# Regressão Logística

## Eletiva Biomedicina

James Hunter

25 de setembro 2019

Nesta apostila, 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 predizer 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 mudar o modelo de regressão linear para acomodar o limite de `TRUE` ou `FALSE`, `1` ou `0`, `infetado` ou `não infetado`? Como na regressão polinomial, onde aumentamos um ou mais termos da parte independente do modelo, nós transformamos uma curva não linear numa *expressão* linear que satisfaz 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). A `logit` expressa o modelo logístico 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)

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

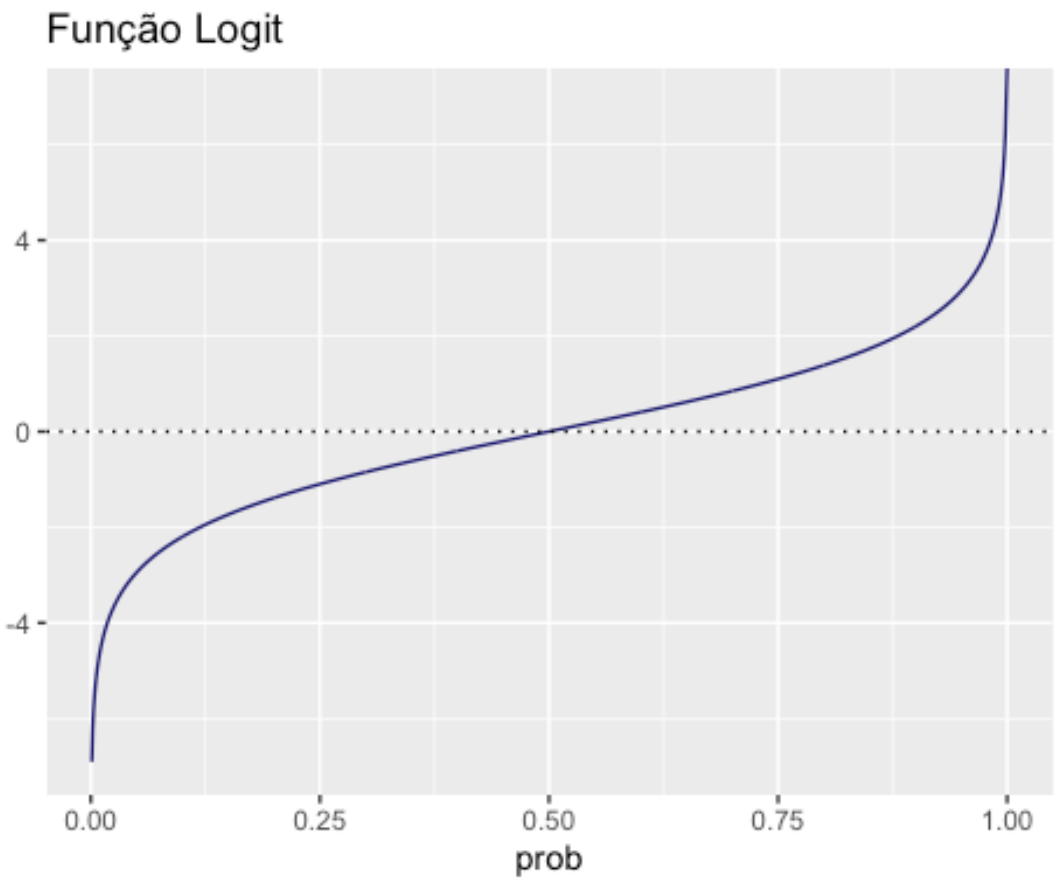
Este quer dizer que um matriz de valores $X$ está sendo multiplicado por um vetor de coeficientes $\beta$. Sabemos isso porque $X$ fica em maiúsculo e $\beta$ em minúsculo, a anotação tradicional para álgebra linear.

## Regressão Logística

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

Esta equação diz que estamos procurando a probabilidade que a variável dependente ter o estado de '1' e que este depende nas 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 curva da função `logit` vai entre os limites de $-\infty$ e $\infty$ e segue o formato seguinte:

```
dados_log <- tibble::tibble(p = seq(0.001, 1, .001),
                            yy = log(p/(1 - p)))
ggplot(dados_log, aes(x = p, y = yy)) +
  geom_line(colour = "midnightblue") +
  geom_hline(yintercept = 0, linetype = "dotted")+
  labs(title = "Função Logit", x = "prob", y = "")
```

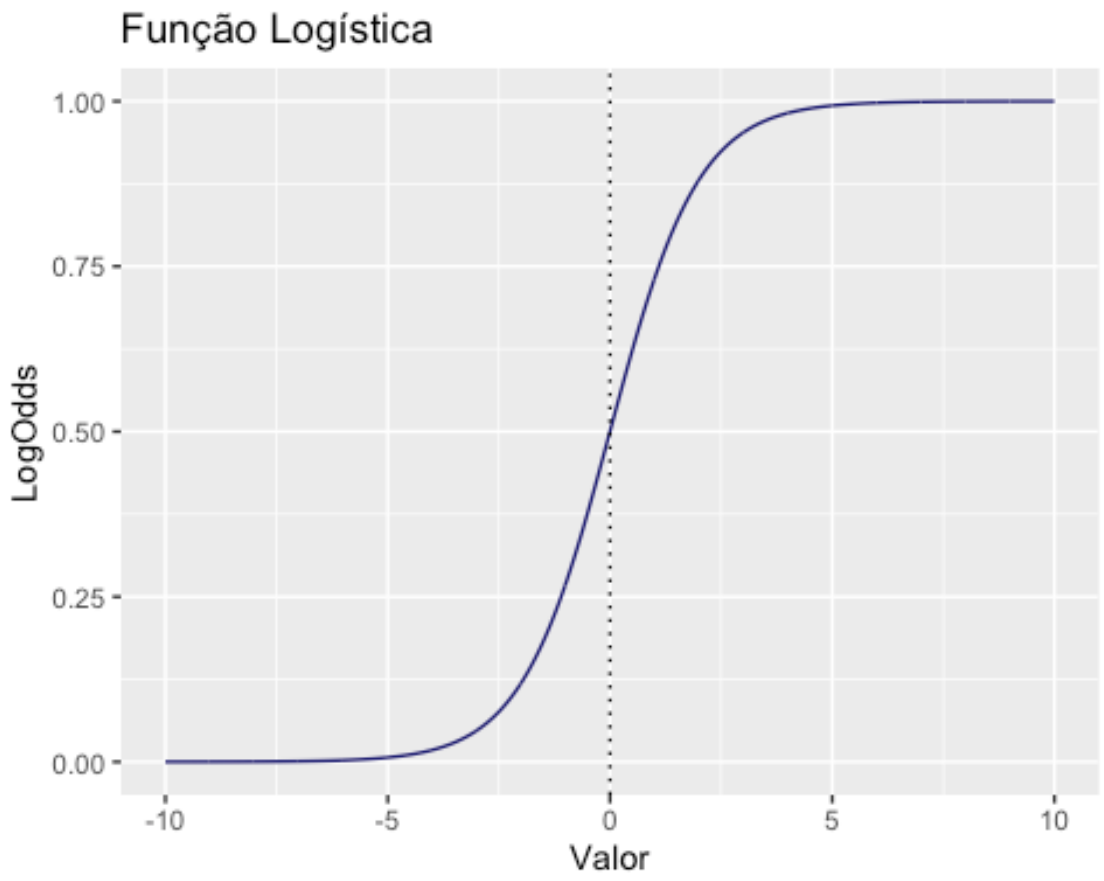



A função `logit` invertido (a versão que usamos; veja o "-1" como o expoente na fórmula) tem a forma que limite os valores para os limites naturais de probabilidades: 0 e 1. Aliás, a função em si pode assumir qualquer valor real, mas a probabilidade sempre cairá no intervalo [0, 1].

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

A função tem a curva seguinte:

```r
dados_logis <- tibble::tibble(xx = seq(-10, 10, by = 0.01),
                              yy = 1/(1 + exp(-xx)))
gr_logistic <- ggplot(dados_logis, aes(x = xx, y = yy)) +
  geom_line(colour = "midnightblue") +
  geom_vline(xintercept = 0, linetype = "dotted") +
  labs(title = "Função Logística", x = "Valor", y = "LogOdds")
gr_logistic
```




## Modelos Lineares Gerais (*General Linear Models*)

Uma regressão logística faz parte de uma classe dos modelos chamados *general linear models (GLM)*. Eles manipulam as matrizes dos parâmetros numa maneira diferente dos modelos lineares simples (como regressão linear). Esses 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 a 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.

## Primeiro Exemplo

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

Para conduzir este exercício da regressão logística, precisamos utilizar alguns dos pacotes de **R**.

- `tidyverse` - suite dos pacotes para ajudar com a organização, limpeza e visualização dos dados
- `knitr` - função `kable()` que produz relatórios visualmente limpos e estéticos
- `car` - funções que ajudam com os detalhes da análise de regressão
- `summarytools` - funções que fornecem resumos das estatísticas descritivas
- `broom` - funções que organizam os resultados de regressão em `tibbles`
- `nortest` - funções que incluem estatísticas sobre a normalidade de uma distribuição
- `coefplot` - uma função que faz uma plotagem dos coeficientes de um modelo de ML
- `janitor` - forma mais avançada de uma tabela de 1 ou 2 dimensões (`tabyl`)

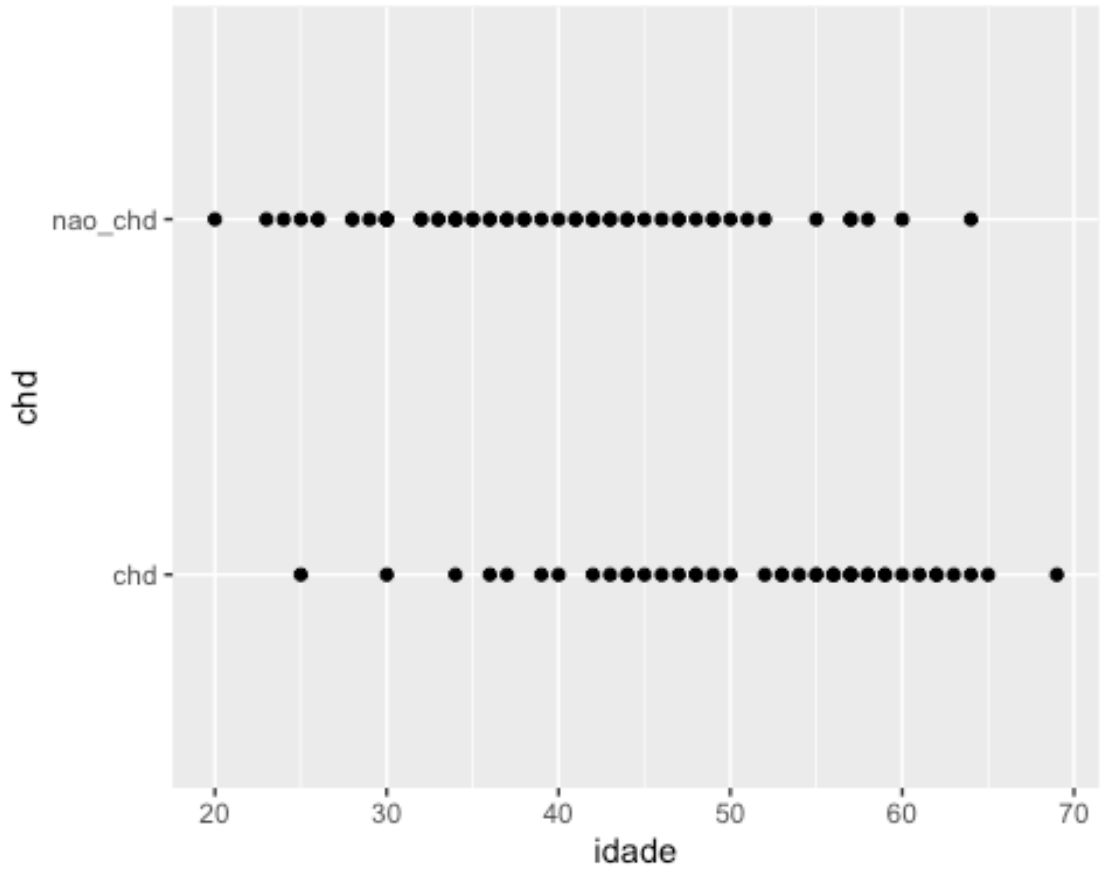
```
chdage <- read_csv("chdage.csv") %>%
  mutate(chd = factor(ifelse(chd == 0, "nao_chd", "chd")))
glimpse(chdage)

## Observations: 100
## Variables: 3
```

```
## $ id    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,…
## $ idade <dbl> 20, 23, 24, 25, 25, 26, 26, 28, 28, 29, 30, 30, 30, 30, 30…
## $ chd   <fct> nao_chd, nao_chd, nao_chd, nao_chd, chd, nao_chd, nao_chd,…
```

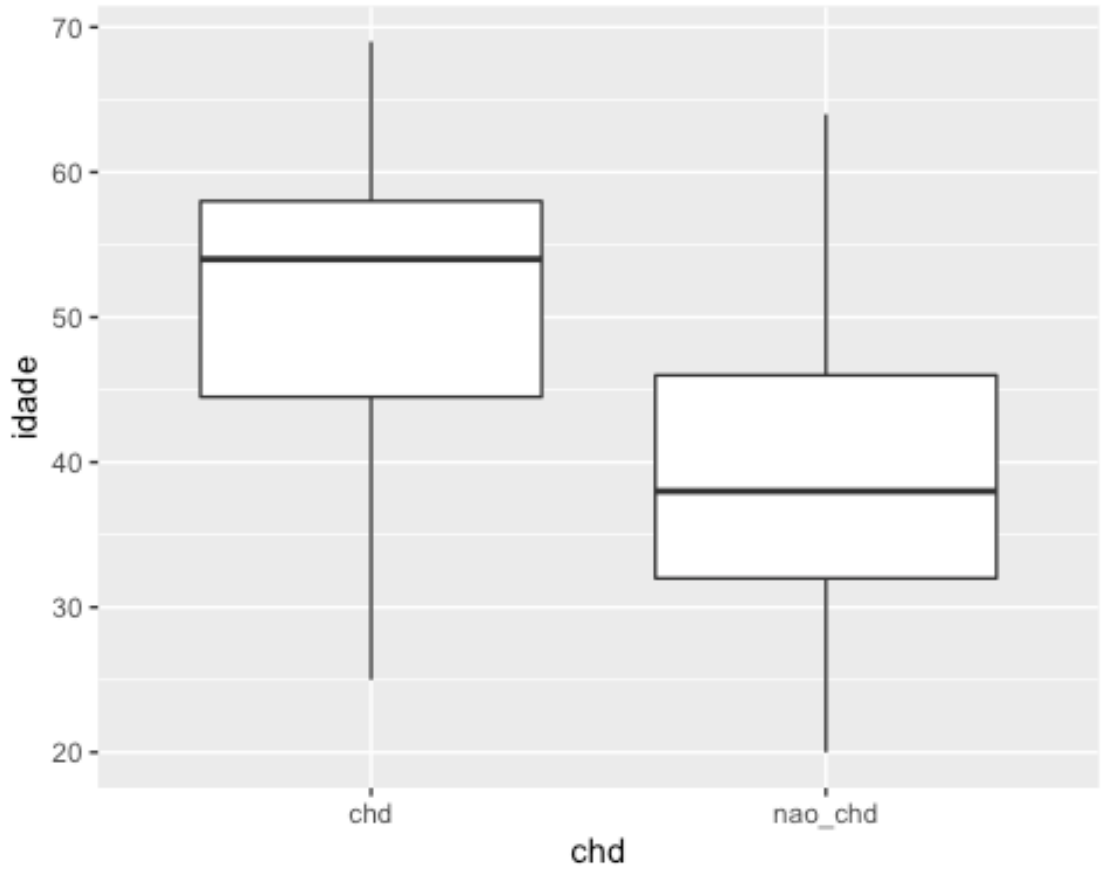
## Estrutura do chdage

Vamos fazer um estudo rápido exploratório dos dados para ver se podemos perceber uma tendência nos dados.

```
chdage %>%
  select(idade) %>%
  descr(transpose = TRUE,
        stats = c("mean", "sd", "min", "q1", "med", "q3",
                  "max", "iqr", "cv"))

## Descriptive Statistics
## chdage$idade
## N: 100
##
##                Mean   Std.Dev     Min      Q1   Median      Q3     Max     IQR      CV
## ----------- ------- --------- ------- ------- -------- ------- ------- ------- ------
##       idade   44.38     11.72   20.00   34.50    44.00   55.00   69.00   20.25   0.26

chdage %>%
  select(chd) %>%
  freq(chd, report.nas = FALSE)

## Frequencies
## chdage$chd
## Type: Factor
##
##                 Freq        %   % Cum.
## ------------- ------ -------- --------
##           chd     43    43.00    43.00
##       nao_chd     57    57.00   100.00
##         Total    100   100.00   100.00

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

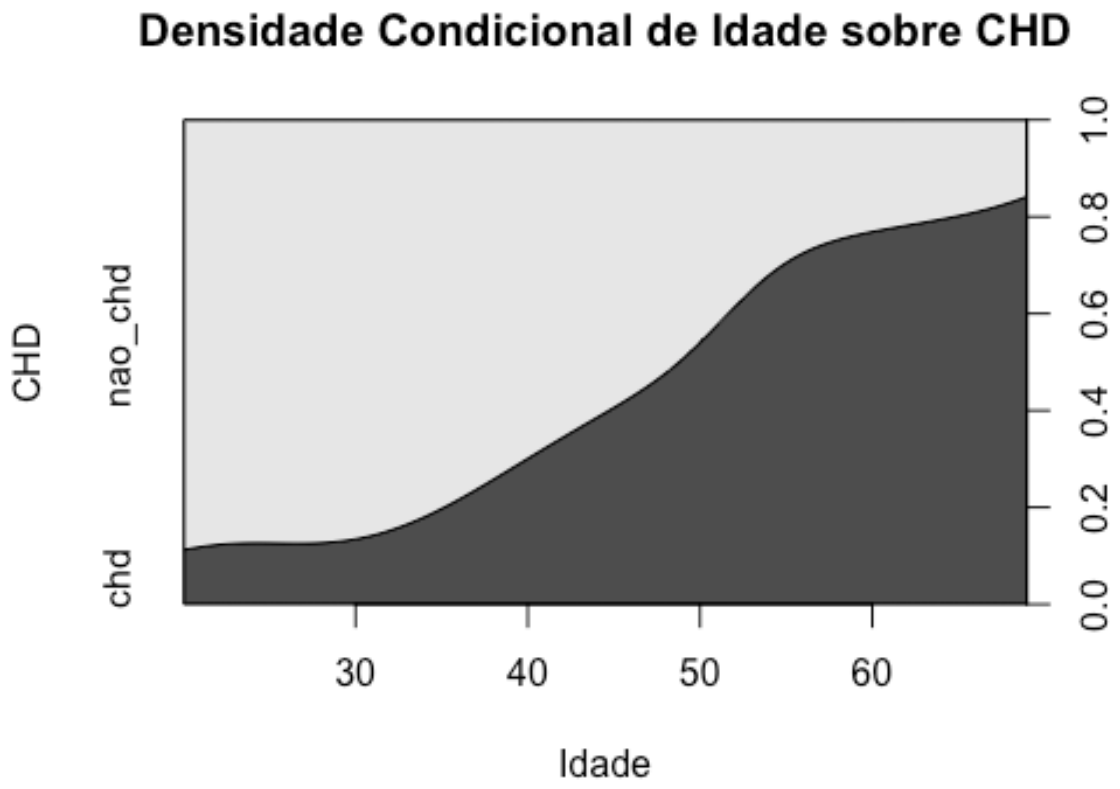



```
chdbox <- ggplot(data = chdage, aes(x = chd, y = idade, group = chd))
chdbox <- chdbox + geom_boxplot()
chdbox
```




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 tiveram 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 máxima dos pacientes na amostra).

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

```r
chdage$idgrp <- factor(Recode(chdage$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'"))

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$chd
## chdage$idgrp |       chd  |    nao_chd  | Row Total |
## -------------|------------|-------------|-----------|
##        20-29 |         1  |          9  |        10 |
##              |    10.000% |     90.000% |    10.000% |
## -------------|------------|-------------|-----------|
##        30-34 |         2  |         13  |        15 |
##              |    13.333% |     86.667% |    15.000% |
## -------------|------------|-------------|-----------|
##        35-39 |         3  |          9  |        12 |
##              |    25.000% |     75.000% |    12.000% |
## -------------|------------|-------------|-----------|
##        40-44 |         5  |         10  |        15 |
##              |    33.333% |     66.667% |    15.000% |
## -------------|------------|-------------|-----------|
##        45-49 |         6  |          7  |        13 |
##              |    46.154% |     53.846% |    13.000% |
## -------------|------------|-------------|-----------|
##        50-54 |         5  |          3  |         8 |
##              |    62.500% |     37.500% |     8.000% |
## -------------|------------|-------------|-----------|
##        55-59 |        13  |          4  |        17 |
##              |    76.471% |     23.529% |    17.000% |
## -------------|------------|-------------|-----------|
##        60-69 |         8  |          2  |        10 |
##              |    80.000% |     20.000% |    10.000% |
## -------------|------------|-------------|-----------|
## Column Total |        43  |         57  |       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 tabela é um `CrossTable` do pacote `gmodels`. Ela 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 fórmula 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)
coefplot(chdfit1)
```

## Os Coeficientes

A apresentação dos coeficientes é parecida 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 significativa. Mas, como vamos interpretá-la?

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*. O resultado será uma probabilidade. Nós vamos criar uma função para fazer este cálculo para os coeficientes.

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

##     idade
## 0.4722981
```

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`

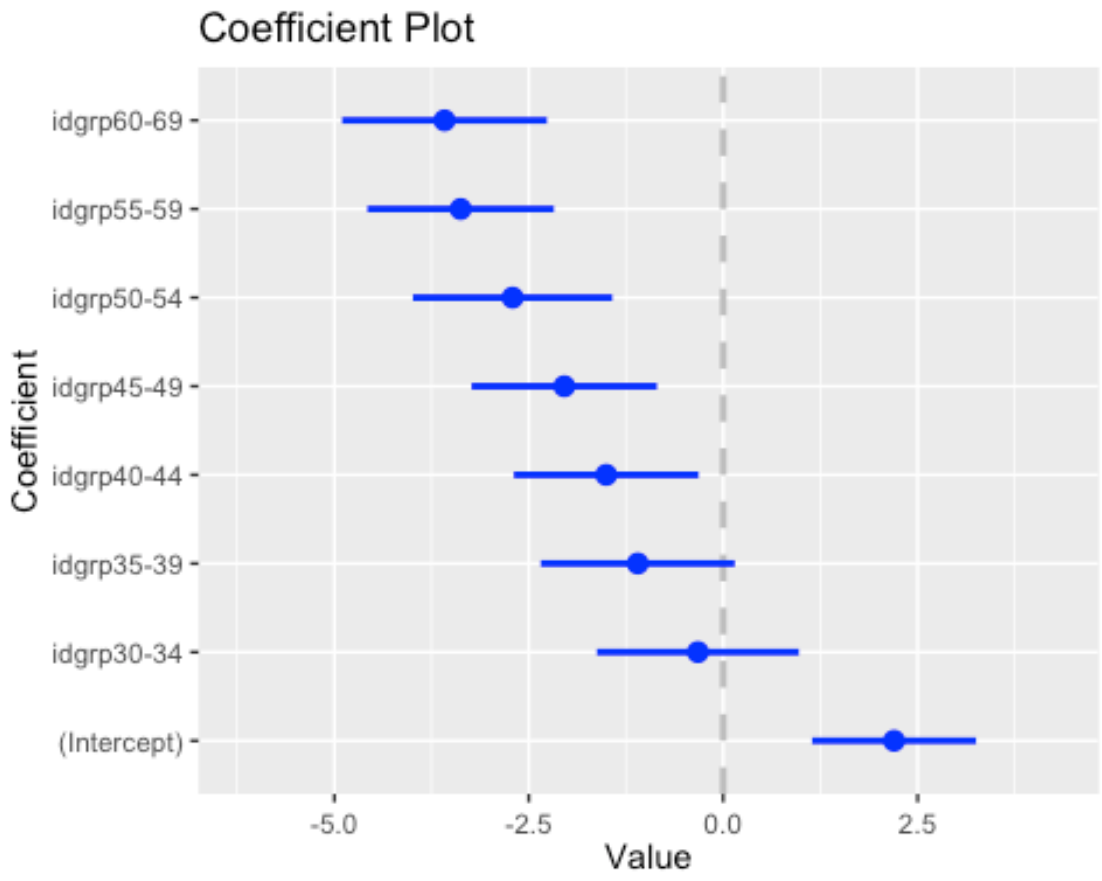
```
chdfit2 <- glm(chd ~  idgrp, data = chdage,
               family = binomial(link = "logit"))
summary(chdfit2)

##
## Call:
## glm(formula = chd ~ idgrp, family = binomial(link = "logit"),
##     data = chdage)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1460  -0.7325   0.4590   0.9005   1.7941
##
## 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 $\alpha$ 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
## 0.06250000 0.03305785 0.02702703
```

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 nosso modelo, o desvio residual e o AIC são basicamente igual nos dois casos porque os modelos estão considerando os mesmos dados. No próximo exemplo, nós podemos ver o que acontece quando acrescentamos novas variáveis ao modelo.

## 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 memoria. Os dados ficam num arquivo de R, `riscochd.RData`. Depois, fazermos um pequeno estudo exploratório.

## Análise Exploratória

```
load("riscochd.RData")
riscochd <- riscochd %>%
  mutate(chd = fct_recode(factor(chd), nao_chd = "0", chd = "1"),
         genero = fct_recode(factor(genero), masculino = "0", feminino =
"1"))
glimpse(riscochd)
```

```
## Observations: 65
## Variables: 5
## $ id     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17…
## $ idade  <int> 75, 98, 91, 88, 56, 86, 93, 74, 56, 95, 64, 99, 68, 66, 9…
## $ bmi    <dbl> 36.38134, 27.65790, 26.47878, 35.70601, 33.71147, 32.1208…
```

```
## $ genero <fct> masculino, feminino, feminino, masculino, feminino, mascu…
## $ chd    <fct> chd, chd, chd, chd, nao_chd, chd, chd, chd, nao_chd, chd,…
```

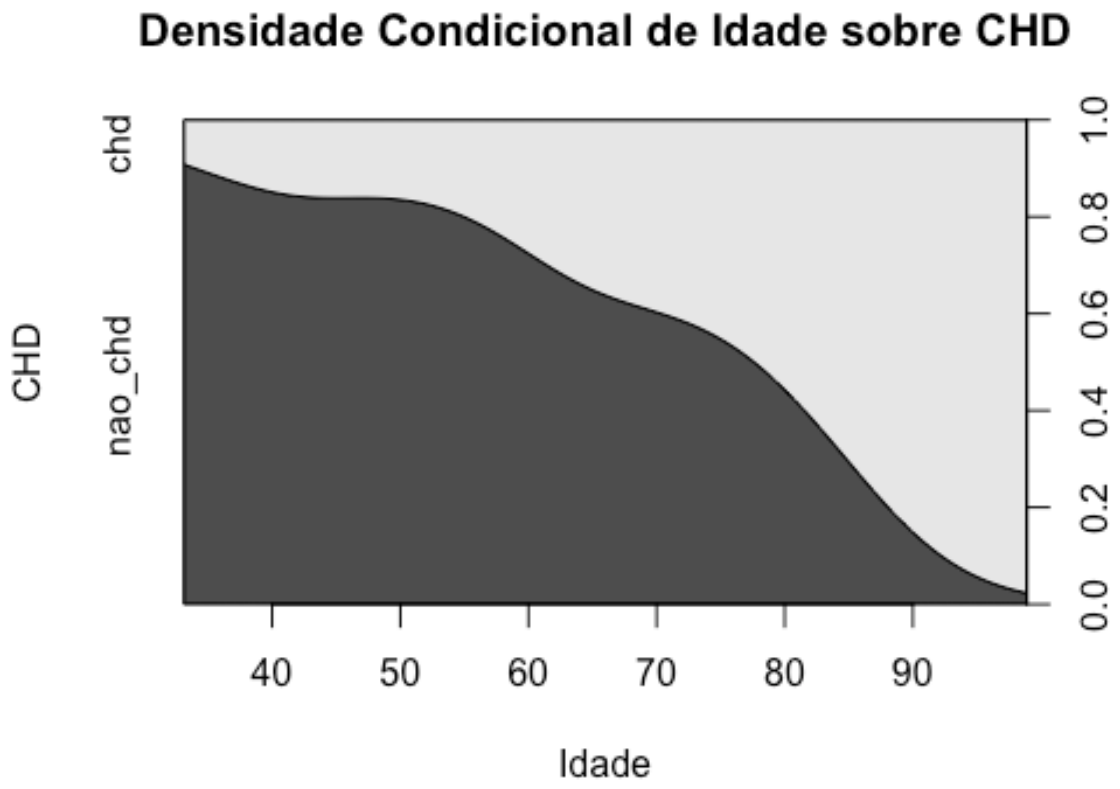
```
## EDA
riscochd %>%
  select(idade, bmi) %>%
  descr(transpose = TRUE,
        stats = c("mean", "sd", "min", "q1", "med", "q3",
                  "max", "iqr", "cv"))
```

```
## Descriptive Statistics
## riscochd
## N: 65
##
##                Mean   Std.Dev     Min      Q1   Median      Q3     Max     IQR      CV
## ----------- ------- --------- ------- ------- -------- ------- ------- ------- ------
##         bmi   28.42      5.36   16.78   25.18    28.06   31.47   44.94    6.30   0.19
##       idade   71.38     17.67   33.00   56.00    74.00   84.00   99.00   28.00   0.25
```

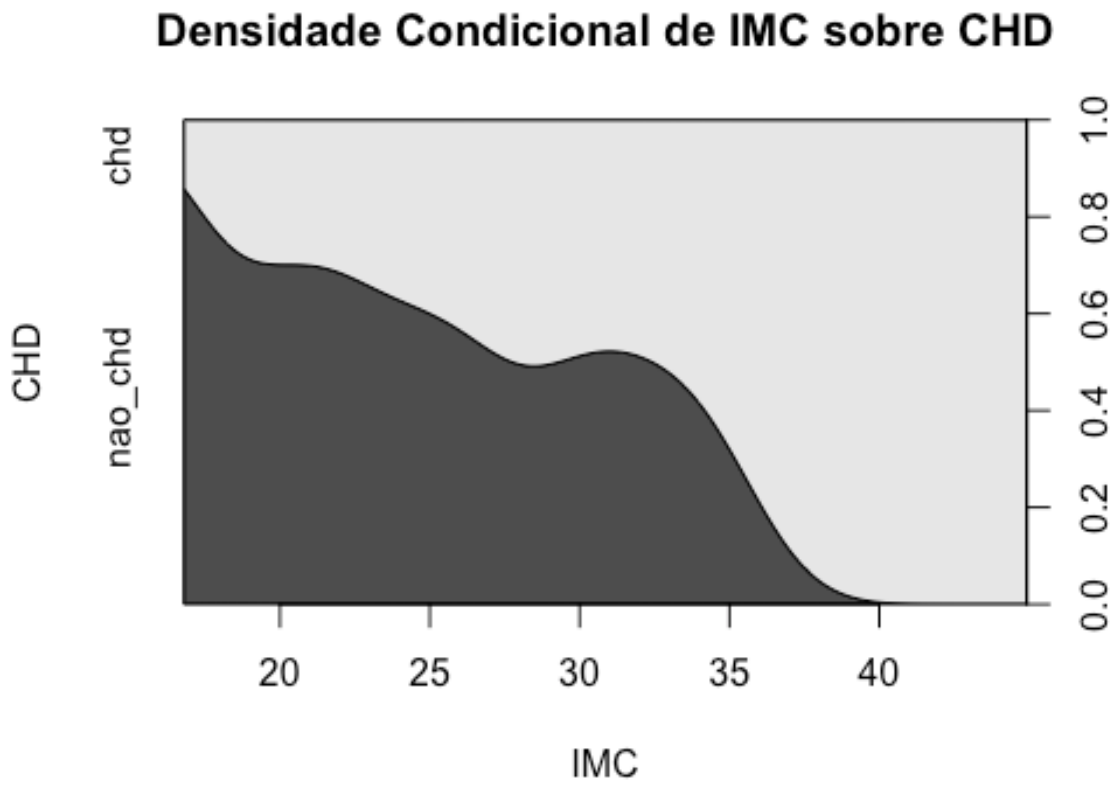
```
riscochd %>%
  select(genero, chd) %>%
  freq()
```

```
## Frequencies
## riscochd$genero
## Type: Factor
##
##                   Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## --------------- ------ --------- -------------- --------- --------------
##       masculino     41     63.08          63.08     63.08          63.08
##        feminino     24     36.92         100.00     36.92         100.00
##            <NA>      0                               0.00         100.00
##           Total     65    100.00         100.00    100.00         100.00
##
## riscochd$chd
## Type: Factor
##
##                  Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ------------- ------ --------- -------------- --------- --------------
##       nao_chd     33     50.77          50.77     50.77          50.77
##           chd     32     49.23         100.00     49.23         100.00
##          <NA>      0                               0.00         100.00
##         Total     65    100.00         100.00    100.00         100.00
```

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




```r
cdplot(factor(chd) ~ bmi, data = riscochd,
       main = "Densidade Condicional de IMC sobre CHD",
       xlab = "IMC", ylab = "CHD")
```




## Modelo 1 – Todas as Variáveis Independentes

```
chdfit3 <- glm(chd ~ idade + bmi + genero, data = riscochd,
               family = binomial(link = "logit"))
summary(chdfit3)

##
## Call:
## glm(formula = chd ~ idade + bmi + genero, family = binomial(link =
"logit"),
##     data = riscochd)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.84596  -0.48371  -0.05345   0.48149   2.46001
##
## Coefficients:
##                 Estimate Std. Error z value  Pr(>|z|)
## (Intercept)    -20.64336    5.06903  -4.072 0.0000465 ***
## idade            0.14814    0.03822   3.876  0.000106 ***
## bmi              0.34613    0.10189   3.397  0.000681 ***
## generofeminino   0.45202    0.77568   0.583  0.560069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 90.094  on 64  degrees of freedom
## Residual deviance: 43.886  on 61  degrees of freedom
## AIC: 51.886
##
## Number of Fisher Scoring iterations: 6
```

## Modelo 2 – Usando Somente a Variável `idade`

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

```
chdfit4 <- glm(chd ~ idade, data = riscochd,
               family = binomial(link = "logit"))
summary(chdfit4)
```

```
##
## Call:
## glm(formula = chd ~ idade, family = binomial(link = "logit"),
##     data = riscochd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6471  -0.7813  -0.2121   0.7718   2.4418
##
## Coefficients:
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -6.91677    1.79219  -3.859  0.000114 ***
## idade        0.09495    0.02393   3.968 0.0000725 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 90.094  on 64  degrees of freedom
## Residual deviance: 64.000  on 63  degrees of freedom
## AIC: 68
##
## Number of Fisher Scoring iterations: 5
```

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.

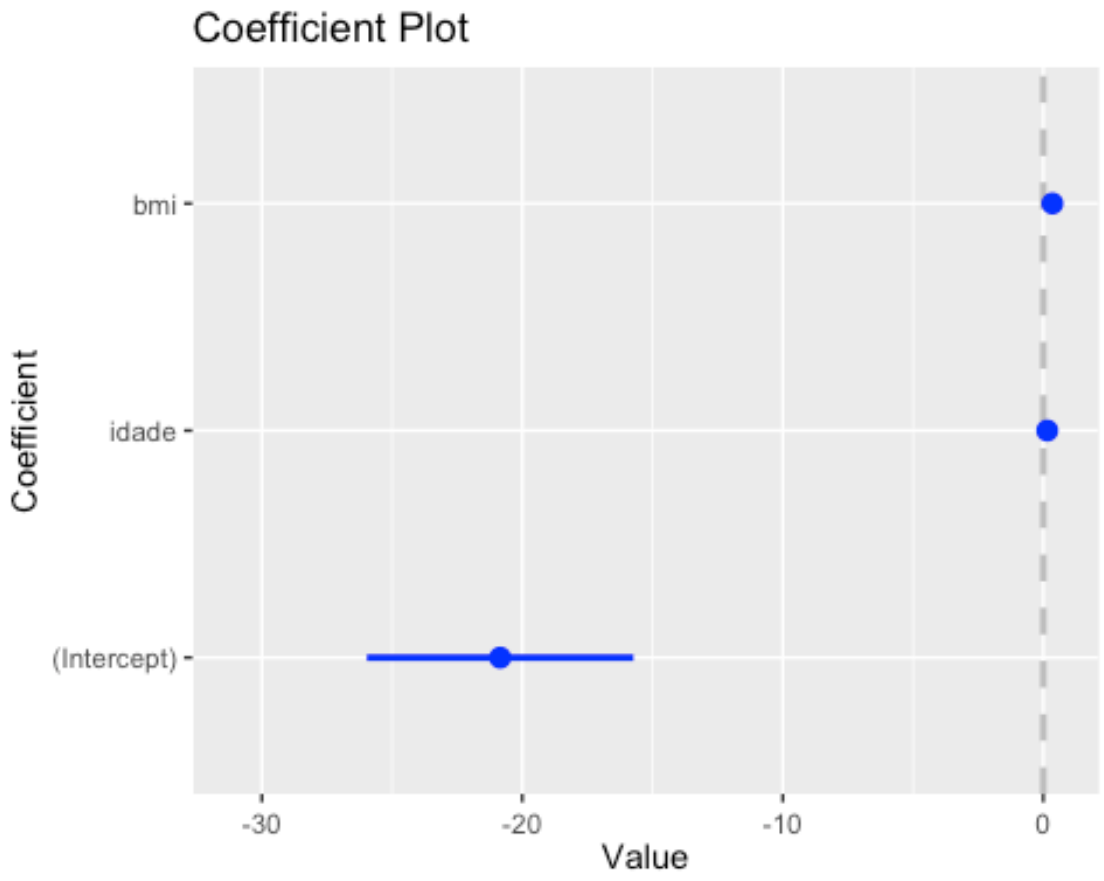
```r
chdfit5 <- glm(chd ~ idade + bmi, data = riscochd,
               family = binomial(link = "logit"))
summary(chdfit5)
```

```
##
## Call:
## glm(formula = chd ~ idade + bmi, family = binomial(link = "logit"),
##     data = riscochd)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.94448  -0.51392  -0.05453   0.52326   2.40266
##
## Coefficients:
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -20.84877    5.11434  -4.077 0.0000457 ***
## idade         0.15229    0.03819   3.988 0.0000667 ***
## bmi           0.35020    0.10196   3.435  0.000593 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 90.094  on 64  degrees of freedom
## Residual deviance: 44.225  on 62  degrees of freedom
## AIC: 50.225
##
## Number of Fisher Scoring iterations: 6
```

De todos os três modelos, este tem o melhor desempenho. O AIC fica abaixo daquela do primeiro e o desvio residual fica muito perto (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.

```r
coefplot(chdfit5)
```




```
paste("Relação de Odds:")
```

```
## [1] "Relação de Odds:"
```

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

```
##  (Intercept)        idade          bmi
## 8.820497e-10 1.164503e+00 1.419352e+00
```

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

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

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

```
##                    2.5 %         97.5 %
## (Intercept) 6.449713e-15 0.000004578083
## idade       1.092511e+00 1.272408489774
## bmi         1.192024e+00 1.794849303037
```

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

```
##  (Intercept)        idade          bmi
## 8.820497e-10 5.380002e-01 5.866661e-01
```

Esses números contam uma história 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.

## 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.