Trabalho de dados Binários

Acidentes de carro

Laís Hoffmam, Simone Matsubara, Yasmin Fernandes, Willian Meira

2020-09-16

1. Base de Dados

1.1 Descrição dos dados

Os dados foram retirados do pacote "DAAG", sendo dados dos EUA, entre 1997-2002, de acidentes de carro relatados pela polícia nos quais há um evento prejudicial (pessoas ou propriedade) e do qual pelo menos um veículo foi rebocado. Os dados são restritos aos ocupantes do banco da frente, incluem apenas um subconjunto das variáveis registradas e são restritos de outras maneiras também.

A base original possui uma base de dados com 26.217 observações nas 15 variáveis a seguir.

1 - **dvcat**: velocidades estimadas do impacto do acidente: 1-9km/h, 10-24, 25-39, 40-54, 55+
2 - **weight**: Pesos de observação
3 - **dead**: Classificação se sobreviveu ao acidente: 1 = sobreviveu ou 0 = morreu
4 - **airbag**: Se o carro possui airbag: com ou sem airbag
5 - **seatbelt**: uso do cinto de segurança: com ou sem cinto
6 - **frontal**: impacto do acidente: 0 = não frontal, 1 = impacto frontal
7 - **sex**: Sexo: 0 = Feminino ou 1 = Masculino
8 - **ageOFocc**: Idade dos ocupantes do veículo
9 - **yearacc**: Ano do acidente (1997-2002)
10 - **yearVeh**: Ano do veículo (1953-2003)
11 - **abcat**: Se Airbags foram acionados: deploy, nodeploy, unavail
12 - **occRole**: Posição do airbag acionado: driver, pass
13 - **deploy**: Airbag acionados: 0: Se não possuia airbag ou não foi acionado, 1: Um ou mais airbags foram acionados
14 - **injSeverity**: Gravidade do acidente: 0:none, 1 = Possível Lesão, 2:no incapacity, 3:incapacity, 4:killed; 5:unknown, 6:prior death
15 - **caseid**: Número do caso.

O objetivo da análise foi modelar a probalidade de sobrevivencia dos acidentes de carro da base "nassCD" sob a influência do airbag e outros elementos.

O escopo da análise tem como respota a variável "dead", que informa se o ocupante do veículo sobreviveu ou não após o acidente. As demais covariáveis serão explicativas, no entando, foi constatado, verificando os dados da base, que algumas delas são irrelevantes ou estão mal explicadas ("weight" e "caseid"). Foram retiradas também as variáveis "airbag", "ageOFocc", "yearVeh", "abcat" e "injSeverity", pois não convergiram ao rodar os modelos.

Outro detalhe, devido ao número elevado de registros (mais de 26000 linhas), pegamos uma amostra de aproximadamente 10% do total de registros.

```
##    veloc sobrev cinto frontal sexo idade ocupantes abfunc
```

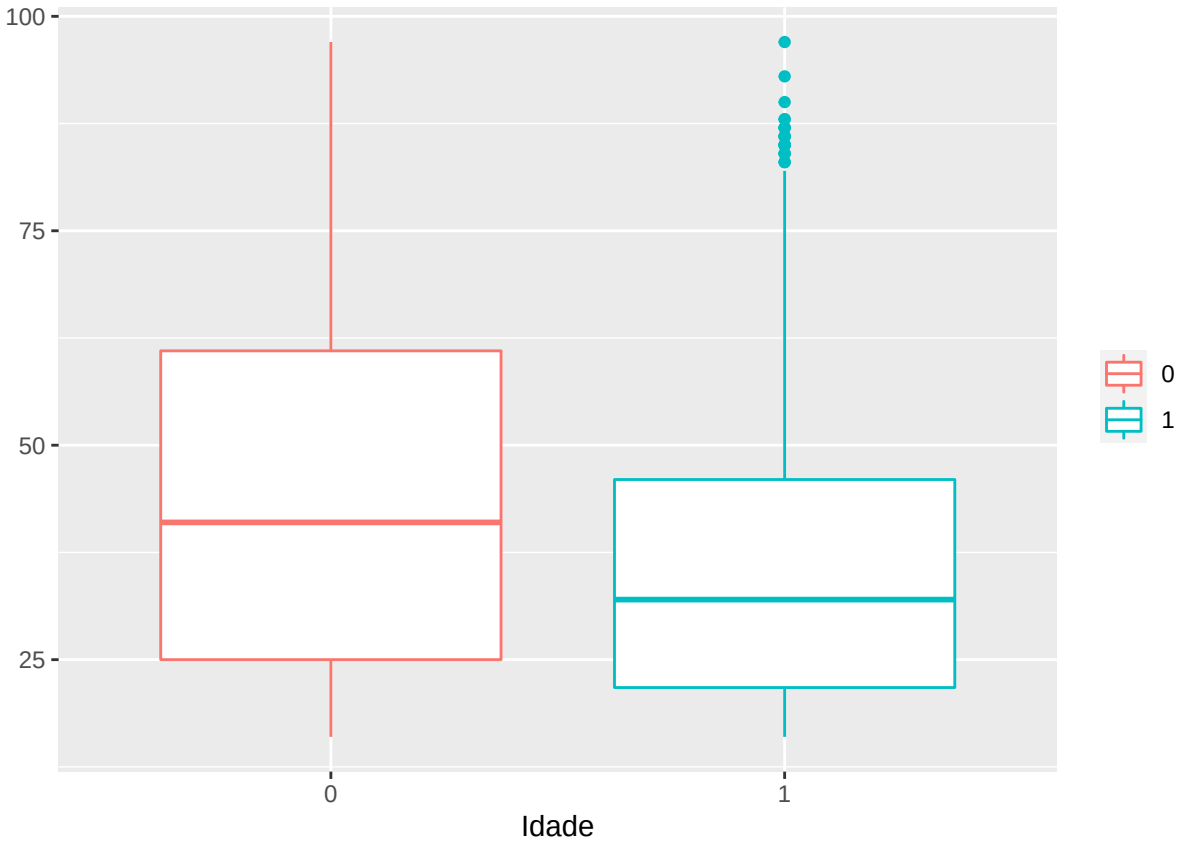
```
## 1  25/39      1   Sim     Sim  Fem    40    driver    Sim
## 2  25/39      1   Sim     Sim Masc    41    driver    Nao
## 3  40/54      1   Sim     Nao Masc    35    driver    Sim
## 4  25/39      1   Sim     Sim Masc    26      pass    Sim
## 5  25/39      1   Sim     Sim  Fem    32    driver    Sim
## 6  25/39      1   Sim     Sim  Fem    41    driver    Sim
## 7  10/24      1   Sim     Sim  Fem    24    driver    Nao
## 8  25/39      1   Nao     Sim Masc    25    driver    Sim
## 9  40/54      1   Nao     Nao  Fem    22    driver    Nao
## 10 10/24      1   Sim     Sim  Fem    47      pass    Nao
```

2 Análise Descritiva

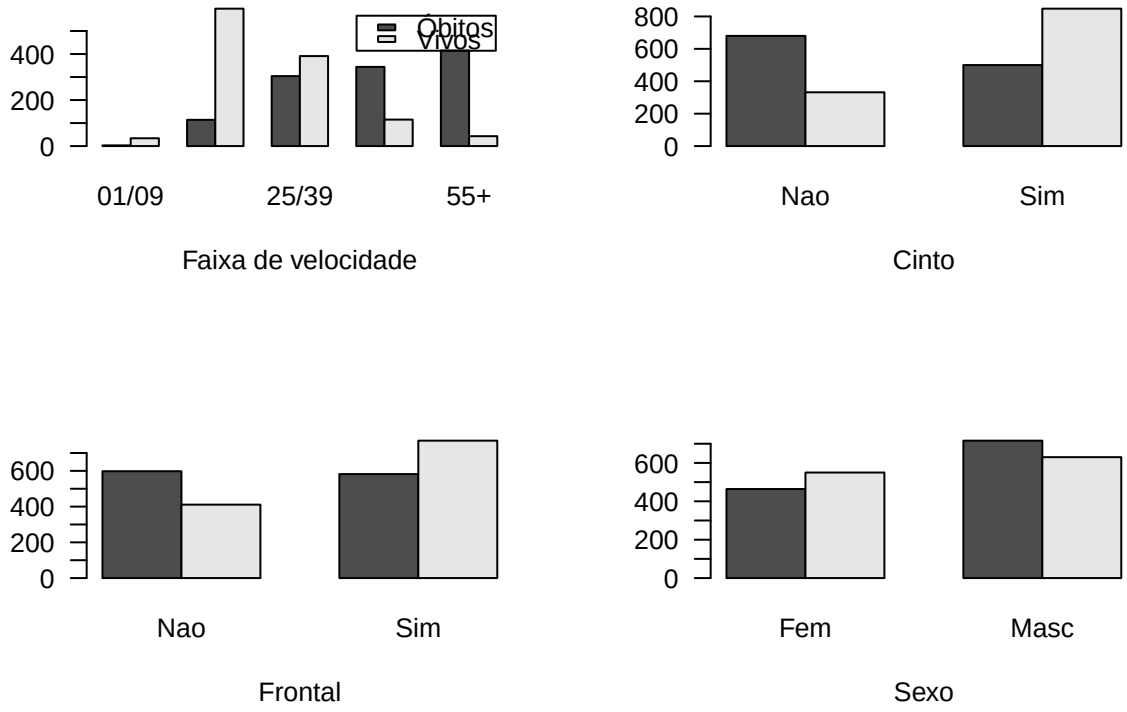
2.1 Medidas de Resumo

Nota-se na varável velocidade uma frequência maior de acidentes na faixa de 25-39 milhas. A maioria estava com cinto de segurança e os acidentes foram a maioria frontais.

2.2 Boxplot



2.3 Gráficos de frequencia



###Intuitivamente sabemos que para nosso escopo a variável idade não é significativa para o nosso modelo porém para comprovar adiante faremos um teste para evidênciar a irrelevancia da variável no modelo.

3. Ajuste do Modelo de Regressão

3.1 Ligação Logito

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Logito. A expressão do modelo é dada por:

$ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 Veloc_i + \beta_2 Sobrev_i + \beta_3 AbFunc_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 Grav_i$

No R, o modelo é declarado da seguinte forma:

3.2 Ligação Probito

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Probito. A expressão do modelo é dada por:

$\phi^{-1}(\pi_i) = \beta_0 + \beta_1 Veloc_i + \beta_2 Sobrev_i + \beta_3 AbFunc_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 Grav_i$

No R, o modelo é declarado da seguinte forma:

```
ajuste2 <- glm(sobrev ~ .,family=binomial(link = 'probit'),data = dados)
```

3.3 Ligação Complemento log-log

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Complemento Log Log. A expressão do modelo é dada por:

$ln[-ln(1-\pi_i)] = \beta_0 + \beta_1 Veloc_i + \beta_2 Sobrev_i + \beta_3 AbFunc_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 Grav_i$

No R, o modelo é declarado da seguinte forma:

```
ajuste3 <- glm(sobrev ~ .,family=binomial(link='cloglog'),data = dados)
```

##3.4 Ligação Cauchy

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Cauchy. A expressão do modelo é dada por:

$tan[\pi_i(\mu_i - 0,5)] = \beta_0 + \beta_1 Veloc_i + \beta_2 Sobrev_i + \beta_3 AbFunc_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 Grav_i$

No R, o modelo é declarado da seguinte forma:

```
ajuste4 <- glm(sobrev ~ .,control=glm.control(epsilon = 1e-8, maxit = 42,
                                              trace = FALSE), family=binomial(link='cauchit'),data = dados)
```

4. Escolha do Modelo

O critério de informação AIC pode também ser utilizado, porém o AIC penaliza o número de parâmetros do modelo. Como os modelos tem o mesmo número de parâmetros, o critério aponta para a mesma direção da verossimilhança pois todos são penalizados da mesma forma.

O modelo que apresentou menor AIC e maior verossimilhança foi o modelo Binomial com função de ligação Cauchy.

5. Análise do Modelo Ajustado Selecionado

##5.1 Resumo do Modelo

LIneu O resumo do modelo ajustado indica que as variáveis adesão marginal, nucléolos nus, cromatina branda, nucléolo normal e espessura do aglomerado estão associadas a uma maior probabilidade de tumor maligno, enquanto as demais variáveis não apresentam relação com a resposta.

##5.2 Reajuste do Modelo

Lineu Como as covariáveis são altamente correlacionadas, é válido inserir as covariáveis uma a uma no modelo para verificar sua significância na presença das outras, tal como o realizado pelo algoritmo stepwise.

Sendo assim, o novo modelo fica da seguinte forma:

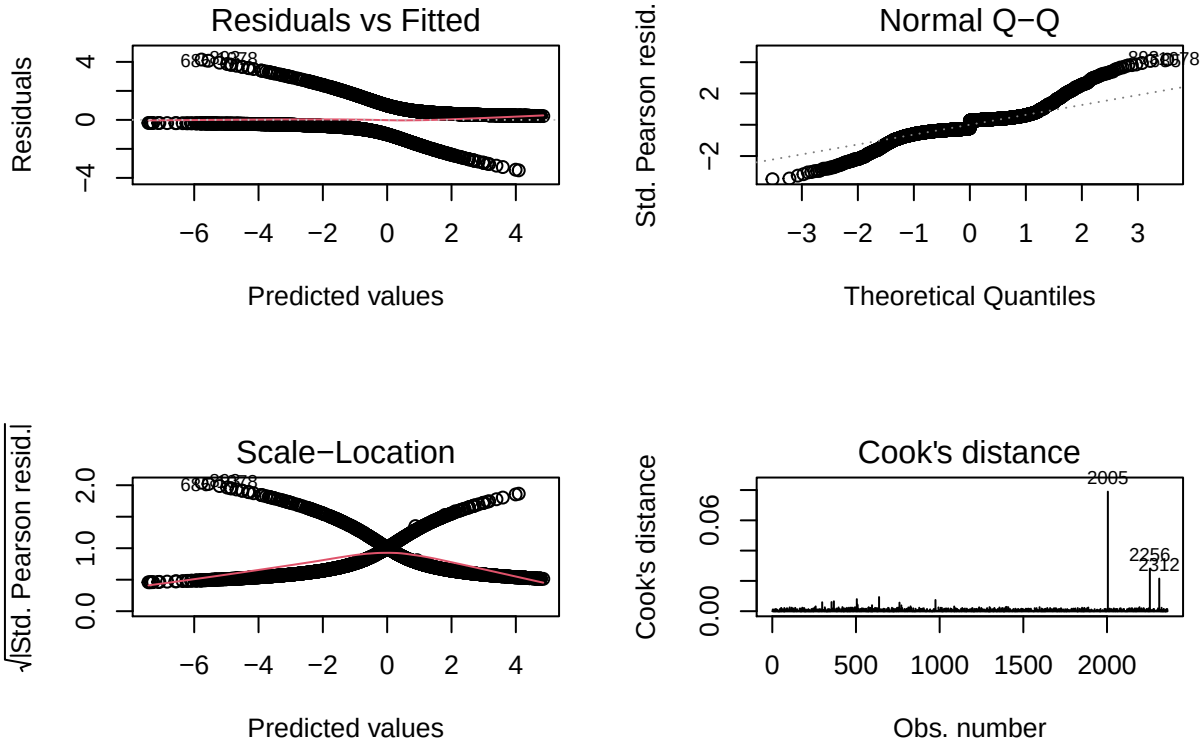
```
ajuste4.1 <- step(ajuste4, direction = "both")
```

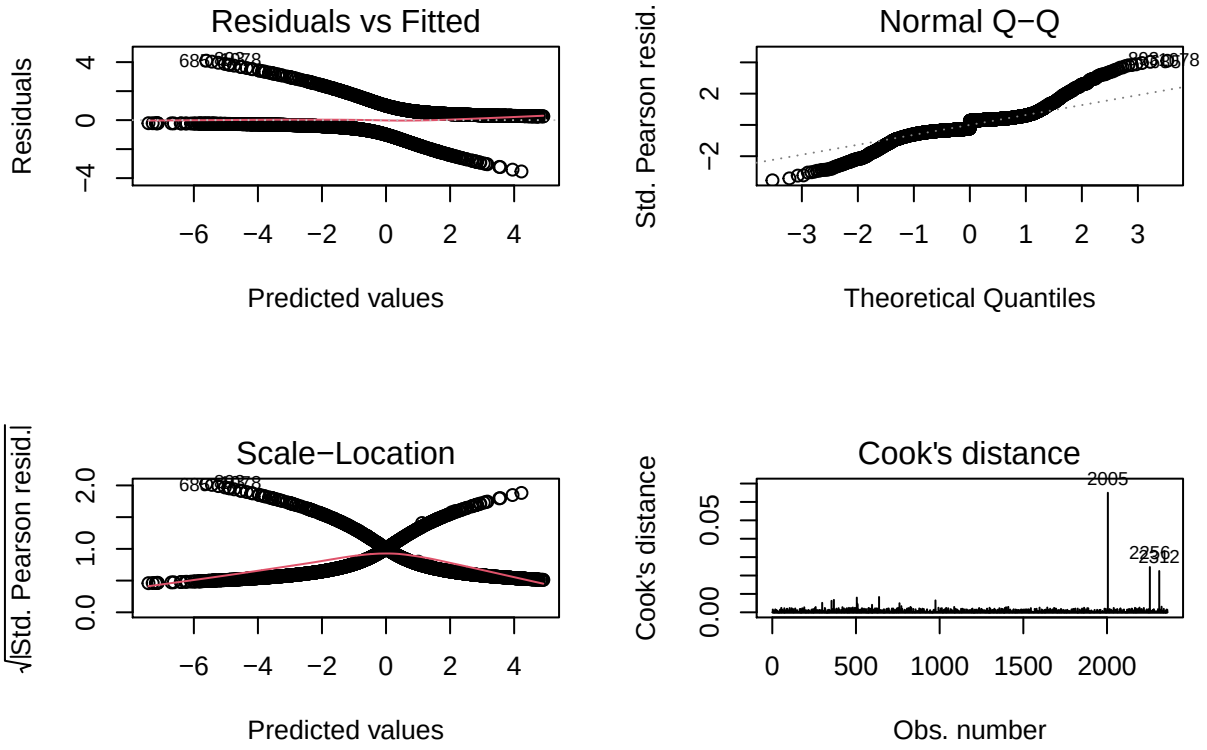
O resumo do novo modelo ajustado:

```
anova(ajuste4, ajuste4.1, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: sobrev ~ veloc + cinto + frontal + sexo + idade + ocupantes +
##     abfunc
## Model 2: sobrev ~ veloc + cinto + frontal + sexo + idade + ocupantes
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      2349     2018.1
## 2      2350     2019.4 -1  -1.2619   0.2613
```

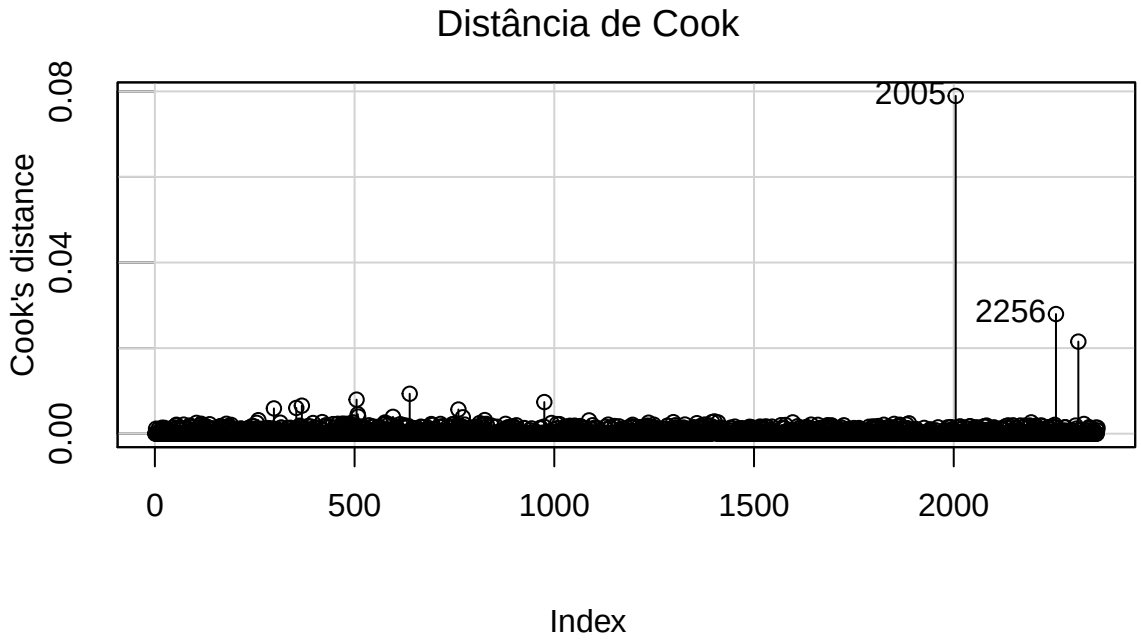
##5.3 Análise de Resíduos



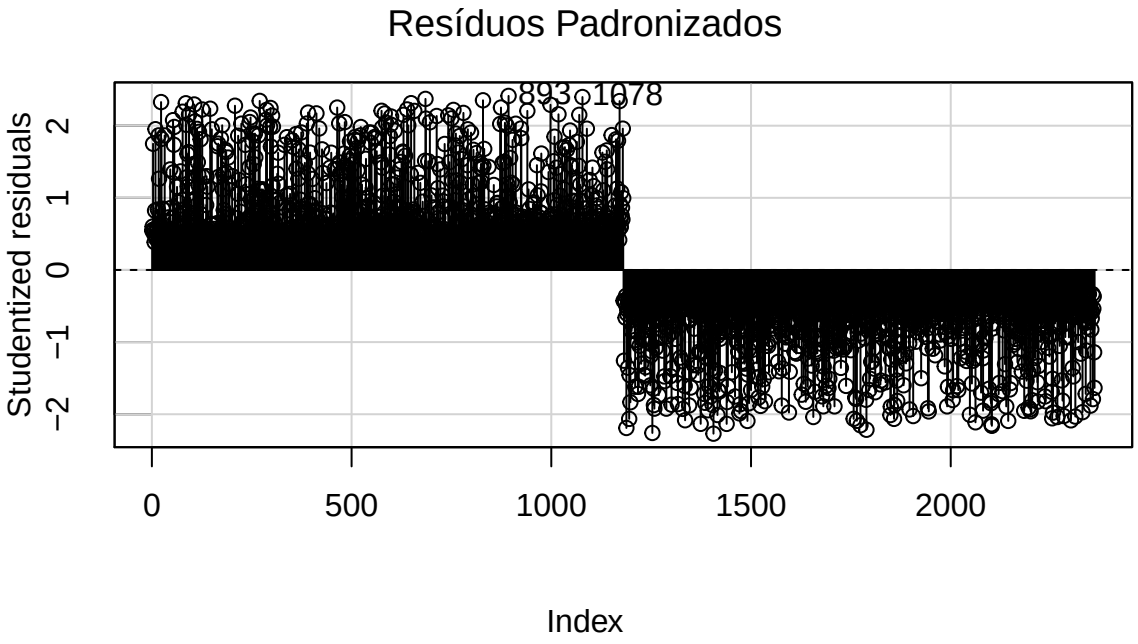


5.4 Medidas de Influencia

```
influenceIndexPlot(ajuste4.1, vars=c("Cook"), main="Distância de Cook")
```




```
influenceIndexPlot(ajuste4.1, vars=c("Studentized"), main="Resíduos Padronizados")
```



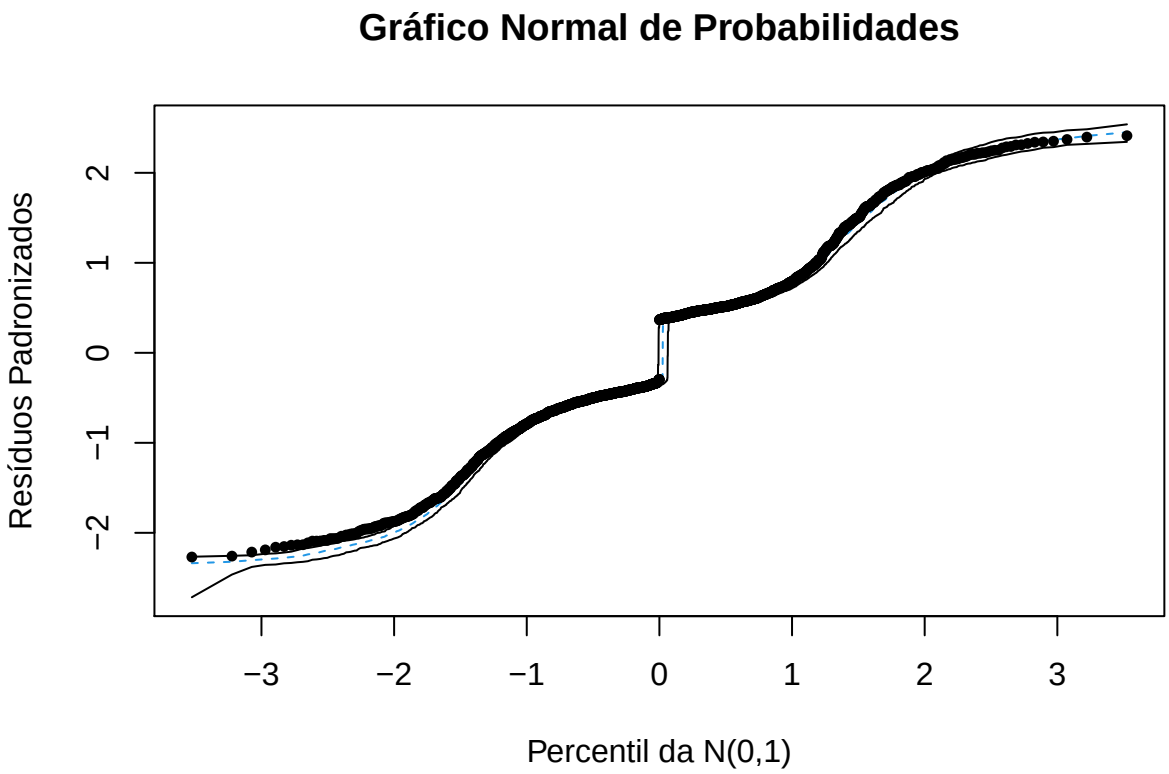
5.5 Resíduos Quantílicos Aleatoriazados

##5.6 Gráfico Normal de Probabilidades com Envelope Simulado

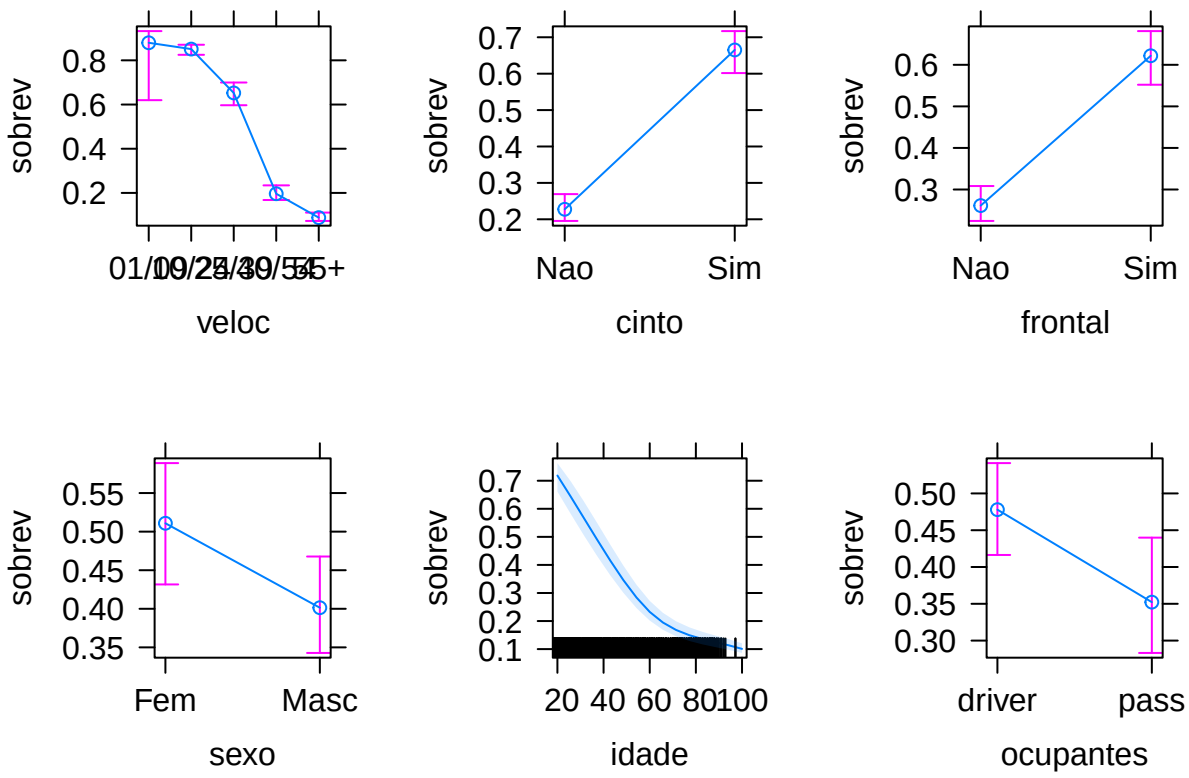
Lineu O gráfico de resíduos simulados permite verificar a adequação do modelo ajustado mesmo que os resíduos não tenham uma aproximação adequada com a distribuição Normal. Neste tipo de gráfico espera-se, para um modelo bem ajustado, os pontos (resíduos) dispersos aleatoriamente entre os limites do envelope.

Deve-se ficar atento à presença de pontos fora dos limites do envelope ou ainda a pontos dentro dos limites porém apresentando padrões sistemáticos.

Vamos utilizar a função envelope implementada pelo professor Cesar Augusto Taconeli :



5.7 Gráficos de Efeitos



6. PREDIÇÃO

7. AVALIAÇÃO DO PODER PREDITIVO DO MODELO

Como temos uma base de tamanho razoável para fins preditivos, uma alternativa é separar a base em duas: uma para o ajuste do modelo, com 70% dos dados (com 477 observações) e outra para validação, com 30% (com 203 observações).

7.1 Divisão da Base de dados

```
set.seed(1909)
indices <- sample(1:680, size = 477)
dadosajuste <- dados[indices,]
dadosvalid <- dados[-indices,]
```

7.2 Ponto de Corte

Como estamos modelando a probabilidade de tumor maligno, vamos estabelecer o ponto de corte 0.5, isso é, se a probabilidade estimada for maior que este valor o tumor será classificado como maligno. Vamos armazenar os valores preditos do modelo para os dados de validação:

```r
pred <- predict(ajuste4.1, newdata = dadosvalid, type = 'response')
corte <- ifelse(pred > 0.5, 'maligno', 'benigno')
```

7.3 Sensibilidade e Especificidade

Para fazer uso dos dados de validação, dois conceitos são necessários: sensibilidade e especificidade.

Define-se por sensibilidade a capacidade do modelo de detectar tumores malignos, ou seja, de classificar como malignos os tumores que de fato o são .

Já a especificidade é a capacidade do modelo de detectar classificar como benignos tumores verdadeiramente benignos.

A sensibilidade é dada por

```r
#sens <- dados[2,2]/sum(dados[,2])
#sens
```

7.4 Curva ROC

7.5 Outra Alternativa de validação

8. REFERÊNCIAS