UNIJORGE
MEMBRO DA REDE ILUMNO

PROGRAMA DE PÓS-GRADUAÇÃO

CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA APLICADA

MODELOS PARA DADOS CATEGORIZADOS

(Avaliação da associação entre o baixo peso ao nascer de crianças ao comportamento da mãe durante a gravidez em um Hospital de Salvador)

SALVADOR

2020

ANA PAULA DE SOUZA VANDERLEY

MODELOS PARA DADOS CATEGORIZADOS

(Avaliação da associação entre o baixo peso ao nascer de crianças ao comportamento da mãe durante a gravidez em um Hospital de Salvador)

Trabalho apresentado ao Curso de Especialização em Estatística Aplicada do Centro Universitário Jorge Amado, como requisito para a conclusão da disciplina Modelos para dados Categorizados, sob a orientação do Profº Jackson Conceição .

.

SALVADOR

2020

SUMÁRIO

1 INTRODUÇÃO

Bebê nascer com baixo peso é uma característica bastante comum em crianças de todo o mundo. Várias pesquisas levantaram as causas deste problema e tentam identificar as principais consequências. É um problema sério ligado diretamente a mortalidade infantil, e uma das principais causas do nascimento prematuro.

Um bebê é considerado abaixo do peso quando nasce com um peso inferior a 2.500 gramas. Em alguns casos é porque ele nasceu antes do tempo ideal, ou em outros casos esta relacionado ao comportamento da mãe durante a gravidez.

Este trabalho visa apresentar uma avaliação entre a associação entre o baixo peso ao nascer de crianças ao comportamento da mãe durante a gravidez em um hospital de Salvador. O banco de dados foi disponibilizado pelo Professor Me. Jackson Conceição na matéria de dados categorizados avaliação 6.

Foi feita uma análise descritiva dos dados a princípios para verificar algumas características do banco de dados fornecido, em seguida foi feito um modelo para comparar essas variáveis explicativas com a variável resposta que é o baixo peso ao nascer das crianças para que assim possamos verificar quais causas estão mais correlacionadas com o baixo peso do bebê.

2 ANÁLISE DESCRITIVA

O banco de dados possui 189 observações e 11 variáveis sendo uma variável resposta LOW (baixo peso ao nascer), As variáveis explicativas analisadas foram:

LOW: Baixo peso ao nascer (1:> = 2500, 2: <2500)

AGE: Idade da mãe (anos)

LWT: Peso da mãe no último período menstrual (Libras)

RACE: Raça (1: Branco, 2: Preto, 3: Outro)

SMOKE: Status do tabagismo durante a gravidez (0: Não, 1: Sim)

PTL: História de trab. de parto prematuro (0: Nenhuma, 1: Uma, 2: Duas ou +)

HT: História de hipertensão (0: Não, 1: Sim)

UI: Presença de irritabilidade uterina (0: Não, 1: Sim)

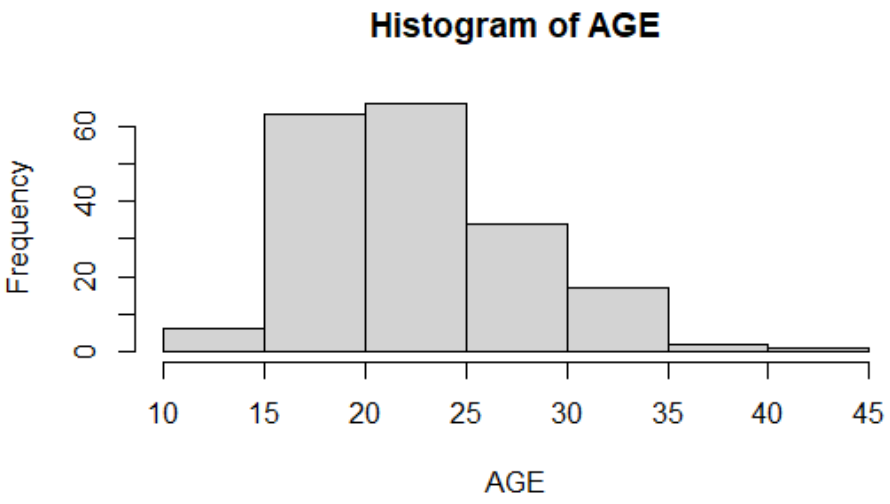
FTV: Número de consultas médicas durante o primeiro trimestre (0: Nenhuma, 1: Uma, 2: Duas ou +)

BWT: Peso ao nascer (gramas).

Primeiramente foi feita uma análise descritiva dos dados, foi feita uma conversão de algumas variáveis de quantitativas para binaria com a finalidade de dividir por categorias foram essas **AGE** (idade) e **LWT** (peso da mãe). A variável **AGE** ficou 0 para <=23 e 1 para >23, a variável **LWT** ficou 0 <=121 e 1 para >121.

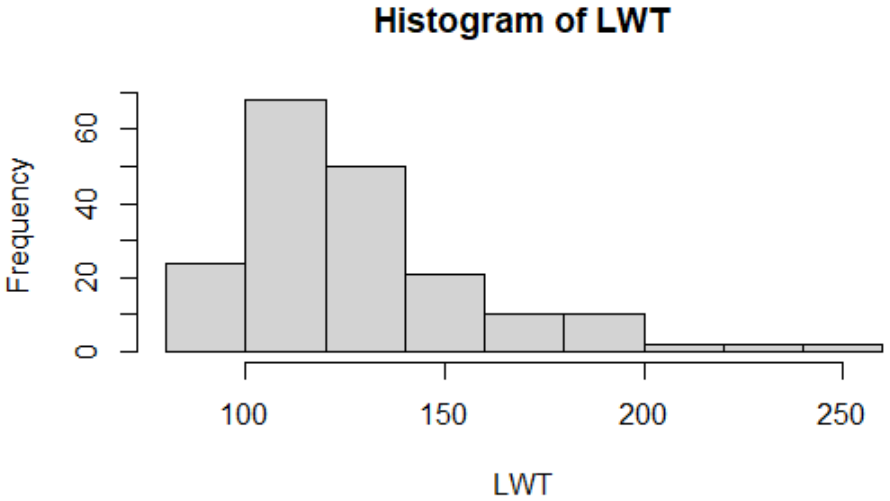
	ID	LOW	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FTV	BWT
1	4	1	28	120	3	1	1	0	1	0	709
2	10	1	29	130	1	0	0	0	1	2	1021
3	11	1	34	187	2	1	0	1	0	0	1135
4	13	1	25	105	3	0	1	1	0	0	1330
5	15	1	25	85	3	0	0	0	1	0	1474
6	16	1	27	150	3	0	0	0	0	0	1588

Abaixo temos as estatísticas descritivas das variáveis em estudo para termos algumas características iniciais do nosso banco de dados.




```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
14.00   19.00   23.00   23.24   26.00   45.00
```

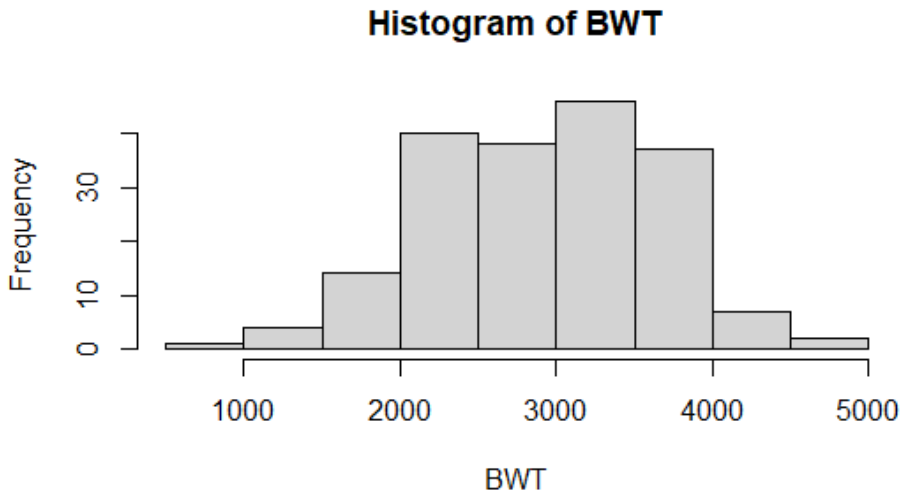
Podemos verificar que a média de idade das mães é de 23 anos sendo que a maioria tem entre 15 e 25 anos.



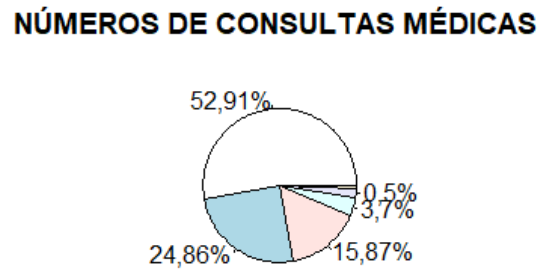
Já o peso da mãe tem média de 64,26 kg , minima de 38,77 kg e máxima de 83,25.

Min	1Q	Median	3Q	Max
-1.3877	-0.6426	-0.6426	0.9808	1.8325

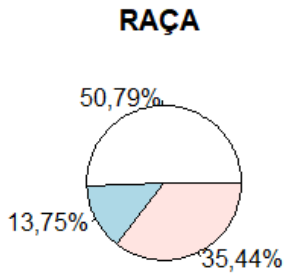
Já o Peso do bebê ao nascer se comportou a maioria dentro de um intervalo de 2kg a 4kg. Ou seja a maioria dos bebês se comportaram dentro de um peso considerado normal.



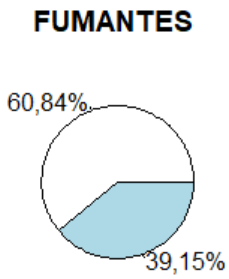
Abaixo veremos o comportamento das mães durante a gravidez.



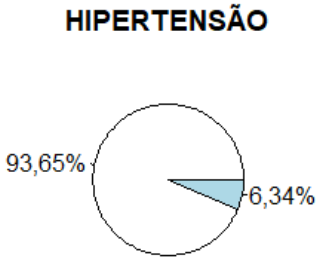
52% das mães não fizeram nenhuma consulta ao médico durante a gravidez, 24% fizeram 1, 15% fizeram 2, e assim por diante.



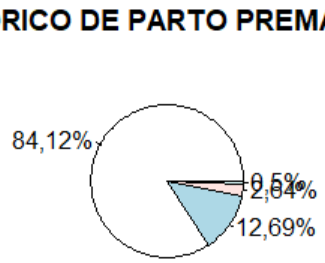
50,79% das mães se declararam brancas, 13,755 negros e 35,445% outros.



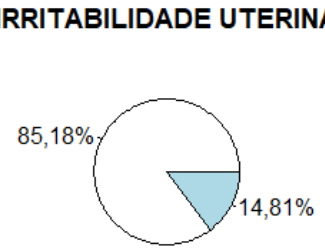
60% das mães não são fumantes.



93,65% das mães não tem hipertensão.



84,12% das mães não apresentaram nenhum histórico de parto prematuro, 12,69% apresentaram 1, 2,64% apresentaram 2 e 0,5% apresentaram 3 .



85,18% não apresentaram irritabilidade uterina durante a gravidez.

A principio não foi identificada nenhuma variável que tenha um valor significativo de comportamento que explique a variável resposta, foi feito um modelo de regressão logística binaria para correlacionar essas variáveis com a variável resposta e através dele medir a significância dessas variáveis.

3 AJUSTE DO MODELO

Foi feito a cruzamento dessas variáveis com a variável resposta e através da análise do *p-valor* podemos identificas quais variáveis apresentam maior significância para o modelo. *p-valor* inferior ao alpha de 0,05 indica que existe evidência suficiente para afirmar que a estimativa é diferente de zero. Os valores encontrados estão representados na tabela abaixo.

	Estimate	Std. Error	z value	Pr(>|z|)	
AGE	-0.1611	0.3184	-0.506	0.612925	
PTL	0.8018	0.3172	2.528	0.0115	*
LWT	-0.5237	0.3171	-1.652	0.0986	.
SMOKE	0.7041	0.3196	2.203	0.0276	*
HT	1.2135	0.6083	1.995	0.0461	*
UI	0.9469	0.4168	2.272	0.0231	*
FTV	-0.1351	0.1567	-0.862	0.388527	
BWT	-1.186	87.251	-0.014	0.989	

Verificamos que as variáveis **SMOKE, PTL, HT** e **UI** apresentaram valores signi ficantes abaixo de 0,05 ou seja há indícios que essas variáveis podem contribui r para o baixo peso ao nascer das crianças. Abaixo faremos um cruzamento no vamente somente com as variáveis que apresentaram alguma significância par a o modelo.

```
formula = Y ~ PTL + SMOKE + HT + UI
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.4175	0.2464	-5.753	8.76e-09	***
PTL	0.6069	0.3356	1.808	0.0705	.
SMOKE	0.5822	0.3365	1.730	0.0836	.
HT	1.4167	0.6221	2.277	0.0228	*
UI	0.8838	0.4429	1.996	0.0460	*

Observa-se que as variáveis HT e UI apresentaram valores mais significantes p ara o modelo final. aplicaremos o modelo proposto e testaremos as razões das chances.

```
formula = Y ~ HT + UI
```

4 INFERÊNCIA DOS PARAMETROS DO MODELO

Após escolha das variáveis do modelo ajustado com base no *p-valor*, foram testadas as razões de chances (*odds ratio* - OR) e o intervalo de confiança de 95%.

O modelo composto com as variáveis **PTL, SMOKE, HT** e **UI** apresentaram uma significância menor do que o segundo modelo contendo somente as variáveis **HT** e **UI.**

	OR	2.5 %	97.5 %
(Intercept)	0.2423196	0.1461299	0.3854593
PTL	1.8347979	0.9604967	3.6272142
SMOKE	1.7899273	0.9240999	3.4720937
HT	4.1234305	1.2281275	14.8606939
UI	2.4201630	1.0048669	5.7828895

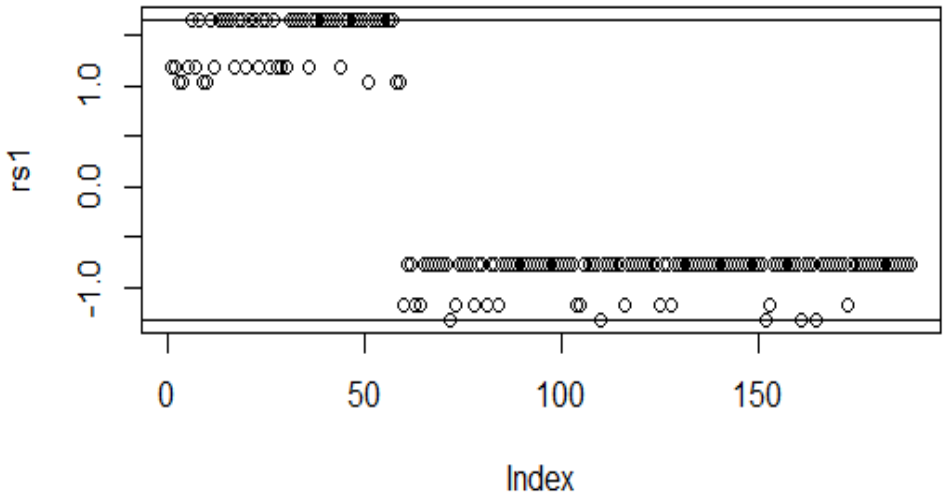
	OR	2.5 %	97.5 %
(Intercept)	0.3423423	0.2339676	0.489891
HT	4.0894737	1.2343010	14.545492
UI	2.9210526	1.2717694	6.737588

Podemos concluir que as variáveis que tem maior ligação com o peso do bebê ao nascer é a hipertensão e a irritabilidade uterina.

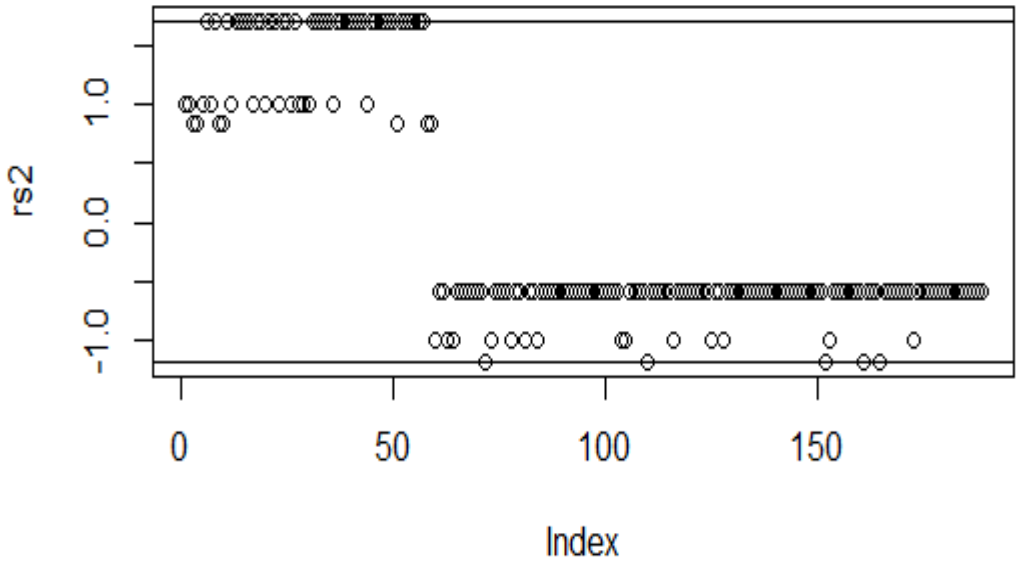
5 DIAGNOSTICO DO MODELO

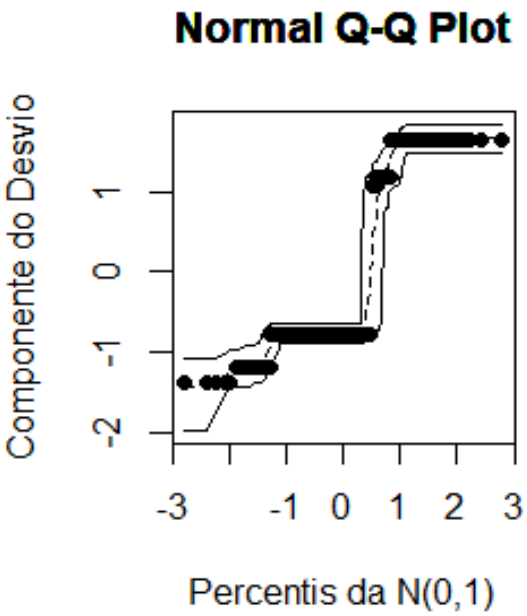
Com base nas análises dos gráficos a seguir podemos assegurar a validade do modelo pois os resíduos se comportaram dentro de um intervalo de confiança de 95% e os resíduos se comportaram dentro do envelope.

Resíduo Deviance



Resíduo Pearson





ROTEIRO NO R

```
# LEITURA DO BANCO DE DADOS

getwd()
setwd("C:/Users/user/Documents")

dados <- read.csv("C:/Users/user/Downloads/AV6.txt", sep="")
attach(dados)
names(dados)
head(dados)


# TRANSFORMANDO AS VARIAVEIS "AGE", "LWT"  EM BINARIA
```

```
summary(AGE)
sd(AGE)
hist(AGE)
dados$AGE[AGE<=23]=0
dados$AGE[AGE>=23]=1
head(dados)



summary(LWT)
sd(LWT)
hist(LWT)
dados$LWT[LWT<=121]=0
dados$LWT[LWT>=121]=1
head(dados)


# VERIFICANDO O P VALOR DE CADA VARIAVEL


Y = dados$LOW


modAGE = glm(Y~AGE, family=binomial(link="logit"), data=dados)
summary(modAGE)

modPTL = glm(Y~PTL, family=binomial(link="logit"), data=dados)
summary(modPTL)

modLWT = glm(Y~LWT, family=binomial(link="logit"), data=dados)
summary(modLWT)

modSMOKE = glm(Y~SMOKE, family=binomial(link="logit"), data=dados)
```

```
summary(modSMOKE)

modHT = glm(Y~HT, family=binomial(link="logit"), data=dados)
summary(modHT)

modUI = glm(Y~UI, family=binomial(link="logit"), data=dados)
summary(modUI)

modFTV = glm(Y~FTV, family=binomial(link="logit"), data=dados)
summary(modFTV)

modBWT = glm(Y~BWT, family=binomial(link="logit"), data=dados)
summary(modBWT)



# MODELO

mod = glm(Y~SMOKE+HT, family=binomial(link="logit"),data=dados)
summary(mod)

# RAZÃO DE CHANCES

OR = exp(mod$coefficients)
OR

cbind(OR, exp(confint(mod)))

# DIAGNÓSTICO

#RESÍDUO DEVIANCE

rs1 = resid(mod, type = "deviance")
```

```
quantile(rs1, probs = seq(0,1,0.025))
plot(rs1, main="Resíduo Deviance")
abline(h=-1.5138324)
abline(h=1.7011641)

#RESÍDUO DEVIANCE

rs2 = resid(mod, type = "pearson")
quantile(rs2, probs = seq(0, 1, 0.025))
plot(rs2, main="Resíduo Pearson")
abline(h=-1.4646144)
abline(h=1.8028471)

#ENVELOPE PARA OS RESÍDUOS

par(mfrow=c(1,1))
X <- model.matrix(mod)
n <- nrow(X)
p <- ncol(X)
w <- mod$weights
W <- diag(w)
H <- solve(t(X)%*%W%*%X)
H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
h <- diag(H)
td <- resid(mod,type="deviance")/sqrt(1-h)
e <- matrix(0,n,100)
for(i in 1:100){
  dif <- runif(n) - fitted(mod)
  dif[dif >= 0 ] <- 0
  dif[dif<0] <- 1
  nresp <- dif
  fit <- glm(nresp ~ X, family=binomial)
  w <- fit$weights
  W <- diag(w)
```

```
  H <- solve(t(X)%*%W%*%X)
  H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
  h <- diag(H)
  e[,i] <- sort(resid(fit,type="deviance")/sqrt(1-h))}
#
e1 <- numeric(n)
e2 <- numeric(n)
#
for(i in 1:n){
  eo <- sort(e[i,])
  e1[i] <- (eo[2]+eo[3])/2
  e2[i] <- (eo[97]+eo[98])/2}
#
med <- apply(e,1,mean)
faixa <- range(td,e1,e2)
par(pty="s")
qqnorm(td,xlab="Percentis da N(0,1)",
       ylab="Componente do Desvio", ylim=faixa, pch=16)
par(new=T)
#
qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1)
par(new=T)
qqnorm(e2,axes=F,xlab="",ylab="", type="l",ylim=faixa,lty=1)
par(new=T)
qqnorm(med,axes=F,xlab="", ylab="", type="l",ylim=faixa,lty=2)
```