MAD-CB

Regressão Logística

Regressão - O que Fizemos

1. Regressão Simples

Regressão - O que Fizemos

1. Regressão Simples
2. Regressão Polinomial

Regressão - O que Fizemos

1. Regressão Simples
2. Regressão Polinomial
3. Regressão Múltipla

Regressão - O que Fizemos

1. Regressão Simples
2. Regressão Polinomial
3. Regressão Múltipla

- Hoje - Regressão Logística

Recursos

- Jared Lander, **R for Everyone**
- Brian Caffo, **Regression Models for Data Science in R**
- Hosmer & Lemeshow, **Applied Logistic Regression**
- Diez, Barr & Cetinkaya-Rundel, **OpenIntro Statistics (3a Ed.)**
- Everitt & Hothorn, **A Handbook of Statistical Analyses Using R**

Regressão Logística

- Tipo usado freqüentemente em bioestatística
- Extensão do conceito básico da regressão linear
  - como regressão polinomial
- Variável dependente (`Y`) agora é **binomial**
  - Tem 2 estados:
    - `TRUE`; `FALSE`
    - 1 ; 0
    - “R5” ; “X4”
    - “infetado” ; “não infetado”
- As variáveis independentes podem ser numéricas ou categóricas

Função logit

- Aplicamos função para as variáveis independentes (X)
- Resultado: Variável dependente fica no intervalo entre 0 e 1
  - intervalo de probabilidades

Comparar RSL com Regressão Logística

- Regressão Linear (usando notação de matrizes)

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

- Regressão Logística

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

Função `logit`

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

- Anote: este formato da função é o inverso da função original

Modelos Lineares Gerais (General Linear Models)

- Regressão logística faz parte de uma classe dos modelos: GLM
- Eles manipulam os matrizes diferente do modelo linear simples
  - que é um caso especial dos GLM
- Outros modelos GLM: poisson (dados de contagem)
- Output seria semelhante com o output do regressão simples

Exemplo Simples

- Estudo de 100 pacientes que têm ou não têm doença cardíaca coronária (CHD)
- Estudo interessado na relação entre a idade do paciente e a CHD
- Dados vêm de Hosmer & Lemeshow, *Applied Logistic Regression* (2a Ed.)
  - No arquivo `chdage.csv`

Carregar os Dados

```
chdage <- read_csv("chdage.csv")
```

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

```
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,...
```

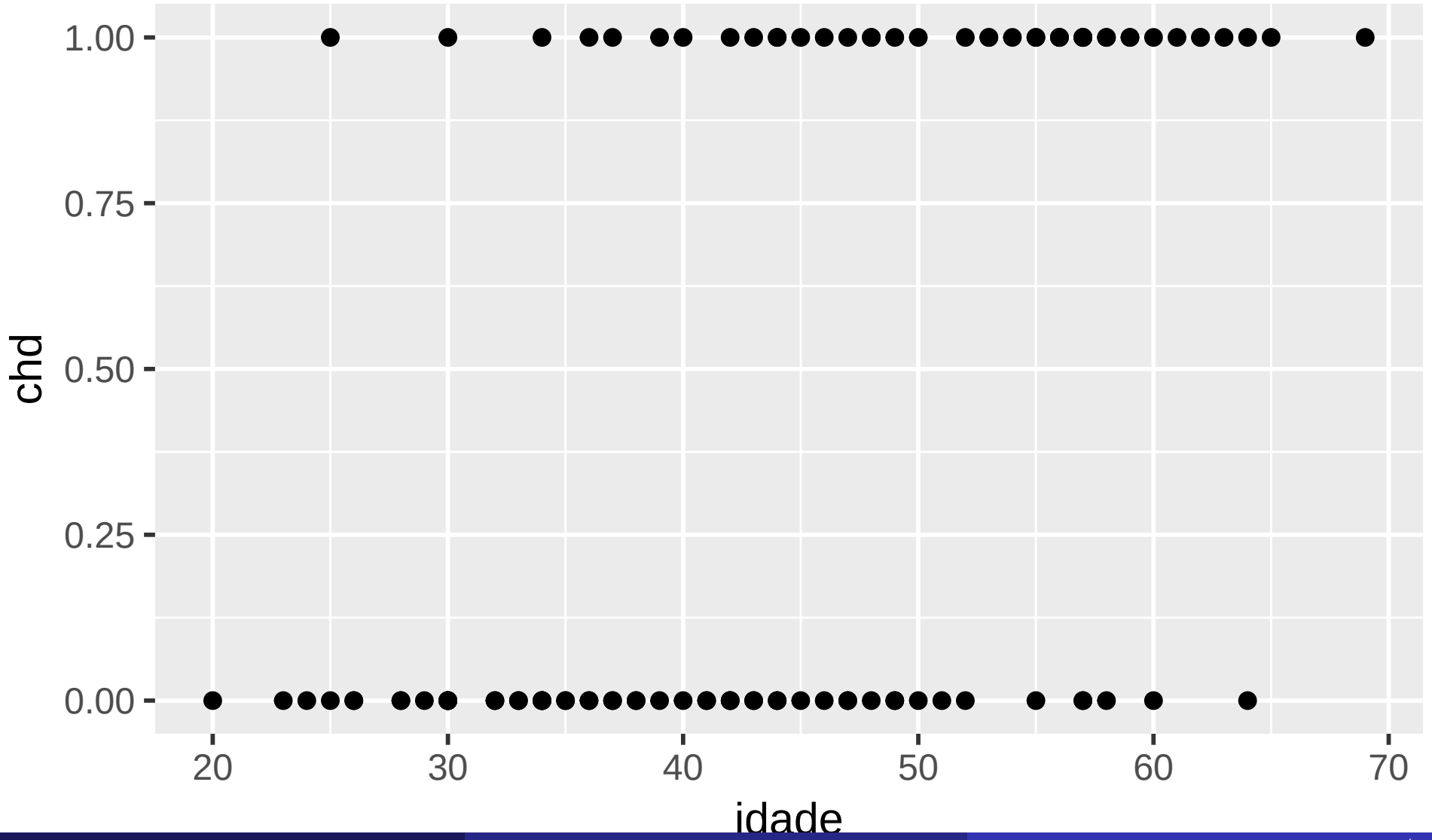
Analise Básica Exploratória

```
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
## 0s              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%
```

ScatterPlot de CHD e Idade

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



Boxplot da Idade

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

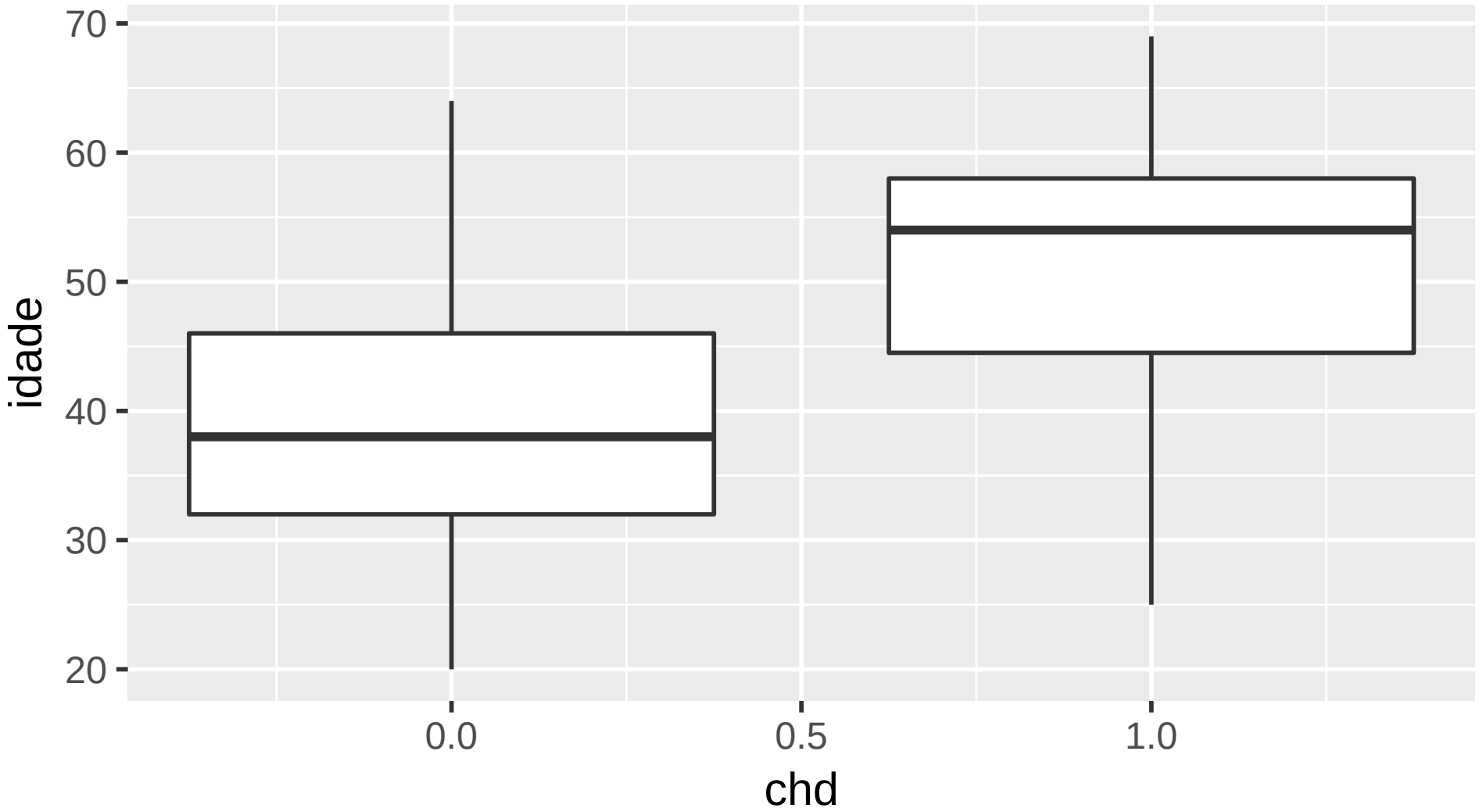
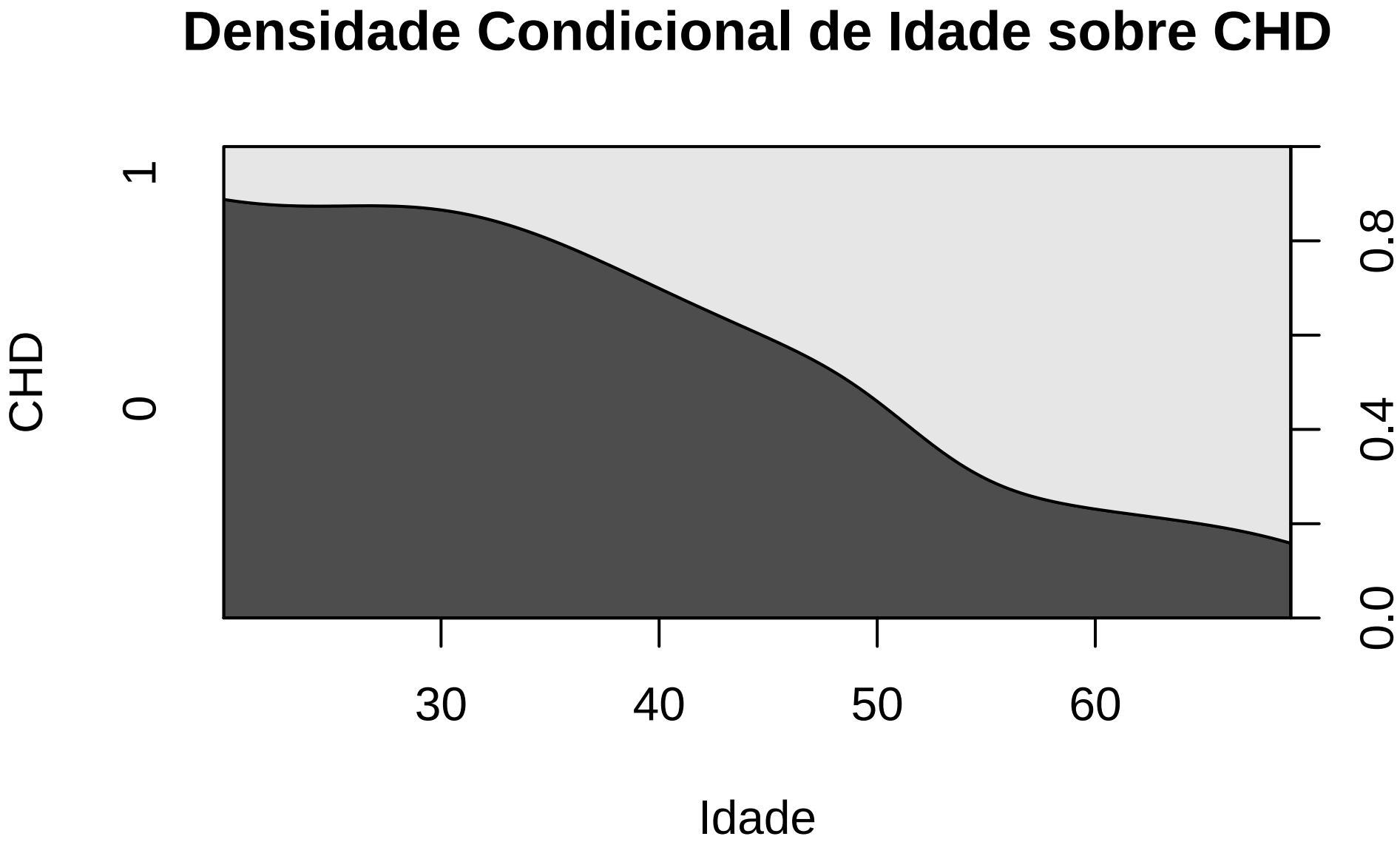


Gráfico de Densidade Condicional

- Também útil para entender como idade muda nas 2 categorias de CHD
- Mostra o número daqueles com a doença (chd $= 1$) para todos as idades
- Numa forma continua

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



O Modelo

- Como o pacote `lm`, `glm` usa o formato de formula para especificar o modelo
  - variável dependente ~ variáveis independentes
  - variáveis independentes separados com +
- Fonte dos dados (`data =`)
- `Family` dos modelos (neste caso, `binomial`)
- Função `link` (neste caso, `logit`)

Resultados

- Obter os resultados como no `lm`, com `summary`
- Também podemos olhar nos coeficientes com um gráfico chamada `coefplot`
- Vem de pacote de mesmo nome

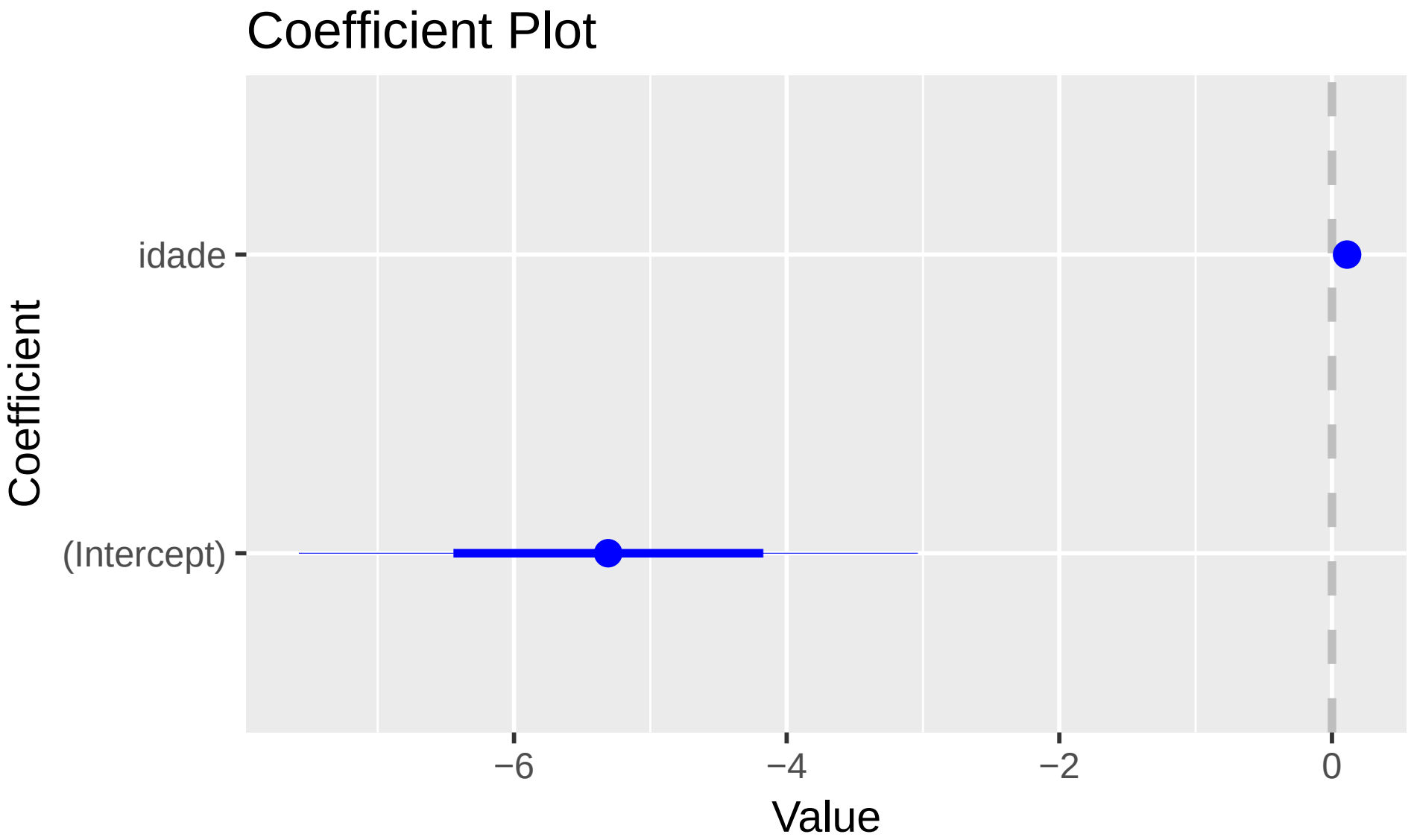
Coeficientes do Modelo

```
summary(chdfit1)
```

```
##
## Call:
## glm(formula = chd ~ idade, family = binomial(link = "logit"),
##     data = chdage)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9718  -0.8456  -0.4576   0.8253   2.2859
##
## Coefficients:
##             Estimate Std. Error z value   Pr(>|z|)
## (Intercept) -5.30945    1.13365  -4.683 0.00000282 ***
## idade        0.11092    0.02406   4.610 0.00000402 ***
## ---
## 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: 107.35  on 98  degrees of freedom
## AIC: 111.35
##
## Number of Fisher Scoring iterations: 4
```

Plotagem dos Coeficientes

```
coefplot(chdfit1)
```



Entender os Coeficientes

- Parecido com o que nós conhecemos da regressão linear
- Os coeficientes em si representam os log odds que o resultado $Y = 1$.
- Pode ver no gráfico quais são positivos e quais negativos
- Gráfico indica também o tamanho do erro padrão para cada variável independente
- Para entender os coeficientes melhor, precisa calcular o *logit inverso*
- Este põe os coeficientes no intervalo entre 0 e 1
  - ou seja, probabilidade

Logit Inverso

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

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

- Assim, podemos interpretar os resultados como probabilidades
- Com uma probabilidade acima de 50%, podemos dizer que idade tem uma relação positiva com a ocorrência de CHD

Desvio e AIC

- 2a parte dos resultados são os equivalentes de R^2, medidas de qualidade do modelo
- Invés da variância, com `glm` falamos de desvio
- Queremos minimizar o *desvio residual*
- AIC = Akaike's Information Criterion
- AIC útil para comparar modelos
  - Nota menor melhor

Este Modelo

- Desvio Residual = 107.3530927
- AIC = 111.3530927

Segundo Modelo para Comparação

- Modelo com Idade categorica – grupos de idade
- Esperança que podemos entender melhor as probabilidades relacionados aos grupos de idade mais especificos
  - Idosos mais propensos a CHD?
- Vamos usar `recode` do pacote `car`
  - Mais flexível que `recode` de `dplyr`

Grupos de Idade

```
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)
```

Modelo de Grupos

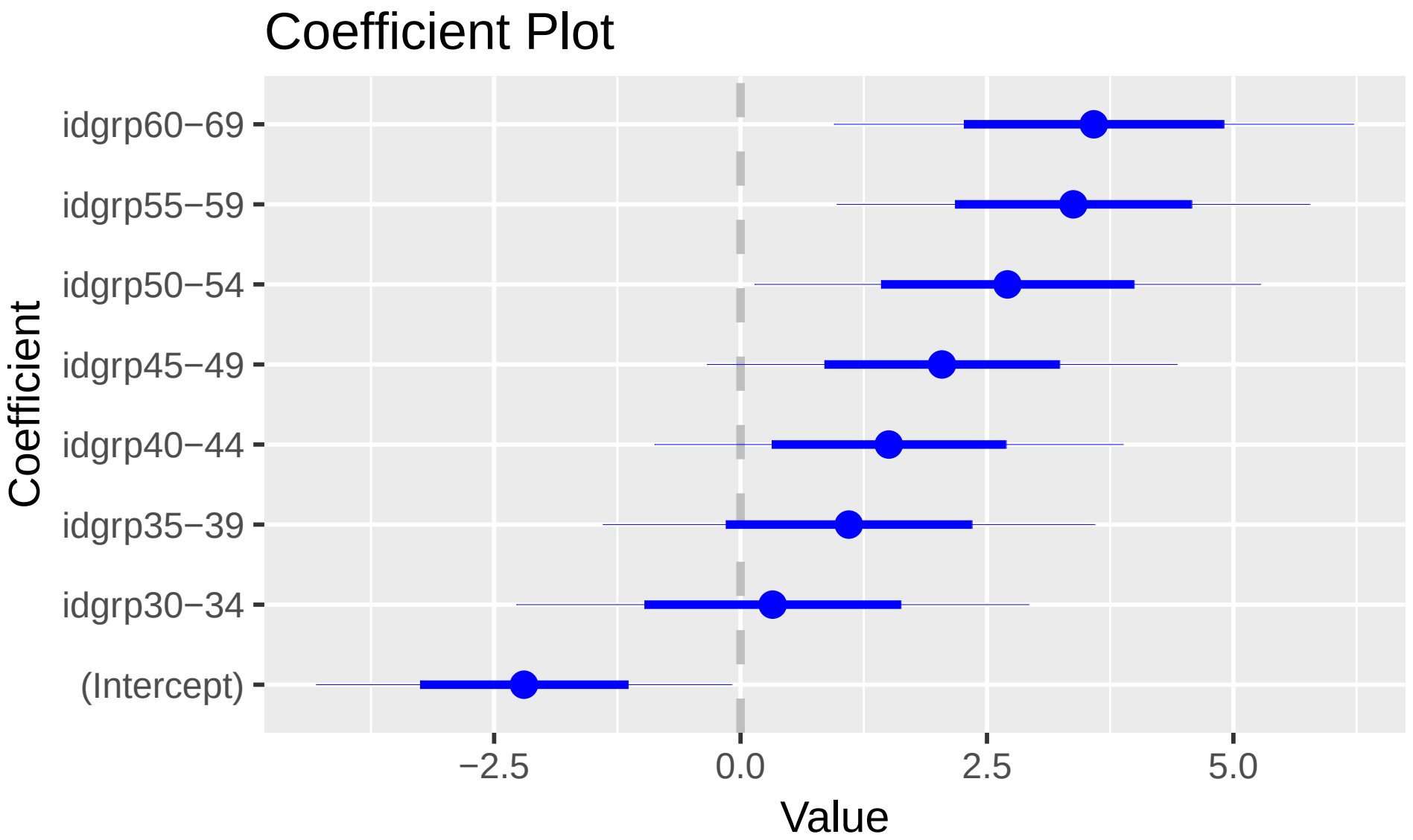
```
chdfit2 <- glm(chd ~  idgrp, data = chdage,
               family = binomial(link = "logit"))
```

Resultados

```
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
```

Gráfico dos Coeficientes do Modelo



Idosos Têm Alta Probabilidade de CHD

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

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

Qual modelo parece melhor?

- Modelo 1 – Idade Númerica
  - Desvio Residual = 107.3530927
  - AIC = 111.3530927
- Modelo 2 – Idade Categorica
  - Desvio Residual = 107.9614654
  - AIC = 123.9614654
- AIC melhor no modelo númerico
- Mas, modelo categorico oferece mais informação sobre grupos de idade de interesse

Exemplo com Múltiplas Variáveis Independentes

Outro Estudo sobre CHD

- Pesquisadores querem identificar fatores causativos para CHD
- Covariates independentes
  - id (Número de identificação do caso)
  - idade (em anos)
  - bmi (índice de massa corporal em kg/m^2)
  - genero ($0 =$ masculino, $1 =$ feminino)
- 65 casos
- Dados - `riscochd.RData`

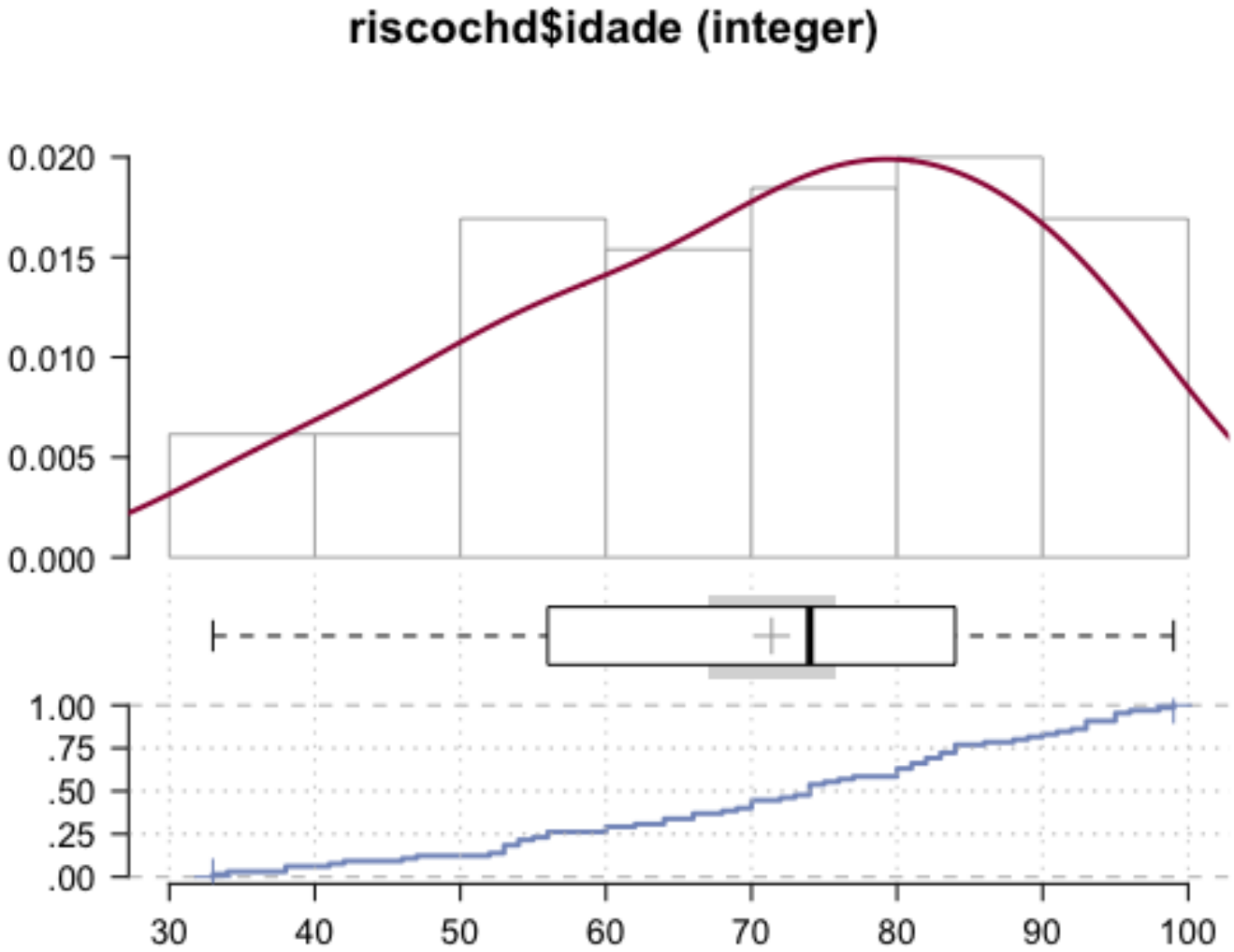
Análise Exploratório

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

```
## ------------------------------------------------------------------------
## riscochd$chd (numeric)
##
##   length       n    NAs  unique     0s  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%
```

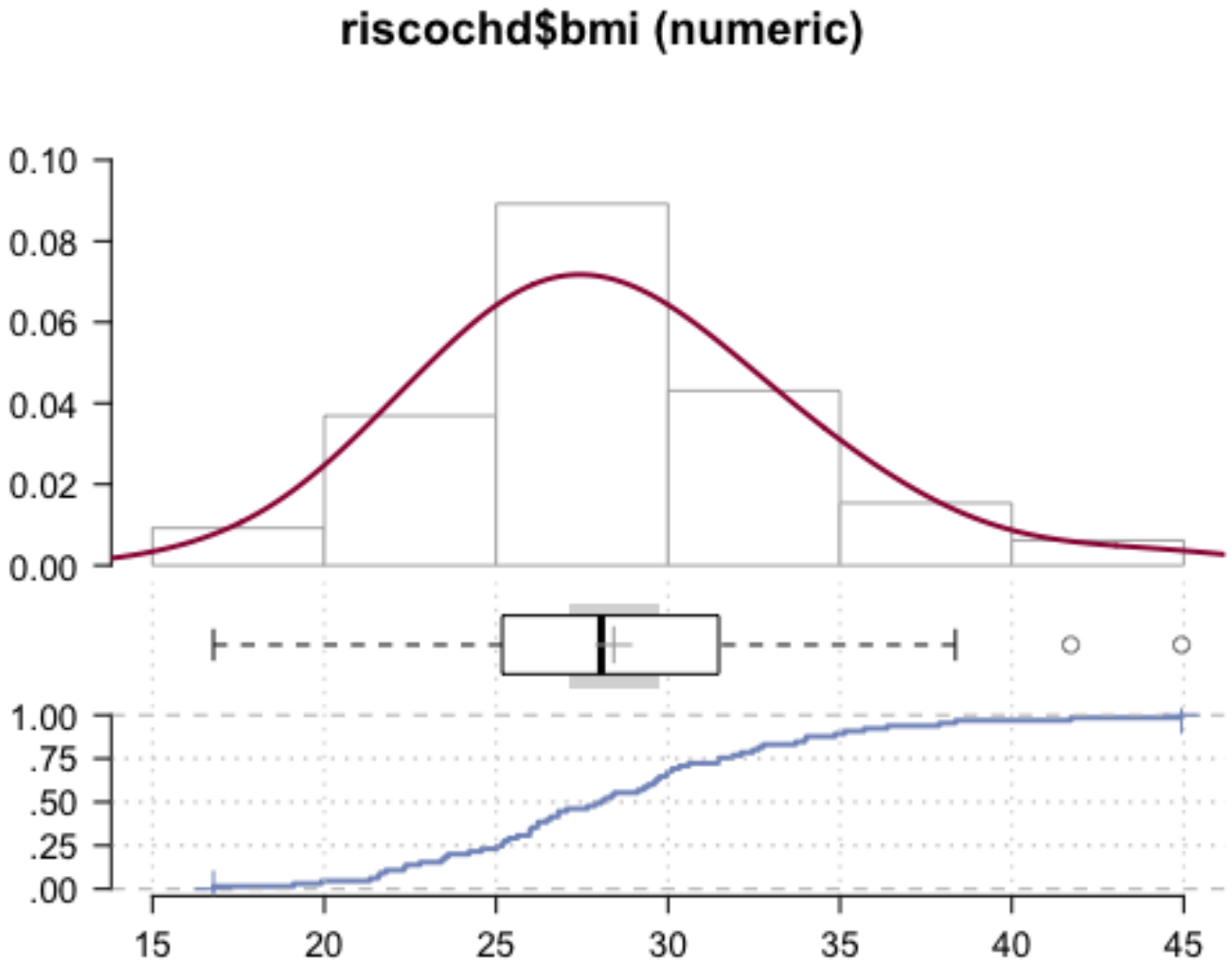
```r
Desc(riscochd$idade, plotit = FALSE)
```

```
## ------------------------------------------------------------------------
## riscochd$idade (integer)
##
##   length       n    NAs  unique     0s   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)
0.020
0.015
0.010
0.005
0.000
1.00
.75
.50
.25
.00
30
40
50
60
70
80
90
100

```r
Desc(riscochd$bmi, plotit = FALSE)
```

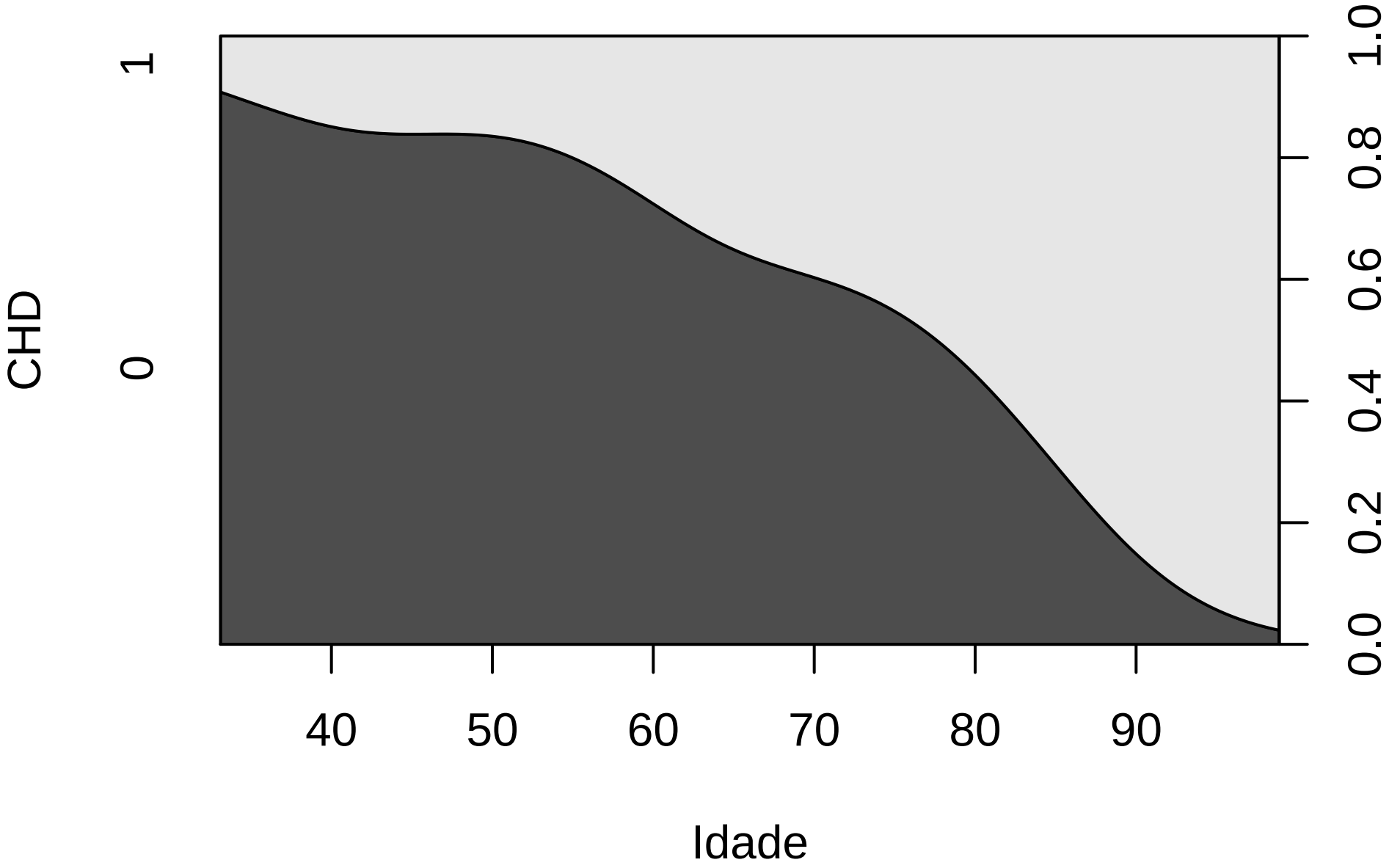
```
## ------------------------------------------------------------------------
## riscochd$bmi (numeric)
##
##     length          n       NAs    unique         0s       mean    meanCI
##         65         65         0        64          0   28.42026  27.09293
##             100.0%      0.0%             0.0%             29.74760
##
##        .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)
0.10
0.08
0.06
0.04
0.02
0.00
1.00
.75
.50
.25
.00
15
20
25
30
35
40
45

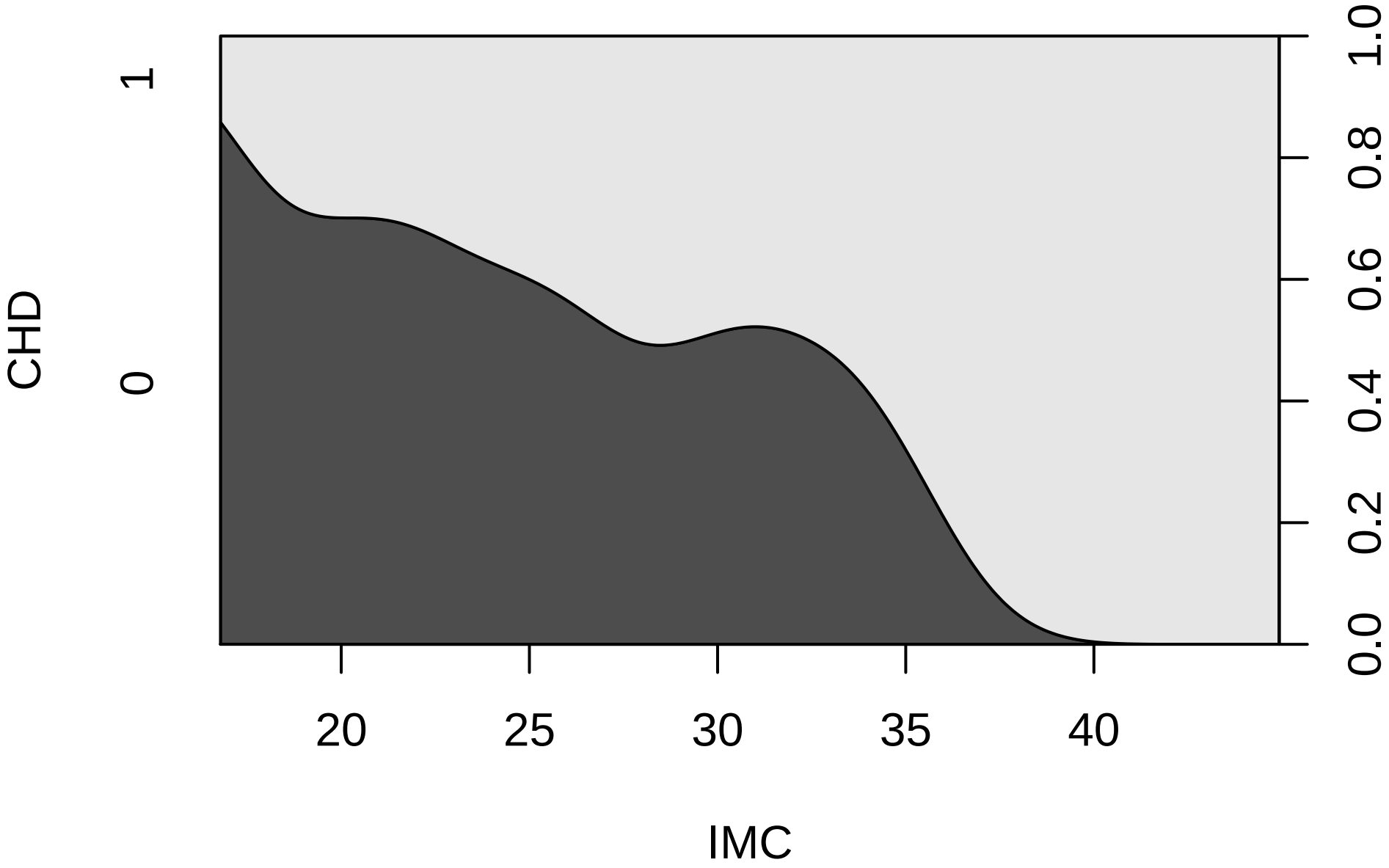
```r
Desc(riscochd$genero, plotit = FALSE)
```

```
## ------------------------------------------------------------------------
## 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
```

Densidade Condicional de Idade sobre 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

```
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 0.00000033 ***
## ---
## 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
```

Segundo Modelo Comparado ao Primeiro

- AIC aumentou com só `idade`
- Modelo piorou em qualidade

Modelo 3 – Usando as Variáveis idade e bmi

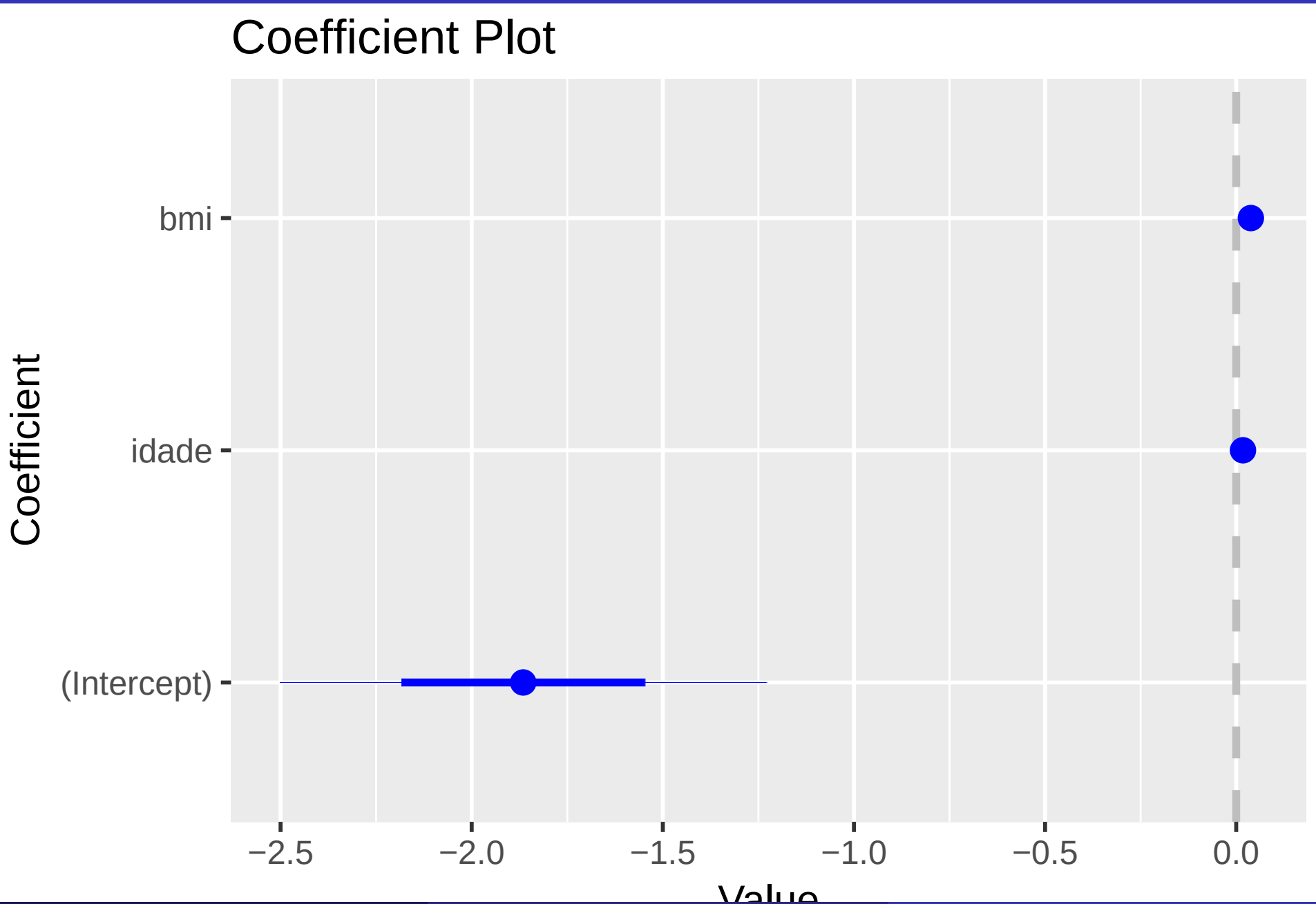
```
chdfit5 <- glm(chd ~ idade + bmi, data = riscochd)
summary(chdfit5)
```

```
##
## 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
```

Desempenho do Novo Modelo

- De todos os três, tem o melhor AIC (56.556668)
- Desvio residual fica muito perto (mais um pouco mais alto) do desvio do primeiro

Gráfico de Coeficientes do Modelo Final



Resultados Traduzidos em Probabilidade e Odds

```r
paste("Relação de Odds:", exp(coef(chdfit5))) # Calculate the odds
```

```
## [1] "Relação de Odds: 0.154851838017071"
## [2] "Relação de Odds: 1.01792154102839"
## [3] "Relação de Odds: 1.0390832748153"
```

```r
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
```

```r
paste("Probabilidade de Ocorrência:", invlogit(chdfit5$coefficients))
```

```
## [1] "Probabilidade de Ocorrência: 0.134088056077356"
## [2] "Probabilidade de Ocorrência: 0.504440594112309"
## [3] "Probabilidade de Ocorrência: 0.509583540627794"
```

Último Exemplo do Dia

Projeto de Mestrado

- Projeto de uma colega
- As vezes, regressão logística não produz resultados claros.
- Prever o efeito dos ativadores sobre tropismo com CD4+ como controle

Modelo

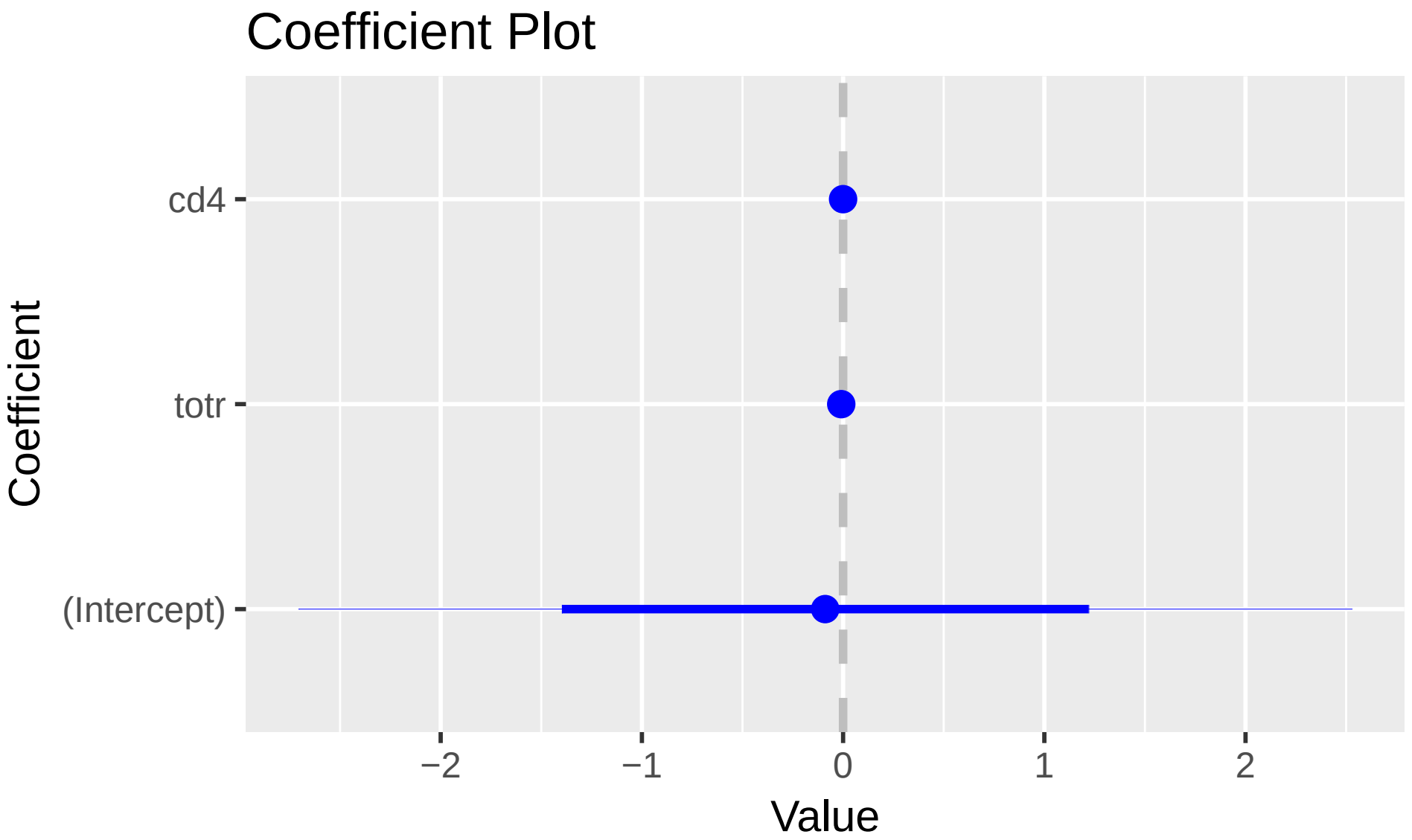
```
load(file = "studdat.Rda")
actmodfit <- glm(tropismo ~ totr + cd4,
                 data = dat2, family = "binomial")
```

Resultados

```
summary(actmodfit)
```

```
##
## Call:
## glm(formula = tropismo ~ totr + cd4, family = "binomial", data = dat2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8815  -0.6508  -0.5810  -0.4505   2.0706
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.0890136  1.3084158  -0.068    0.946
## totr        -0.0092666  0.0084563  -1.096    0.273
## cd4         -0.0002108  0.0011971  -0.176    0.860
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 58.352  on 62  degrees of freedom
## Residual deviance: 56.906  on 60  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 62.906
##
## Number of Fisher Scoring iterations: 4
```

Gráfico


Coefficient Plot
cd4
totr
(Intercept)
Coefficient
−2
−1
0
1
2
Value

Coeficientes Traduzidos

```
paste("Relação de Odds:")
```

```
## [1] "Relação de Odds:"
```

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

```
## (Intercept)        totr         cd4
##   0.9148332   0.9907762   0.9997892
```

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

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

```
exp(confint(actmodfit))
```

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

```
##                    2.5 %    97.5 %
## (Intercept) 0.07195555 12.864715
## totr        0.97304494  1.005770
## cd4         0.99694887  1.001836
```

Próxima Vez

- Programação em R