Problemas – Regressão & Programação

Soluções

James R. Hunter

18 de abril de 2017

Nestes exercícios, vamos experimentar com regressão e programação. Os exercícios terão várias partes. Não esqueça responder a todas! Tem no Github um arquivo `probsRegress.RData` com os dados que você precisa para completar os exercícios. Fazer o download dele e `load("probsRegress.RData")`.

1. Expectativa de Vida em Europa

Neste exercício, tirei dados do pacote `gapminder` sobre expectativa da vida (`lifeExp`) e PIB por capita (`gdpPercap`) para os anos 2002 e 2007 para os países de Europa. O código para reproduzir os dados para o problema segue. Você deve copiar e colar ele no seu trabalho. Faça uma regressão simples linear que mostra qual é o efeito que PIB tem sobre expectativa da vida. Países ricas têm expectativa de vida maior? Responde às partes a - f. Como sempre, não esqueça de fazer um pequeno estudo exploratório dos dados.

Dados do problema

```
library(gapminder)
vidaExp <- gapminder %>%
           filter(year > 2000 & continent == "Europe") %>%
           select(year, lifeExp, gdpPercap)
```

Perguntas

a. A variável `lifeExp` tem uma distribuição normal segunda o teste Shapiro-Wilks?
b. Uma transformação logarítmica pode fazer ela normal? Por que?
c. Reconhecendo que a variável dependente não é puramente normal, você pode confiar em qual regra de estatística para usar regressão linear? Por que?
d. O que é a equação linear que determina a relação entre as variáveis no formato de $y = \beta_0 + \beta_1 x$
e. Qual proporção de variância no modelo esta equação descreve?
f. Mostre e examine os quatro gráficos que pode usar para entender melhor a regressão. Essa regressão é confiável? Por que?

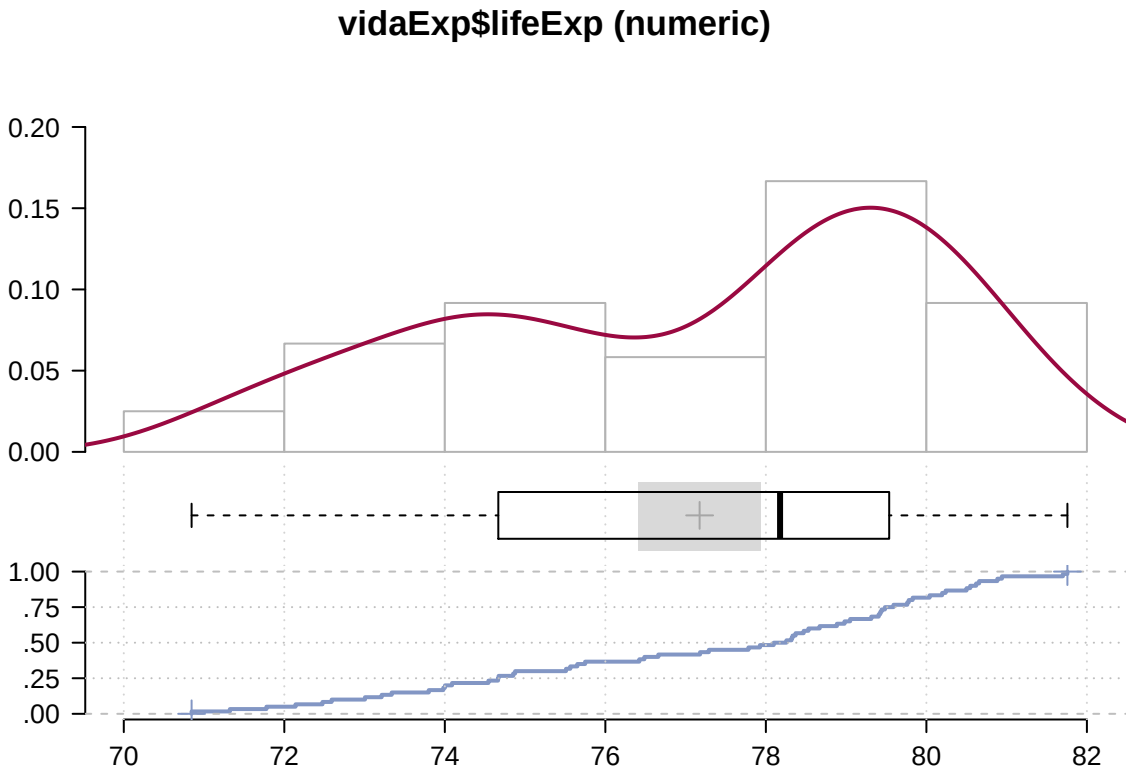
1. Solução

a.

```
Desc(vidaExp$lifeExp)
```

```
## ----------------------------------------------------------------------
## vidaExp$lifeExp (numeric)
##
```

```
##     length         n       NAs    unique        0s      mean    meanCI
##         60        60         0       = n         0  77.17460  76.40871
##               100.0%      0.0%                0.0%            77.94049
##
##        .05       .10       .25    median       .75       .90       .95
##   72.12185  72.96350  74.66825  78.17700  79.50975  80.55340  80.88685
##
##      range        sd     vcoef       mad       IQR      skew      kurt
##   10.91200   2.96481   0.03842   3.25134   4.84150  -0.42494  -1.06304
##
## lowest : 70.845, 71.322, 71.777, 72.14, 72.476
## highest: 80.657, 80.884, 80.941, 81.701, 81.757
```




```
shapiro.test(vidaExp$lifeExp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  vidaExp$lifeExp
## W = 0.93935, p-value = 0.005063
```

O teste de normalidade de Shapiro-Wilk tem um valor-p muito abaixo do nível tradicional de $\alpha = 0.05$. Assim, provavelmente a distribuição **não** está normal.

b.

```
vidaExp <- vidaExp %>% mutate(lifeExplog = log10(lifeExp))
shapiro.test(vidaExp$lifeExplog)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  vidaExp$lifeExplog
## W = 0.93567, p-value = 0.003486
```

Neste caso, a transformação **não** ajuda por causa dos 2 modos na distribuição.

c. Qual regra de estatística:

Teorema de Limite Central: Com um n alto (> 35), podemos assumir que a distribuição aproxima à normal

d. Equação de Regressão

```
vidafit <- lm(lifeExp ~ gdpPercap, data = vidaExp)
summary(vidafit)
```

```
##
## Call:
## lm(formula = lifeExp ~ gdpPercap, data = vidaExp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81053 -1.26704  0.05817  1.21515  3.08633
##
## Coefficients:
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) 72.03059980  0.45519976  158.24   <2e-16 ***
## gdpPercap    0.00021999  0.00001749   12.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.549 on 58 degrees of freedom
## Multiple R-squared:  0.7318, Adjusted R-squared:  0.7271
## F-statistic: 158.2 on 1 and 58 DF,  p-value: < 2.2e-16
```

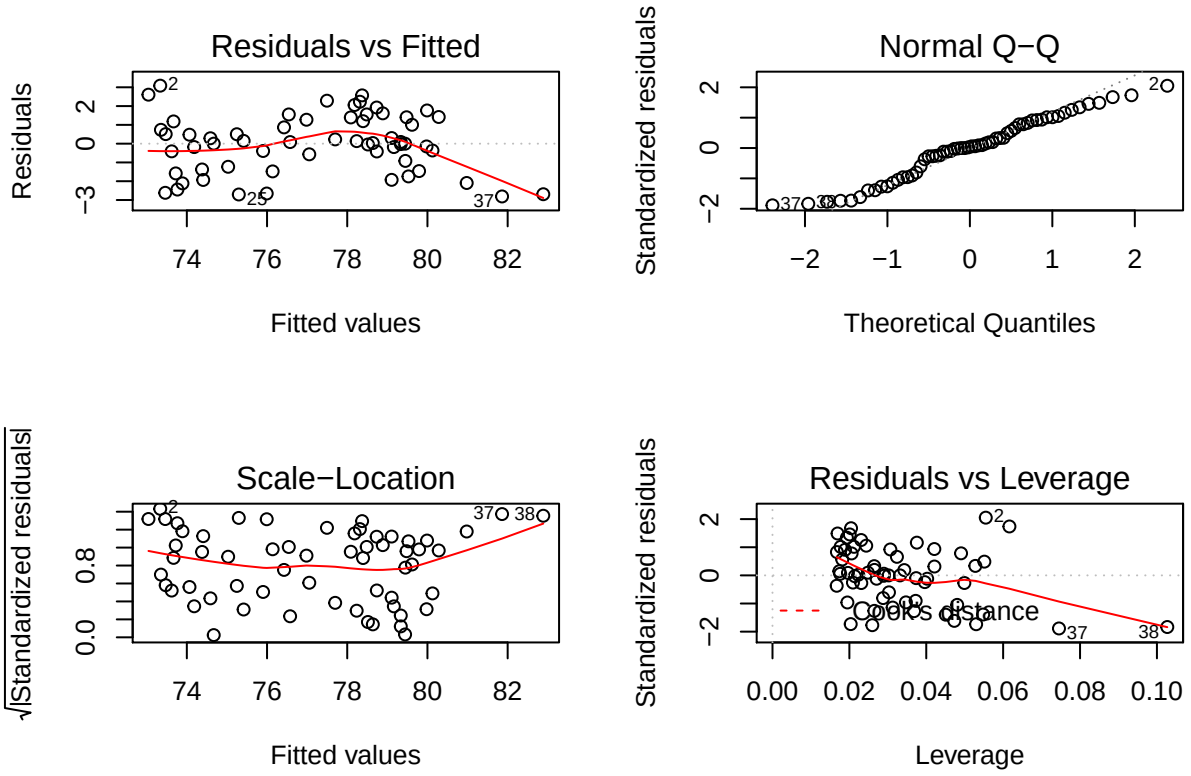
Equação: $y = 72.031 + 0.002x$

e.

$R^2 = 0.7318$ da variância

f.

```
par(mfrow=c(2,2))
plot(vidafit)
```




```
par(mfrow=c(1,1))
```

A curva no gráfico dos residuais (#1) indica que a relação entre gdpPercap e lifeExp é provávelmente de segundo grau. Esta conclusão seria apoiada pelas 2 pequenas curvas no meio da linha Q-Q e a curva no gráfico de Scale-Location. Assim, a solução em si merece pouco confiança no formato atual.

2. Loops, if . . . then

No conjunto de dados `vidaExp`, você quer criar uma nova variável categórica que expressa `gdpPercap` em duas categorias: "alto", "baixo". Você vai dividir a variável ao ponto da média da `gdpPercap`.

a. Escreva e execute um bloco de código usando `ifelse()` que cria a nova variável `pibcat`.
b. Use uma combinação de um loop e uma construção condicional ("if . . . then") para conseguir esta tarefa.

2. Solução

a.

```
vidaExp$gdpcat <- ifelse(vidaExp$gdpPercap > mean(vidaExp$gdpPercap), "alto", "baixo")
```

b.

```r
vidaExp$gdpcat2 <- 0
mediagdp <- mean(vidaExp$gdpPercap) # só quero calcular 1 vez, não cada vez que o loop roda
for (i in seq_along(vidaExp)) { # pode ser também (i in 1:nrow(vidaExp))
  if (vidaExp$gdpPercap[i] > mediagdp) {
    vidaExp$gdpcat2[i] <- "alto"
  }
  else {
    vidaExp$gdpcat2[i] <- "baixo"
  }
}
```

NB: Outras soluções são possíveis. Este não é a única possibilidade.

3. Kilometragem dos Carros

Uma sondagem sobre carros em 1970 listou 392 modelos de carros e a economia de combustível eles tiveram. Teve vários indicadores de que seria a quilometragem de combustível, como horsepower (cavalos). Para este problema, nós vamos trabalhar com `auto1`.

Perguntas e Tarefas

a. Faça uma análise exploratória dos duas variáveis (`mpg` e `horsepower`)
b. Faça um scatterplot de mpg (eixo-y) e horsepower (eixo-x). Mostra alguma tendência?
c. Tendência é linear ou não-linear? Se for não-linear, qual poder melhor expressa esta relação
d. Faça uma regressão linear simples entre mpg e horsepower. Escreva a equação da regressão e o R^2

e. Mostre os 4 gráficos para o modelo simples. Mostra uma tendência nos resíduos?
f. Faça uma regressão linear polinomial de segundo grau entre mpg e horsepower. Escreva a equação da regressão e o R^2
g. Qual modelo teve a melhor R^2?
h. Mostre os 4 gráficos para modelo polinomial.

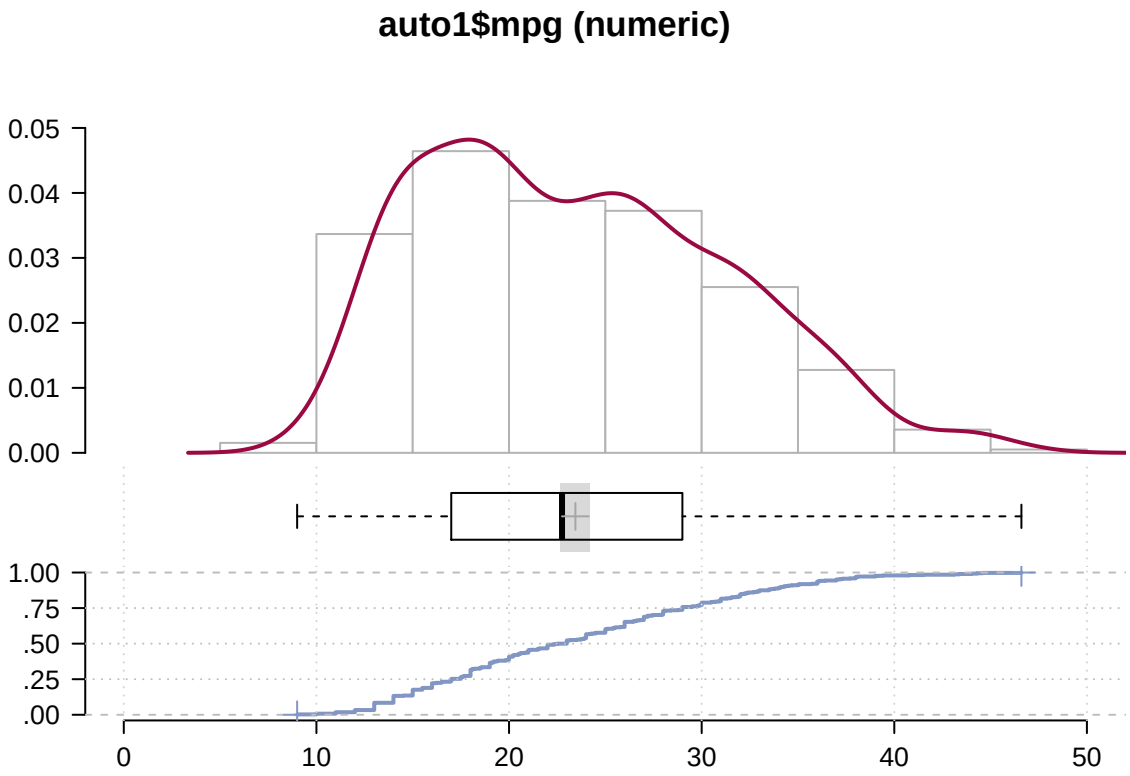
3. Solução

a. Análise Exploratória

```r
Desc(auto1$mpg)
```

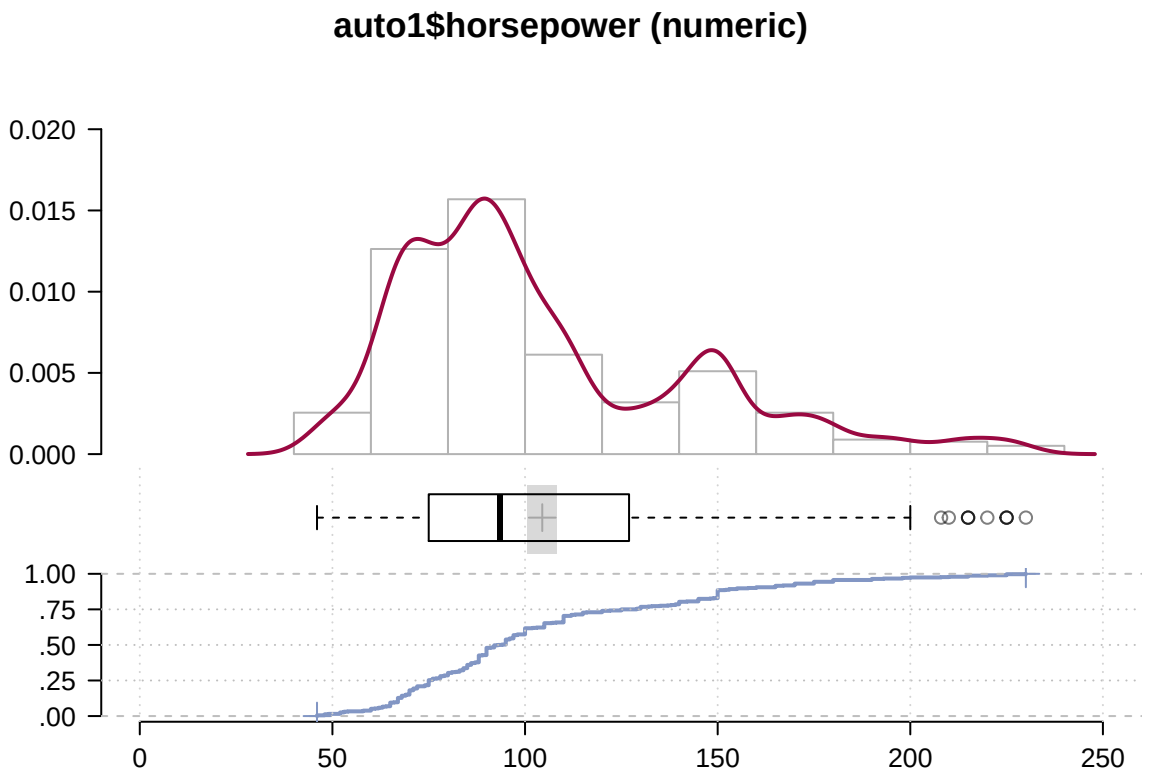
```
## ------------------------------------------------------------------------
## auto1$mpg (numeric)
##
##   length       n    NAs  unique     0s   mean  meanCI
##      392     392      0     127      0  23.45   22.67
##           100.0%   0.0%          0.0%          24.22
##
##      .05     .10    .25  median    .75    .90     .95
##    13.00   14.00  17.00   22.75  29.00  34.19   37.00
##
##    range      sd  vcoef     mad    IQR   skew    kurt
```

```
##    37.60    7.81   0.33    8.60  12.00   0.45   -0.54
##
## lowest : 9.0, 10.0 (2), 11.0 (4), 12.0 (6), 13.0 (20)
## highest: 43.4, 44.0, 44.3, 44.6, 46.6
```




```
Desc(auto1$horsepower)
```

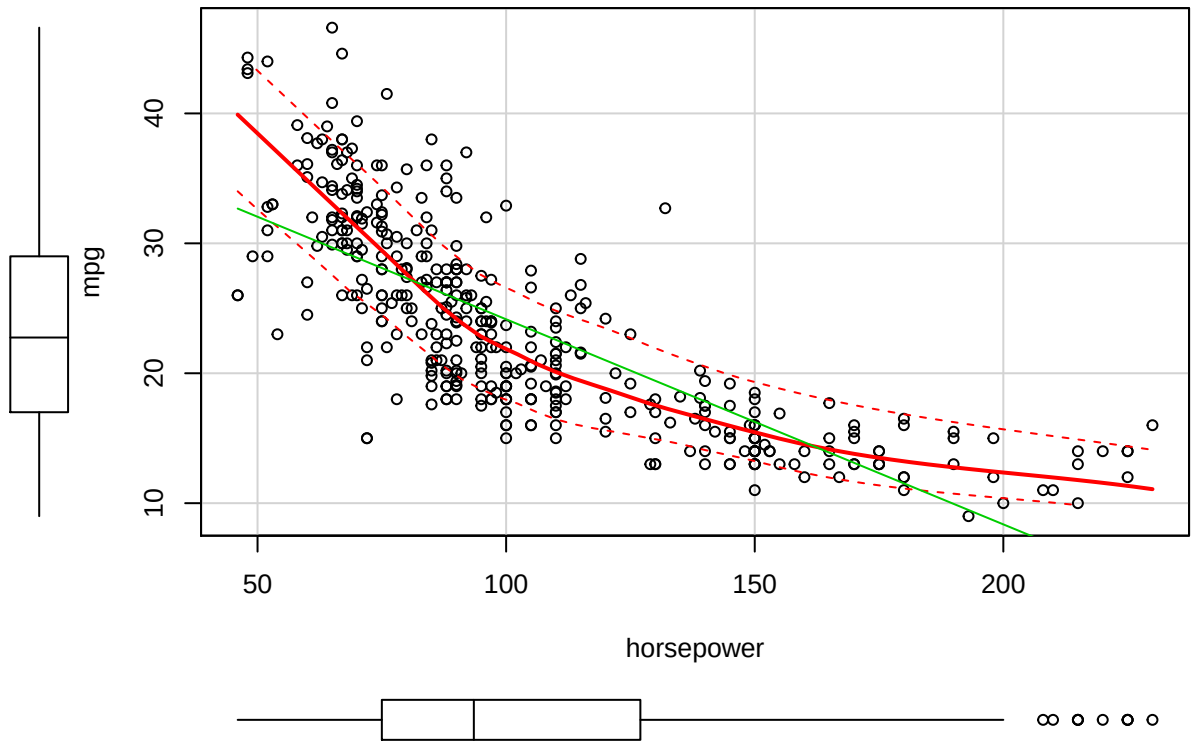
```
## ------------------------------------------------------------------------
## auto1$horsepower (numeric)
##
##   length       n    NAs  unique      0s    mean  meanCI
##      392     392      0      93       0  104.47  100.65
##           100.0%   0.0%           0.0%          108.29
##
##      .05     .10    .25  median     .75     .90     .95
##    60.55   67.00  75.00   93.50  126.00  157.70  180.00
##
##    range      sd  vcoef     mad     IQR    skew    kurt
##   184.00   38.49   0.37   28.91   51.00    1.08    0.65
##
## lowest : 46.0 (2), 48.0 (3), 49.0, 52.0 (4), 53.0 (2)
## highest: 210.0, 215.0 (3), 220.0, 225.0 (3), 230.0
```



b. Scatterplot

Tem vários que pode usar. Vou usar a função do pacote `car`

```
scatterplot(mpg ~ horsepower, data = auto1)
```



c. Tendências

A tendência não é linear. Parece de pertencer a uma equação de segundo grau.

d. Regressão Simples

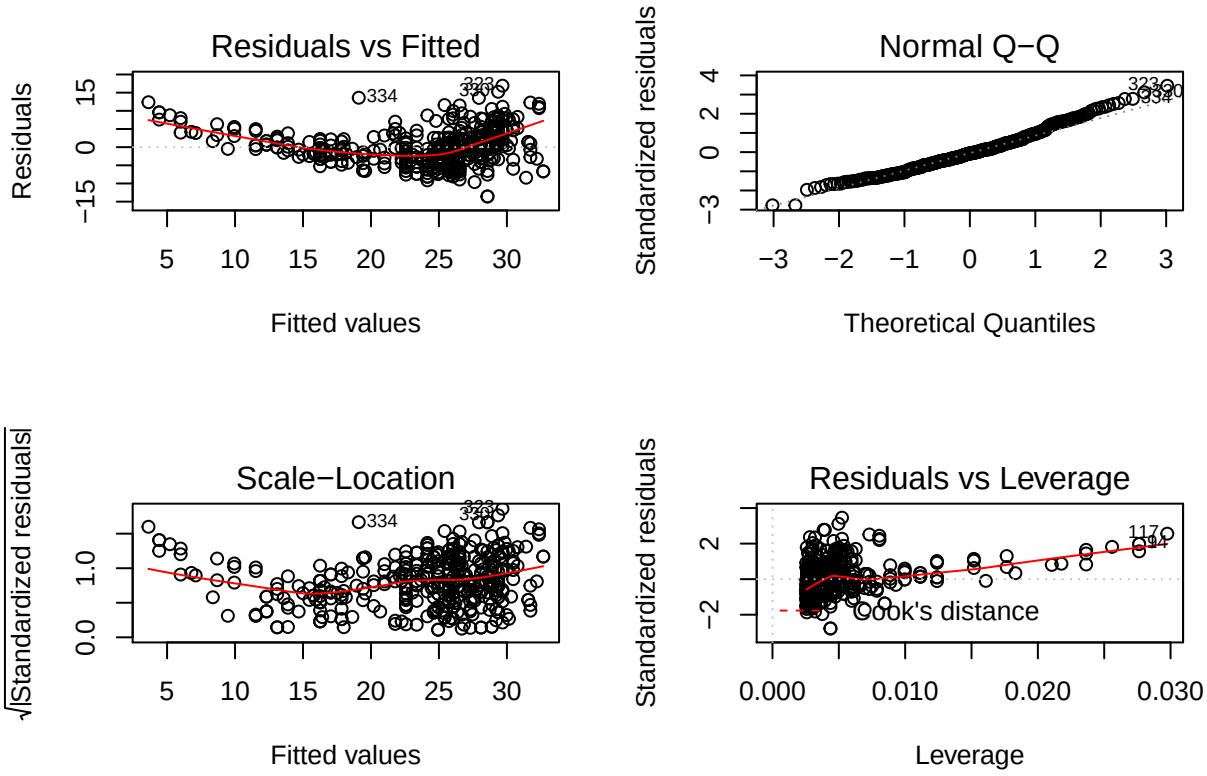
```
mpgfit1 <- lm(mpg ~ horsepower, data = auto1)
summary(mpgfit1)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = auto1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

Equação: $y = 39.936 - 0.158x$
$R^2 = 0.606$

e. 4 Gráficos

```
par(mfrow=c(2,2))
plot(mpgfit1)
```




```
par(mfrow=c(1,1))
```

Os resíduos mostram uma tendência clara de 2 grau

f. Regressão Polinomial

```
mpgfitpoli <- lm(mpg ~ horsepower + I(horsepower^2), data = auto1)
summary(mpgfitpoli)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + I(horsepower^2), data = auto1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7135  -2.5943  -0.0859   2.2868  15.8961
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     56.9000997  1.8004268   31.60   <2e-16 ***
## horsepower      -0.4661896  0.0311246  -14.98   <2e-16 ***
## I(horsepower^2)  0.0012305  0.0001221   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.374 on 389 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.686
## F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```

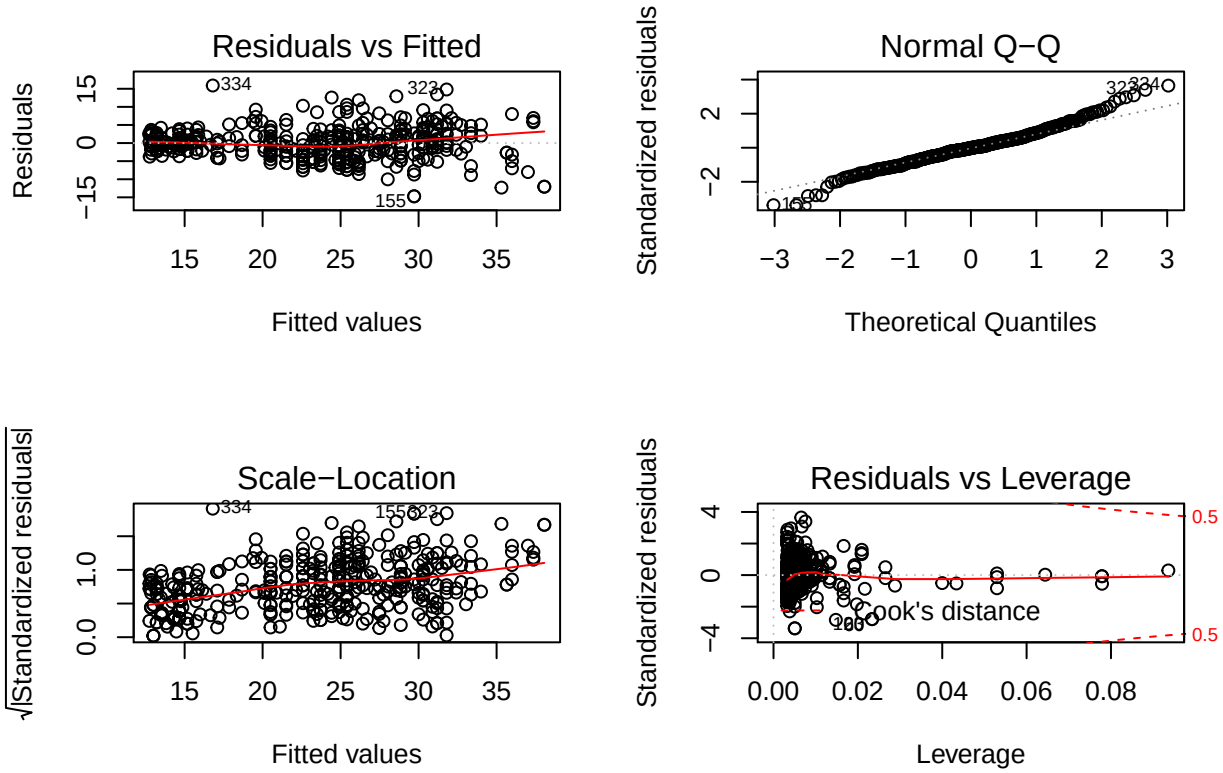
Equação: $y = 56.900 - 0.466x + 0.001x^2$
$R^2 = 0.688$

g. Melhor Modelo

Modelo polinomial

h. 4 Gráficos

```
par(mfrow=c(2,2))
plot(mpgfitpoli)
```




```
par(mfrow=c(1,1))
```

4. auto2 – Regressão Múltipla

Esta vez, nós vamos usar outras variáveis relacionados aos motores dos carros para ver se elas têm influência sobre economia de combustível. O conjunto `auto2` tem esses dados.

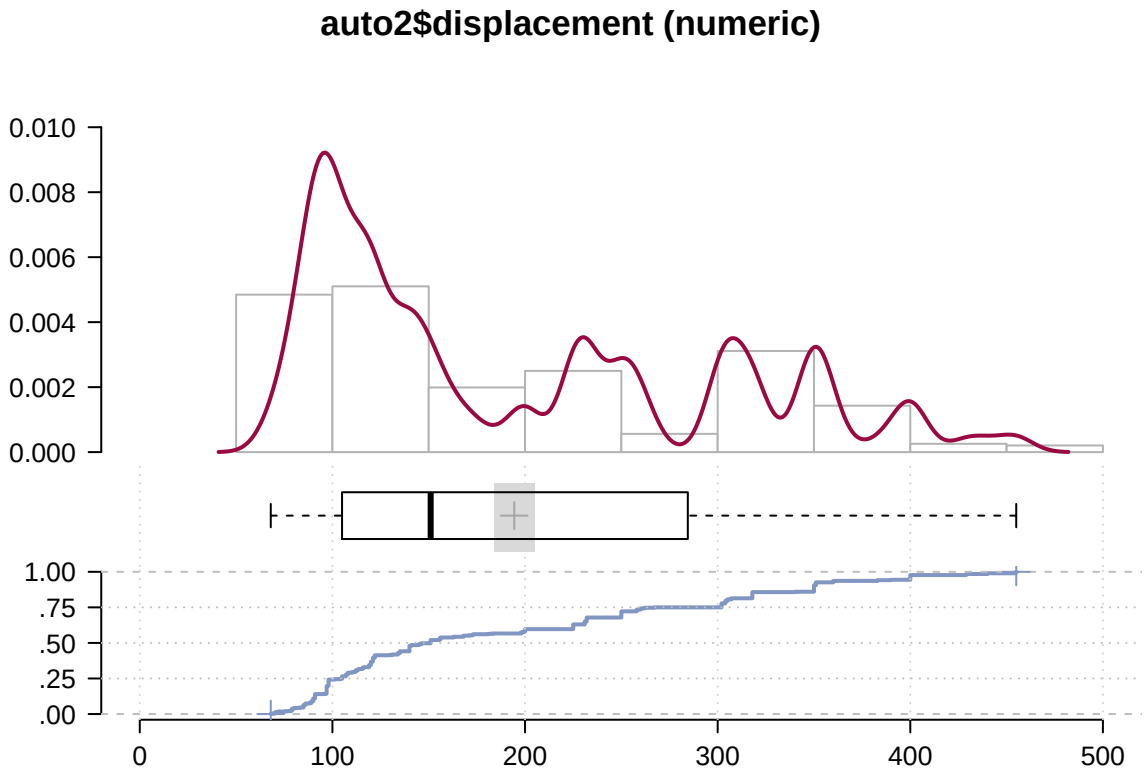
a. Faça uma análise exploratória sobre as variáveis novas (`displacement`, `weight`, `acceleration`)
b. Faça uma regressão múltipla usando todas as variáveis independentes.
c. Mostre o resultado (`summary()`)
d. Qual porcentagem da variância dos dados em total este modelo descreve?
e. Quais variáveis parecem não ter uma relação significante com a `mpg`? Porque, você acha?

4. Solução

a. Análise Exploratória

```
Desc(auto2$displacement)
```

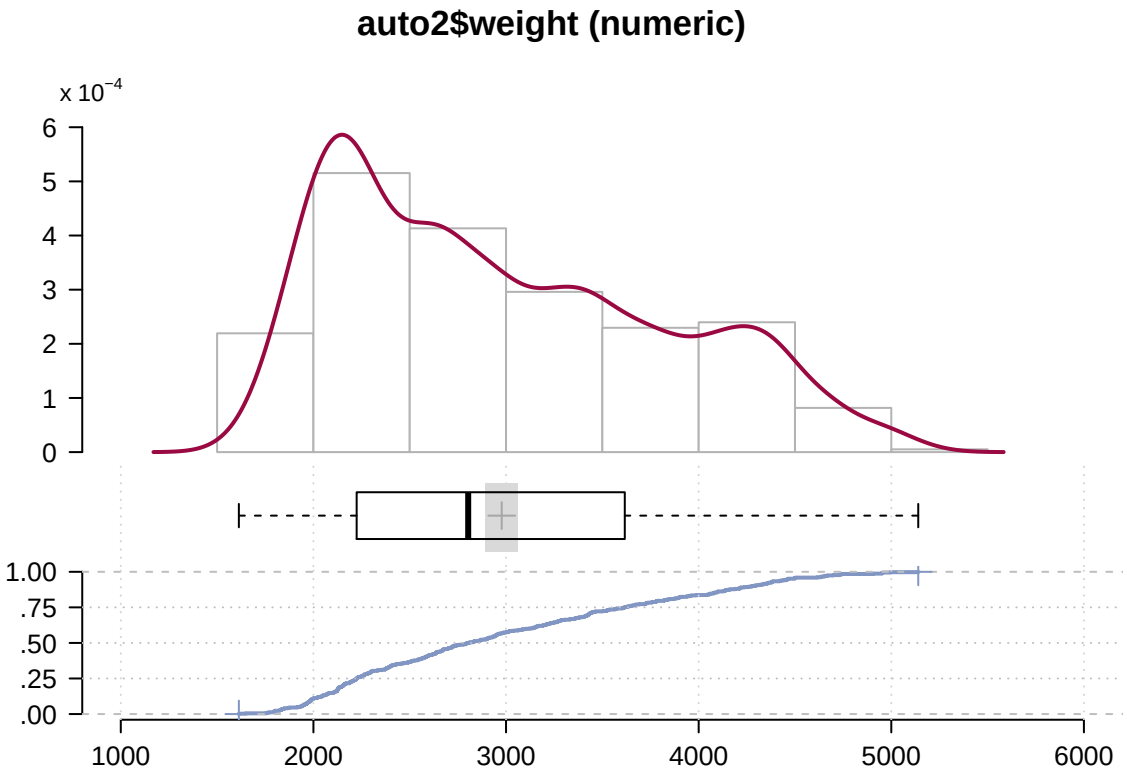
```
## ------------------------------------------------------------------------------
## auto2$displacement (numeric)
##
##   length       n     NAs  unique      0s    mean  meanCI
##      392     392       0      81       0  194.41  184.02
##           100.0%    0.0%            0.0%          204.80
##
##      .05     .10     .25  median     .75     .90     .95
##    85.00   90.00  105.00  151.00  275.75  350.00  400.00
##
##    range      sd   vcoef     mad     IQR    skew    kurt
##   387.00  104.64    0.54   90.44  170.75    0.70   -0.79
##
## lowest : 68.0, 70.0 (3), 71.0 (2), 72.0, 76.0
## highest: 400.0 (13), 429.0 (3), 440.0 (2), 454.0, 455.0 (3)
```




```
Desc(auto2$weight)
```

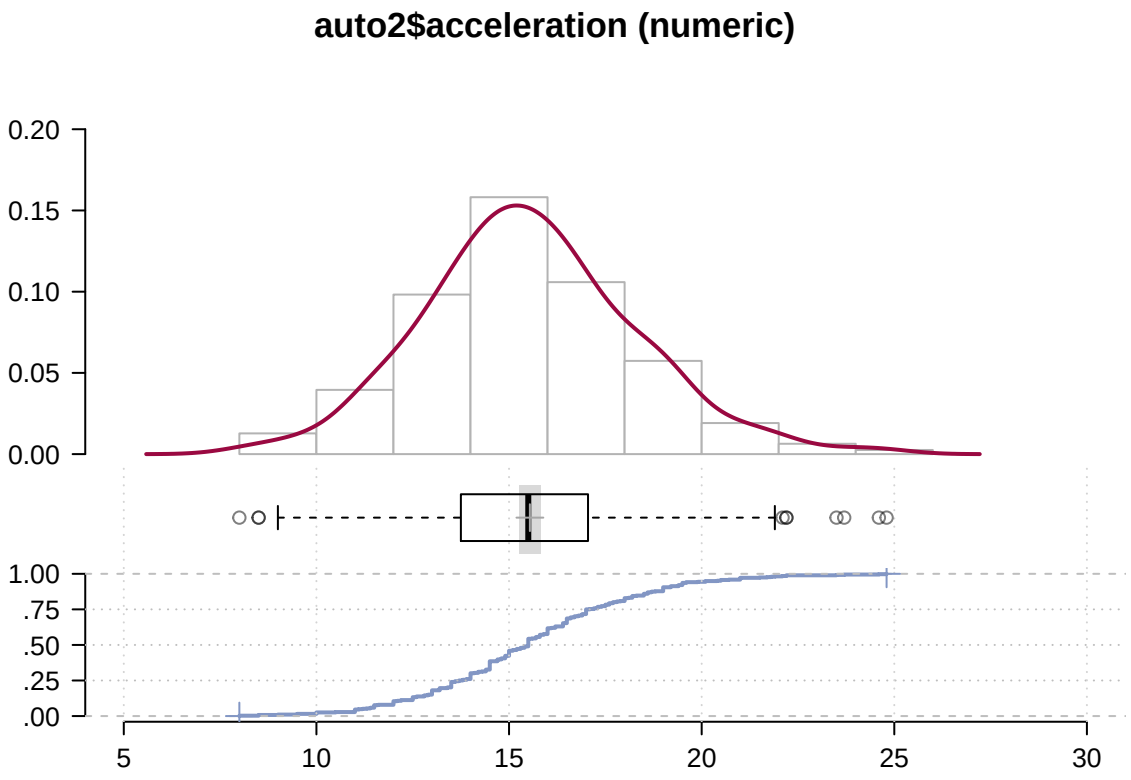
```
## ------------------------------------------------------------------------------
## auto2$weight (numeric)
##
##     length         n       NAs    unique        0s      mean    meanCI
##        392       392         0       346         0  2'977.58  2'893.24
```

```
##               100.0%      0.0%                0.0%           3'061.93
##
##        .05       .10       .25    median       .75       .90       .95
##   1'931.60  1'990.00  2'225.25  2'803.50  3'614.75  4'277.60  4'464.00
##
##      range        sd     vcoef       mad       IQR      skew      kurt
##   3'527.00    849.40      0.29    948.12  1'389.50      0.52     -0.83
##
## lowest : 1'613.0, 1'649.0, 1'755.0, 1'760.0, 1'773.0
## highest: 4'951.0, 4'952.0, 4'955.0, 4'997.0, 5'140.0
```




```
Desc(auto2$acceleration)
```

```
## ------------------------------------------------------------------------
## auto2$acceleration (numeric)
##
##   length       n     NAs  unique      0s    mean  meanCI
##      392     392       0      95       0  15.541  15.267
##           100.0%    0.0%            0.0%          15.815
##
##      .05     .10     .25  median     .75     .90     .95
##   11.255  12.000  13.775  15.500  17.025  19.000  20.235
##
##    range      sd   vcoef     mad     IQR    skew    kurt
##   16.800   2.759   0.178   2.520   3.250   0.289   0.406
##
## lowest : 8.0, 8.5 (2), 9.0, 9.5 (2), 10.0 (4)
## highest: 22.2 (2), 23.5, 23.7, 24.6, 24.8
```



b. Regressão Multiplá

```
auto2fit <- lm(mpg ~ ., data = auto2)
```

c. Resumo

```
summary(auto2fit)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = auto2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.378  -2.793  -0.333   2.193  16.256
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.2511397  2.4560447  18.424  < 2e-16 ***
## horsepower   -0.0436077  0.0165735  -2.631  0.00885 **
## displacement -0.0060009  0.0067093  -0.894  0.37166
## weight       -0.0052805  0.0008109  -6.512  2.3e-10 ***
## acceleration -0.0231480  0.1256012  -0.184  0.85388
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.247 on 387 degrees of freedom
## Multiple R-squared:  0.707,  Adjusted R-squared:  0.704
## F-statistic: 233.4 on 4 and 387 DF,  p-value: < 2.2e-16
```

d. % de Variância

$R^2 = .707$

e. Variáveis não-significativas

`displacement` e `acceleration` não parecem significativas, possivelmente porque `horsepower` também descreve a mesma caracteristica dos carros.

5. Regressão Lógistica

Vamos agora olhar num estudo sobre câncer de próstata. A questão aqui é de entender melhor se o câncer espalhou para os linfonodos em volta da próstata. O estudo tenta avaliar se cinco indicadores podem substituir uma cirurgia exploratória. As cinco variáveis no conjunto de `proscan` são

1. `raioX`: leitura de um raio X; valores binários 1 = positivo, 0 = negativo
2. `grau`: leitura patológica como resultado de uma biopsia de agulha fina; valores binários 1 = positivo, 0 = negativo
3. `estagio`: tamanho do tumor obtido pela palpação com os dedos; valores binários 1 = positivo, 0 = negativo
4. `idade`: idade do paciente em anos
5. `acido`: nível x 100 de fosfatase ácida sérica

A variável `linfonodos` tem o resultado determinado pela cirurgia se o câncer tinha espalhado ou não

Tarefas

a. Faça uma análise exploratória dos dados, inclusive com `cplot()` para entender o problema melhor
b. Construa um modelo logístico de linfonodos contra as outras variáveis
c. Todas as variáveis são significativas? Quais são e quais não são
d. Construa um segundo modelo logístico usando `raioX`, `estagio` e `acido`
e. Este modelo descreve mais da deviança nos dados?
f. Construa um terceiro modelo com só as variáveis significativas.
g. Faça uma comparação entre os três modelos. Qual é o melhor? Com este modelo, calcule os odds, um intervalo de confiança para os odds e a probabilidade de ocorrência da presença de tecido maligno nos linfonodos.

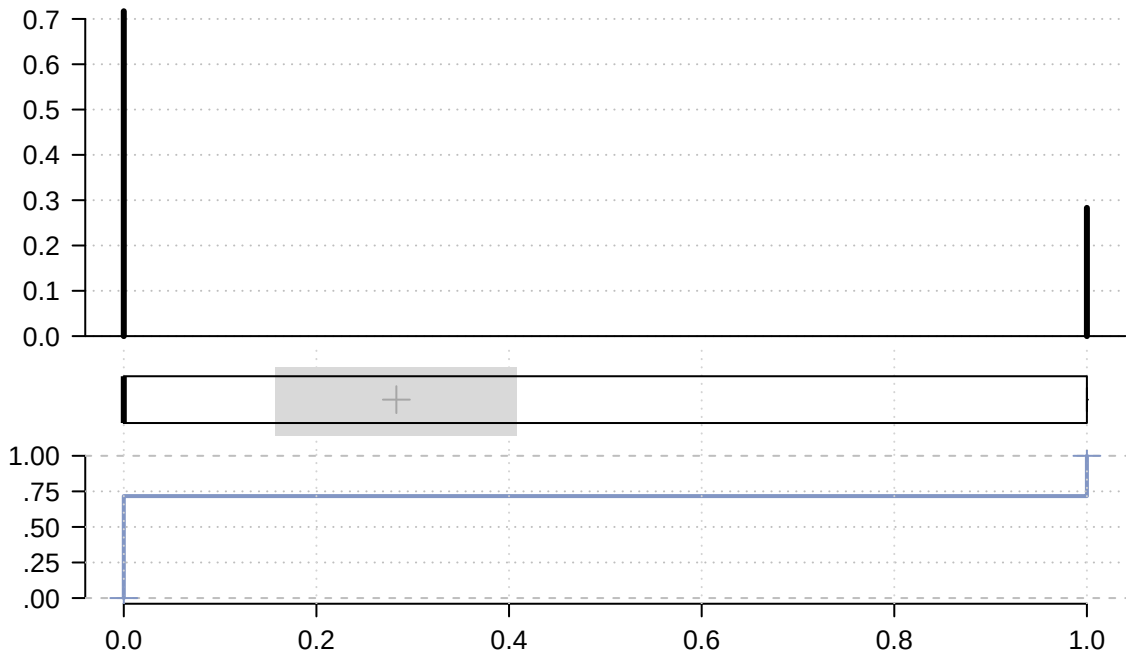
5. Solução

a. Análise Exploratória

```
Desc(proscan)
```

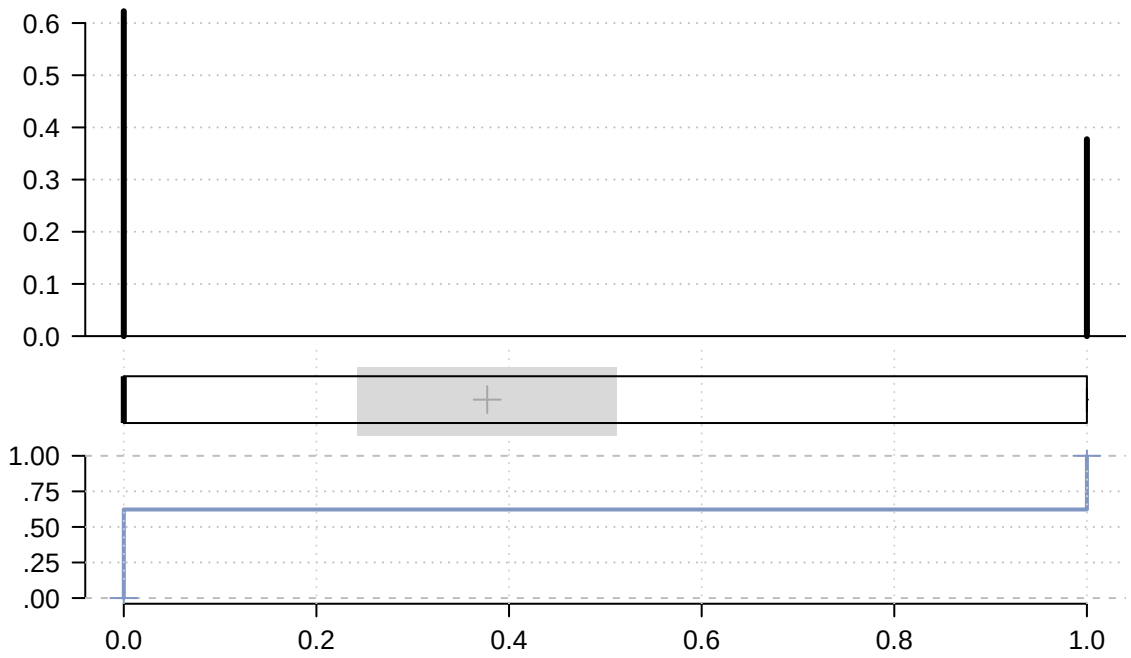
```
## ------------------------------------------------------------------------
## Describe proscan (data.frame):
##
## data.frame:  53 obs. of  6 variables
##
##   Nr  ColName     Class    NAs  Levels
##   1   raioX       numeric  .
##   2   grau        numeric  .
##   3   estagio     numeric  .
##   4   idade       numeric  .
##   5   acido       numeric  .
##   6   linfonodos  integer  .
##
##
## ------------------------------------------------------------------------
## 1 - raioX (numeric)
##
##   length       n    NAs  unique     0s  mean  meanCI
##       53      53      0       2     38  0.28    0.16
##           100.0%   0.0%          71.7%          0.41
##
##      .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    0.45   1.61    0.00   1.00  0.94   -1.14
##
##
##    level  freq   perc  cumfreq  cumperc
## 1      0    38  71.7%       38    71.7%
## 2      1    15  28.3%       53   100.0%
```

1 − raioX (numeric)



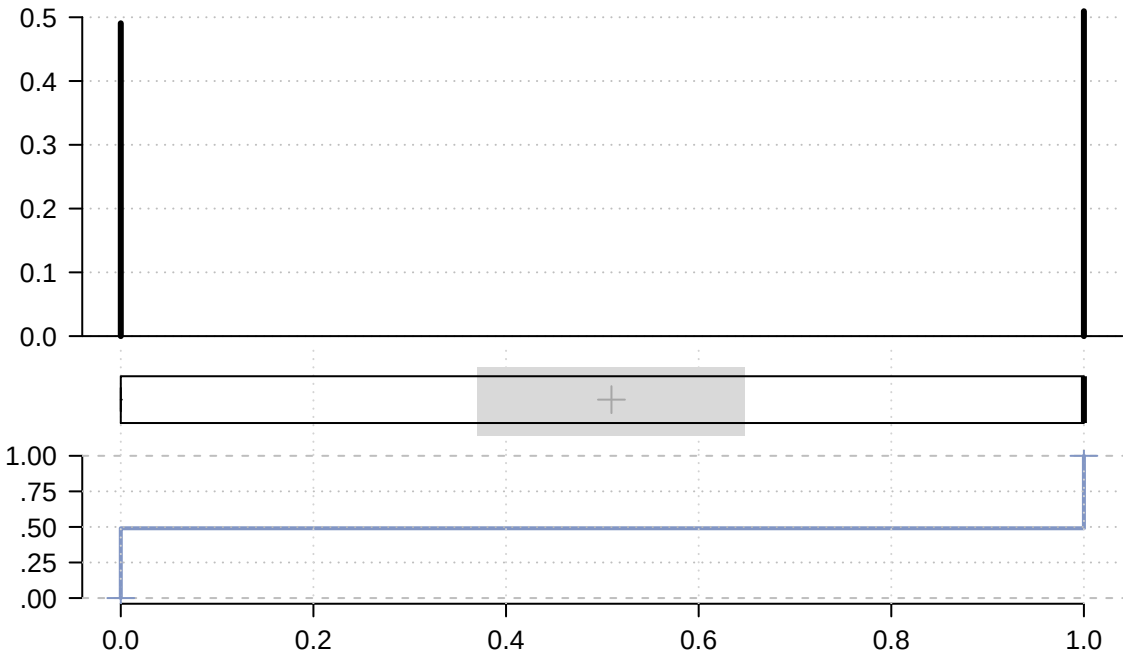

```
## ------------------------------------------------------------------------
## 2 - grau (numeric)
##
##   length       n    NAs  unique     0s  mean  meanCI
##       53      53      0       2     33  0.38    0.24
##           100.0%   0.0%          62.3%          0.51
##
##      .05     .10    .25  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    0.49   1.30    0.00   1.00  0.49   -1.79
##
##
##    level  freq   perc  cumfreq  cumperc
## 1      0    33  62.3%       33    62.3%
## 2      1    20  37.7%       53   100.0%
```

2 − grau (numeric)



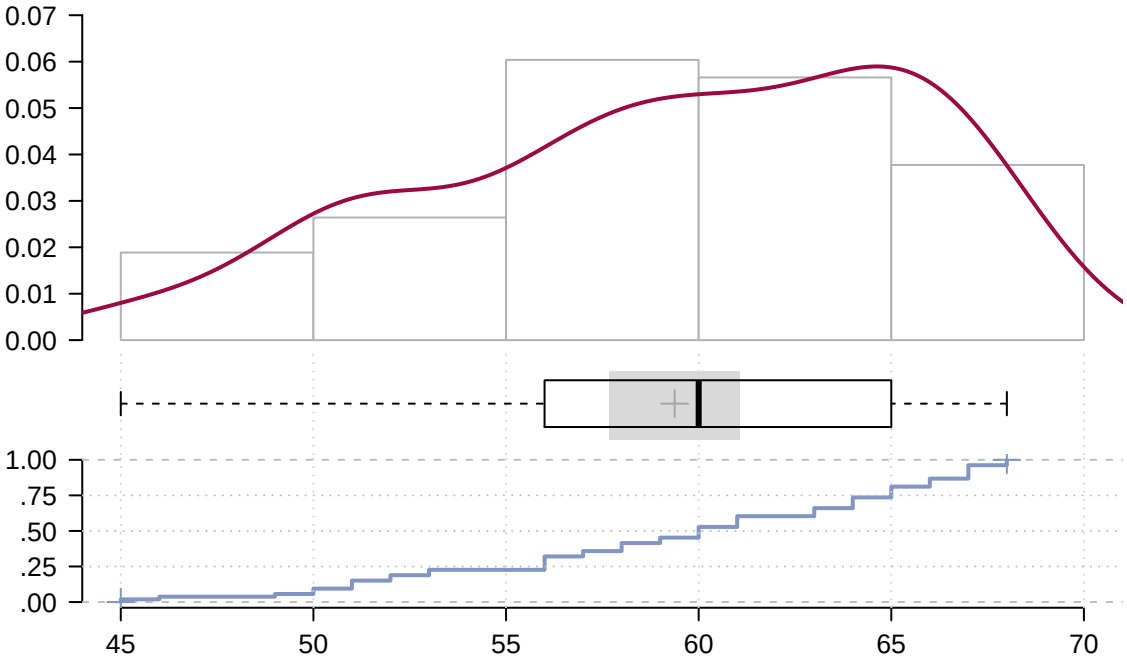

```
## ------------------------------------------------------------------------
## 3 - estagio (numeric)
##
##   length       n    NAs  unique     0s   mean  meanCI
##       53      53      0       2     26   0.51    0.37
##           100.0%   0.0%          49.1%           0.65
##
##      .05     .10    .25  median    .75    .90     .95
##     0.00    0.00   0.00    1.00   1.00   1.00    1.00
##
##    range      sd  vcoef     mad    IQR   skew    kurt
##     1.00    0.50   0.99    0.00   1.00  -0.04   -2.04
##
##
##    level  freq   perc  cumfreq  cumperc
## 1      0    26  49.1%       26    49.1%
## 2      1    27  50.9%       53   100.0%
```

3 − estagio (numeric)



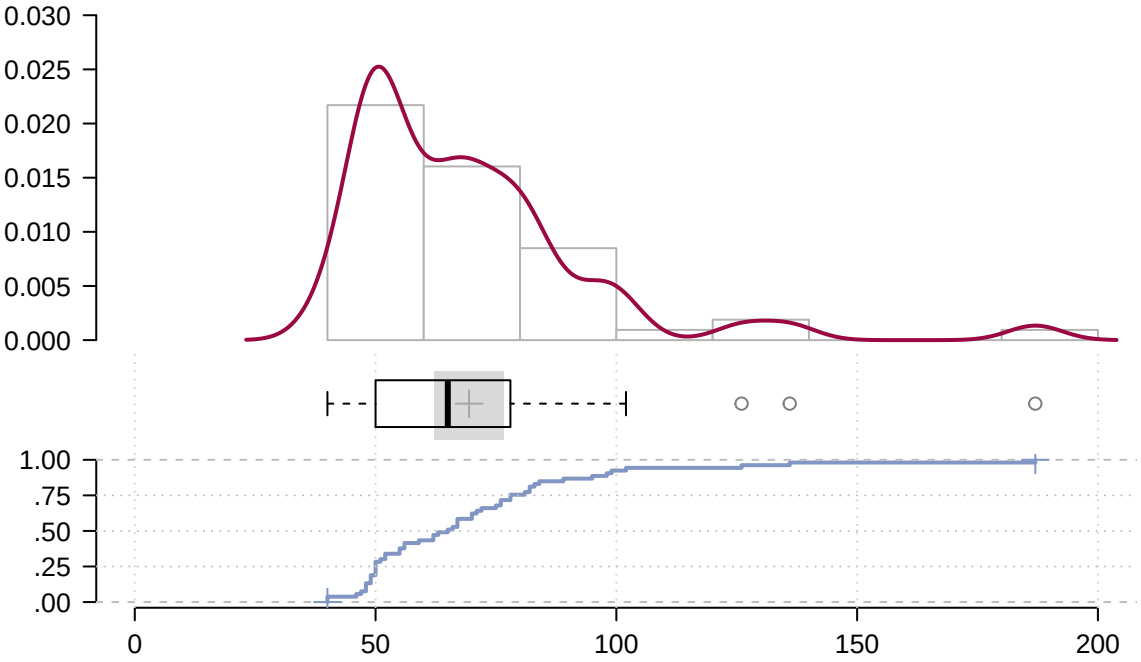

```
## ------------------------------------------------------------------------
## 4 - idade (numeric)
##
##   length       n    NAs  unique     0s   mean  meanCI
##       53      53      0      19      0  59.38   57.68
##           100.0%   0.0%           0.0%          61.08
##
##      .05     .10    .25  median    .75    .90     .95
##    49.60   51.00  56.00   60.00  65.00  67.00   67.00
##
##    range      sd  vcoef     mad    IQR   skew    kurt
##    23.00    6.17   0.10    5.93   9.00  -0.47   -0.83
##
## lowest : 45.0, 46.0, 49.0, 50.0 (2), 51.0 (3)
## highest: 64.0 (4), 65.0 (4), 66.0 (3), 67.0 (5), 68.0 (2)
```

4 − idade (numeric)



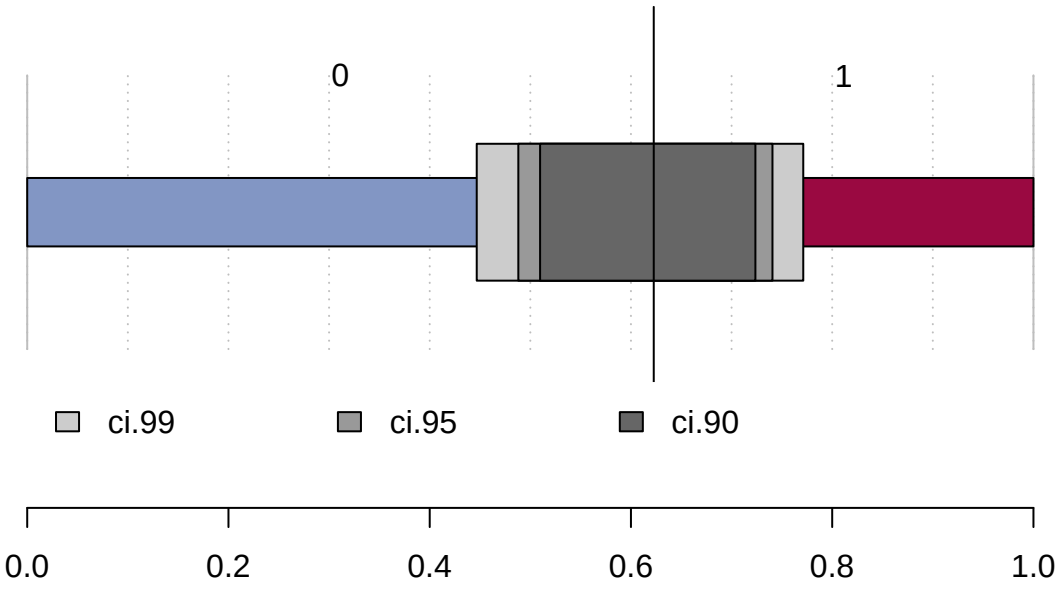

```
## ------------------------------------------------------------------------
## 5 - acido (numeric)
##
##   length       n    NAs  unique     0s   mean  meanCI
##       53      53      0      34      0  69.42   62.19
##           100.0%   0.0%           0.0%          76.64
##
##      .05     .10    .25  median    .75    .90     .95
##    46.60   48.00  50.00   65.00  78.00  97.40  111.60
##
##    range      sd  vcoef     mad    IQR   skew    kurt
##   147.00   26.20   0.38   22.24  28.00   2.13    6.16
##
## lowest : 40.0 (2), 46.0, 47.0, 48.0 (3), 49.0 (3)
## highest: 99.0, 102.0, 126.0, 136.0, 187.0
```

5 − acido (numeric)

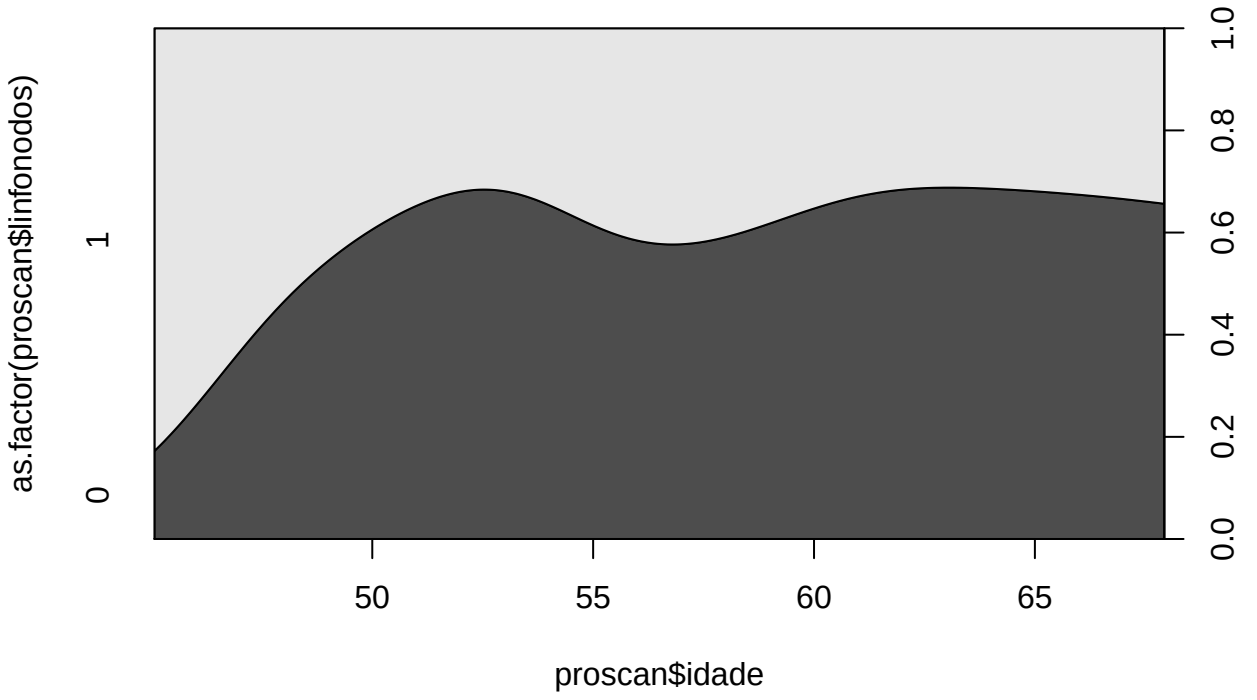



```
## ------------------------------------------------------------------------
## 6 - linfonodos (integer - dichotomous)
##
##   length      n    NAs unique
##       53     53      0      2
##          100.0%   0.0%
##
##    freq   perc  lci.95  uci.95'
## 0    33  62.3%   48.8%   74.1%
## 1    20  37.7%   25.9%   51.2%
##
## ' 95%-CI Wilson
```

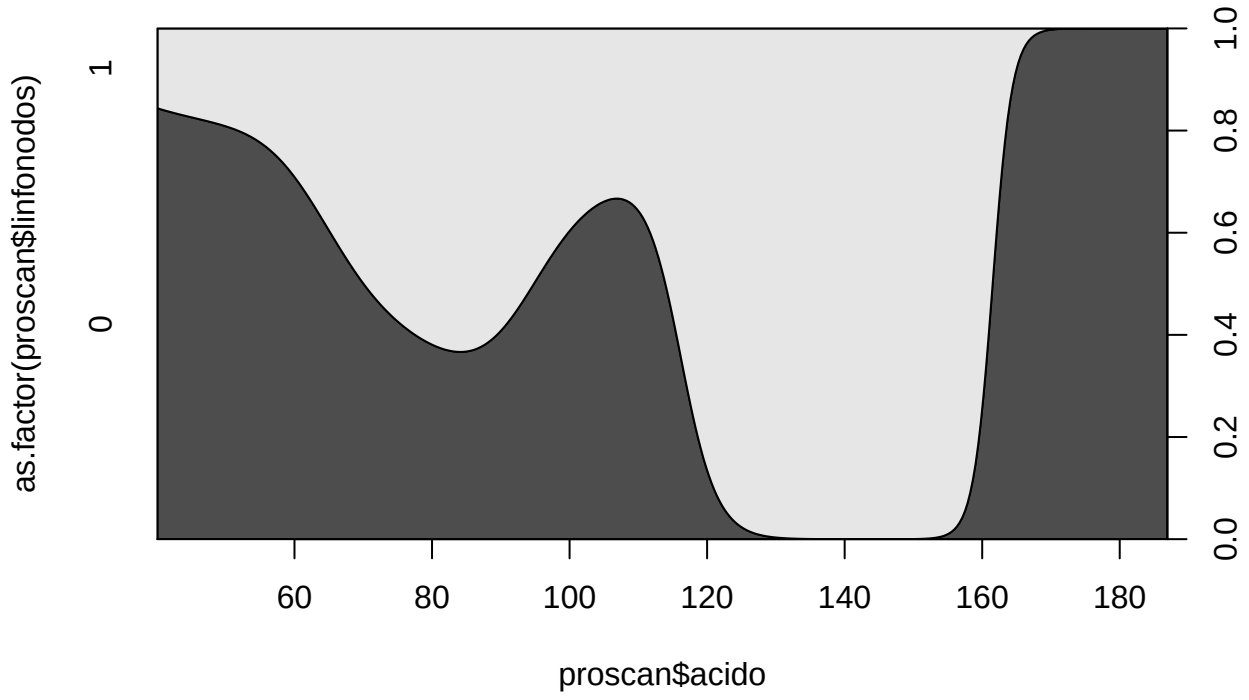
6 − linfonodos (integer − dichotomous)




```
cdplot(proscan$idade, as.factor(proscan$linfonodos))
```




```
cdplot(proscan$acido, as.factor(proscan$linfonodos))
```



b. Modelo Logístico 1

```
linffit1 <- glm(linfonodos ~ ., data = proscan, family = binomial(link = "logit"))
summary(linffit1)
```

```
##
## Call:
## glm(formula = linfonodos ~ ., family = binomial(link = "logit"),
##     data = proscan)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0110  -0.7020  -0.3654   0.5723   1.9852
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.06180    3.45992   0.018   0.9857
## raioX        2.04534    0.80718   2.534   0.0113 *
## grau         0.76142    0.77077   0.988   0.3232
## estagio      1.56410    0.77401   2.021   0.0433 *
## idade       -0.06926    0.05788  -1.197   0.2314
## acido        0.02434    0.01316   1.850   0.0643 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.252  on 52  degrees of freedom
## Residual deviance: 48.126  on 47  degrees of freedom
## AIC: 60.126
```

```
##
## Number of Fisher Scoring iterations: 5
```

c. Variáveis Significativos

Não todas são significativas. Significativos: `raioX` e `estagio`. Outras: não

d. Segundo Modelo

```
linffit2 <- glm(linfonodos ~ raioX + estagio + acido, data = proscan,
                family = binomial(link = "logit"))
summary(linffit2)
```

```
##
## Call:
## glm(formula = linfonodos ~ raioX + estagio + acido, family = binomial(link = "logit"),
##     data = proscan)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8630  -0.8508  -0.3889   0.5721   2.2386
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.57565    1.18115  -3.027  0.00247 **
## raioX        2.06179    0.77767   2.651  0.00802 **
## estagio      1.75556    0.73902   2.376  0.01752 *
## acido        0.02063    0.01265   1.631  0.10291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.252  on 52  degrees of freedom
## Residual deviance: 50.660  on 49  degrees of freedom
## AIC: 58.66
##
## Number of Fisher Scoring iterations: 4
```

e. Modelo descreve mais de deviância

Apesar da melhora no AIC, este modelo tem um leve aumento no desvio residual (de 48.126 até 50.660). Então formalmente, piorou o desvio.

f. Modelo 3 - Variáveis Significativos

```
linffit3 <- glm(linfonodos ~ raioX + estagio, data = proscan,
                family = binomial(link = "logit"))
summary(linffit3)
```

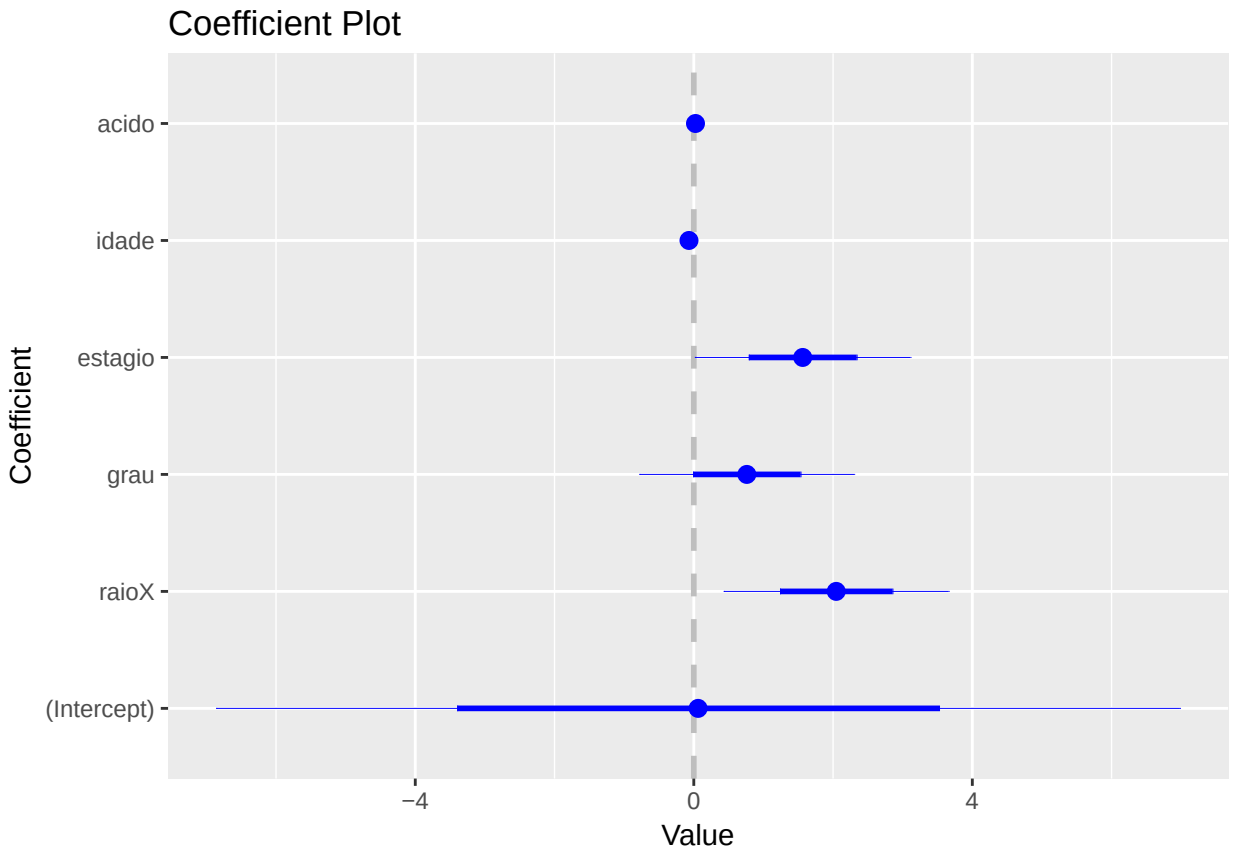
```
##
## Call:
## glm(formula = linfonodos ~ raioX + estagio, family = binomial(link = "logit"),
##     data = proscan)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9166  -0.9907  -0.4934   0.5892   2.0815
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.0446     0.6100  -3.352 0.000802 ***
## raioX         2.1194     0.7468   2.838 0.004541 **
## estagio       1.5883     0.7000   2.269 0.023274 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.252  on 52  degrees of freedom
## Residual deviance: 53.353  on 50  degrees of freedom
## AIC: 59.353
##
## Number of Fisher Scoring iterations: 4
```

g. Comparação

O melhor modelo parece de ser o 1º porque tem o desvio residual mínimo.

```
invlogit <- function(x) { ## função para calcular invlogit
   1/(1 + exp(-x))
}
coefplot(linffit1)
```

```
## Warning: Ignoring unknown aesthetics: xmin, xmax
```




```
paste("Relação de Odds:")
```

```
## [1] "Relação de Odds:"
```

```
exp(coef(linffit1)) # Calculate the odds
```

```
## (Intercept)       raioX        grau     estagio       idade       acido
##   1.0637501   7.7318248   2.1413054   4.7783782   0.9330843   1.0246432
```

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

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

```
exp(confint(linffit1))
```

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

```
##                   2.5 %      97.5 %
## (Intercept) 0.001034489 1069.888640
## raioX       1.716422071   43.757848
## grau        0.463285378   10.063372
## estagio     1.111673710   24.644741
## idade       0.826708615    1.042456
## acido       0.998575134    1.054816
```

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

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

```
invlogit(linffit3$coefficients)
```

```
## (Intercept)       raioX     estagio
##   0.1145964   0.8927786   0.8303733
```