; 1.313)	0.1621
β_2 (HF)	5.459	(1.433 ; 33.553)	0.0282
β_3 (ATF)	0.496	(0.126 ; 1.738)	0.2785

Regressão logística não condicional

Efeito	*OR*	IC 95%	Valor *p*
β_1 (IMC)	1.069	(0.962 ; 1.200)	0.2296
β_2 (HF)	3.119	(1.029 ; 10.171)	0.0494
β_3 (ATF)	0.423	(0.127 ; 1.332)	0.1465

EXEMPLO – MEDEIROS ET AL. (2003)
EJEMPLO – MEDEIROS ET AL. (2003)
EXAMPLE – MEDEIROS ET AL. (2003)

Estudo caso-controle sobre exposição precoce ao leite de vaca e ocorrência de Diabetes Mellitus tipo 1 em Campina Grande, Paraíba

Case-control study on early exposure to cow's milk and the occurrence of Diabetes Mellitus type 1 in Campina Grande in the State of Paraíba

Josimar dos Santos Medeiros [1]
Maria Amélia Amado Rivera [2]
Maria José Cariri Benigna [3]
Maria Aparecida Alves Cardoso [4]
Maria José de Carvalho Costa [5]

[1.4] Departamento de Farmácia e Biologia. Centro de Ciências Biológicas e da Saúde. Universidade Estadual da Paraíba. Av. das Baraúnas, 351. Bodocongó. Campina Grande, PB, Brasil. CEP: 58.109-753
[2] Curso de Nutrição. Faculdade de Ciências Médicas da Paraíba. João Pessoa, PB
[3] Departamento de Enfermagem. Centro de Ciências Biológicas e da Saúde. Universidade Estadual da Paraíba. Campina Grande, PB
[5] Departamento de Nutrição. Centro de Ciências da Saúde. Universidade Federal da Paraíba. João Pessoa, PB

Rev. Bras. Saúde Matern. Infant., Recife, 3 (3): 271-280, jul. / set., 2003

EXEMPLO – MEDEIROS ET AL. (2003)
EJEMPLO – MEDEIROS ET AL. (2003)
EXAMPLE – MEDEIROS ET AL. (2003)

Métodos

O estudo realizado foi do tipo caso-controle. A amostra foi constituída por 128 indivíduos de ambos os sexos, dos quais 64 eram portadores de Diabetes Mellitus tipo 1 (casos) e 64 indivíduos eram normais (controles). Os casos selecionados foram todos os diabéticos tipo 1 menores de 18 anos atendidos no ambulatório de endocrinologia do Hospital Universitário Alcides Carneiro da Universidade Federal da Paraíba, Campus II, em Campina Grande, Brasil, durante o ano de 1999. Os casos foram selecionados utilizando-se os seguintes critérios de inclusão: indivíduos de ambos os sexos portadores de Diabetes Mellitus tipo 1, com idade igual ou inferior a 18 anos na época do diagnóstico definitivo da doença, com residência em Campina Grande e possibilidade de participação da mãe (natural) durante a entrevista.

O grupo controle foi selecionado entre os vizinhos dos pacientes e emparelhados por idade, sexo e cor. O instrumento utilizado para coleta de dados foi um formulário adaptado de Barros e Victora.[16] O formulário foi testado, revisado e aperfeiçoado. A entrevista foi realizada com as mães dos pacientes e dos controles, que deram consentimento para participação na pesquisa. O formulário foi aplicado por um dos autores da pesquisa, que estava ciente do grupo a que a mãe dos participantes pertencia; portanto, para se evitar tendenciosidades na aferição das variáveis em estudo, o mesmo foi composto por 23 questões fechadas. Todos os cuidados foram tomados para que não houvesse diferença na abordagem de mães de casos e controles de modo a garantir a fidedignidade e evitar vícios de informação que pudessem comprometer a validade dos resultados.[16]

EXEMPLO – MEDEIROS ET AL. (2003)
EJEMPLO – MEDEIROS ET AL. (2003)
EXAMPLE – MEDEIROS ET AL. (2003)

Foram realizadas entrevistas com as mães de todos os indivíduos estudados para levantar dados relacionados com a identificação dos pacientes e seus pais, incluindo idade, cor, renda familiar e escolaridade materna. Foram também obtidas informações sobre antecedentes familiares, enfatizando história familiar de Diabetes Mellitus (tipos 1 ou 2), número de gestações da mãe, peso ao nascer e tipo de parto. A exposição precoce ao leite de vaca foi avaliada por meio da variável relativa à duração do aleitamento materno exclusivo em dias. Foram considerados expostos todos os indivíduos que consumiram leite de vaca antes dos quatro meses de idade. Com relação à história familiar de diabetes, foi verificado se os indivíduos estudados tinham qualquer parente em primeiro grau portador da doença.

Os dados foram processados utilizando-se o *software* Epi-info versão 6.04b. As análises univariadas e multivariadas, ajustando-se por escolaridade materna, renda familiar e peso ao nascer, e história familiar de Diabetes Mellitus, foram realizadas usando-se modelos de regressão logística condicional por meio do *software* Stata, versão 7.0.

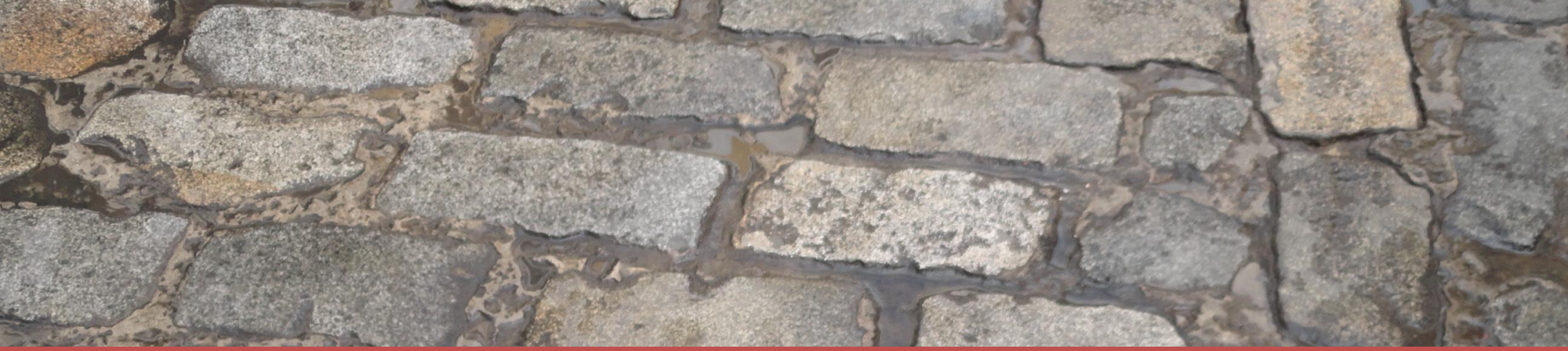
EXEMPLO – MEDEIROS ET AL. (2003)

Tabela 6

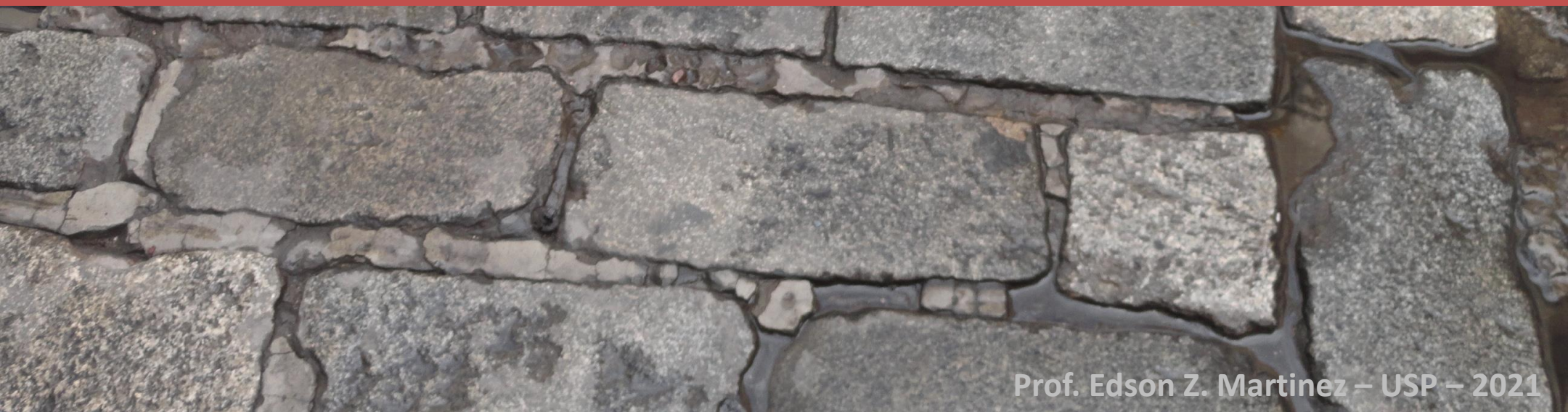
Razões de chances não-ajustadas e ajustadas e seus respectivos intervalos de confiança 95%, comparando casos de Diabetes Mellitus e controles. Campina Grande, PB, 2000.

Características	RC (IC95%)	RCajust (IC95%)	p*
Amamentação exclusiva até 4 meses			0,01
Sim	1,00	1,00	
Não	3,17 (1,26 - 7,93)	4,09 (1,19 - 14,04)	
Escolaridade materna**			0,01
0 (Analfabeta)	1,00	1,00	
Até 4	0,33 (0,03 - 3,20)	0,11 (<0,01 - 1,93)	
Até 8	0,06 (<0,01 - 1,00)	0,02 (<0,01 - 0,69)	
Até 11	0,20 (0,01 - 3,66)	0,07 (<0,01 - 2,82)	
maior que 11	0,58 (0,02 - 15,3)	0,24 (<0,01 - 11,9)	
Renda familiar***	0,66 (0,44 - 1,01)	0,68 (0,41 - 1,13)	0,13
Peso ao nascer***	1,00 (0,99 - 1,00)	0,99 (0,99 - 1,00)	0,65
História familiar de diabetes			0,07
Não	1,00	1,00	
Sim	2,22 (1,01 - 4,92)	2,57 (0,85 - 7,73)	

* Teste da razão de verossimilhança; ** Em anos de estudo; *** Variável contínua



SELEÇÃO DE VARIÁVEIS
SELECCIÓN DE VARIABLES
VARIABLE SELECTION



Prof. Edson Z. Martinez – USP – 2021

SELEÇÃO DE VARIÁVEIS
SELECCIÓN DE VARIABLES
VARIABLE SELECTION

Estratégia usual (porém inadequada): selecionar para compor um modelo múltiplo as variáveis independentes que mostrarem evidência de associação com a dependente em um modelo univariado, considerando um valor-p menor que 0,10 ou 0,20. Críticas:

Sun, G. W., Shook, T. L., & Kay, G. L. (1996). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. Journal of Clinical Epidemiology, 49(8), 907-916.

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection - a review and recommendations for the practicing statistician. Biometrical Journal, 60(3), 431-449.

Heinze, G., & Dunkler, D. (2017). Five myths about variable selection. Transplant International, 30, 6-10.

Greenland, S. (1989). Modeling and variable selection in epidemiologic analysis. American Journal of Public Health, 79(3), 340-349.

SELEÇÃO DE VARIÁVEIS
SELECCIÓN DE VARIABLES
VARIABLE SELECTION

Estratégia usual (porém inadequada): seleção *stepwise*. Críticas:

Malek, M. H., Berger, D. E., & Coburn, J. W. (2007). On the inappropriateness of stepwise regression analysis for model building and testing. European Journal of Applied Physiology, 101(2), 263-264.

Sainani, K. L. (2013). Multivariate regression: the pitfalls of automated variable selection. PM&R, 5(9), 791-794.

Smith, G. (2018). Step away from stepwise. Journal of Big Data, 5(1), 1-12.

Flom, P. (2018). Stopping stepwise: Why stepwise selection is bad and what you should use instead. Available at: https://towardsdatascience.com/stopping-stepwise-why-stepwise-selection-is-bad-and-what-you-should-use-instead-90818b3f52df

SELEÇÃO DE VARIÁVEIS
SELECCIÓN DE VARIABLES
VARIABLE SELECTION

Estratégias válidas:

- Entender a função de cada variável no estudo (confundimento, moderação, mediação, causa remota)
- Entender as estruturas de associação (cadeia, garfo, garfo invertido)
- Gráficos acíclicos direcionados (DAG)
- Modelos de equações estruturais
- Modelos LASSO

REFERÊNCIAS
REFERENCIAS
REFERENCES

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons.

Lee, E. T., & Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis*, Third Edition. John Wiley, New York.

Medeiros, J. D. S., Rivera, M. A. A., Benigna, M. J. C., Cardoso, M. A. A., & Costa, M. J. D. C. (2003). Estudo caso-controle sobre exposição precoce ao leite de vaca e ocorrência de Diabetes Mellitus tipo 1 em Campina Grande, Paraíba. *Rev. Bras. Saúde Matern. Infant*, *3*(3), 271-280.

Pagano, M., & Gauvreau, K. (2004). *Princípios de Bioestatística*. Thomson Learning.

Roberts, W. F., Perry, K. G., Naef, R. W., Washburne, J. F., & Morrison, J. C. (1995). The irritable uterus: a risk factor for preterm birth?. *American Journal of Obstetrics and Gynecology*, *172*(1), 138-142.

Schlesselman, J. J. (1982). *Case-control studies: design, conduct, analysis*. Oxford University Press.



MODELOS DE REGRESSÃO LOGÍSTICA
ANÁLISIS DE REGRESIÓN LOGÍSTICA
LOGISTIC REGRESSION MODELS

Prof. Edson Zangiacomi Martinez

Faculdade de Medicina de Ribeirão Preto
Universidade de São Paulo (USP)
Ribeirão Preto, Brasil
edson@fmrp.usp.br