

Matéria de Análise de Dados – Ciências Biomédicas

Aula 12 – Regressão Logística

James Hunter

28 de março de 2017

Nesta aula, examinaremos **regressão logística**, uma forma de regressão que usamos frequentemente em bioestatística. Como regressão polinomial, regressão logística é uma extensão do conceito da regressão linear. Porém, em regressão logística, reduzimos o extenso da variável dependente para forma **binomial**.

Ou seja, desenvolvemos um modelo que tenta prever se uma condição existe ou não existe. Baseado nas condições genotípicos e fenotípicos, nós tentamos prever se um paciente tem ou não tem uma doença. Um exemplo específico: uma aluna quer entender se um paciente terá o tropismo R5 ou X4 baseado nos níveis de vários fatores. Em outras palavras, trabalhamos com uma variável dependente com dois estados, 0 ou 1. O modelo em si medirá a probabilidade que o estado 1 aconteceria.

Como podemos torcer o modelo de regressão linear para acomodar o limite de TRUE ou FALSE, 1 ou 0, **infetado** ou **não infetado**? Lembrando regressão polinomial, onde aumentamos um ou mais termos para a parte independente do modelo, nós transformávamos uma curva não linear numa *expressão* linear que satisfazia as premissas de regressão, especialmente o requisito que o modelo seja linear.

Em regressão logística, nós fazemos uma coisa parecida. Aplicamos uma função “link” para converter probabilidades em uma linha. Esta função, chamada a **logit** está aplicada para os valores da variável Y (a dependente). Este expressa o modelo logística na forma do inverso do logaritmo da relação de odds (“inverted log odds ratio”) que o evento dependente ocorrerá.

Regressão Linear (usando a notação de matrizes)

$$y = X\beta + \epsilon_i$$

Este quer dizer que um matriz de valores X está sendo multiplicado por um vetor de coeficientes β . Sabemos isso porque X fica em maiúsculo e β em minúsculo, a anotação tradicional para álgebra linear.

Regressão Logística

$$p(y_i = 1) = \text{logit}^{-1}(X_i\beta) + \epsilon_i$$

Esta equação diz que estamos procurando a probabilidade que a variável dependente ter o estado de ‘1’ e que este depende nos variáveis independentes (no matriz X), transformados pela função **logit**. O **logit** de uma probabilidade é o logaritmo dos odds da variável assumindo o valor 1.

A função **logit** invertido (a versão que usamos; veja o “-1” como o expoente na formula) tem a forma que limite os valores para os limites naturais de probabilidades: 0 e 1. Alias, a função em si pode assumir qualquer valor real, mas a probabilidade sempre sempre cairá no intervalo $[0, 1]$.

$$\text{logit}^{-1}(x) = \frac{1}{1 + e^{-x}}$$

Modelos Lineares Gerais (*General Linear Models*)

Uma regressão logística faz parte de uma classe dos modelos chamados *general linear models (GLM)*, ou seja, eles manipulam os matrizes dos parâmetros numa maneira diferente dos modelos lineares simples, que são um caso específico de GLM. Como em regressão linear múltipla, o modelo usa uma combinação linear de variáveis independentes (também chamadas “covariates”).

Esses modelos usam a função `glm` invés de `lm` para os cálculos, mas o output dos modelos parece quase parecido com o output dos modelos lineares que vimos até agora.

Cálculos dos Coeficientes nos GLM

Lembramos que a regressão linear usou o método de *mínimos quadrados* para determinar os coeficientes dos modelos. Com regressão logística, precisamos utilizar outro método porque agora nosso objetivo não é minimizar a diferenças entre os valores de Y calculados e os observados. Agora, queremos maximizar a probabilidade de obter os valores da variável dependente observados. O software avalia a contribuição de cada caso para o probabilidade (*likelihood*) que Y ficaria igual a 1. Porque os valores dependentes são binomiais e são determinados independentemente, a probabilidade ($l(\beta)$) é o produto (\prod) das probabilidades dos casos:

$$l(\beta) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

Usando esta função, o software maximiza as probabilidades fazendo iterações até que chega numa probabilidade máxima.

Nosso Primeiro Problema para Hoje

Primeiro, nós vamos considerar um caso simples. Este é um estudo de 100 pacientes que ou têm ou não têm doença cardíaca coronária – “coronary heart disease” (CHD). O estudo está interessado na relação entre a idade do paciente e a CHD. (Esses dados vêm de Hosmer & Lemeshow, *Applied Logistic Regression* (2a Ed.), 2000, p.2)

Carregar os Pacotes Necessários

```
suppressMessages(library(tidyverse))
suppressPackageStartupMessages(library(DescTools))
suppressPackageStartupMessages(library(knitr))
suppressPackageStartupMessages(library(car))
suppressPackageStartupMessages(library(psych))
suppressPackageStartupMessages(library(broom))
suppressPackageStartupMessages(library(nortest))
suppressPackageStartupMessages(library(coefplot))
suppressMessages(library(mosaic))
options(scipen = 5)

chdage <- read_csv("chdage.csv")

## Parsed with column specification:
## cols(
##   id = col_integer(),
```

```
## idade = col_integer(),
## chd = col_integer()
## )
```

Estrutura do chdage

Vamos ver quais variáveis são disponíveis para julgar se precisa-se desta renda para ocupar uma casa listado na sondagem. Podemos também fazer um estudo rápido exploratório dos dados para ver se podemos perceber uma tendência nos dados.

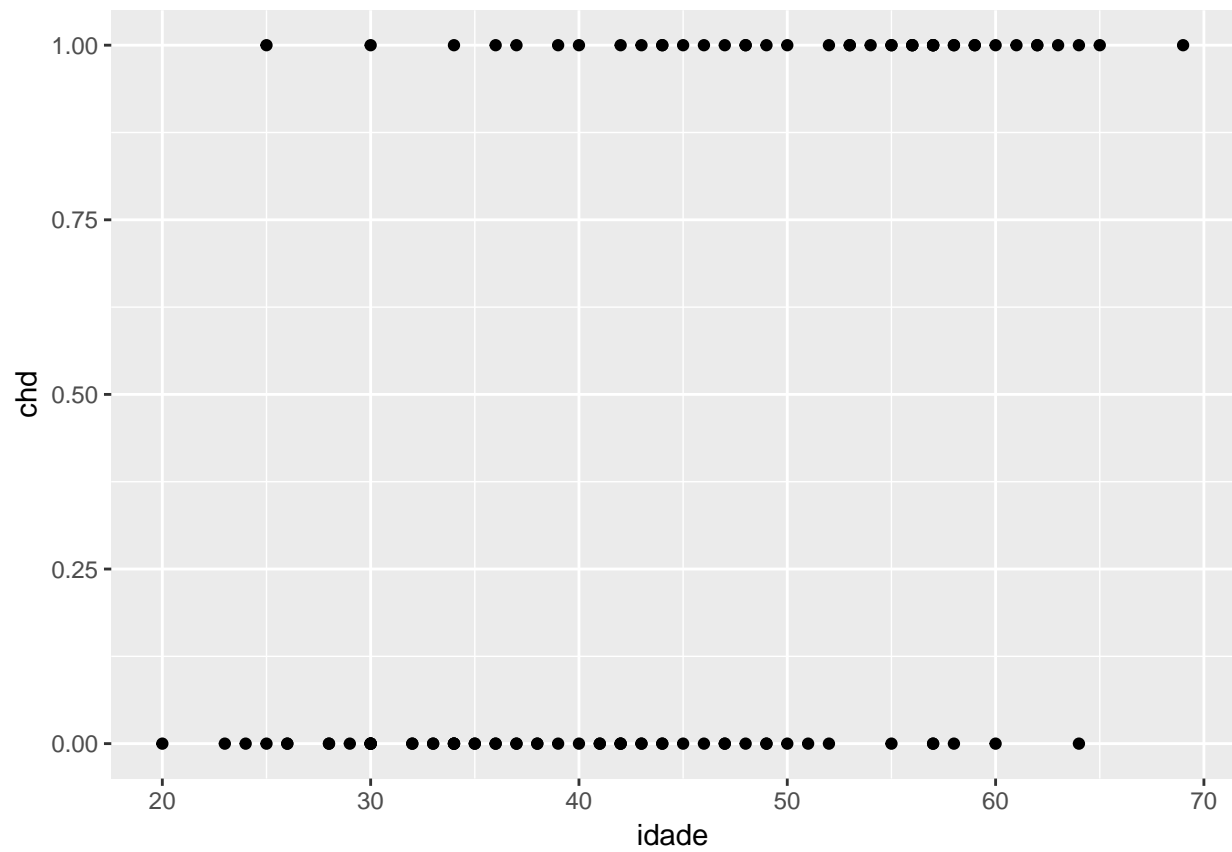
```
glimpse(chdage)
```

```
## Observations: 100
## Variables: 3
## $ id      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ idade <int> 20, 23, 24, 25, 25, 26, 26, 28, 28, 29, 30, 30, 30, 30, ...
## $ chd    <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,...
```

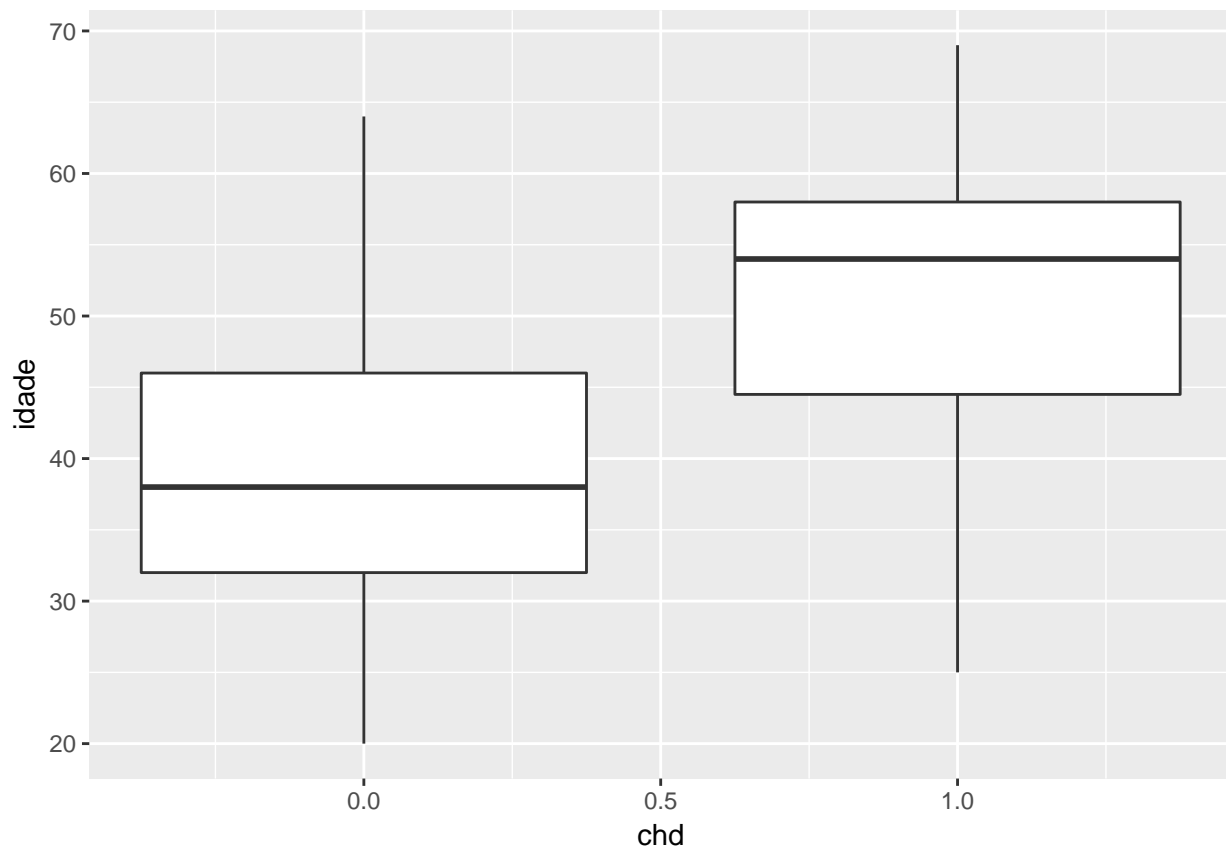
```
Desc(chd ~ idade, data = chdage, plotit = FALSE)
```

```
## -----
## chd ~ idade
##
## Summary:
## n pairs: 100, valid: 100 (100.0%), missings: 0 (0.0%), groups: 2
##
##
##          0          1
## mean    39.175    51.279
## median   38.000    54.000
## sd       10.202     9.979
## IQR      14.000    13.500
## n         57         43
## np       57.000%   43.000%
## NAs        0         0
## Os         0         0
##
## Kruskal-Wallis rank sum test:
##   Kruskal-Wallis chi-squared = 26.213, df = 1, p-value = 0.0000003057
##
##
## Proportions of chd in the quantiles of idade:
##
##          Q1          Q2          Q3          Q4
## 0    88.0%    70.4%    45.8%    20.8%
## 1    12.0%    29.6%    54.2%    79.2%
```

```
chdscat <- ggplot(data = chdage, aes(y = chd, x = idade)) + geom_point()
chdscat
```



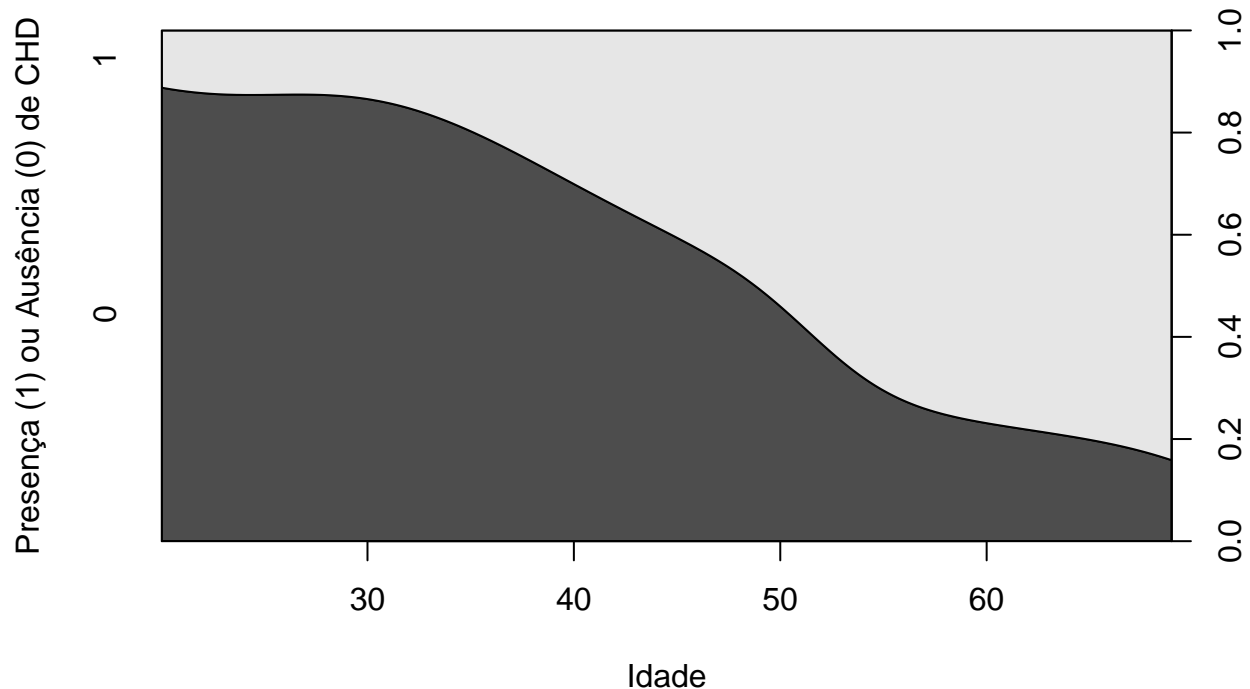
```
chdbox <- ggplot(data = chdage, aes(x = chd, y = idade, group = chd))  
chdbox <- chdbox + geom_boxplot()  
chdbox
```



Nós podemos também usar um gráfico de densidade condicional (“Conditional Density Plot”) para entender como a CHD varia com idade. O gráfico seguinte mostra que começando com mais ou menos 35 anos, os pacientes teve mais ocorrências de CHD e depois de 50 anos a proporção dos pacientes sofrendo CHD supera 50%, aumentando para 80% antes de 67 (a idade máximo dos pacientes na amostra).

```
cdplot(factor(chd) ~ idade, data = chdage,  
       main = "Densidade Condicional de Idade sobre CHD",  
       xlab = "Idade", ylab = "Presença (1) ou Ausência (0) de CHD")
```

Densidade Condicional de Idade sobre CHD



A análise indica que a idade média com CHD parece mais alta que a idade que não sofrem da doença. O scatterplot tradicional não mostra isso claramente porque todos os pontos são agrupados em 0 e 1 no eixo Y, os únicos valores que existem. Então um boxplot mostra melhor a diferença em idade. Mas, também a grande variabilidade em CHD entre as idades atrapalha uma visão clara da relação entre idade e CHD.

Uma maneira que podemos controlar essa variabilidade melhor é criar intervalos (grupos de idade) para variável independente e olhar na proporção em cada grupo que sofre CHD. Nós vamos criar uma variável `idgrp` que vai agrupar idades nas categorias seguintes utilizando a função `Recode` de pacote `car` que oferece mais flexibilidade na especificação das substituições que `recode` de `dplyr`:

- 20 - 29 anos
- 30 - 34 anos
- 35 - 39 anos
- 40 - 44 anos
- 45 - 49 anos
- 50 - 54 anos
- 55 - 59 anos
- 60 - 69 anos

```
chdage$idgrp <- Recode(chdage$idade, "20:29 = '20-29'; 30:34 = '30-34';
  35:39 = '35-39'; 40:44 = '40-44'; 45:49 = '45-49';
  50:54 = '50-54'; 55:59 = '55-59'; 60:69 = '60-69'",
  as.factor.result = TRUE)
```

```
kable(table(chdage$idgrp, chdage$chd))
```

	0	1
20-29	9	1
30-34	13	2
35-39	9	3
40-44	10	5

	0	1
45-49	7	6
50-54	3	5
55-59	4	13
60-69	2	8

```
gmodels::CrossTable(chdage$idgrp, chdage$chd, chisq = TRUE,
  prop.c = FALSE, prop.t = FALSE,
  prop.chisq = FALSE, format = "SPSS")
```

```
## Warning in chisq.test(t, correct = FALSE, ...): Chi-squared approximation
## may be incorrect
```

```
##
```

```
## Cell Contents
```

```
## |-----|
## |                Count |
## |                Row Percent |
## |-----|
```

```
##
```

```
## Total Observations in Table: 100
```

```
##
```

chdage\$idgrp	chdage\$chd		Row Total
	0	1	
20-29	9	1	10
	90.000%	10.000%	10.000%
30-34	13	2	15
	86.667%	13.333%	15.000%
35-39	9	3	12
	75.000%	25.000%	12.000%
40-44	10	5	15
	66.667%	33.333%	15.000%
45-49	7	6	13
	53.846%	46.154%	13.000%
50-54	3	5	8
	37.500%	62.500%	8.000%
55-59	4	13	17
	23.529%	76.471%	17.000%
60-69	2	8	10
	20.000%	80.000%	10.000%
Column Total	57	43	100

```
##
```

```
##
```

```
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## -----
## Chi^2 = 26.63705      d.f. = 7      p = 0.0003873022
##
##
##
##      Minimum expected frequency: 3.44
## Cells with Expected Frequency < 5: 4 of 16 (25%)
```

A segunda tabela, um `CrossTable` do pacote `gmodels` mostra as proporções de cada fileira da tabela no mesmo formato que o SPSS usa.

Podemos agora construir o modelo, que vamos fazer em duas versões, em como idade na forma numérica e outra na forma categórica.

O Modelo

A `glm` usa a mesmo formato de formula para especificar as variáveis que a `lm`. Separamos a variável dependente do independente com um til `~` e os várias variáveis independentes com sinais de mais `+` (que não precisamos neste caso). Depois de avisar o modelo em que data frame para achar as variáveis (`data =`), nós vamos especificar uma família de dos modelos gerais que queremos usar e qual seria a função “link” para determinar como o modelo deve ser calculado. Neste caso, nossa função link é a função `logit` que descrevi antes. O que a função “link” faz é de ligar a variável dependente que tem a forma *binomial* às variáveis independentes.

Versão 1 – idade como uma variável numérica

```
chdfit1 <- glm(chd ~ idade, data = chdage,
              family = binomial(link = "logit"))
```

Versão 1 – Resultados

Olhamos nestes resultados. Na mesma maneira que precisamos imprimir o resumo do modelo para `lm`, assim precisamos fazer com `glm`. Depois, vamos mostrar uma plotagem chamada `coefplot`, que apresenta os coeficientes do modelo na forma gráfica. Esta função vem do pacote `coefplot`. Vamos olhar nesses resultados e explicarei o que é diferente da regressão linear.

```
summary(chdfit1)
# library(coefplot) não está funcionando certo
# coefplot(chdfit1)
```

Os Coeficientes

A apresentação dos coeficientes é parecido com o que já conhecemos. Têm estimativa, erro padrão, valor-z e valor-p. Os valores p indicam que a contribuição da variável `idade` ao modelo foi significativo. Mas, como interpretá-los.

Os coeficientes em si representam o log odds que o resultado $Y = 1$. Em nosso caso, que a o paciente tem CHD. Para entender os coeficientes do modelo melhor, precisamos reverter o logit invertido e calcular o *logit inverso*. Nós vamos criar uma função para fazer este cálculo para os coeficientes.


```
invlogit <- function(x) {
  1/(1 + exp(-x))
}
invlogit(coef(chdfit1)[2])
```

```
##      idade
## 0.5277019
```

Nós podemos agora interpretar os coeficientes em termos de probabilidades. A idade tem uma probabilidade acima de 0.50. Com uma probabilidade acima de 0.50, podemos dizer que uma relação provavelmente existe entre idade e a presença de CHD ($Y = 1$). Mas, não oferece muito mais informação sobre quais são as probabilidades para cada grupo de idade.

Vamos montar o modelo com os grupos que criávamos antes.

Versão 2 – Modelo com idgrp

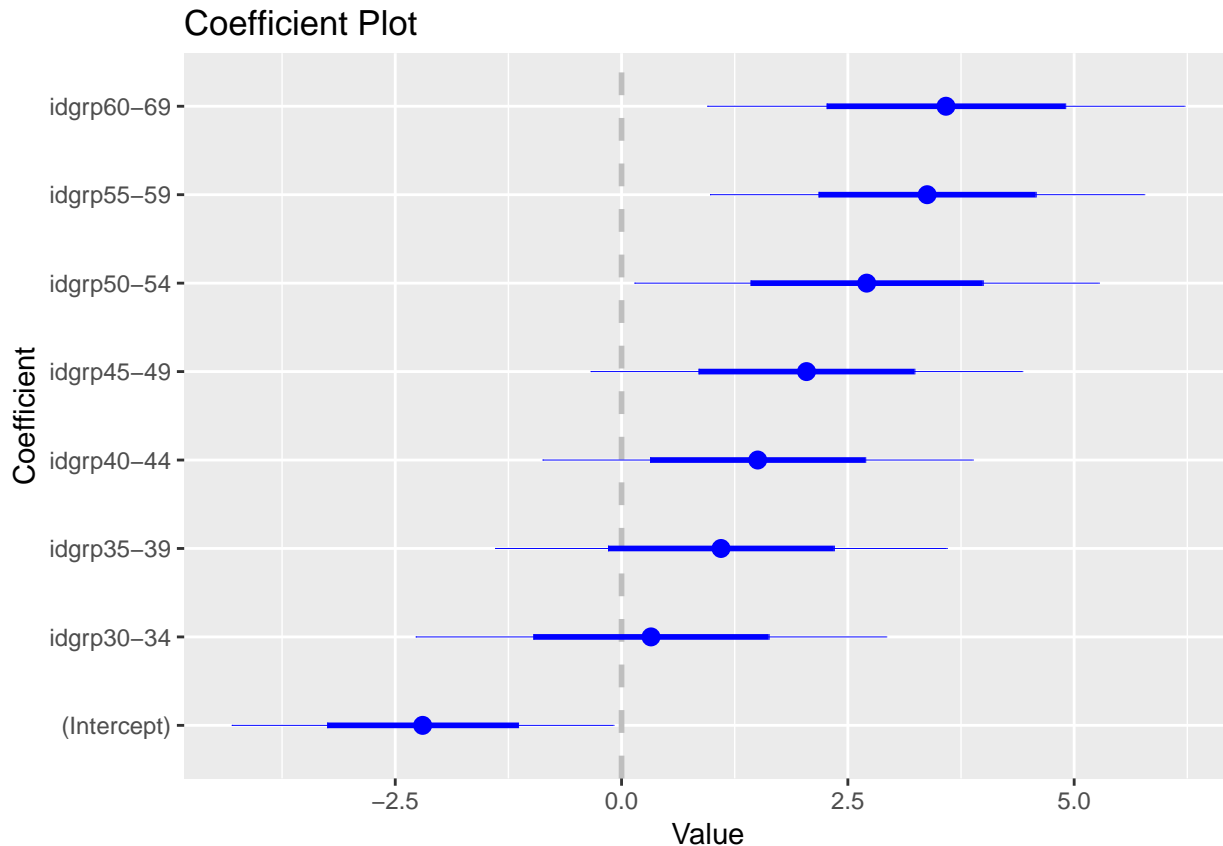
```
chdfit2 <- glm(chd ~ idgrp, data = chdage,
  family = binomial(link = "logit"))
summary(chdfit2)
```

```
##
## Call:
## glm(formula = chd ~ idgrp, family = binomial(link = "logit"),
##      data = chdage)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7941  -0.9005  -0.4590   0.7325   2.1460
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.1972     1.0540  -2.085  0.03710 *
## idgrp30-34    0.3254     1.2992   0.250  0.80221
## idgrp35-39    1.0986     1.2471   0.881  0.37837
## idgrp40-44    1.5041     1.1878   1.266  0.20543
## idgrp45-49    2.0431     1.1918   1.714  0.08649 .
## idgrp50-54    2.7081     1.2823   2.112  0.03470 *
## idgrp55-59    3.3759     1.1991   2.815  0.00487 **
## idgrp60-69    3.5835     1.3175   2.720  0.00653 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 136.66  on 99  degrees of freedom
## Residual deviance: 107.96  on 92  degrees of freedom
## AIC: 123.96
##
## Number of Fisher Scoring iterations: 4
```

Versão 2 – Resultados

Agora, os resultados oferecem mais informação. Os grupos de idade acima de 50 anos todos são significativos. O valor-p deles fica abaixo da α assumido de 0.05. Se nós convertemos os coeficientes desses grupos de idade significantes em probabilidades usando nossa função `invlogit`, podemos ver quais categorias têm uma probabilidade acima de 0.50 de ter CHD.

```
coefplot(chdfit2)
```



```
invlogit(coef(chdfit2)[6:8])
```

```
## idgrp50-54 idgrp55-59 idgrp60-69  
## 0.9375000 0.9669421 0.9729730
```

Como estes valores indicam, a probabilidade é muito alta que pessoas nessas faixas de idade teria CHD, se consideramos só esta variável independente.

Desvio e AIC

Também temos equivalentes ao R^2 . Esses medem o poder explicativo do modelo, neste caso o **desvio residual** (*residual deviance*) e o **AIC** (*Akaike's Information Criterion*). São medidas da qualidade do modelo. Nós queremos um desvio residual menor que possível. O AIC combina vários elementos da qualidade do modelo para criar um valor que pode usar para comparar um modelo contra um outro. Você vai preferir o modelo com o menor AIC.

Em nossos modelos, o desvio residual e o AIC são basicamente iguais nos dois casos porque os modelos estão considerando os mesmos dados. No próximo exemplo, nós podemos ver que

Exemplo com Múltiplas Variáveis Independentes

Vamos considerar um outro dataset que trata de CHD. Neste caso, temos várias variáveis independentes que podemos usar para prever a aparência da doença. Neste caso, temos 65 casos em que os médicos gravaram as variáveis seguintes:

- id (Número de identificação do caso)
- idade (em anos)
- bmi (índice de massa corporal em kg/m^2)
- genero (0 = masculino, 1 = feminino)
- chd (Ocorrência ou não de um evento cardíaco)

A variável dependente é a `chd`. Primeiro, vamos colocar os dados na memória. Os dados ficam num arquivo de R, `riscochd.RData`. Depois, fazemos um pequeno estudo exploratório.

Análise Exploratória

```
load("riscochd.RData")
Desc(riscochd$chd, plotit = FALSE)
```

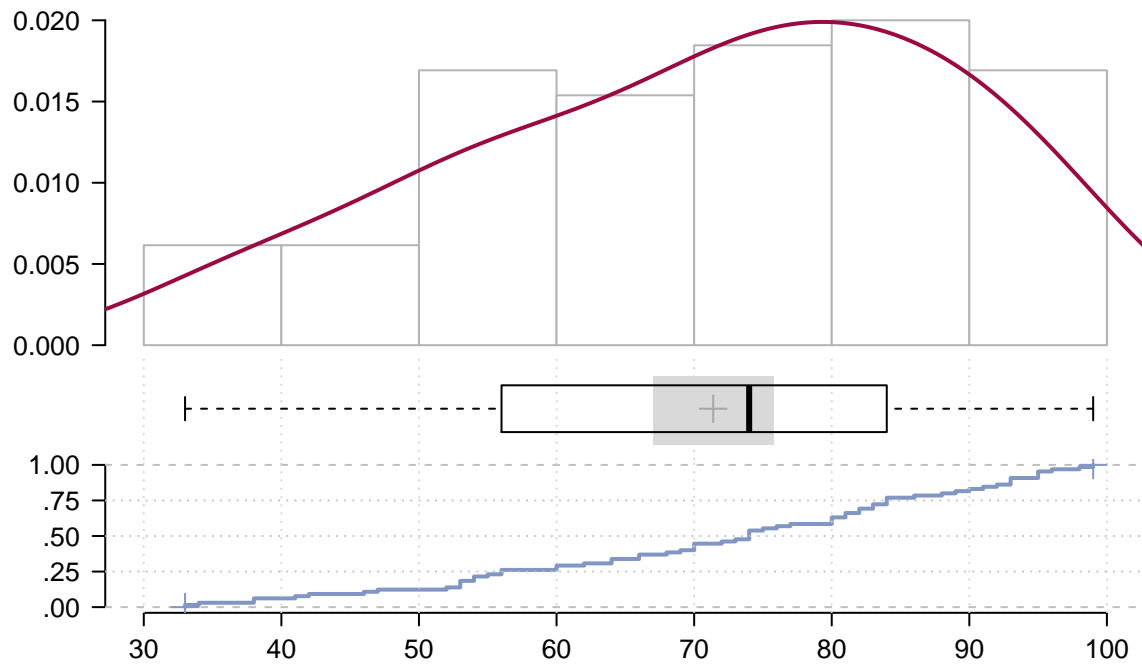
```
## -----
## riscochd$chd (numeric)
##
##   length      n    NAs  unique    Os  mean  meanCI
##      65      65      0      2    33  0.49   0.37
##           100.0%  0.0%           50.8%           0.62
##
##   .05   .10   .25  median   .75   .90   .95
##   0.00   0.00   0.00   0.00   1.00   1.00   1.00
##
##   range    sd  vcoef    mad    IQR  skew   kurt
##    1.00   0.50  1.02    0.00   1.00  0.03  -2.03
##
##
##   level  freq  perc  cumfreq  cumperc
## 1      0    33  50.8%      33   50.8%
## 2      1    32  49.2%      65  100.0%
```

```
Desc(riscochd$idade, plotit = TRUE)
```

```
## -----
## riscochd$idade (integer)
##
##   length      n    NAs  unique    Os  mean  meanCI
##      65      65      0      41      0  71.38   67.01
##           100.0%  0.0%           0.0%           75.76
##
##   .05   .10   .25  median   .75   .90   .95
##  38.60  46.40  56.00  74.00  84.00  93.00  95.00
##
##   range    sd  vcoef    mad    IQR  skew   kurt
##   66.00  17.67  0.25   20.76  28.00 -0.40  -0.83
##
##
## lowest : 33, 34, 38 (2), 41, 42
```

```
## highest: 93 (3), 95 (3), 96, 98, 99
```

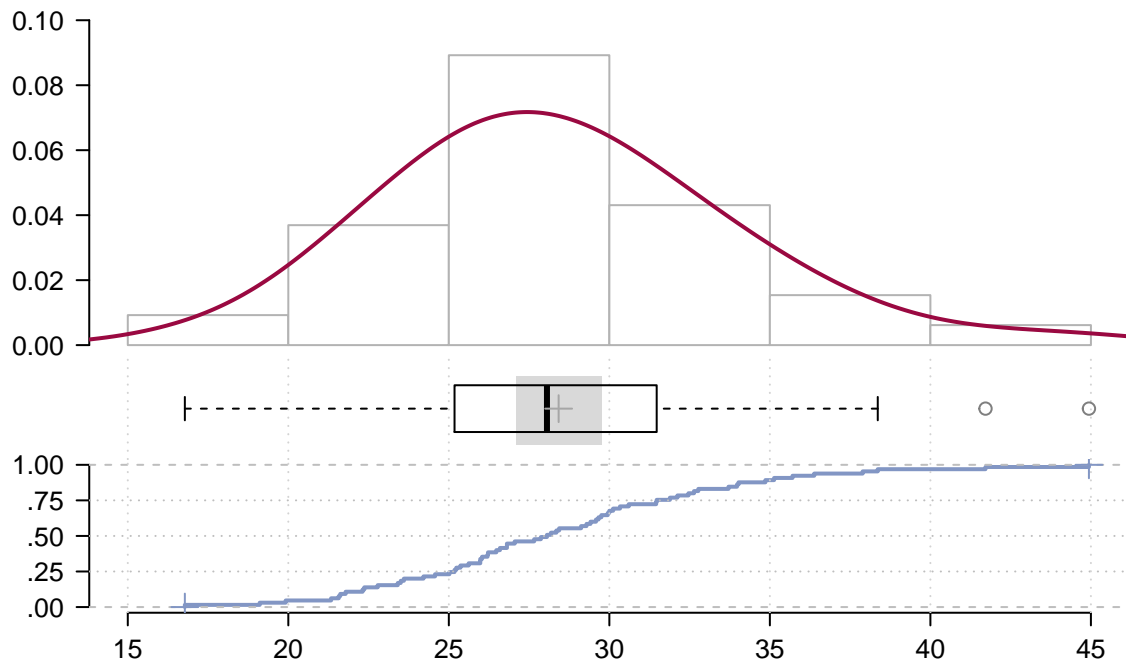
riscochd\$idade (integer)



```
Desc(riscochd$bmi, plotit = TRUE)
```

```
## -----
## riscochd$bmi (numeric)
##
##      length      n      NAs  unique      0s      mean  meanCI
##         65      65       0      64       0 28.42026 27.09293
##         100.0%   0.0%       0.0%
##
##      .05      .10      .25  median      .75      .90      .95
## 21.38031 21.99739 25.17772 28.05630 31.47445 35.02323 37.58896
##
##      range      sd      vcoef      mad      IQR      skew      kurt
## 28.16117  5.35673  0.18848  5.04155  6.29673  0.54136  0.45599
##
## lowest : 16.77718, 19.10959, 19.91878, 21.33267, 21.57087
## highest: 36.38134, 37.89087, 38.36074, 41.71647, 44.93835
```

riscochd\$bmi (numeric)

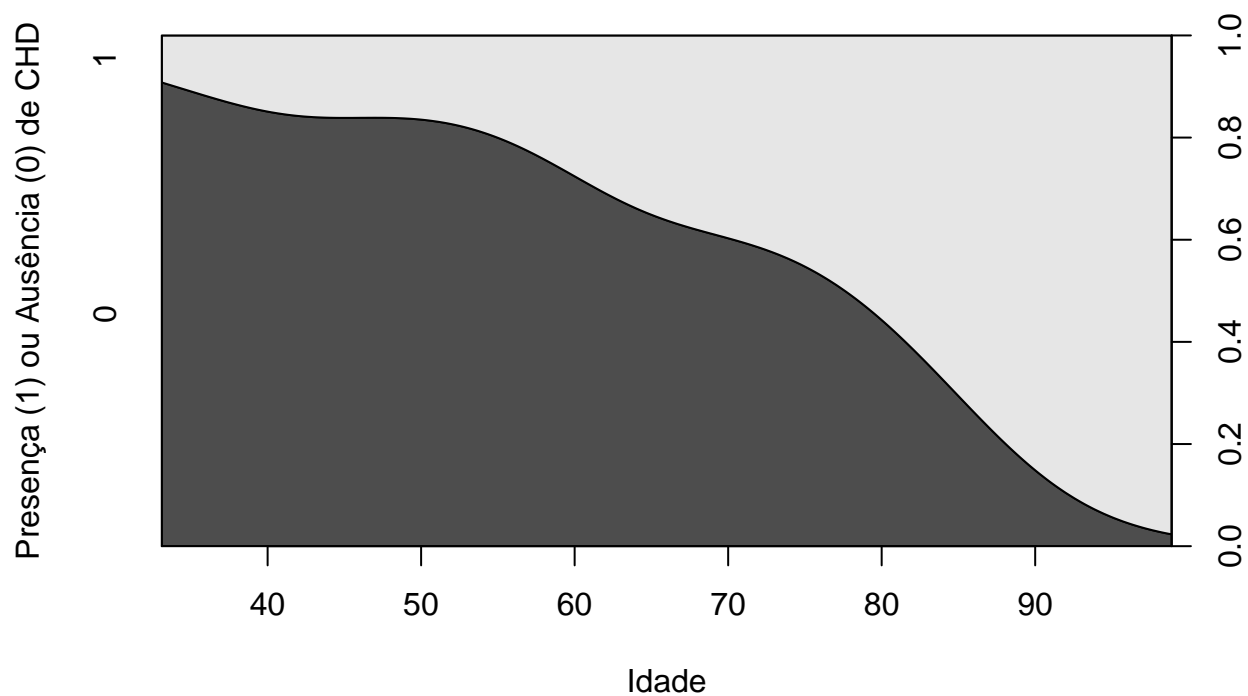


```
Desc(riscochd$genero, plotit = FALSE)
```

```
## -----
## riscochd$genero (integer - dichotomous)
##
##   length      n    NAs unique
##      65      65      0      2
##      100.0%   0.0%
##
##   freq  perc  lci.95  uci.95'
##  0    41 63.1%  50.9%  73.8%
##  1    24 36.9%  26.2%  49.1%
##
## ' 95%-CI Wilson
```

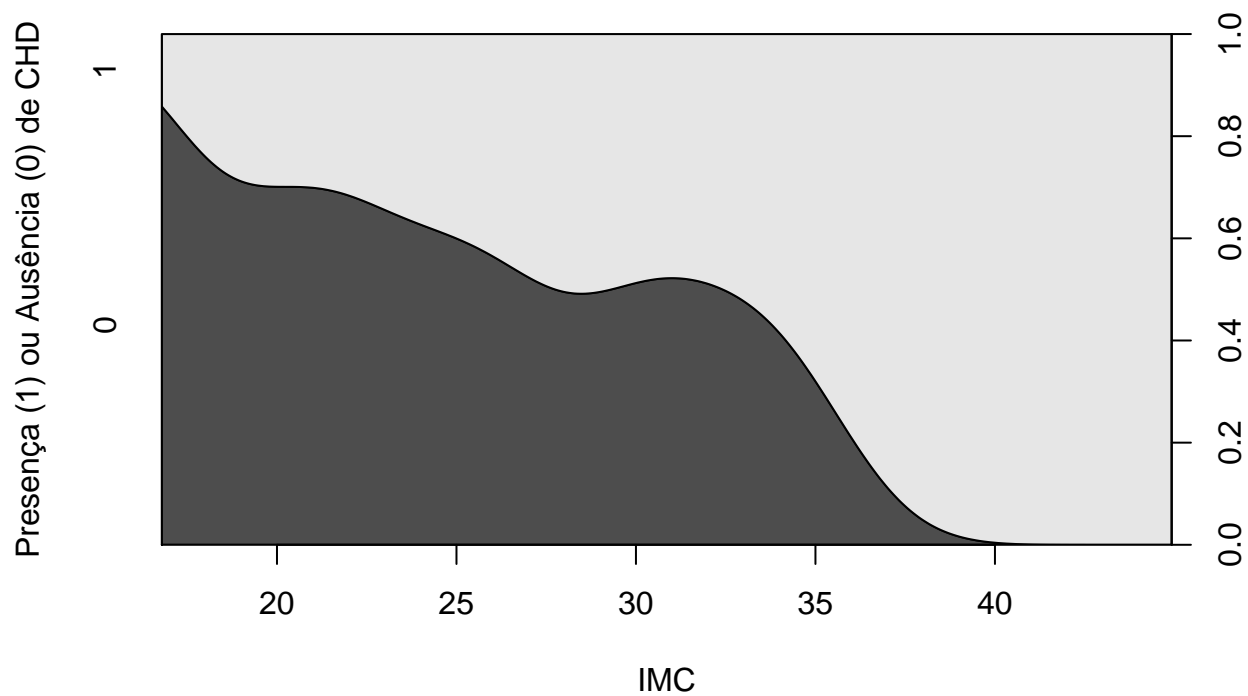
```
cdplot(factor(chd) ~ idade, data = riscochd,
        main = "Densidade Condicional de Idade sobre CHD",
        xlab = "Idade", ylab = "Presença (1) ou Ausência (0) de CHD")
```

Densidade Condicional de Idade sobre CHD



```
cdplot(factor(chd) ~ bmi, data = riscochd,  
       main = "Densidade Condicional de IMC sobre CHD",  
       xlab = "IMC", ylab = "Presença (1) ou Ausência (0) de CHD")
```

Densidade Condicional de IMC sobre CHD



Modelo 1 – Todas as Variáveis Independentes

```
chdfit3 <- glm(chd ~ idade + bmi + genero, data = riscochd)
summary(chdfit3)

##
## Call:
## glm(formula = chd ~ idade + bmi + genero, data = riscochd)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68623  -0.25981   0.02615   0.25221   0.85005
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept) -1.862754   0.319161  -5.836 0.0000002197 ***
## idade        0.017236   0.002650   6.505 0.0000000163 ***
## bmi          0.038581   0.008469   4.556 0.0000255814 ***
## genero       0.076313   0.095999   0.795      0.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1303354)
##
##      Null deviance: 16.2462  on 64  degrees of freedom
## Residual deviance:  7.9505  on 61  degrees of freedom
## AIC: 57.887
##
## Number of Fisher Scoring iterations: 2
```

Modelo 2 – Usando Somente a Variável idade

Idade é a variável mais importante no primeiro modelo. O que aconteceria se construíssemos um modelo com somente esta variável.

```
chdfit4 <- glm(chd ~ idade, data = riscochd)
summary(chdfit4)

##
## Call:
## glm(formula = chd ~ idade, data = riscochd)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70233  -0.33607   0.06459   0.29767   0.99690
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|)
## (Intercept) -0.696135   0.214369  -3.247   0.00187 **
## idade        0.016648   0.002916   5.709 0.000000033 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian family taken to be 0.1699578)
##
## Null deviance: 16.246 on 64 degrees of freedom
## Residual deviance: 10.707 on 63 degrees of freedom
## AIC: 73.237
##
## Number of Fisher Scoring iterations: 2
```

Este modelo tem um AIC acima daquele do primeiro modelo (73.237 vs. 57.887). Também o desvio residual fica mais alto. Então podemos concluir que precisamos mais variáveis que **idade** para formar um modelo bom.

Modelo 3 – idade e bmi

No primeiro modelo, **genero** não foi significativa. No último modelo, vamos eliminar esta variável e calcular o modelo.

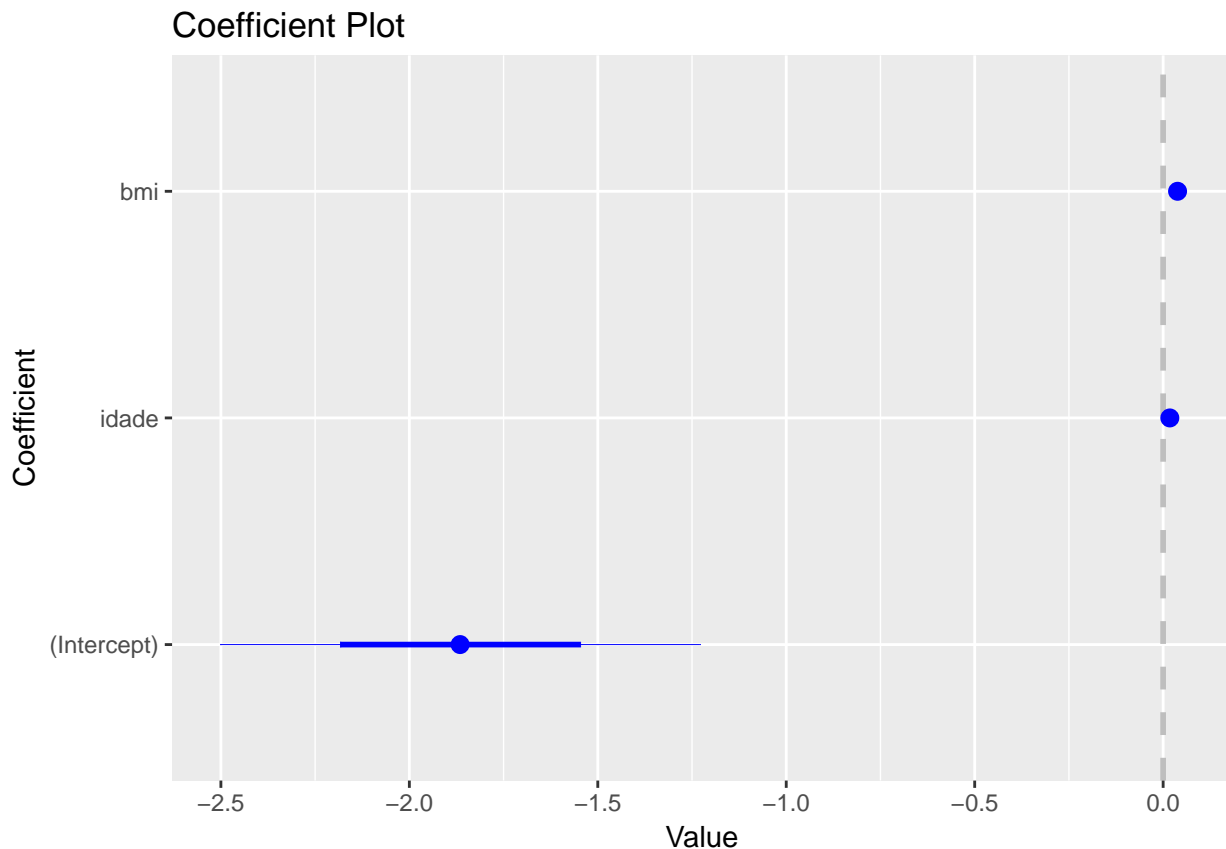
```
chdfit5 <- glm(chd ~ idade + bmi, data = riscochd)
summary(chdfit5)
```

```
##
## Call:
## glm(formula = chd ~ idade + bmi, data = riscochd)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68253  -0.27915   0.01656   0.27133   0.82713
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.865287    0.318196  -5.862 0.00000019020 ***
## idade        0.017763    0.002558   6.944 0.00000000269 ***
## bmi          0.038339    0.008438   4.543 0.00002615319 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1295616)
##
## Null deviance: 16.2462 on 64 degrees of freedom
## Residual deviance: 8.0328 on 62 degrees of freedom
## AIC: 56.557
##
## Number of Fisher Scoring iterations: 2
```

De todos os três modelos, este tem o melhor desempenho. O AIC fica abaixo daquela do primeiro e o desvio residual fica muito perto (mais um pouco mais alto) do desvio do primeiro. Então, um pesquisador pode ficar contente usando este modelo final para fazer previsões e afirmar que idade e IMC são importante para determinar o risco de CHD.

Agora que decidimos qual modelo queremos usar, podemos ver os resultados traduzidos em odds e probabilidades.

```
coefplot(chdfit5)
```

```
paste("Relação de Odds:")
```

```
## [1] "Relação de Odds:"
```

```
exp(coef(chdfit5)) # Calculate the odds
```

```
## (Intercept)      idade      bmi
##  0.1548518    1.0179215    1.0390833
```

```
paste("Intervalo de Confiança dos Odds:")
```

```
## [1] "Intervalo de Confiança dos Odds:"
```

```
exp(confint(chdfit5))
```

```
## Waiting for profiling to be done...
```

```
##           2.5 %    97.5 %
## (Intercept) 0.08299794 0.2889119
## idade      1.01283077 1.0230379
## bmi        1.02203948 1.0564113
```

```
paste("Probabilidade de Ocorrência:")
```

```
## [1] "Probabilidade de Ocorrência:"
```

```
invlogit(chdfit5$coefficients)
```

```
## (Intercept)      idade      bmi
##  0.1340881    0.5044406    0.5095835
```

Esses números contam uma historia que apesar que o modelo seja significativo, a probabilidade de ocorrência

de CHD dado cada condição (idade ou alto IMC) fica entorno de 0.5, ainda não uma clara indicação que uma ou outra pode causar a CHD. Provavelmente, há outras variáveis que não foram sondadas neste estudo que influenciam a CHD.

Último Exemplo – Uma Historia Mais Triste

As vezes, regressão logística não produz imediatamente resultados claros. Um projeto de uma colega nossa mostra isso. Ela tentou de ver o efeito de vários fatores ativantes sobre o tropismo do vírus HIV. Esses fatores podem causar o tropismo de vírus de estar CCR5 ou CXCR4.

Modelo com Soma de Marcas Activantes com CD4+ como Controle

```
load(file = "studdat.Rda")
actmodfit <- glm(tropismo ~ totr + cd4,
                 data = dat2, family = "binomial")
summary(actmodfit)
coefplot(actmodfit)
paste("Relação de Odds:")
exp(coef(actmodfit)) # Calculate the odds
paste("Intervalo de Confiança dos Odds:")
exp(confint(actmodfit))
paste("Probabilidade de Ocorrência:")
invlogit(actmodfit$coefficients)
```

Bibliografia

Esta apresentação deve muito aos livros seguintes:

- **R for Everyone** de Jared P. Lander. Este livro cobra muitos tópicos analíticos importantes numa forma clara com código para ajudar na aplicação.
- **Regression Models for Data Science in R** de Brian Caffo. Este é um texto avançado sobre os tipos de modelos de regressão e serve como texto do curso do Caffo sobre regressão na Coursera.
- **Applied Logistic Regression** de David Hosmer e Stanley Lemeshow. Além de ser a referência para os estudos relatados nesta palestra, este livro é um dos livros mais importantes sobre regressão logística.
- **OpenIntro Statistics (3a Ed.)** de Diez, Barr e Cetinkaya-Rundel. Um texto excelente introdutório sobre estatística.