Exercício de Métodos Inferenciais Avançados 2/18

William Peixoto

knit mais recente: 31 de outubro 2018

Introdução

O exercício consiste em construir um ou mais modelos lineares em uma base pública de dados sobre nascimentos na Carolina do Norte (EUA) no ano de 2001, que consiste em uma amostra de 1450 registros de nascimento selecionados pelo estatístico John Holcomb.

Carregue a base de dados **NCBirths**, disponível em https://vincentarelbundock.github.io /Rdatasets/datasets.html (https://vincentarelbundock.github.io/Rdatasets/datasets.html).

Desenvolva os itens a seguir aplicando técnicas gráficas e formais, e apresente resultados, explicações e considerações que julgar necessários.

Utilize o presente documento Markdown (referências em http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com), https://bookdown.org/yihui/rmarkdown/ (https://bookdown.org /yihui/rmarkdown/)) para a inserção do código e geração de resultados.

Regressão Múltipla

Estrutura e variáveis

a. Conheça a estrutura dos dados e explore as variáveis quantitativas e qualitativas.

```
# bloco de código - item a


NCbirths = read.csv("NCbirths.csv")
nc_births = NCbirths
str(nc_births)
```

```
## 'data.frame':    1450 obs. of  16 variables:
##  $ X            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Plural       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Sex          : int  1 2 1 1 1 1 2 2 2 2 ...
##  $ MomAge       : int  32 32 27 27 25 28 25 15 21 27 ...
##  $ Weeks        : int  40 37 39 39 39 43 39 42 39 40 ...
##  $ Marital      : int  1 1 1 1 1 1 1 2 1 2 ...
##  $ RaceMom      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ HispMom      : Factor w/ 6 levels "C","M","N","O",..: 3 3 3 3 3
3 3 3 3 3 ...
##  $ Gained       : int  38 34 12 15 32 32 75 25 28 37 ...
##  $ Smoke        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BirthWeightOz: int  111 116 138 136 121 117 143 113 120 124 ...
##  $ BirthWeightGm: num  3147 3289 3912 3856 3430 ...
##  $ Low          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Premie       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ MomRace      : Factor w/ 4 levels "black","hispanic",..: 4 4 4 4
4 4 4 4 4 4 ...
```

```
summary(nc_births)
```

```
##        X                ID             Plural          Sex
##  Min.   :   1.0   Min.   :   1.0   Min.   :1.000   Min.   :1.000
##  1st Qu.: 363.2   1st Qu.: 363.2   1st Qu.:1.000   1st Qu.:1.000
##  Median : 725.5   Median : 725.5   Median :1.000   Median :1.000
##  Mean   : 725.5   Mean   : 725.5   Mean   :1.037   Mean   :1.487
##  3rd Qu.:1087.8   3rd Qu.:1087.8   3rd Qu.:1.000   3rd Qu.:2.000
##  Max.   :1450.0   Max.   :1450.0   Max.   :3.000   Max.   :2.000
##
##      MomAge          Weeks          Marital         RaceMom      Hi
spMom
##  Min.   :13.00   Min.   :22.00   Min.   :1.000   Min.   :1.000   C:
2
##  1st Qu.:22.00   1st Qu.:38.00   1st Qu.:1.000   1st Qu.:1.000   M:
128
##  Median :26.00   Median :39.00   Median :1.000   Median :1.000   N:
1283
##  Mean   :26.76   Mean   :38.62   Mean   :1.345   Mean   :1.831   O:
3
##  3rd Qu.:31.00   3rd Qu.:40.00   3rd Qu.:2.000   3rd Qu.:2.000   P:
9
##  Max.   :43.00   Max.   :45.00   Max.   :2.000   Max.   :8.000   S:
25
##                  NA's   :1
##      Gained          Smoke        BirthWeightOz   BirthWeightGm
##  Min.   : 0.0   Min.   :0.0000   Min.   : 12.0   Min.   : 340.2
##  1st Qu.:20.0   1st Qu.:0.0000   1st Qu.:106.0   1st Qu.:3005.1
##  Median :30.0   Median :0.0000   Median :118.0   Median :3345.3
##  Mean   :30.6   Mean   :0.1446   Mean   :116.2   Mean   :3295.6
##  3rd Qu.:40.0   3rd Qu.:0.0000   3rd Qu.:130.0   3rd Qu.:3685.5
##  Max.   :95.0   Max.   :1.0000   Max.   :181.0   Max.   :5131.4
##  NA's   :40     NA's   :5
##       Low              Premie           MomRace
##  Min.   :0.00000   Min.   :0.0000   black   :332
##  1st Qu.:0.00000   1st Qu.:0.0000   hispanic:164
##  Median :0.00000   Median :0.0000   other   : 48
##  Mean   :0.08621   Mean   :0.1317   white   :906
##  3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000
##
```

Uma inspeção visual já permite identificar NAs nos campos `Weeks` , `Gained` e `Smoke` , mas os campos que a descrição diz serem categóricos foram tratados como numéricos, e seus sumários não fazem sentido.

Possíveis variáveis dependentes

Não foi especificado um problema para ser estudado. Vejamos como cada campo poderia ser interpretado se fosse a variável dependente. Há dois grupos: Características da mãe e as da gestação e bebê:

Maternas

- `MomAge` : Embora se apresente como numérica discreta, pode ser considerada contínua.
- `Marital` : Categórica com apenas dois valores (Casada ou não)
- Dados "raciais": `MomRace` , `RaceMom` e `HispMom` , todos categóricos
- `Smoke` , categórico
- `Gained` , apesar de ser o peso ganho pela mãe, foi considerado parte da gestação e comentado na seção seguinte

Gestação e bebê

- `Plural` : Categórica, pode ser uma logística por Single/Not single
- `Sex` : Categórica, também pode ser uma logística por Male/Female
- `Gained` : Contínua
- Se for possível prever algum valor, ele pode ser usado para preencher os 40 NAs encontrados.
- Tempo de gestação:
  - `Weeks` : Numérica discreta
  - `Premie` : Categórica, depende de `Weeks`
- Peso do bebê:
  - `BirthWeightOz` ou `BirthWeightGm` , numéricas. Em Oz, parece ser apenas inteiro enquanto em g, é float.
    - Se a conversão estiver correta, basta eliminar uma das colunas. ~~Mas estou curioso para ver se o modelo linear vai mostrar alguma diferença inexistente.~~
    - A documentação não diz como interpretar esses valores no caso de gravidez múltipla (`Plural` != "Single")
- `Low` : Categórica. Adequada para logística. Depende integralmente do peso do bebê.

Prováveis Correlações entre as colunas

- A mais gritante é a do peso dos bebês, a mesma informação em duas unidades.
- `MomRace` , `HispMom` e `MomRace` (ô povo racista!): Há grande sobreposição entre os conceitos.
  - Pode valer a pena converter os três em uma única coluna
- `Weeks` e `Premie` : A segunda depende completamente da primeira
- `BirthWeight{Oz,Gm}` e `Low` : Mesmo caso: A segunda também depende completamente das primeiras
  - TODO: ~~Verificar se todos os~~ ~~Low~~ ~~realmente têm peso menor que 2500g~~

Escolhas disponíveis

Todas as colunas poderiam ser sujeitas a tentativas de explicação, mas suas características são diferentes:

Colunas candidatas a variável dependente, e suas características

	Quantitativas	**Categóricas**
Mãe	`Gained` [1], `MomAge`	`Smoke` , `Marital` , "raças"[3]
Gestação	`Gained` [1], `Weeks`	`Plural` , `Premie`
Bebê	Peso[2]	`Sex` , `Low`

As "raças" estão espalhadas por três variáveis: `MomRace` , `RaceMom` e `HispMom`

As variáveis quantitativas podem ser objeto de estudo de modelos lineares, enquanto as categóricas precisam ser explicadas por outros modelos. Em particular, as binárias (`Sex` , `Smoke` , `Marital` , `Premie` e `Low`) poderiam ser estudadas com a regressão logística.

Num contexto humano, não faz sentido prever grupo étnico, estado civil ou idade da mãe com base nos dados disponíveis, então as variáveis `MomAge` , `Marital` e as "raças" não serão objeto de estudo dessa natureza.

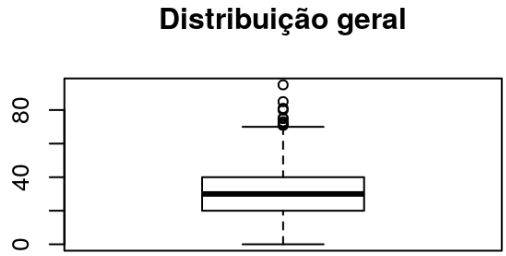
1 - Quem ganha o peso é a mãe, mas isso decorre da gravidez
2 - O peso é dado em duas unidades, basta ignorar uma delas
3 - Geneticamente, o conceito de raça humana não existe (https://www.eurekalert.org/pub_releases/1998-10/WUiS-GSRD-071098.php), mas culturalmente (https://en.wikipedia.org/wiki/Race_%28human_categorization%29) a distinção pode ser importante, como no caso do local da origem desses dados.

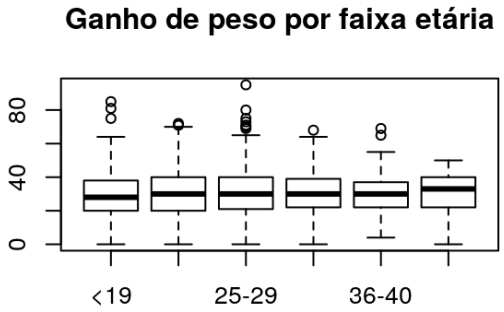
Exploração de dados

As três variáveis quantitativas que podem (talvez) serem estudadas com regressão linear múltipla são:

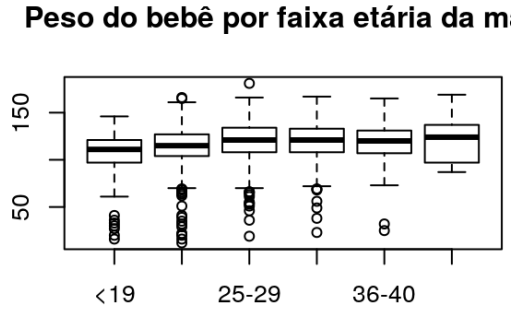
- Variação de peso da mãe durante a gestação (`Gained`)
- Peso do bebê ao nascer (`BirthWeightOz`)
- Duração da gravidez (`Weeks`)

Peso da mãe

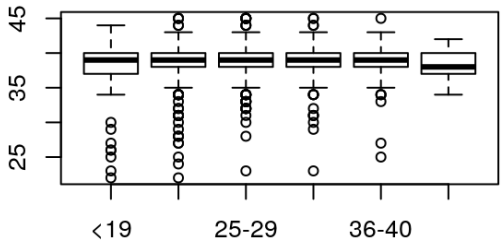




Há pouca variação aparente

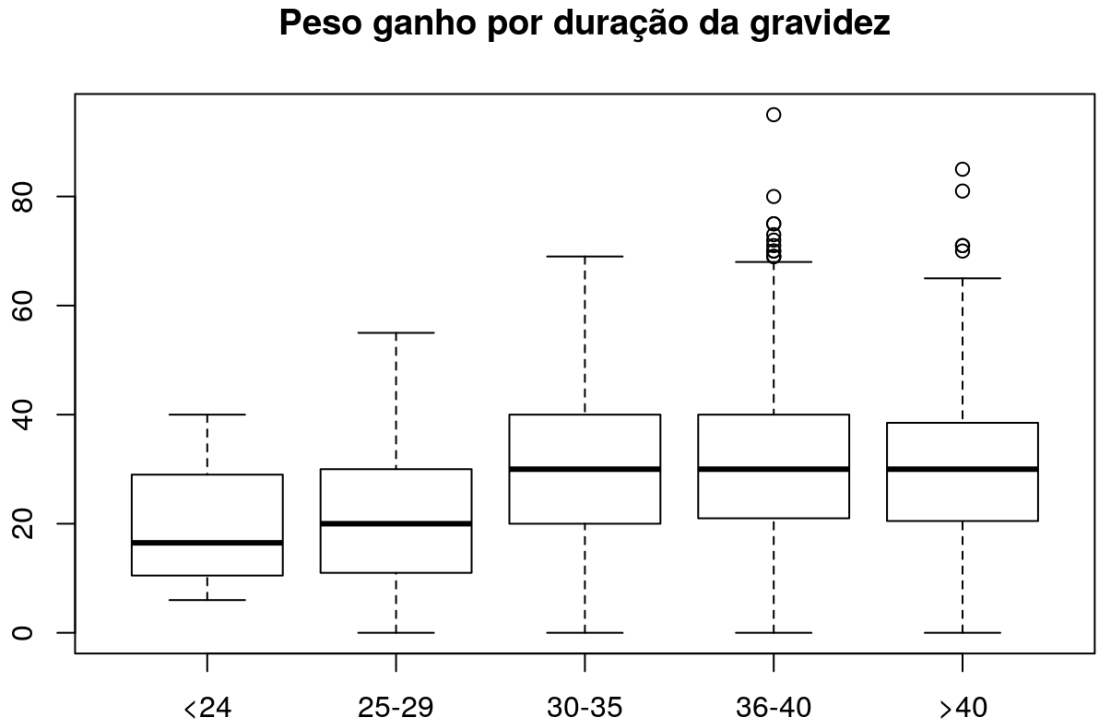


Não faz sentido

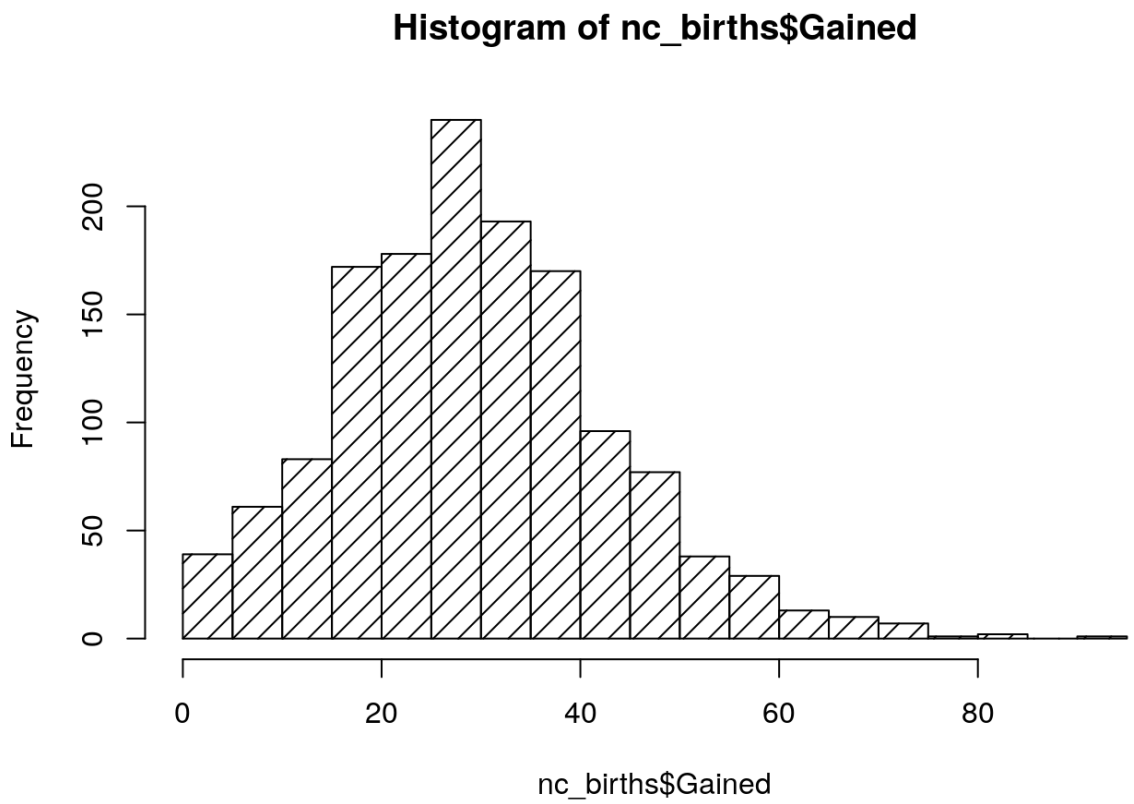


Há pouca variação aparente

Correlações simples



```
hist(nc_births$Gained, density = 10, breaks = 30)
```



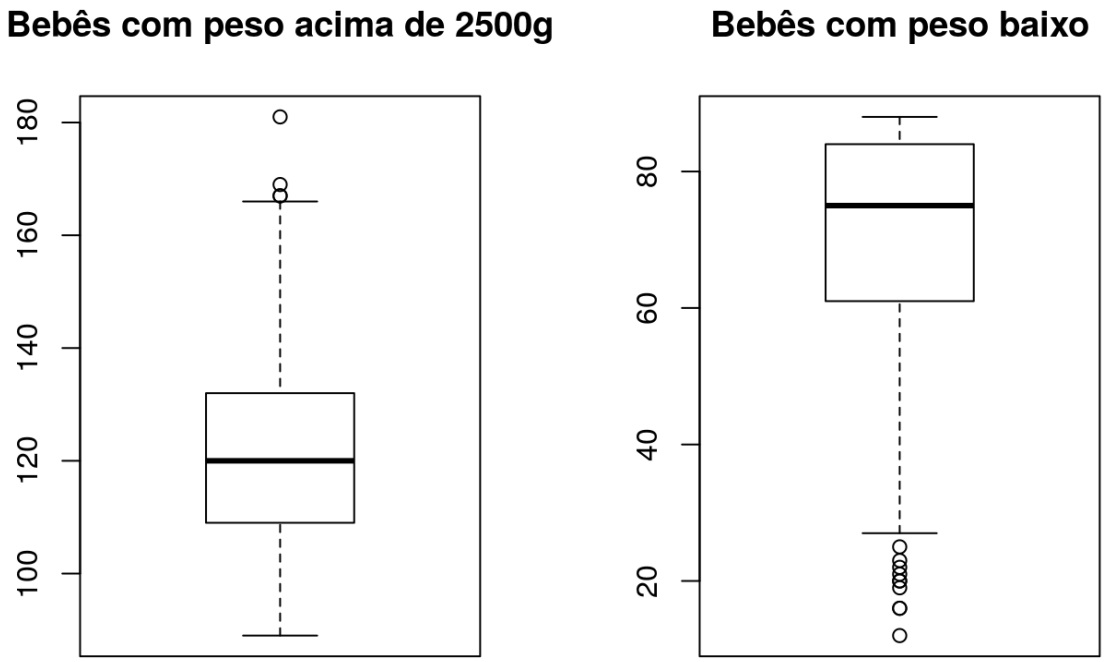

```
# density(nc_births$Gained)
sw_gained = shapiro.test(nc_births$Gained)

gained_mean = mean(nc_births$Gained, na.rm = T)
```

A distribuição das frequências de pesos ganhos parece ser normal, com média 30.6014184 e desvio padrão 13.8774929 (apenas os valores presentes. 40 estão faltando), o resultado de um teste Shapiro-Wil é 0.9843275 com $p < 5\%$ (3.12516610^{-11}).

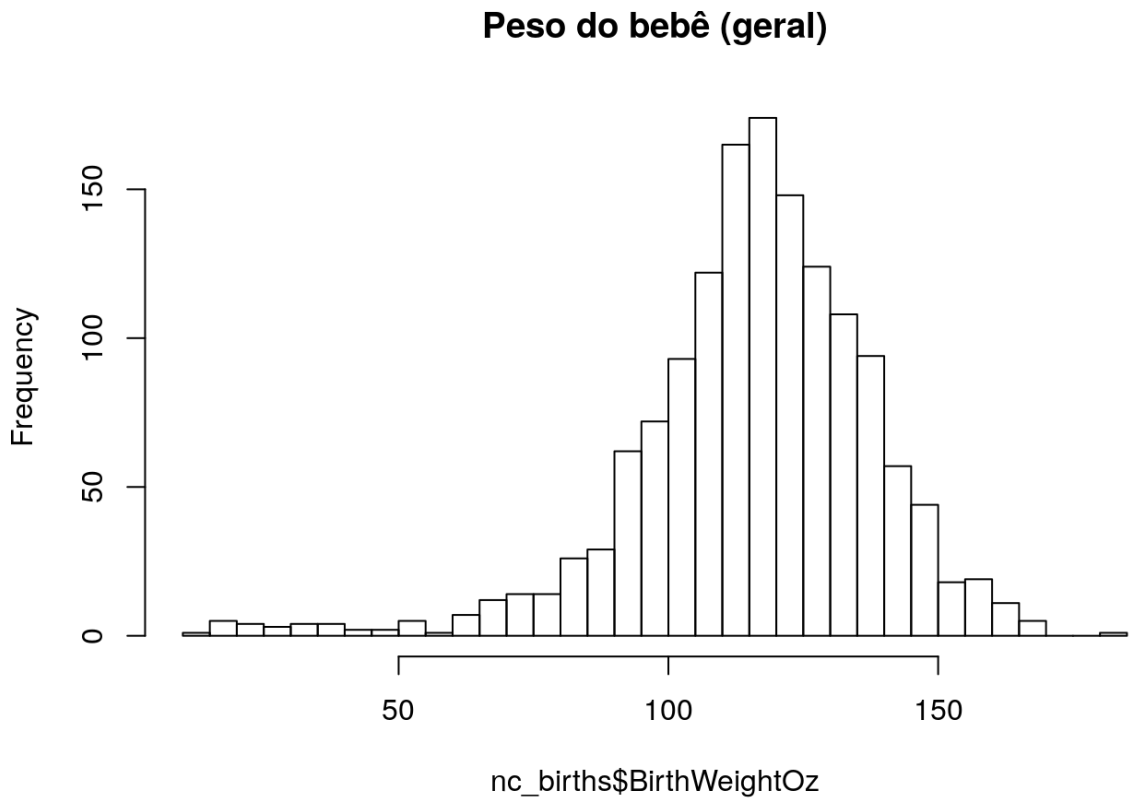
Peso do bebê

TODO: adasad

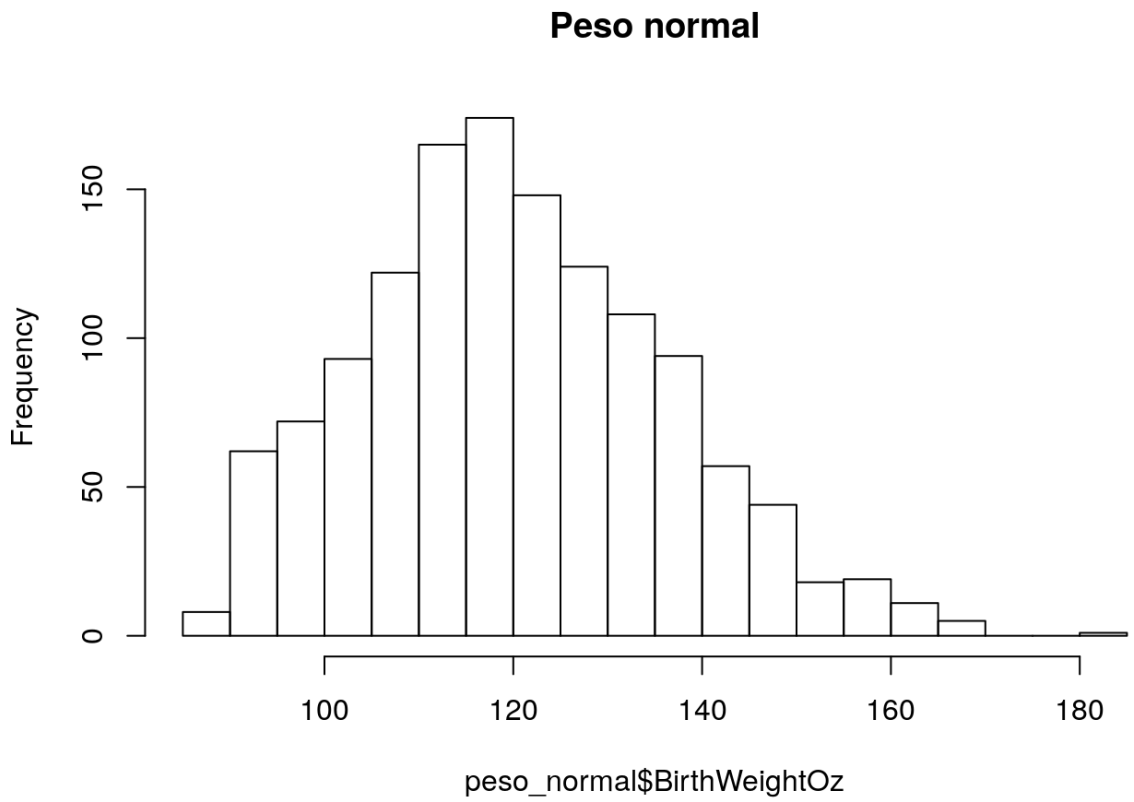
```
peso_normal = nc_births[nc_births$Low==0,]
baixo_peso = nc_births[nc_births$Low==1,]

par(mfrow = c(1, 2))
boxplot(peso_normal$BirthWeightOz,
        main="Bebês com peso acima de 2500g")

boxplot(baixo_peso$BirthWeightOz,
        main="Bebês com peso baixo")
```



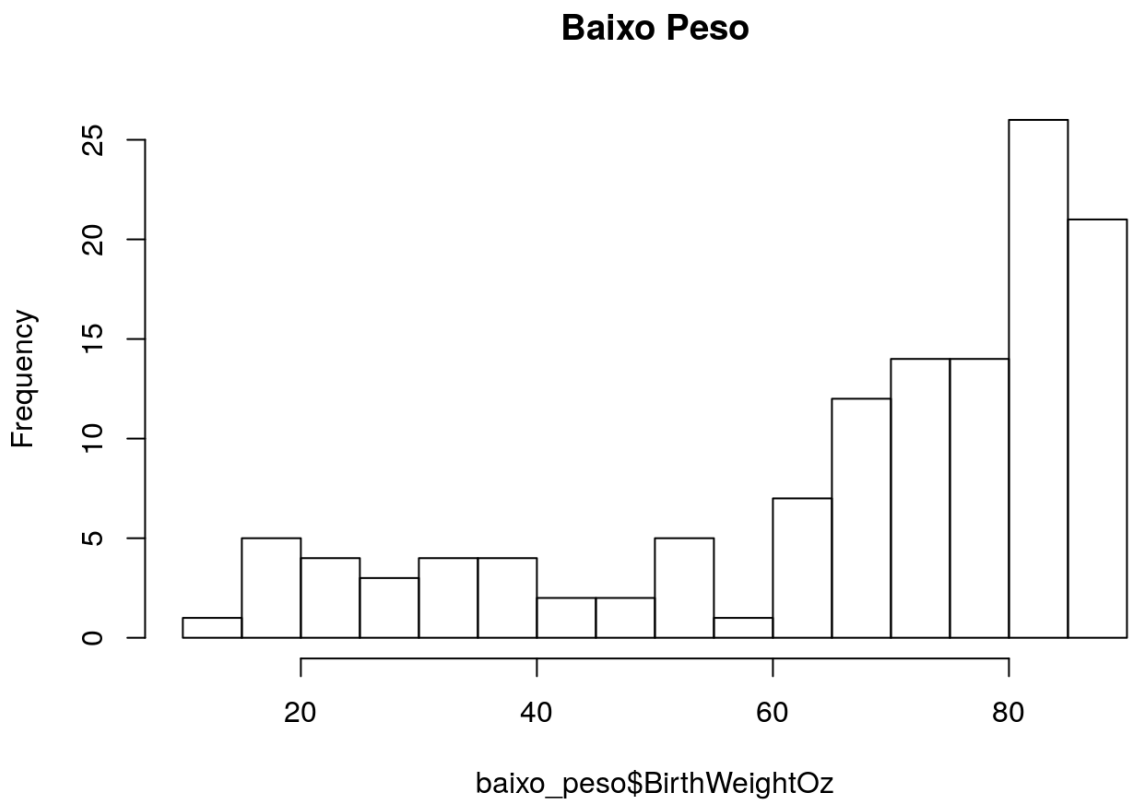

```
par(mfrow = c(1, 1))
hist(nc_births$BirthWeightOz, breaks=25,
#     xlab="Semanas de gestação",
     main="Peso do bebê (geral)")
```




```
hist(peso_normal$BirthWeightOz, breaks=25,
     main="Peso normal")
```

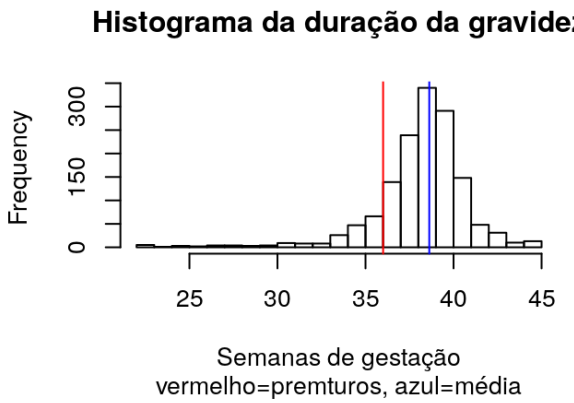


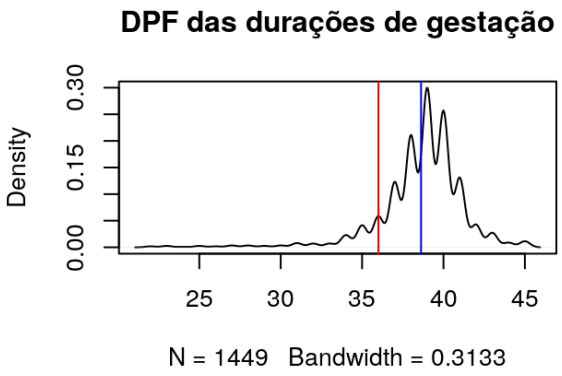

```
hist(baixo_peso$BirthWeightOz, breaks=25,
     main="Baixo Peso")
```

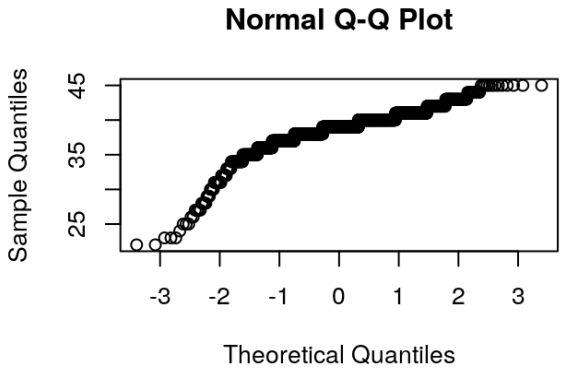


Duração da gravidez

Distribuição de frequências







A distribuiçã das frequências de duração da gravidez é enviesada em direção à duração esperada da gestação a termo: 40 semanas após a última menstruação, 38 após a fertilização, e não pode ser considerada uma distribuição Normal (Teste Shapiro-Wilk: W=0.8506, $p = 6.3253773 \times 10^{-35}$).

A média é 38.621 e o desvio padrão é de 2.699.

Verificação das observações

Peso do bebê: a conversão está correta?

O peso do bebÊ é dado tanto em onças (Oz) quanto em gramas (g). Se não houver erros de conversão de unidade de massa, basta eliminar uma delas, pois têm rigorosamente a mesma informação.

Uma eventual divergência, no entanto, pode indicar erros graves na coleta e/ou registro dos dados, e uma escolha mais complexa precisa ser feita.

```
# bloco de código - item b

summary(nc_births$BirthWeightGm/nc_births$BirthWeightOz)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.35   28.35   28.35   28.35   28.35   28.35
```

```
if (mean(nc_births$BirthWeightGm/nc_births$BirthWeightOz)/max(nc_birth
s$BirthWeightGm/nc_births$BirthWeightOz) != 1 ) {
  error("Há um erro de conversão entre Oz e g")
} else {
  if (max(nc_births$BirthWeightGm/nc_births$BirthWeightOz) == 28.35)
{
    print("Todas as conversões de unidade estão corretas")
  } else {
    print("Embora consistentes, as conversões não estão corretas")
  }
}
```

```
## [1] "Todas as conversões de unidade estão corretas"
```

Conclui-se que as colunas `BirthWeightGm` e `BirthWeightOz` contêm informação idêntica, e uma delas pode ser ignorada com segurança.

Erros aparentes

Registro incorreto de etnia

Parece haver uma confusão no registro de raças. Há três campos para isso, e encontrei pelo menos uma inconsistência no aparente registro de mães hispânicas como japonesas:

```
# for (race in unique(nc_births$MomRace)) { print(paste(race, nrow(nc_
births[nc_births$MomRace == race,])))}  # TODO Um Summary não seria me
lhor?
summary(nc_births$MomRace)
```

```
##    black hispanic    other    white
##      332      164       48      906
```

```
summary(nc_births[nc_births$MomRace=="hispanic", c("RaceMom", "HispMom
")])  # Mostra que `RaceMom` é 5 para todas com `MomRace == hispanic`
```

```
##     RaceMom  HispMom
##  Min.   :5   C:  2
##  1st Qu.:5   M:128
##  Median :5   N:  0
##  Mean   :5   O:  3
##  3rd Qu.:5   P:  8
##  Max.   :5   S: 23
```

```
JapasSoQueNao = nrow(nc_births[nc_births$RaceMom==5,])  # "Japanese",
só que não
print(paste("Registros marcados como mães japonesas: ", JapasSoQueNao)
)
```

```
## [1] "Registros marcados como mães japonesas:  164"
```

```
print("Mães com RaceMom == 5")
```

```
## [1] "Mães com RaceMom == 5"
```

```
summary(nc_births[nc_births$RaceMom==5, c("MomRace", "HispMom")])
```

```
##      MomRace    HispMom
##  black   :  0   C:  2
##  hispanic:164   M:128
##  other   :  0   N:  0
##  white   :  0   O:  3
##                 P:  8
##                 S: 23
```

Aparentemente, os 164 registros de mães “japonesas” na verdade são hispânicas com a marca incorreta no campo `RaceMom` mas correto em `MomRace` . O que me leva a crer que o correto seria considerar todas como hispânicas é a variedade do campo `HispMom` , que tem apenas 3 marcadas como “Other hispanic” (e que ainda não sei se há alguma chance de serem japonesas), e a ausência de uma marca “Not hispanic”.

O campo `Low` está correto?

O campo `Low` , segundo a descrição, chama a atenção para bebês cujo peso ao nascer é inferior a 2500g.

```
# summary(nc_births[nc_births$Low == "Y", c("BirthWeightOz", "BirthWei
ghtGm")])
# summary(nc_births[nc_births$Low == "N", c("BirthWeightOz", "BirthWei
ghtGm")])

# summary(nc_births[nc_births$BirthWeightGm > 2500, c("Low", "BirthWei
ghtOz", "BirthWeightGm")])
nrow(nc_births[nc_births$BirthWeightGm > 2500 & nc_births$Low=='Y', c(
"Low", "BirthWeightOz", "BirthWeightGm")])  # Deve ser zero
```

```
## [1] 0
```

```
nrow(nc_births[nc_births$BirthWeightGm <= 2500 & nc_births$Low=='N', c
("Low", "BirthWeightOz", "BirthWeightGm")])  # Deve ser zero
```

```
## [1] 0
```

Foi confirmado que, realmente, os registros marcados como `Low==Y` todos têm peso abaixo de 2500g, e os marcados com `Low==N` todos têm peso acima do informado na documentação. De forma simétrica, nenhum registro inconsistente (isto é, com marca incorreta) foi encontrado, o que caracteriza uma dependência completa.

Por isso, esse campo pode seguramente ser ignorado nas regressões.

Valores ausentes

Por inspeção visual, foi possível detectar valores ausentes nas colunas `Weeks` (1), `Smoke` (5) e `Gained` (40). Faremos avaliação de cada caso assim que der.

```
col.sums = colSums(is.na(nc_births))
col.sums[col.sums>0]  # Verificação programática de quais colunas têm
NAs
```

```
##  Weeks Gained  Smoke
##      1     40      5
```

```
nc_births<- cbind(nc_births, somaNulos = rowSums(is.na(nc_births)))
NCB_NAs = nc_births[nc_births$somaNulos > 0,]
NCBnoNAs = nc_births[complete.cases(nc_births),]
```

Parece preocupante a grande quantidade de ausências em `Gained` , que corresponde a 2.7586207% das linhas. A estratégia ideal depende da alavancagem desses registros.

Fatores

Segundo a documentação (NCbirths.html) do DataSet, várias dessas colunas podem ser convertidas em fatores:

```
# Fatorizar o que der
source("factorize.R")
useEtn = TRUE  # Flag para unificar campos étnicos

nc_births = NCB_noId(nc_births)
nc_births = NCB_factorize(nc_births, useEtn)

ncb_births_noGainedNA = nc_births[!is.na(nc_births$Gained), ]
ncb_births_noGainedNA = ncb_births_noGainedNA[ncb_births_noGainedNA$so
maNulos == 0, ]

ncb_births_noNAsAtAll = nc_births[nc_births$somaNulos == 0, ]

if (useEtn) {
  COLS_W = c("Plural", "Sex",    "MomAge", "Weeks",  "Marital",
           "Gained", "Smoke", "BirthWeightOz", "Premie", "Etnicidade")
} else {
  COLS_W = c("Plural", "Sex",    "MomAge", "Weeks",  "Marital",
           "RaceMom",  "HispMom", "Gained", "Smoke",
           "BirthWeightOz", "Premie", "MomRace")
}
ncb_weight = nc_births[, COLS_W]
# ncb_weight_noNAs = ncb_births_noGainedNA[, COLS_W]  # ncb_births_noN
AsAtAll
ncb_weight_noNAs = ncb_births_noNAsAtAll[, COLS_W]  #

ncb_weigh_patchedNA = ncb_weight
# ncb_weigh_patchedNA[ncb_weigh_patchedNA$somaNulos >0, "Gained"] = me
an(ncb_weigh_patchedNA$Gained, na.rm = T)

ncb_weigh_patchedNA[is.na(ncb_weigh_patchedNA$Gained), "Gained"] = mea
n(ncb_weigh_patchedNA$Gained, na.rm = T)
ncb_weigh_patchedNA[is.na(ncb_weigh_patchedNA$Smoke), "Smoke"] = "N"

# $Gained[is.na(ncb_weight$Gained)] = mean(ncb_weight$Gained)


# Remover coluna desnecessária
ncb_births_noGainedNA$somaNulos <- NULL
nc_births$somaNulos <- NULL
```

```
# boxplot(nc_births$Gained ~ nc_births$MomRace + nc_births$RaceMom, ma
in='Peso ganho por "Raça"') # Pouco útil sem fatores
```

Correlação entre colunas

```
alias(lm(BirthWeightOz ~ . - BirthWeightGm, data = new_ncbirths(NCbirt
hs)))
```

```
## Model :
## BirthWeightOz ~ (Plural + Sex + MomAge + Weeks + Marital + RaceMom
+
##     HispMom + Gained + Smoke + BirthWeightGm + Low + Premie +
##     MomRace) - BirthWeightGm
##
## Complete :
##                 (Intercept) PluralTwins PluralTriplets SexF MomAge
Weeks
## MomRacehispanic  0           0           0              0    0
0
## MomRaceother     0           0           0              0    0
0
## MomRacewhite     1           0           0              0    0
0
##                 MaritalNot Married RaceMomBlack RaceMomAmericanIndi
an
## MomRacehispanic  0                  0            0
## MomRaceother     0                  0            1
## MomRacewhite     0                 -1           -1
##                 RaceMomChinese RaceMomHispanic RaceMomFilipino
## MomRacehispanic  0              1               0
## MomRaceother     1              0               1
## MomRacewhite    -1             -1              -1
##                 RaceMomOtherAsianOrPacific HispMomM HispMomN HispMo
mO
## MomRacehispanic  0                          0        0        0
## MomRaceother     1                          0        0        0
## MomRacewhite    -1                          0        0        0
##                 HispMomP HispMomS Gained SmokeY LowY PremieY
## MomRacehispanic  0        0        0      0      0    0
## MomRaceother     0        0        0      0      0    0
## MomRacewhite     0        0        0      0      0    0
```

```
alias(lm(BirthWeightOz ~ . - BirthWeightGm, data = new_ncbirths(NCbirt
hs, T)))
```

```
## Model :
## BirthWeightOz ~ (Plural + Sex + MomAge + Weeks + Marital + Gained +
##     Smoke + BirthWeightGm + Low + Premie + Etnicidade) - BirthWeigh
tGm
```

A sobreposição de informações dadas nas colunas étnicas aparece no relatório criado pela função `alias()`. Isso pôde ser corrigido ao se unificar as informações em uma única coluna.

Curiosamente, há combinações inesperadas ali, como mães negras oriundas de Porto Rico e da América do Sul sendo consideradas também hispânicas. Por outro lado, distinguir índios de continentes distintos parece compreensível.

Análises marginais e multivariadas

b. Faça análises marginais e multivariadas.

Peso da mãe

Tentativa de explicar o peso ganho em função das outras variáveis.

Modelo 1: Naïve

```
# summary(lm(Gained ~ ., data=nc_births))  # Ficou igual porque lm() e
liminou as linhas sem valores para Gained
fit_gained1 = lm(Gained ~ ., data=ncb_births_noGainedNA)
summary(fit_gained1)
```

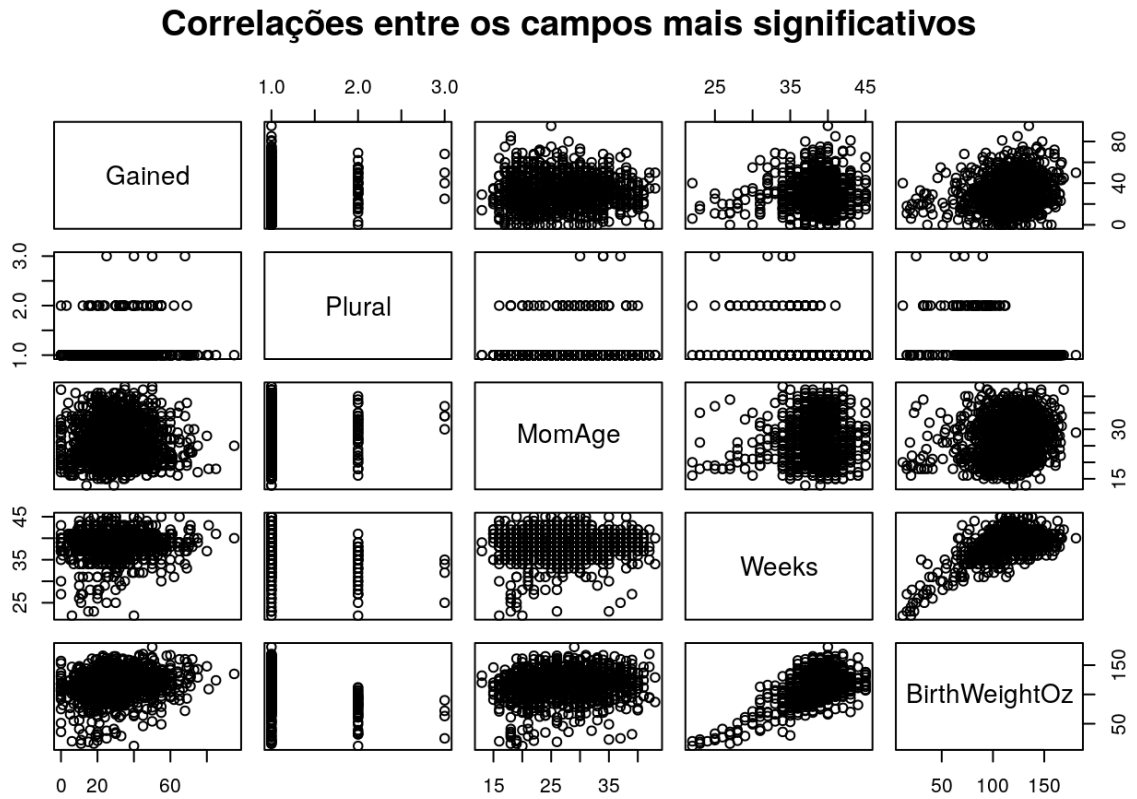
```
##
## Call:
## lm(formula = Gained ~ ., data = ncb_births_noGainedNA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.624  -8.566  -0.883   7.691  61.092
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t
|)
## (Intercept)                    19.88006    8.69879   2.285 0.0224
41 *
## PluralTwins                    12.59628    2.26949   5.550 3.41e-
08 ***
## PluralTriplets                 24.00970    6.80442   3.529 0.0004
32 ***
## SexF                            0.21819    0.71576   0.305 0.7605
42
## MomAge                         -0.14900    0.06712  -2.220 0.0265
89 *
## Weeks                          -0.10314    0.20964  -0.492 0.6227
95
## MaritalNot Married              0.79885    0.92547   0.863 0.3881
88
## SmokeY                          0.87423    1.06189   0.823 0.4104
91
## BirthWeightOz                   0.17957    0.02407   7.461 1.50e-
13 ***
## BirthWeightGm                        NA         NA      NA
NA
## LowY                           -0.78395    1.81751  -0.431 0.6662
93
## PremieY                         1.75629    1.55438   1.130 0.2587
16
## Etnicidadeblack Black N        -4.60985    3.00060  -1.536 0.1246
91
## Etnicidadeblack portoriq      -18.38166   13.66639  -1.345 0.1788
37
## Etnicidadeblack south american -12.87652   13.62680  -0.945 0.3448
54
## Etnicidadecentro-south american  0.76697    4.07141   0.188 0.8506
07
## Etnicidadechinese              10.92400    9.86176   1.108 0.2681
78
## Etnicidadecuban               -12.99595    9.84362  -1.320 0.1869
```

```
73
## Etnicidadefilipino                5.12258   13.65153   0.375 0.7075
40
## Etnicidadehispanic other         -6.97672    8.23266  -0.847 0.3968
94
## Etnicidademexican                -9.24698    3.16379  -2.923 0.0035
26 **
## EtnicidadeOtherAsianOrPacific    -1.00753    4.08095  -0.247 0.8050
34
## Etnicidadeportoriq               -0.22612    5.83565  -0.039 0.9690
96
## Etnicidadesouth americanIndian   19.21733   13.62796   1.410 0.1587
23
## Etnicidadewhite                  -2.36231    2.96472  -0.797 0.4256
99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.29 on 1385 degrees of freedom
## Multiple R-squared:  0.0976, Adjusted R-squared:  0.08261
## F-statistic: 6.513 on 23 and 1385 DF,  p-value: < 2.2e-16
```

Curiosamente, vários campos que eu marquei para exclusão aparecem como NA no sumário, mas não todos: Low foi considerado, mas ficou terrivelmente insignificante.

Esse modelo explica apenas 8% do resultado. Será que existem modelos melhores?

```
# pairs(Gained ~ ., data=nc_births)
pairs(Gained ~ Plural + MomAge + Weeks + BirthWeightOz,
      data=nc_births,
      main = "Correlações entre os campos mais significativos"
      )
```




```
# fit_gained1_noNAs = lm(Gained ~ ., data=nc_births, na.action = na.om
it)
# summary(fit_gained1_noNAs)

# NCBnoNAs = nc_births[]
```

Não parece haver correlação nenhuma entre esses campos! O único padrão relevante é a já conhecida relação entre a idade gestacional do nascituro e o peso ao nascer: quanto mais próximo do termo (40 semanas), mais pesado; depois do termo, nem tanto.

Modelo 2: Menos é mais

Com base no sumário do modelo naïve, pode-se ver que os campos `Plural` e `BirthWeightOz` foram os únicos significativos.

```
fit_gained2 = lm(Gained ~ Plural + BirthWeightOz, data = ncb_births_no
GainedNA)
summary(fit_gained2)
```

```
##
## Call:
## lm(formula = Gained ~ Plural + BirthWeightOz, data = ncb_births_noG
ainedNA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.559  -8.976  -0.862   7.714  61.922
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.71609    2.06272   5.680 1.64e-08 ***
## PluralTwins    12.46352    2.19662   5.674 1.69e-08 ***
## PluralTriplets 24.14434    6.79263   3.554 0.000391 ***
## BirthWeightOz   0.15823    0.01723   9.184  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.43 on 1405 degrees of freedom
## Multiple R-squared:  0.06504,    Adjusted R-squared:  0.06304
## F-statistic: 32.58 on 3 and 1405 DF,  p-value: < 2.2e-16
```

```
anova(fit_gained1, fit_gained2)
```

```
## Analysis of Variance Table
##
## Model 1: Gained ~ Plural + Sex + MomAge + Weeks + Marital + Smoke +
BirthWeightOz +
##     BirthWeightGm + Low + Premie + Etnicidade
## Model 2: Gained ~ Plural + BirthWeightOz
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1   1385 244626
## 2   1405 253451 -20     -8825 2.4982 0.0002696 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como resultado, todos os coeficientes são altamente significativos, e a análise de variância informa uma queda de mais de 8000 pontos na soma dos quadrados dos resíduos. Entretanto, o poder de explicação caiu para 6,34%.

Concluo que não vale a pena tentar prever o peso ganho durante a gestação com base nssses dados.

Peso do bebê

Modelo 3: Naïve Omitindo valores ausentes Da mesma forma que no modelo 1, vejamos Tomando os valores ausentes como a média A mãe filipina Análise de inflação com o que se parece um modelo com “tudo” dentro. Entretanto, como há campos com forte correlação entre si e com a variável dependente, precisamos nos livrar antes de:

- BirthWeightGm
- Low

Além disso, os 40 casos de ganho de peso ausentes precisam ser avaliados para uma tomada de decisão.

```
# ncb_weight = nc_births[,c("Plural", "Sex",    "MomAge", "Weeks",  "M
arital",
#                          "RaceMom",  "HispMom", "Gained", "Smoke",
#                          "BirthWeightOz", "Premie", "MomRace")]
#
# ncb_weight_noNAs = ncb_births_noGainedNA[c("Plural", "Sex",    "MomA
ge", "Weeks",  "Marital",
#                          "RaceMom",  "HispMom", "Gained", "Smoke",
#                          "BirthWeightOz", "Premie", "MomRace"),]

# summary(lm(BirthWeightOz ~ ., data=ncb_weight, na.action = na.omit))
# fit_weight_1 é o fModelo 3                                    ====
fit_weight_1 = lm(BirthWeightOz ~ ., data=ncb_weight, na.action = na.o
mit)
```

Modelo 4: significativas

Menos é mais (de novo)

Baseado no contexto e nos valores observados das significâncias dos coeficientes do modelo 3, seleciono um conjunto menor de colunas para incluir no modelo.

Contexto

- Há estudos confirmando que o tabagismo introduz riscos para a gravidez. Um deles é o nascimento de bebês com menor peso.
- Getação múltipla é considerada de alto risco, com maior índice de partos prematuros e consequente peso menor ao nascer.
- O índice de massa corporal da gestante, segundo um estudo realizado pelo Cincinnati Children’s Hospital Medical Center do estado de Ohio, nos Estados Unidos, pode estar ligado a 25% dos nascimentos prematuros.

Modelo

```
#COLSET_01 = c("Plural", "Sex",    "MomAge", "Weeks",
#              "Gained", "Smoke",
#              "BirthWeightOz")

# fit_weight_2 é o fModelo 4                                    ====
fit_weight_2 = lm(
  BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + Smoke,
  data = ncb_weight
  )
summary(fit_weight_2)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ Plural + Sex + MomAge + Weeks +
##     Gained + Smoke, data = ncb_weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.672 -10.236  -0.157  10.486  49.522
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -63.14127    7.06144  -8.942  < 2e-16 ***
## PluralTwins    -25.54412    2.71720  -9.401  < 2e-16 ***
## PluralTriplets -33.92719    8.45212  -4.014 6.28e-05 ***
## SexF            -3.42455    0.88540  -3.868 0.000115 ***
## MomAge           0.50662    0.07338   6.904 7.65e-12 ***
## Weeks            4.16276    0.17797  23.390  < 2e-16 ***
## Gained           0.28536    0.03215   8.875  < 2e-16 ***
## SmokeY          -7.24103    1.26076  -5.743 1.14e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.59 on 1401 degrees of freedom
##   (41 observations deleted due to missingness)
## Multiple R-squared:  0.4408, Adjusted R-squared:  0.438
## F-statistic: 157.8 on 7 and 1401 DF,  p-value: < 2.2e-16
```

```
fit_weight_2b = lm(
  BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + Smoke,
  data = ncb_weight_NF)

summary(fit_weight_2b)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ Plural + Sex + MomAge + Weeks +
##     Gained + Smoke, data = ncb_weight_NF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.654 -10.185  -0.172  10.478  49.537
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -63.40360    7.05644  -8.985  < 2e-16 ***
## PluralTwins    -25.56954    2.71478  -9.419  < 2e-16 ***
## PluralTriplets -33.98035    8.44453  -4.024 6.03e-05 ***
## SexF            -3.38139    0.88489  -3.821 0.000139 ***
## MomAge           0.51259    0.07339   6.985 4.39e-12 ***
## Weeks            4.16531    0.17782  23.425  < 2e-16 ***
## Gained           0.28563    0.03213   8.891  < 2e-16 ***
## SmokeY          -7.25583    1.25964  -5.760 1.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.57 on 1400 degrees of freedom
##   (41 observations deleted due to missingness)
## Multiple R-squared:  0.4419, Adjusted R-squared:  0.4391
## F-statistic: 158.3 on 7 and 1400 DF,  p-value: < 2.2e-16
```

```
anova(fit_weight_1, fit_weight_2)
```

```
## Analysis of Variance Table
##
## Model 1: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Marital +
Gained +
##     Smoke + Premie + Etnicidade
## Model 2: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + S
moke
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1   1386 372384
## 2   1401 385454 -15    -13069 3.2429 2.512e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_weight_1b, fit_weight_2b)
```

```
## Analysis of Variance Table
##
## Model 1: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Marital +
Gained +
##     Smoke + Premie + Etnicidade
## Model 2: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + S
moke
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1   1386 372384
## 2   1400 384483 -14    -12098 3.2164 4.908e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Inesperadamente, o a soma dos quadrados dos resíduos aumentou quando as outras variáveis foram retiradas.

Outros modelos Topo

Modelo 5: Etnias

Dá pra colocar algo de volta?

```
# fit_weight_5 é o fModelo 5                                    ====
if (useEtn) {
  fit_weight_5 = lm(
    BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + Smoke + E
tnicidade,
    data=ncb_weight)

  sf5 = summary(fit_weight_5)
  print(sf5)

  anova(fit_weight_1, fit_weight_2, fit_weight_5)

} else {
  fit_weight_5 = lm(
    BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + Smoke + R
aceMom,
    data = ncb_weight)
  sf5 = summary(fit_weight_5)
  print(sf5)

# fit_weight_5{b,c} são variações do fModelo 3
====
  fit_weight_5b = lm(
    BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + Smoke + M
omRace,
    data = ncb_weight)
  summary(fit_weight_5b)

  fit_weight_5c = lm(
    BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + Smoke + H
ispMom,
    data = ncb_weight)
  summary(fit_weight_5c)
  anova(fit_weight_1, fit_weight_2, fit_weight_5, fit_weight_5b, fit_w
eight_5c)
}
```

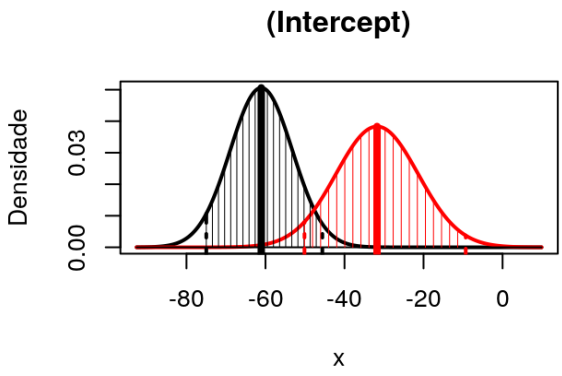
```
##
## Call:
## lm(formula = BirthWeightOz ~ Plural + Sex + MomAge + Weeks +
##     Gained + Smoke + Etnicidade, data = ncb_weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.805 -10.247  -0.376  10.312  52.716
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t
|)
## (Intercept)                   -61.07108    7.88766  -7.743 1.87e-
14 ***
## PluralTwins                   -26.01446    2.70672  -9.611  < 2e-
16 ***
## PluralTriplets                -35.03222    8.40840  -4.166 3.29e-
05 ***
## SexF                           -3.33780    0.88213  -3.784 0.0001
61 ***
## MomAge                          0.44590    0.07632   5.842 6.40e-
09 ***
## Weeks                           4.12488    0.17780  23.199  < 2e-
16 ***
## Gained                          0.27935    0.03246   8.606  < 2e-
16 ***
## SmokeY                         -7.66939    1.28020  -5.991 2.66e-
09 ***
## Etnicidadeblack Black N        -2.18985    3.72465  -0.588 0.5566
72
## Etnicidadeblack portoriq       11.56560   16.92197   0.683 0.4944
26
## Etnicidadeblack south american  5.59946   16.90440   0.331 0.7405
11
## Etnicidadecentro-south american  0.66689    5.05104   0.132 0.8949
79
## Etnicidadechinese                5.60070   12.22582   0.458 0.6469
49
## Etnicidadecuban                  4.74089   12.21830   0.388 0.6980
64
## Etnicidadefilipino             -29.30335   16.91408  -1.732 0.0834
10 .
## Etnicidadehispanic other         5.84796   10.20114   0.573 0.5665
58
## Etnicidademexican                2.61558    3.93672   0.664 0.5065
41
```

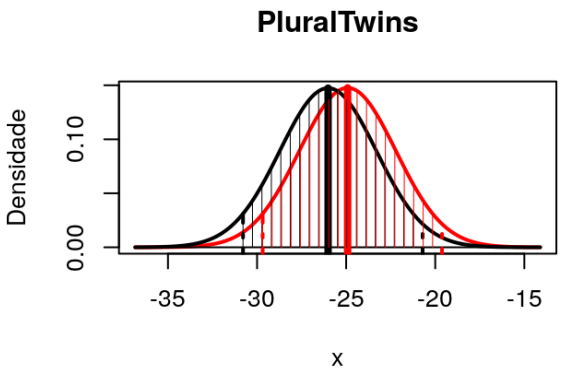
```
## EtnicidadeOtherAsianOrPacific    -1.46203    5.05622  -0.289 0.7725
06
## Etnicidadeportoriq              -12.73123    7.20327  -1.767 0.0773
77 .
## Etnicidadesouth americanIndian    9.05507   16.91154   0.535 0.5924
33
## Etnicidadewhite                   2.50894    3.65806   0.686 0.4929
12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.49 on 1388 degrees of freedom
##   (41 observations deleted due to missingness)
## Multiple R-squared:  0.4524, Adjusted R-squared:  0.4445
## F-statistic: 57.33 on 20 and 1388 DF,  p-value: < 2.2e-16
```

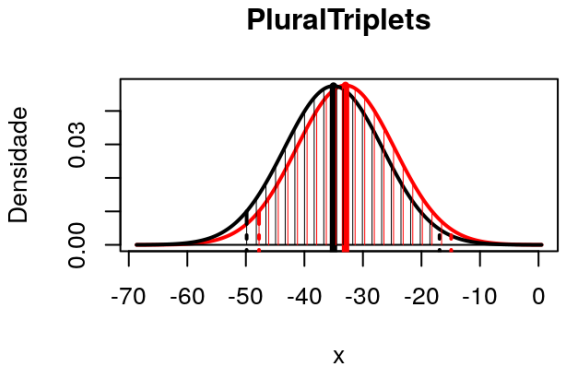
```
## Analysis of Variance Table
##
## Model 1: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Marital +
Gained +
##     Smoke + Premie + Etnicidade
## Model 2: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + S
moke
## Model 3: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + S
moke +
##     Etnicidade
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1   1386 372384
## 2   1401 385454 -15  -13069.3 3.2429 2.512e-05 ***
## 3   1388 377445  13    8008.8 2.2930   0.00536 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

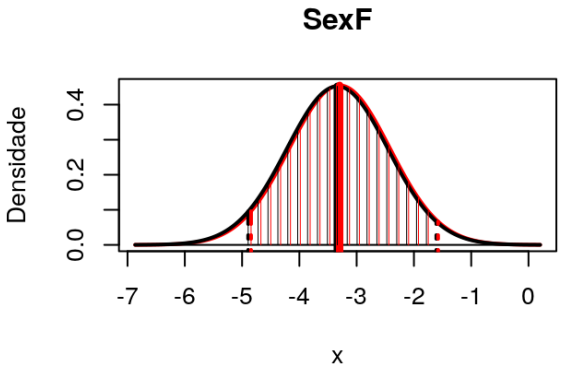
Coeficientes: Modelo 5 x Modelo 3

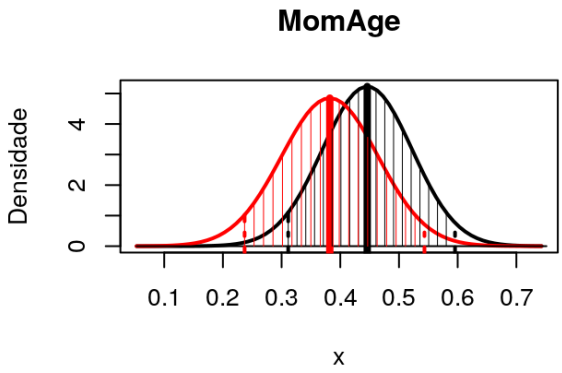
```
comparaCoeficientes(summary(fit_weight_5), sf1)
```

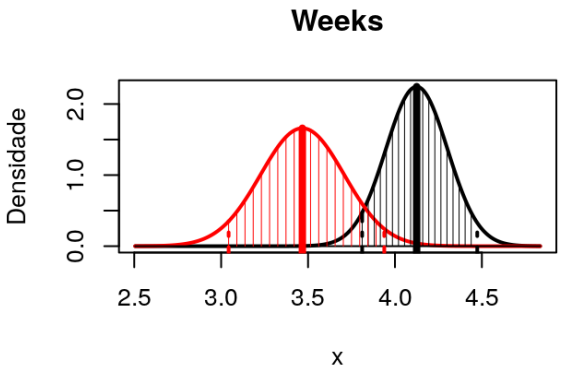

(Intercept)
Densidade
0.00
0.03
-80
-60
-40
-20
0
x

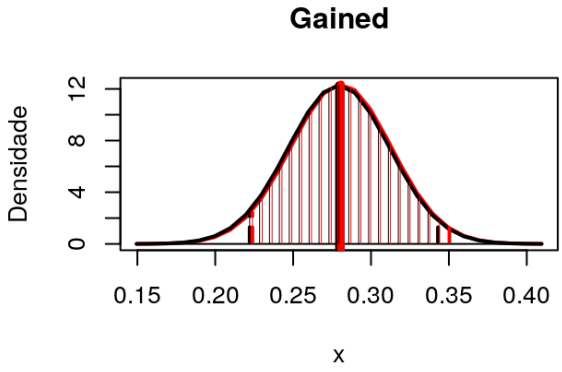

PluralTwins
Densidade
0.00
0.10
-35
-30
-25
-20
-15
x

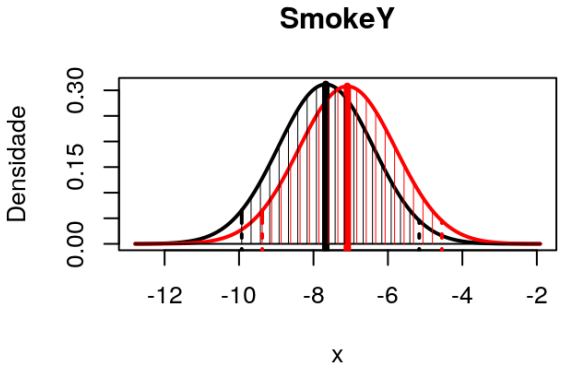

PluralTriplets
Densidade
0.00
0.03
-70
-60
-50
-40
-30
-20
-10
0
x

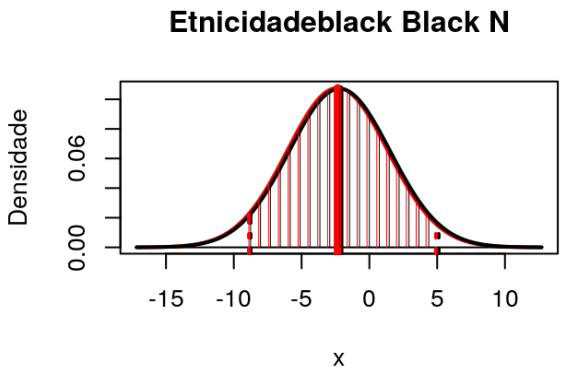

SexF
Densidade
0.0
0.2
0.4
-7
-6
-5
-4
-3
-2
-1
0
x

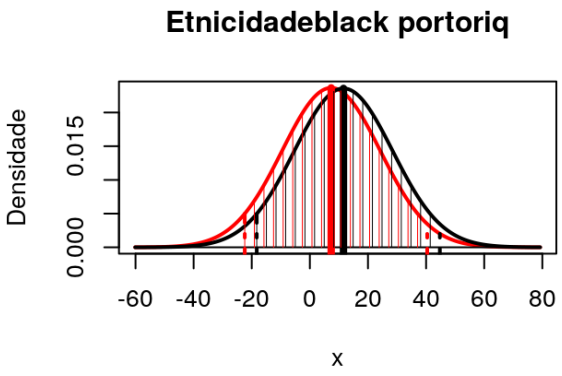

MomAge
Densidade
0
2
4
0.1
0.2
0.3
0.4
0.5
0.6
0.7
x

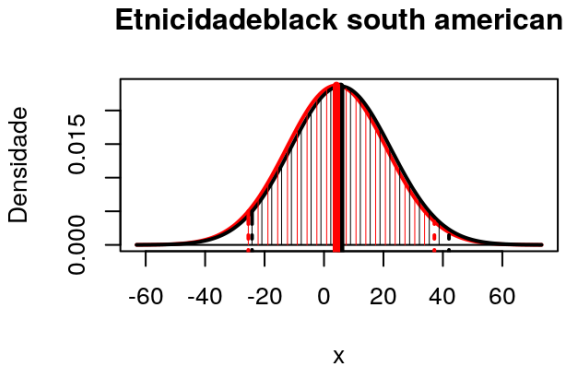

Weeks
Densidade
0.0
1.0
2.0
2.5
3.0
3.5
4.0
4.5
x

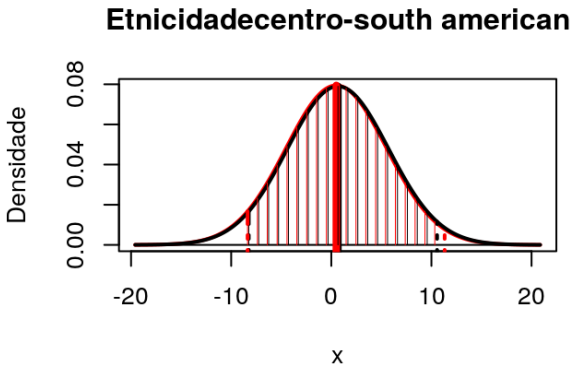

Gained
Densidade
0
4
8
12
0.15
0.20
0.25
0.30
0.35
0.40
x

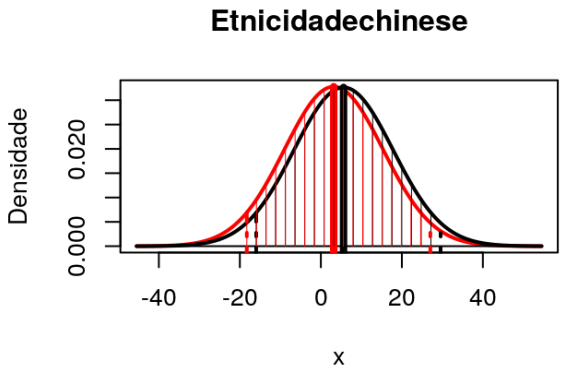

SmokeY
Densidade
0.00
0.15
0.30
-12
-10
-8
-6
-4
-2
x

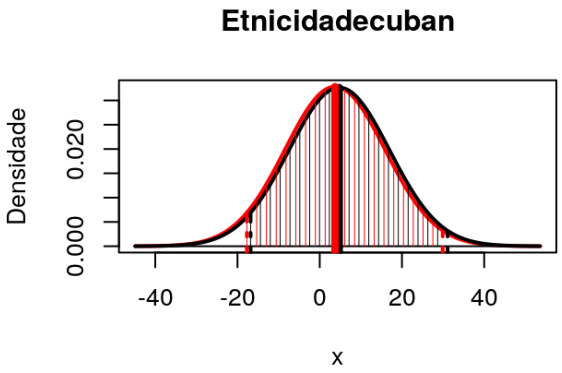

Etnicidadeblack Black N
Densidade
0.00
0.06
-15
-10
-5
0
5
10
x

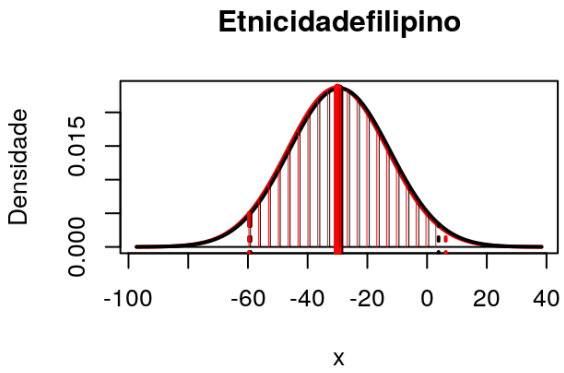

Etnicidadeblack portoriq
Densidade
0.000
0.015
-60
-40
-20
0
20
40
60
80
x

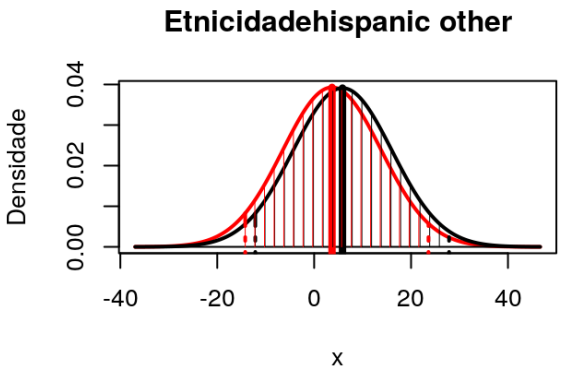

Etnicidadeblack south american
Densidade
0.000
0.015
-60
-40
-20
0
20
40
60
x

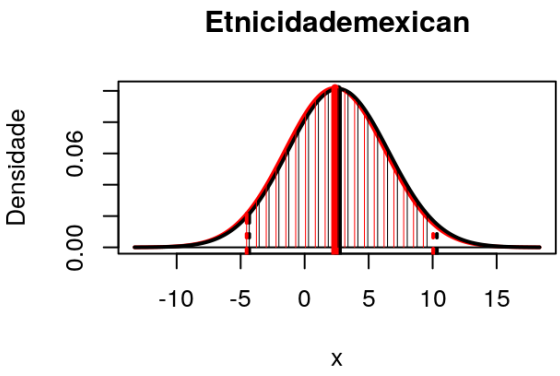

Etnicidadecentro-south american
Densidade
0.00
0.04
0.08
-20
-10
0
10
20
x

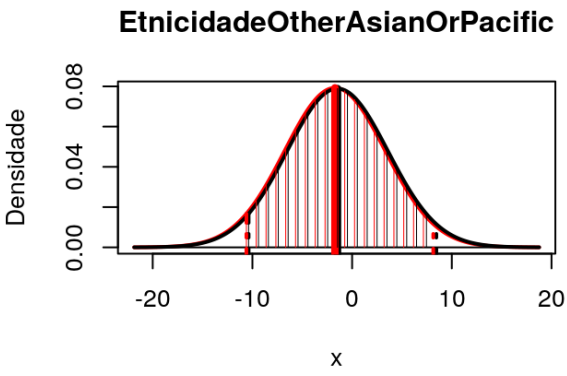

Etnicidadechinese
Densidade
0.000
0.020
-40
-20
0
20
40
x

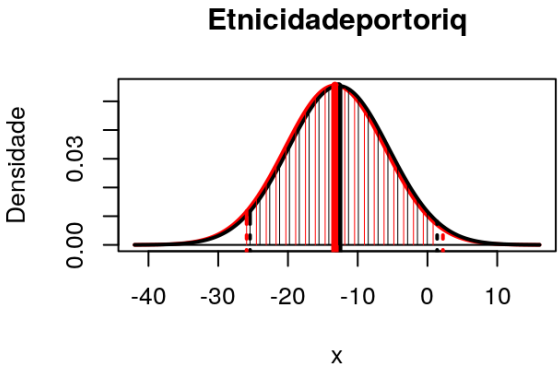

Etnicidadecuban
Densidade
0.000
0.020
-40
-20
0
20
40
x

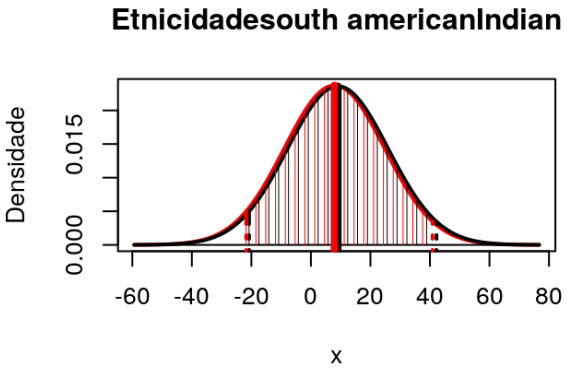

Etnicidadefilipino
Densidade
0.000
0.015
-100
-60
-40
-20
0
20
40
x

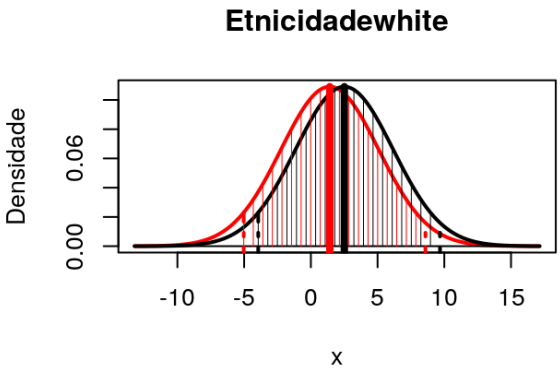

Etnicidadehispanic other
Densidade
0.00
0.02
0.04
-40
-20
0
20
40
x











A inclusão dos campos étnicos prejudicou o modelo, pois os erros cresceram ou ficaram confusos (p>0.05).

Houve diferença significativa entre os coeficientes do intercepto e do campo `Weeks` (mas não sei por que)

Outros modelos Topo

Modelo 6: Tempo em separado

O modelo 5 chamou atenção para o campo `Weeks` . Pelo contexto, espera-se uma forte correlação entre o peso do nascituro e a idade gestacional do parto: quanto mais próximo do termo, maior o peso (o que acontece depois do termo?).

Esta seção tenta isolar a influência desse campo no modelo.

```
# fit_weight_W é um modelo linear simples para comparar com o fModelo
6   ====
fit_weight_W = lm(BirthWeightOz ~ Weeks, data=ncb_weight)
sfW = summary(fit_weight_W)
print(sfW)
```

```
##
## Call:
## lm(formula = BirthWeightOz ~ Weeks, data = ncb_weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.480 -11.994  -0.286  11.908  58.006
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -73.1171     6.7817  -10.78   <2e-16 ***
## Weeks         4.9028     0.1752   27.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.99 on 1447 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3512, Adjusted R-squared:  0.3508
## F-statistic: 783.4 on 1 and 1447 DF,  p-value: < 2.2e-16
```

Um modelo que tem apenas o campo `Weeks` como preditor de `BirthWeighOz` explica pouco mais de `35%` da variância. É bastante, comparado com o modelo 3, com todas as variáveis, que explica 45.9741839%

```
# anova(fit_weight_1, fit_weight_W, fit_weight_2)
print("Cadê o noNA?")
```

```
## [1] "Cadê o noNA?"
```

```
if (useEtn) {  # Campo "etnicidade" unificado no modelo 6     ====
  fit_weight_6 = lm(BirthWeightOz ~ Plural + Sex + MomAge + Gained + S
moke + Etnicidade, data=ncb_weight)
  sf6 = summary(fit_weight_6)
  print(sf6)

  anova(fit_weight_1, fit_weight_6) # compara variância de "1" com "6"

} else {  # Campos étnicos separados (como no original) no modelo 6 ==
==
  fit_weight_6 = lm(BirthWeightOz ~ Plural + Sex + MomAge + Gained + S
moke + RaceMom, data = ncb_weight)
  sf6 = summary(fit_weight_6)
  print(sf6)

  fit_weight_6b = lm(BirthWeightOz ~ Plural + Sex + MomAge + Gained +
Smoke + MomRace, data = ncb_weight)
  sf6b = summary(fit_weight_6b)
  print(sf6b)

  fit_weight_6c = lm(BirthWeightOz ~ Plural + Sex + MomAge + Gained +
Smoke + HispMom, data = ncb_weight)
  sf6c = summary(fit_weight_6c)
  print(sf6c)

  anova(fit_weight_1, fit_weight_2, fit_weight_6, fit_weight_6b, fit_w
eight_6c)
}
```

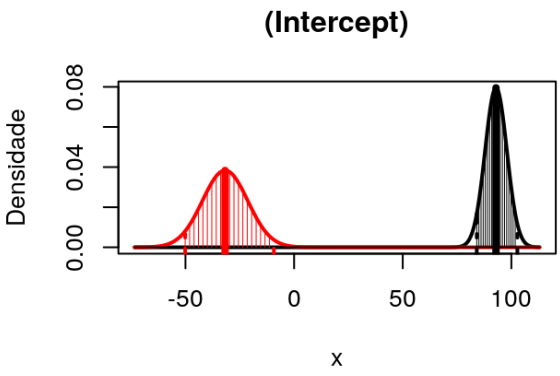
```
##
## Call:
## lm(formula = BirthWeightOz ~ Plural + Sex + MomAge + Gained +
##     Smoke + Etnicidade, data = ncb_weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.525 -10.042   0.959  11.864  52.319
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t
|)
## (Intercept)                     92.92254    5.01749  18.520  < 2e-
16 ***
## PluralTwins                    -45.55676    3.02917 -15.039  < 2e-
16 ***
## PluralTriplets                 -67.39100    9.76461  -6.902 7.80e-
12 ***
## SexF                            -2.27666    1.03741  -2.195   0.02
84 *
## MomAge                           0.54296    0.08974   6.050 1.86e-
09 ***
## Gained                           0.35434    0.03804   9.316  < 2e-
16 ***
## SmokeY                          -8.92184    1.50624  -5.923 3.97e-
09 ***
## Etnicidadeblack Black N         -2.51296    4.38614  -0.573   0.56
68
## Etnicidadeblack portoriq         8.28203   19.92673   0.416   0.67
77
## Etnicidadeblack south american   5.05169   19.90672   0.254   0.79
97
## Etnicidadecentro-south american  -0.57218    5.94780  -0.096   0.92
34
## Etnicidadechinese                1.43880   14.39566   0.100   0.92
04
## Etnicidadecuban                  6.84325   14.38796   0.476   0.63
44
## Etnicidadefilipino             -25.42861   19.91717  -1.277   0.20
19
## Etnicidadehispanic other         4.25154   12.01266   0.354   0.72
35
## Etnicidademexican                5.76263    4.63316   1.244   0.21
38
## EtnicidadeOtherAsianOrPacific    0.72237    5.95321   0.121   0.90
34
```

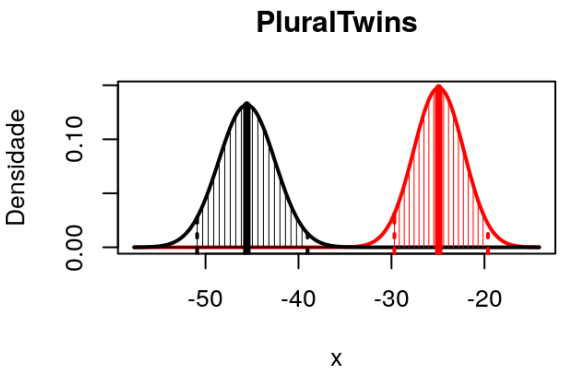
```
## Etnicidadeportoriq               -8.50094    8.47990  -1.002   0.31
63
## Etnicidadesouth americanIndian   12.94574   19.91416   0.650   0.51
58
## Etnicidadewhite                   3.57090    4.30742   0.829   0.40
72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.42 on 1389 degrees of freedom
##   (41 observations deleted due to missingness)
## Multiple R-squared:  0.2401, Adjusted R-squared:  0.2297
## F-statistic: 23.09 on 19 and 1389 DF,  p-value: < 2.2e-16
```

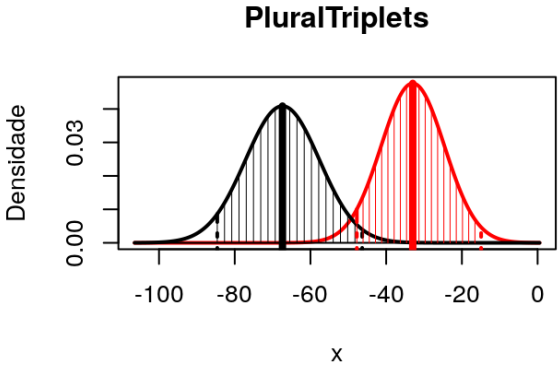
```
## Analysis of Variance Table
##
## Model 1: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Marital +
Gained +
##     Smoke + Premie + Etnicidade
## Model 2: BirthWeightOz ~ Plural + Sex + MomAge + Gained + Smoke + E
tnicidade
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1386 372384
## 2   1389 523802 -3   -151417 187.86 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

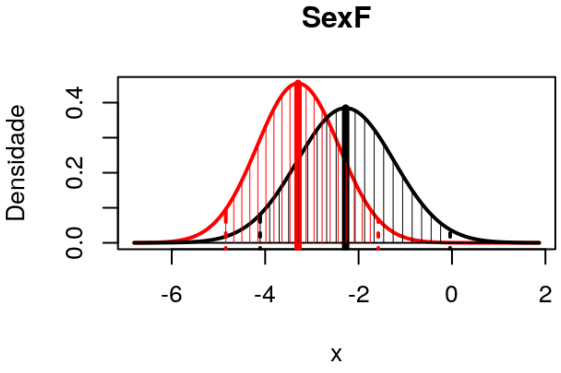
Coeficientes do model 6 x os do modelo 3

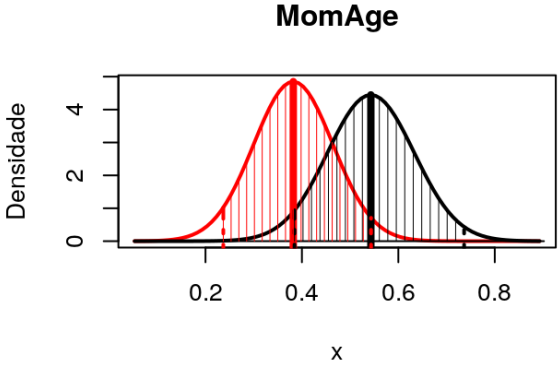
```
comparaCoeficientes(summary(fit_weight_6), sf1)
```

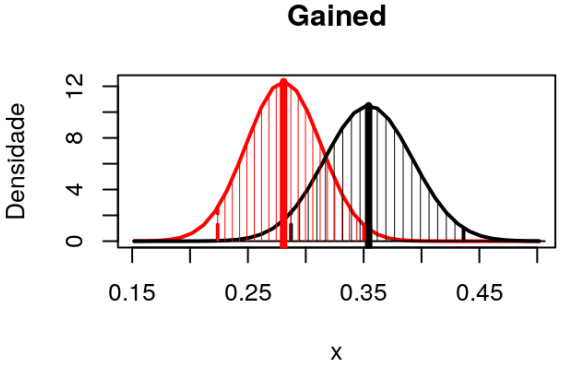

(Intercept)
Densidade
0.00
0.04
0.08
-50
0
50
100
x

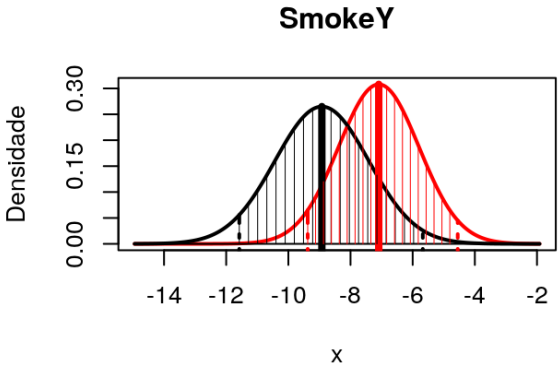

PluralTwins
Densidade
0.00
0.10
-50
-40
-30
-20
x

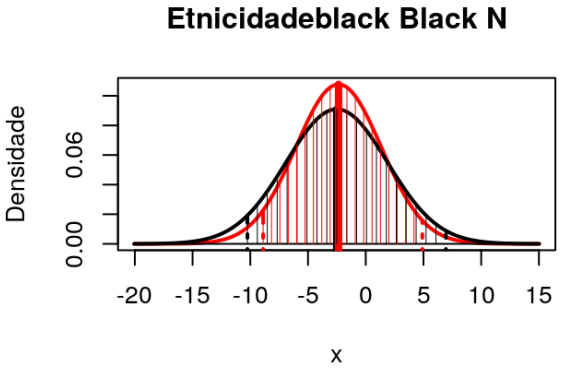

PluralTriplets
Densidade
0.00
0.03
-100
-80
-60
-40
-20
0
x

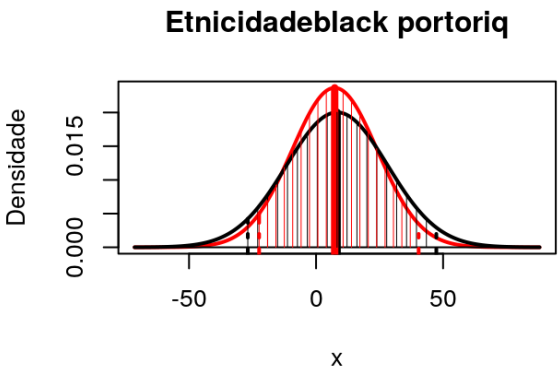

SexF
Densidade
0.0
0.2
0.4
-6
-4
-2
0
2
x

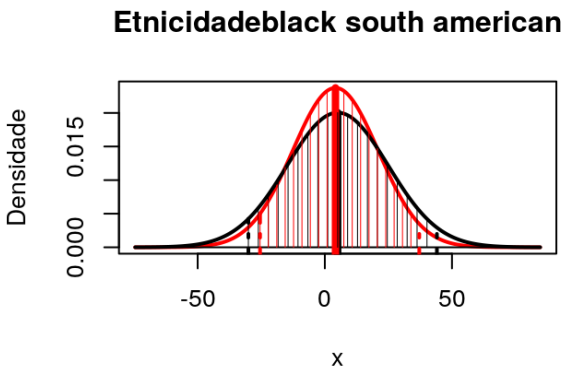

MomAge
Densidade
0
2
4
0.2
0.4
0.6
0.8
x

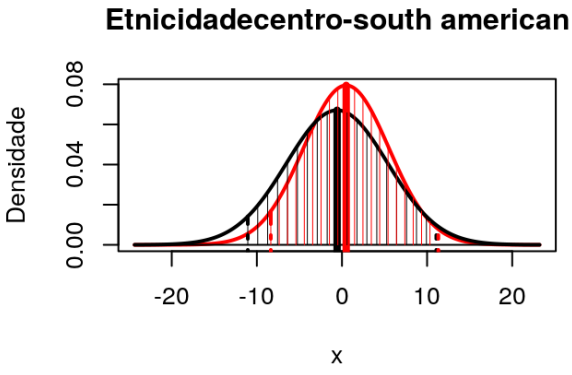

Gained
Densidade
0
4
8
12
0.15
0.25
0.35
0.45
x

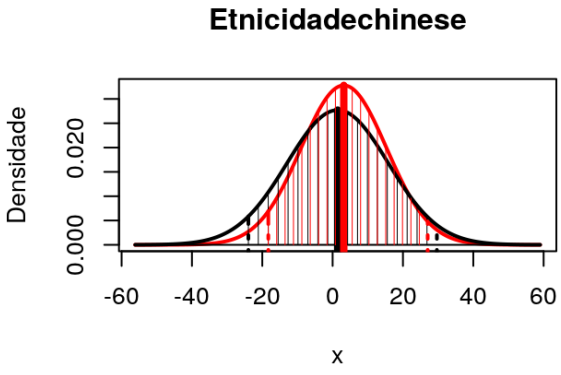

SmokeY
Densidade
0.00
0.15
0.30
-14
-12
-10
-8
-6
-4
-2
x

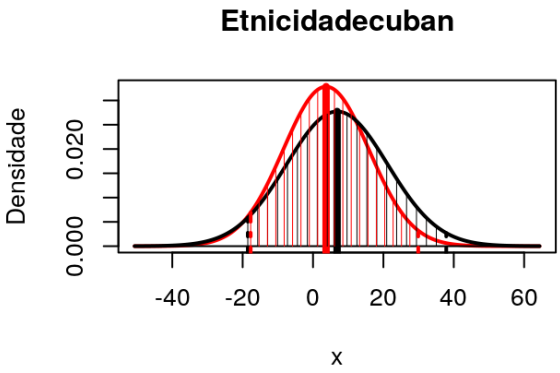

Etnicidadeblack Black N
Densidade
0.00
0.06
-20
-15
-10
-5
0
5
10
15
x

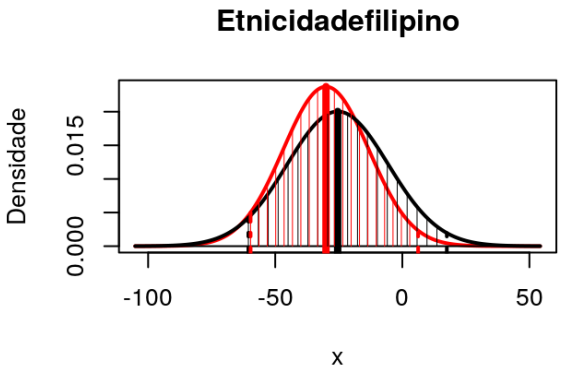

Etnicidadeblack portoriq
Densidade
0.000
0.015
-50
0
50
x

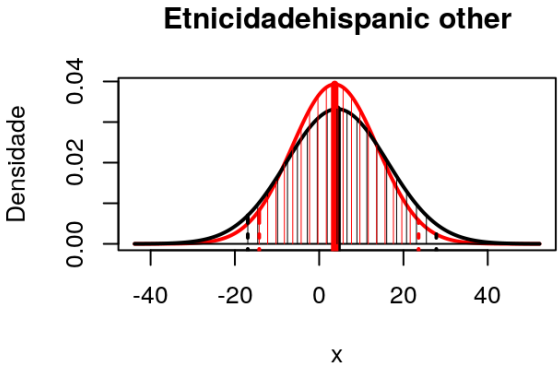

Etnicidadeblack south american
Densidade
0.000
0.015
-50
0
50
x

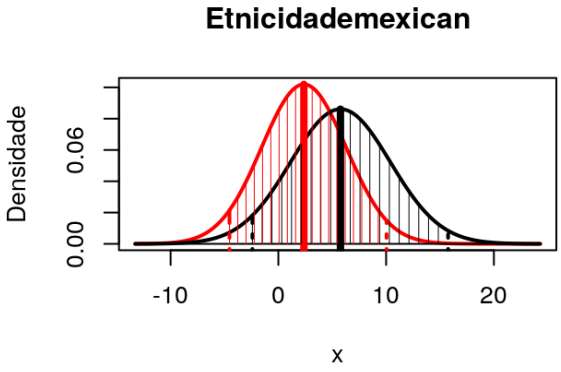

Etnicidadecentro-south american
Densidade
0.00
0.04
0.08
-20
-10
0
10
20
x

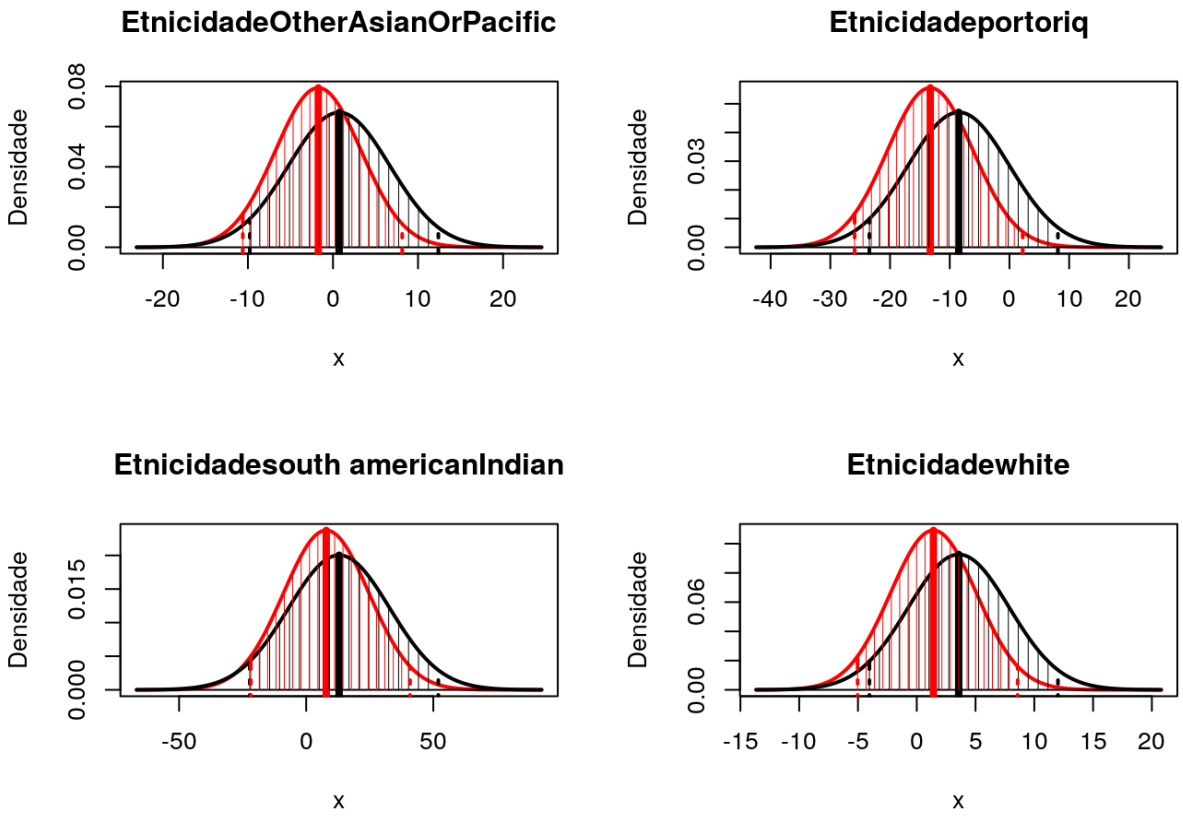

Etnicidadechinese
Densidade
0.000
0.020
-60
-40
-20
0
20
40
60
x


Etnicidadecuban
Densidade
0.000
0.020
-40
-20
0
20
40
60
x


Etnicidadefilipino
Densidade
0.000
0.015
-100
-50
0
50
x


Etnicidadehispanic other
Densidade
0.00
0.02
0.04
-40
-20
0
20
40
x


Etnicidademexican
Densidade
0.00
0.06
-10
0
10
20
x



Outros modelos Topo

Modelo 7: sem gêmeos

Uma informação ausente sobre o peso do nascituro é: de qual bebê é o peso informado em `BirthWeightOz` ? É a média dos dois ou três? É o menor (ou maior) deles? Essa falta de informação pode estar introduzindo alguma distorção no modelo. Façamos um sem gêmeos.

```
# fit_weight_7 é o fModelo 7 (sem gêmeos)
========
if (useEtn) {  # Campo "etnicidade" unificado no modelo 7
  fit_weight_7 = lm(
    BirthWeightOz ~ Sex + MomAge + Gained + Smoke + Etnicidade,
    data=ncb_weight)
  sf7 = summary(fit_weight_7)
  print(sf7)

  anova(fit_weight_1, fit_weight_7, fit_weight_2) # compara variância
de "1" com "2" e "7"   =====

} else {  # Campos étnicos separados (como no original) no modelo 7 ==
=====
  fit_weight_7 = lm(BirthWeightOz ~ Sex + MomAge + Gained + Smoke + Ra
ceMom, data = ncb_weight)
  sf7 = summary(fit_weight_7)
  print(sf7)

  fit_weight_7b = lm(BirthWeightOz ~ Sex + MomAge + Gained + Smoke + M
omRace, data = ncb_weight)
  sf7b = summary(fit_weight_7b)
  print(sf7b)

  fit_weight_7c = lm(BirthWeightOz ~ Sex + MomAge + Gained + Smoke + H
ispMom, data = ncb_weight)
  sf7c = summary(fit_weight_7c)
  print(sf7c)

  anova(fit_weight_1, fit_weight_2, fit_weight_7, fit_weight_7b, fit_w
eight_7c)
}
```

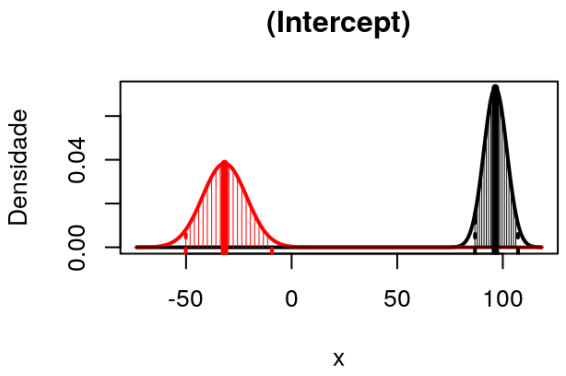
```
##
## Call:
## lm(formula = BirthWeightOz ~ Sex + MomAge + Gained + Smoke +
##     Etnicidade, data = ncb_weight)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.731  -9.860   1.728  12.816  54.469
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t
|)
## (Intercept)                    96.47907    5.47331  17.627  < 2e-
16 ***
## SexF                           -2.13674    1.13286  -1.886   0.05
95 .
## MomAge                          0.45340    0.09778   4.637 3.87e-
06 ***
## Gained                          0.29870    0.04136   7.222 8.43e-
13 ***
## SmokeY                         -7.56524    1.64242  -4.606 4.48e-
06 ***
## Etnicidadeblack Black N        -3.62693    4.78926  -0.757   0.44
90
## Etnicidadeblack portoriq        6.08428   21.76065   0.280   0.77
98
## Etnicidadeblack south american  4.24173   21.73921   0.195   0.84
53
## Etnicidadecentro-south american 0.18372    6.49515   0.028   0.97
74
## Etnicidadechinese               3.09240   15.72049   0.197   0.84
41
## Etnicidadecuban                 6.84812   15.71250   0.436   0.66
30
## Etnicidadefilipino            -23.97521   21.75047  -1.102   0.27
05
## Etnicidadehispanic other        4.87315   13.11846   0.371   0.71
03
## Etnicidademexican               4.93014    5.05936   0.974   0.33
00
## EtnicidadeOtherAsianOrPacific   1.48282    6.50105   0.228   0.81
96
## Etnicidadeportoriq             -8.37477    9.26055  -0.904   0.36
60
## Etnicidadesouth americanIndian 15.19172   21.74690   0.699   0.48
49
```

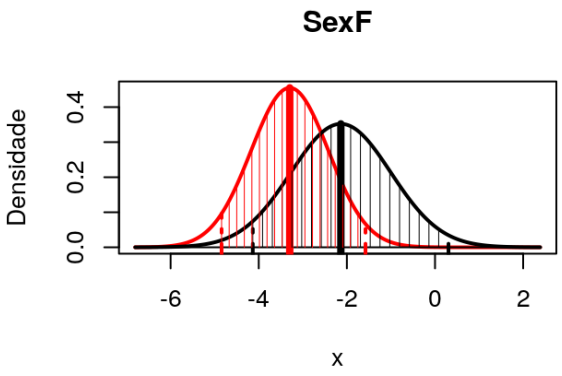
```
## Etnicidadewhite                    1.96807    4.70272   0.418   0.67
56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.21 on 1391 degrees of freedom
##   (41 observations deleted due to missingness)
## Multiple R-squared:  0.0924, Adjusted R-squared:  0.08131
## F-statistic:  8.33 on 17 and 1391 DF,  p-value: < 2.2e-16
```

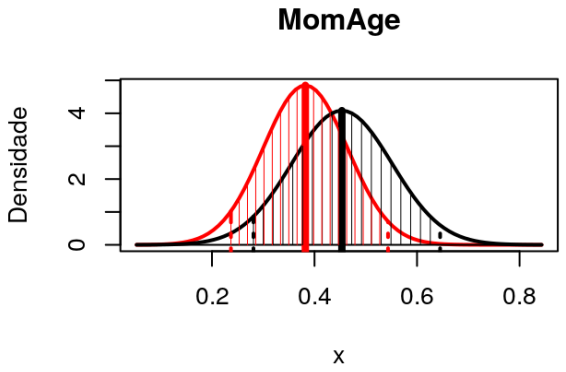
```
## Analysis of Variance Table
##
## Model 1: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Marital +
Gained +
##     Smoke + Premie + Etnicidade
## Model 2: BirthWeightOz ~ Sex + MomAge + Gained + Smoke + Etnicidade
## Model 3: BirthWeightOz ~ Plural + Sex + MomAge + Weeks + Gained + S
moke
##   Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
## 1   1386 372384
## 2   1391 625582  -5   -253197 188.48 < 2.2e-16 ***
## 3   1401 385454 -10    240128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

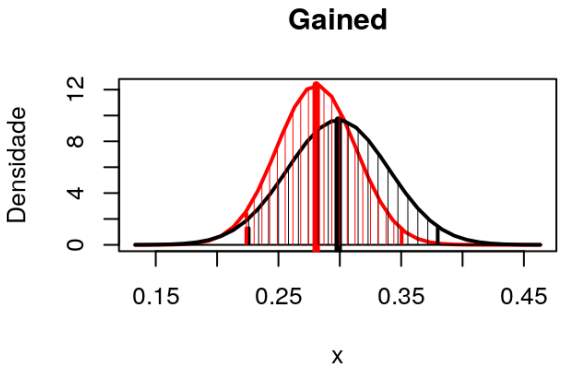
Compara coeficientes do modelo 7 com os do 3

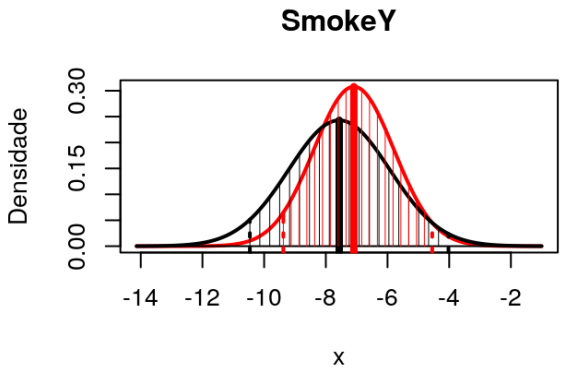
```
comparaCoeficientes(summary(fit_weight_7), sf1)  # , sf2, sf6)
```

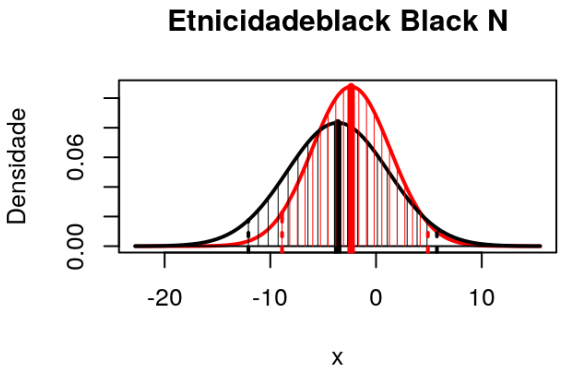

(Intercept)
Densidade
0.00
0.04
-50
0
50
100
x

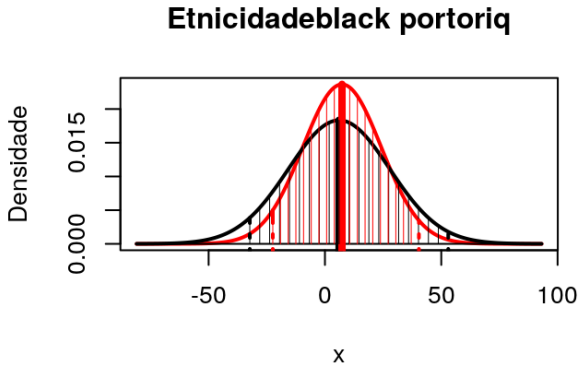

SexF
Densidade
0.0
0.2
0.4
-6
-4
-2
0
2
x

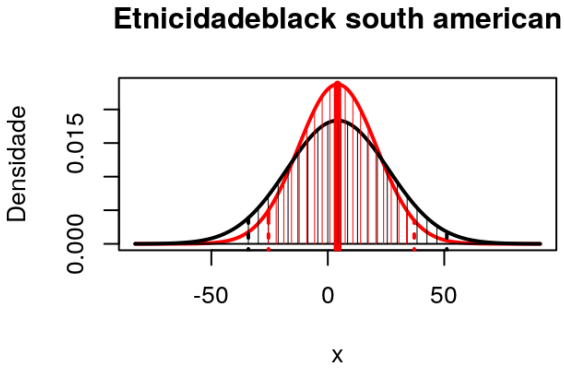

MomAge
Densidade
0
2
4
0.2
0.4
0.6
0.8
x

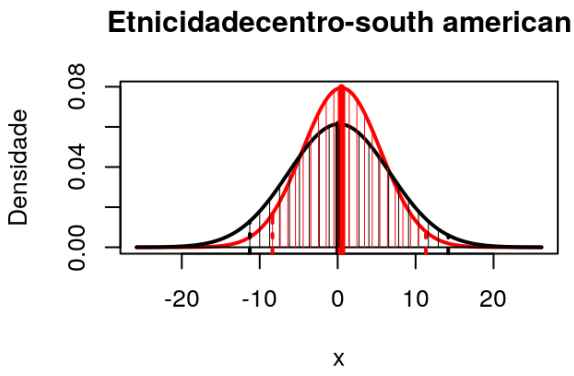

Gained
Densidade
0
4
8
12
0.15
0.25
0.35
0.45
x

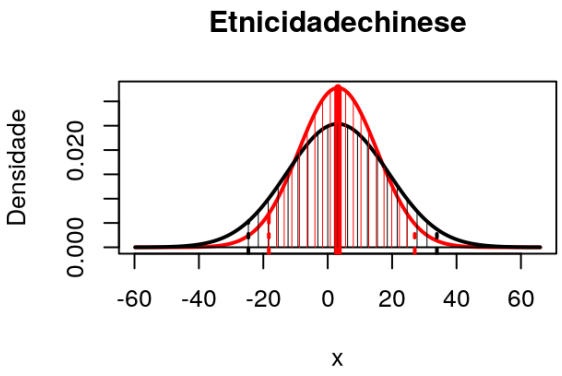

SmokeY
Densidade
0.00
0.15
0.30
-14
-12
-10
-8
-6
-4
-2
x

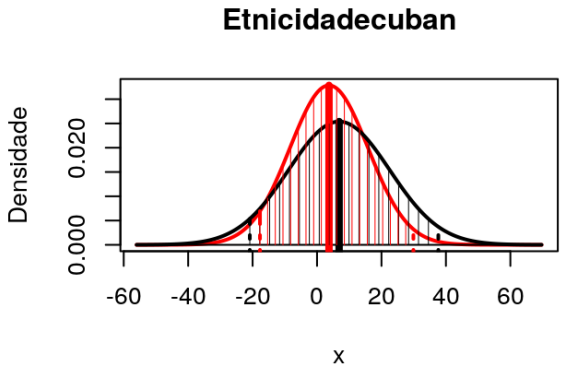

Etnicidadeblack Black N
Densidade
0.00
0.06
-20
-10
0
10
x

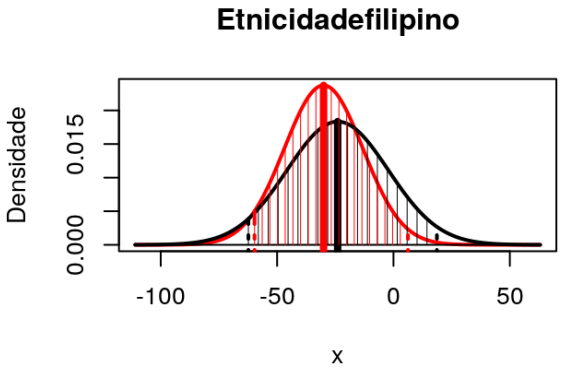

Etnicidadeblack portoriq
Densidade
0.000
0.015
-50
0
50
100
x

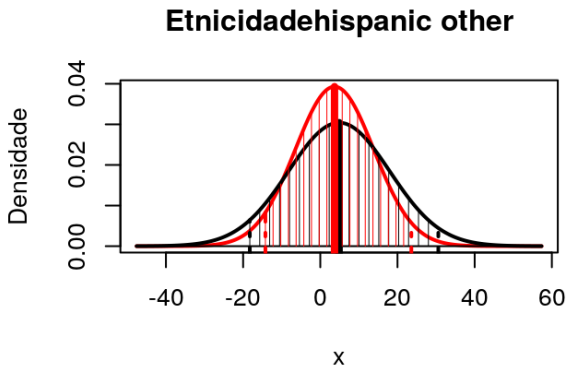

Etnicidadeblack south american
Densidade
0.000
0.015
-50
0
50
x

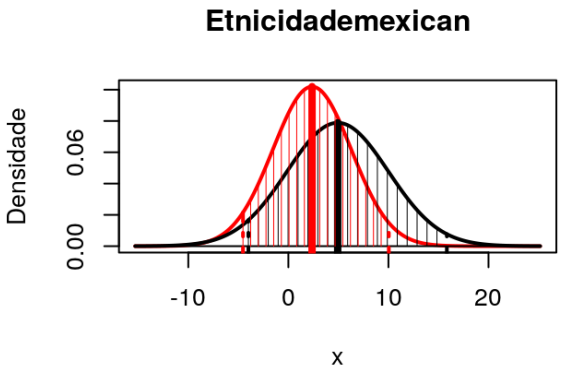

Etnicidadecentro-south american
Densidade
0.00
0.04
0.08
-20
-10
0
10
20
x

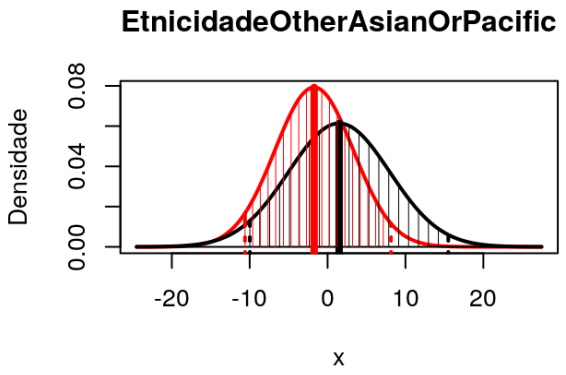

Etnicidadechinese
Densidade
0.000
0.020
-60
-40
-20
0
20
40
60
x

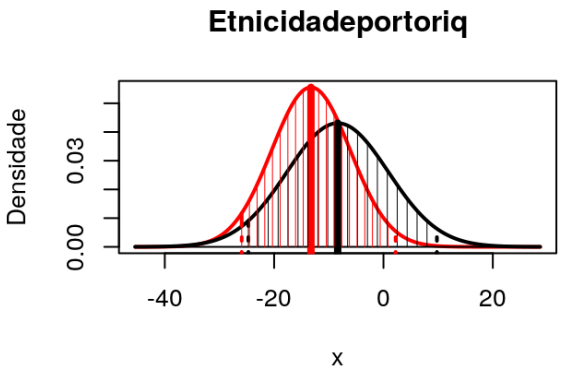

Etnicidadecuban
Densidade
0.000
0.020
-60
-40
-20
0
20
40
60
x

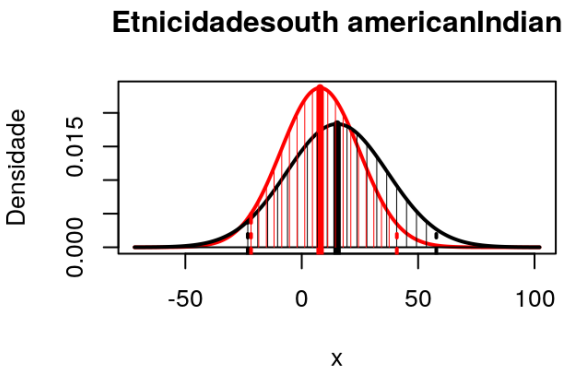

Etnicidadefilipino
Densidade
0.000
0.015
-100
-50
0
50
x

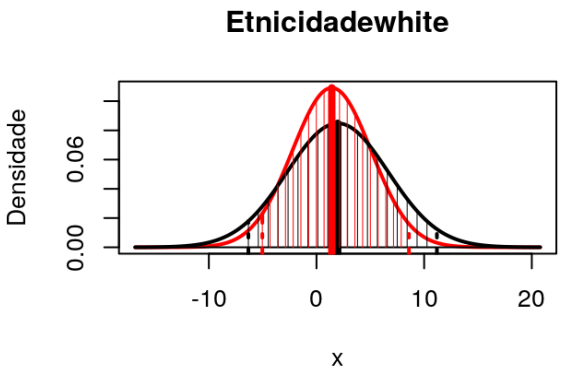

Etnicidadehispanic other
Densidade
0.00
0.02
0.04
-40
-20
0
20
40
60
x


Etnicidademexican
Densidade
0.00
0.06
-10
0
10
20
x


EtnicidadeOtherAsianOrPacific
Densidade
0.00
0.04
0.08
-20
-10
0
10
20
x


Etnicidadeportoriq
Densidade
0.00
0.03
-40
-20
0
20
x





Outros modelos Topo

Tempo de gestação

Não foi feito

Verificação de premissas do modelo linear

c. Verifique as premissas do modelo linear.

- Resíduos
- Independência: os erros são independentes entre si
- Normalidade:
- Homocedasticidade (variância constante)
- Linearidade

Sumário

Normalidade

Os resíduos dos modelos 6 e 7 tem sua estatística W abaixo de 96%. Todos os outros ficara acima de 99%. Seus gráficos Q-Q apresentam deformidades que levantam suspeitas de que não sejam

~N.

Modelos

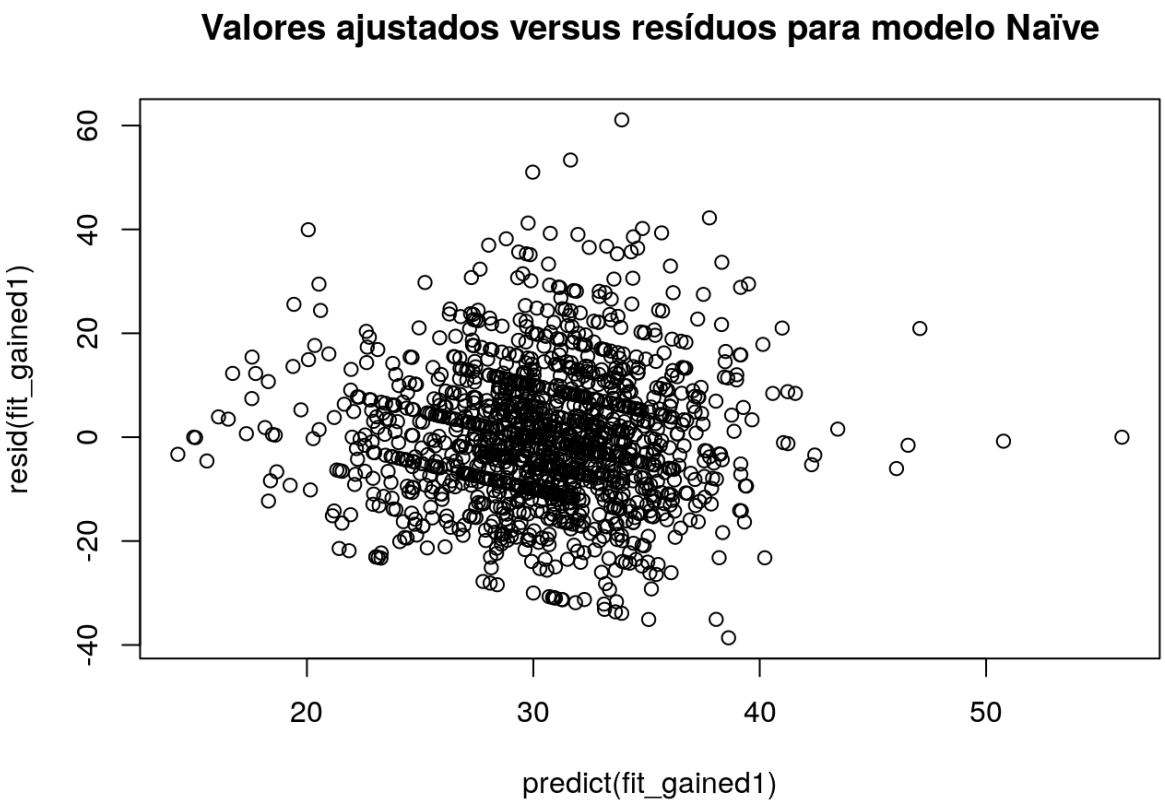
Modelo 1: Naïve

Propriedades dos resíduos

```
source("verifica_residuos.R")

pro_res_1 = verifica_props_residuos(
  fit_gained1$residuals,
  ncb_births_noGainedNA$Gained,
  fit_gained1$fitted.values
  )

if (pro_res_1$is.small) {
    print("Propriedades do modelo 1 OK")
  } else {
  print("Alguma propriedade do modelo 1 não tem o valor esperado:")
  print_props_residuos(pro_res_1)
}
```

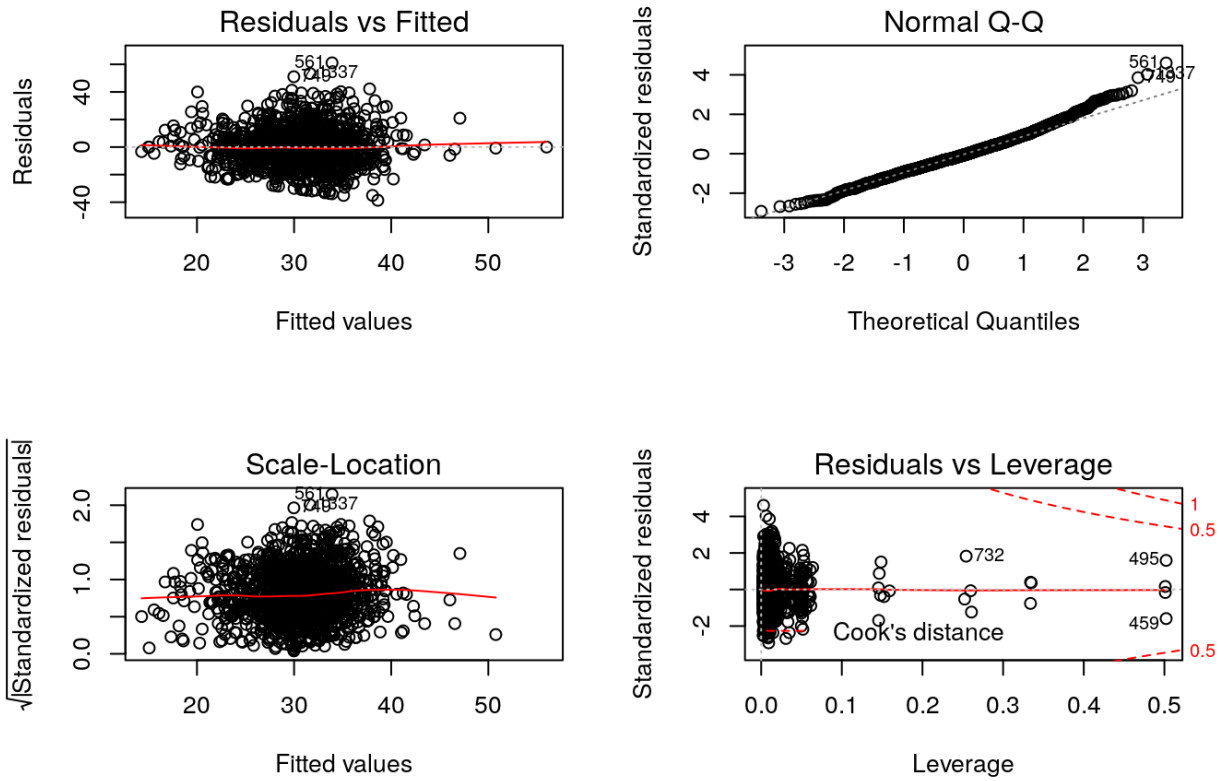
```
## [1] "Propriedades do modelo 1 OK"
```

```
# bloco de código - item c
plot(predict(fit_gained1), resid(fit_gained1),   # data=nc_births,
     main="Valores ajustados versus resíduos para modelo Naïve")
```




```
par(mfrow=c(2,2))
plot(fit_gained1)
```

```
## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383, 1387

## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383, 1387
```




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

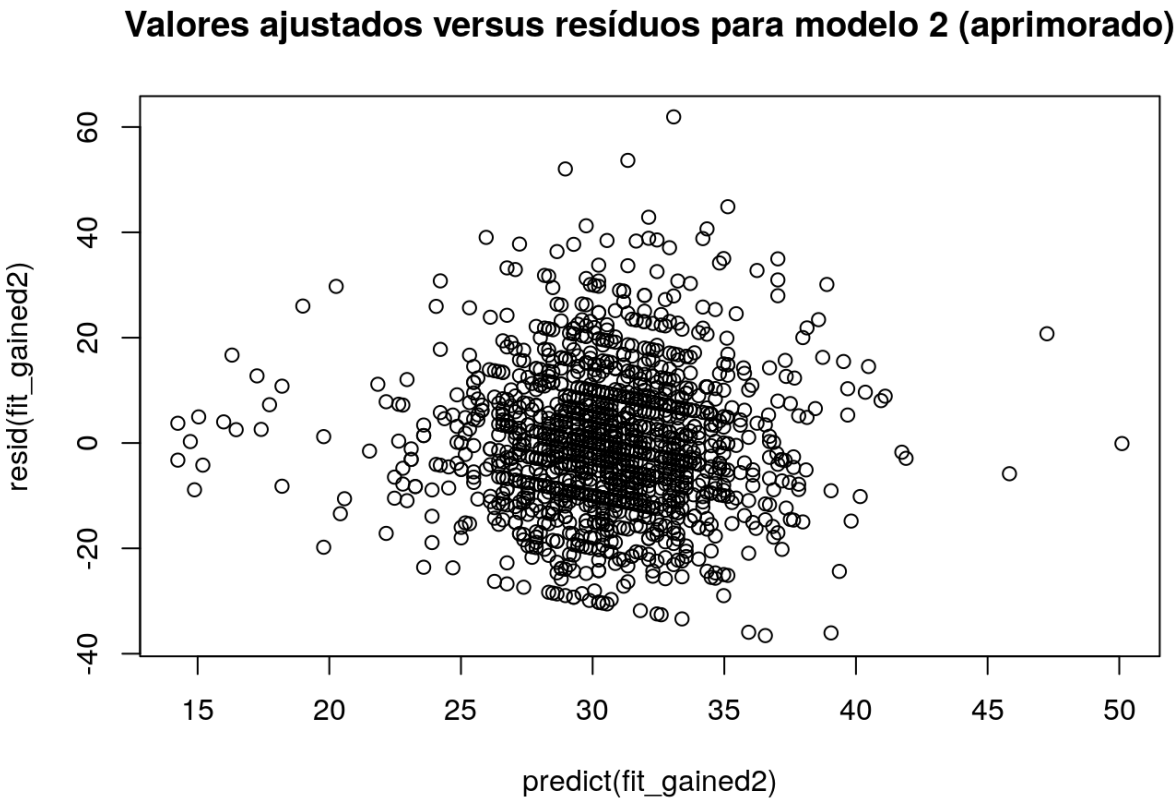
Modelo 2:

Menos é (ou devia ser) mais

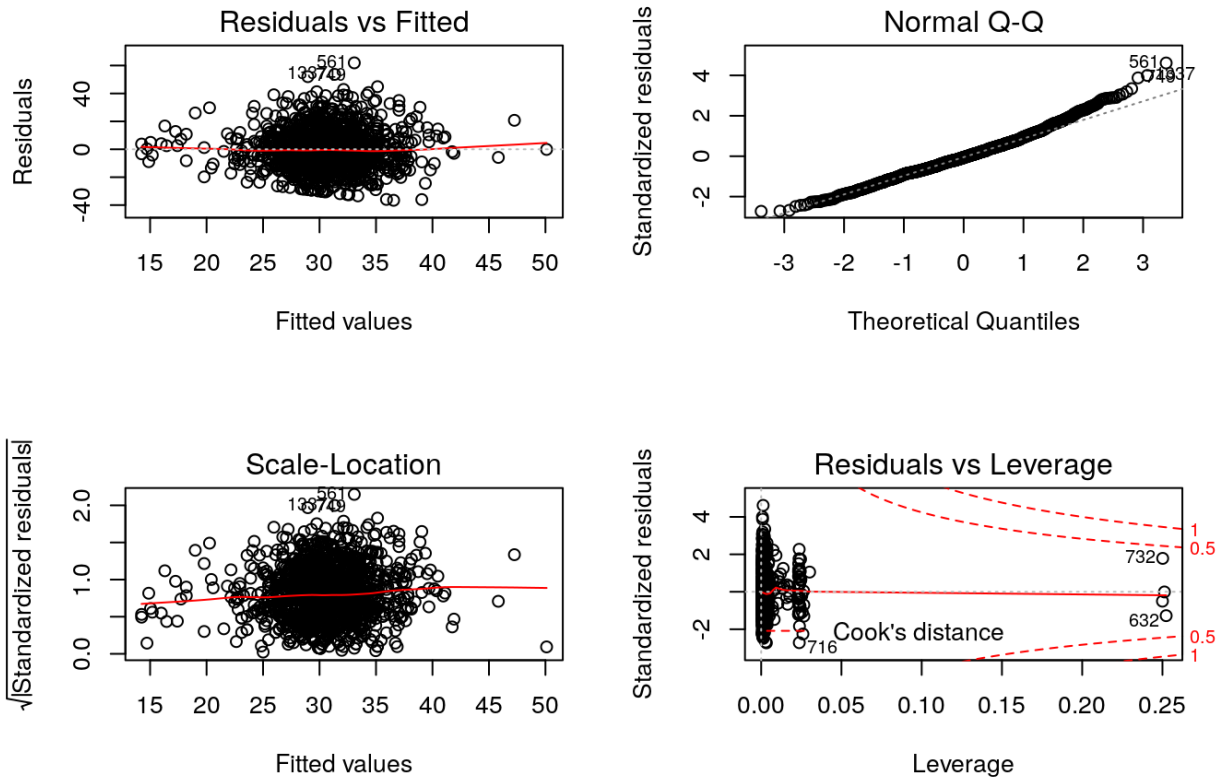
```
pro_res_2 = verifica_props_residuos(
  fit_gained2$residuals,
  ncb_births_noGainedNA$Gained,
  fit_gained2$fitted.values
  )

if (pro_res_2$is.small) {
    print("Propriedades do modelo 2 OK")
  } else {
  print("Alguma propriedade do modelo 2 não tem o valor esperado:")
  print_props_residuos(pro_res_2)
}
```

```
## [1] "Propriedades do modelo 2 OK"
```

```
plot(predict(fit_gained2), resid(fit_gained2),   # data=nc_births,
     main="Valores ajustados versus resíduos para modelo 2 (aprimorado
)")
```




```
par(mfrow=c(2,2))
plot(fit_gained2)
```




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

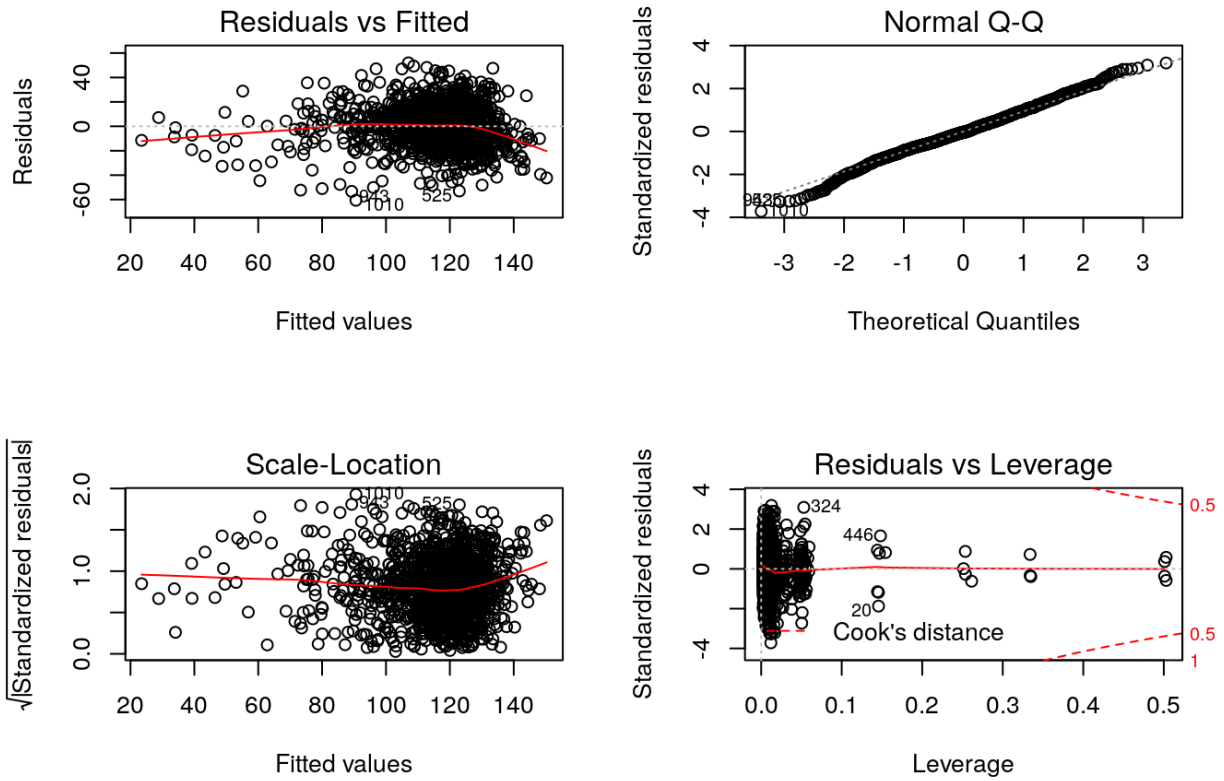
TODO: Interpretação

Modelo 3: Naïve

Com a mãe filipina

```
par(mfrow = c(2, 2))
plot(fit_weight_1)
```

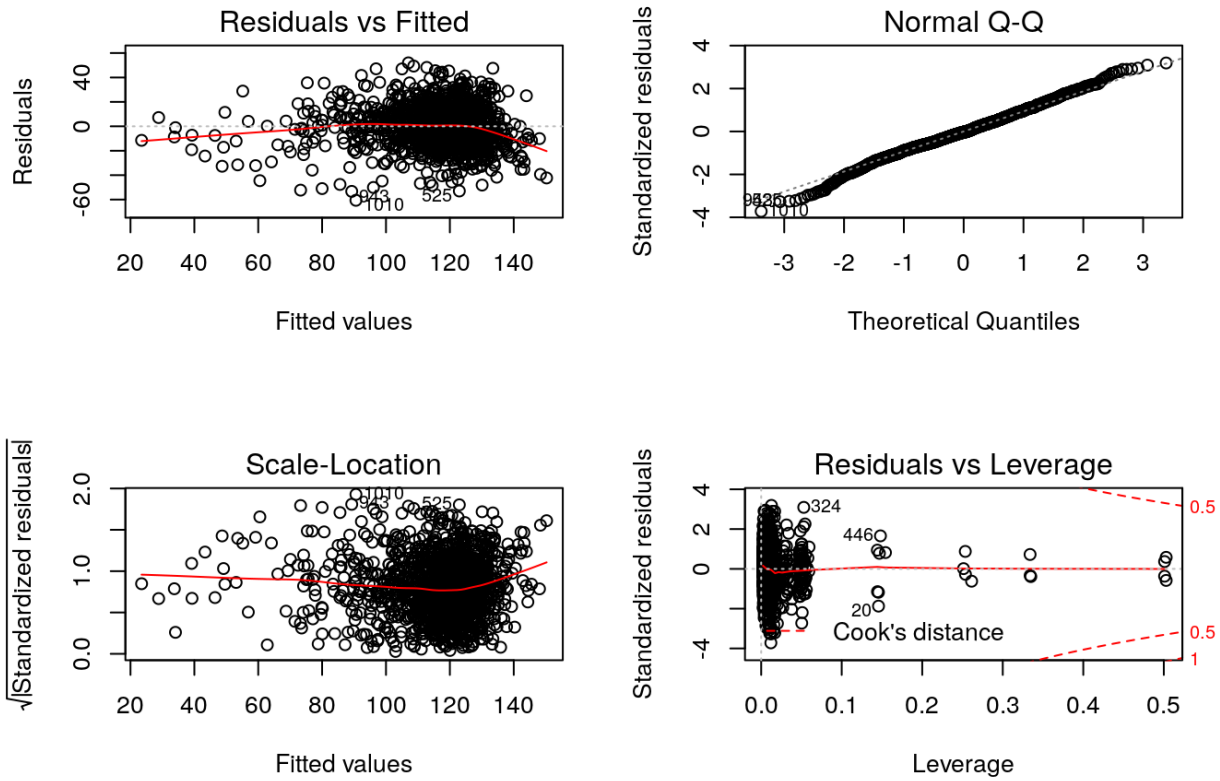
```
## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383, 1387

## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383, 1387
```



Sem a filipina

```
par(mfrow = c(2, 2))
plot(fit_weight_1b)
```

```
## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383

## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383
```



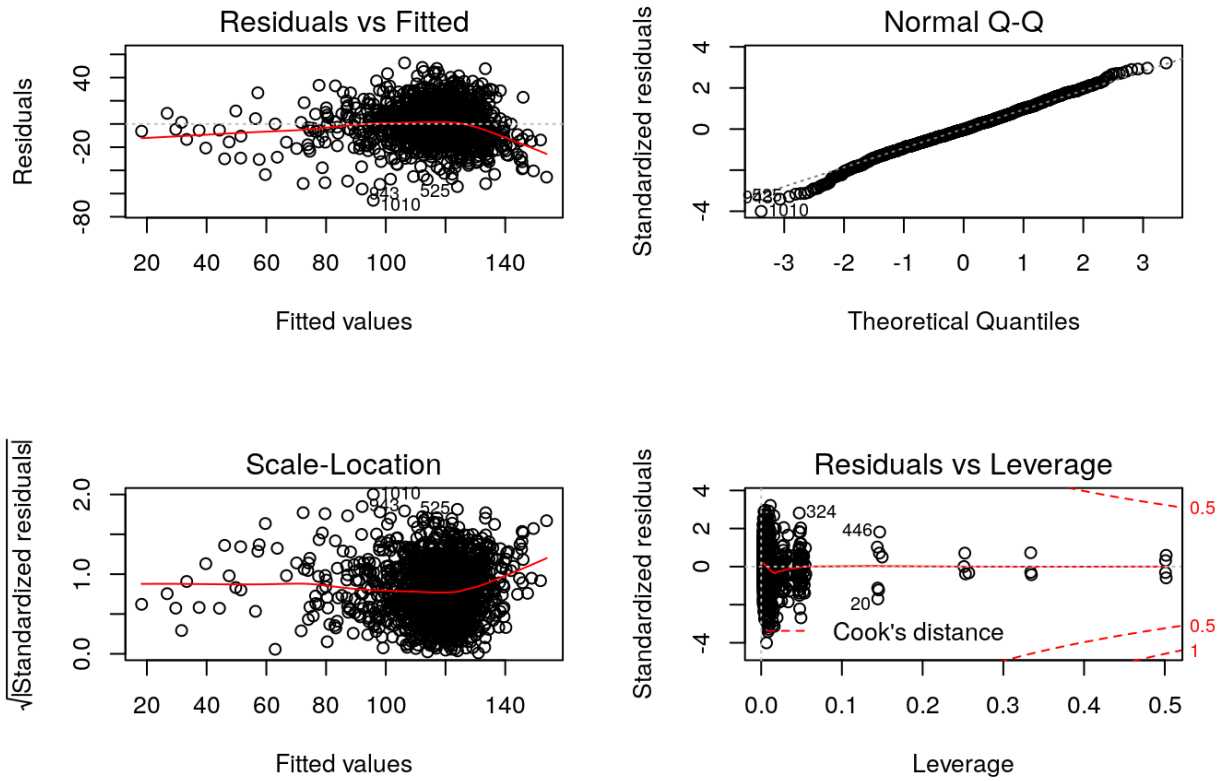
O gráfico de resíduos x valores ajustados apresenta uma forte heterodascidade, com uma grande concentração em volta de 123 Oz.

Modelo 5

Com etnias

```
par(mfrow = c(2, 2))
plot(fit_weight_5)
```

```
## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383, 1387

## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383, 1387
```

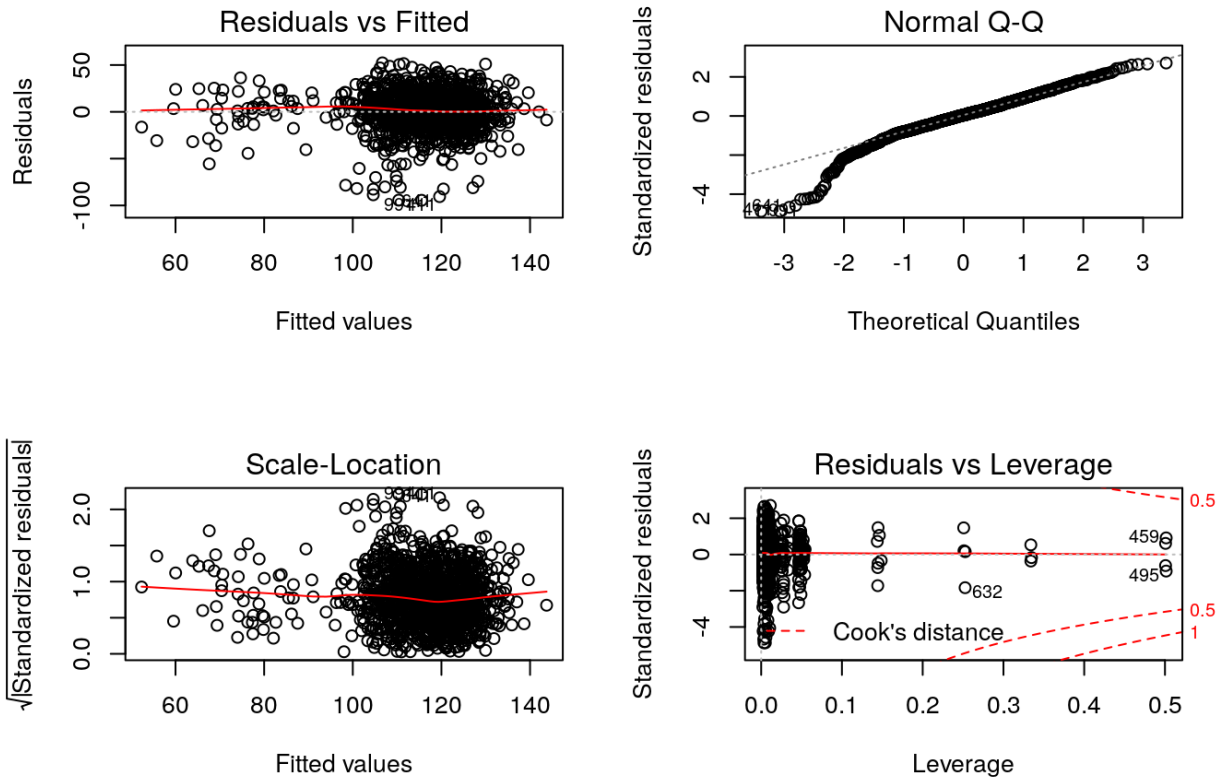


Modelo 6

sem **Weeks**

```
par(mfrow = c(2, 2))
plot(fit_weight_6)
```

```
## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383, 1387

## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383, 1387
```

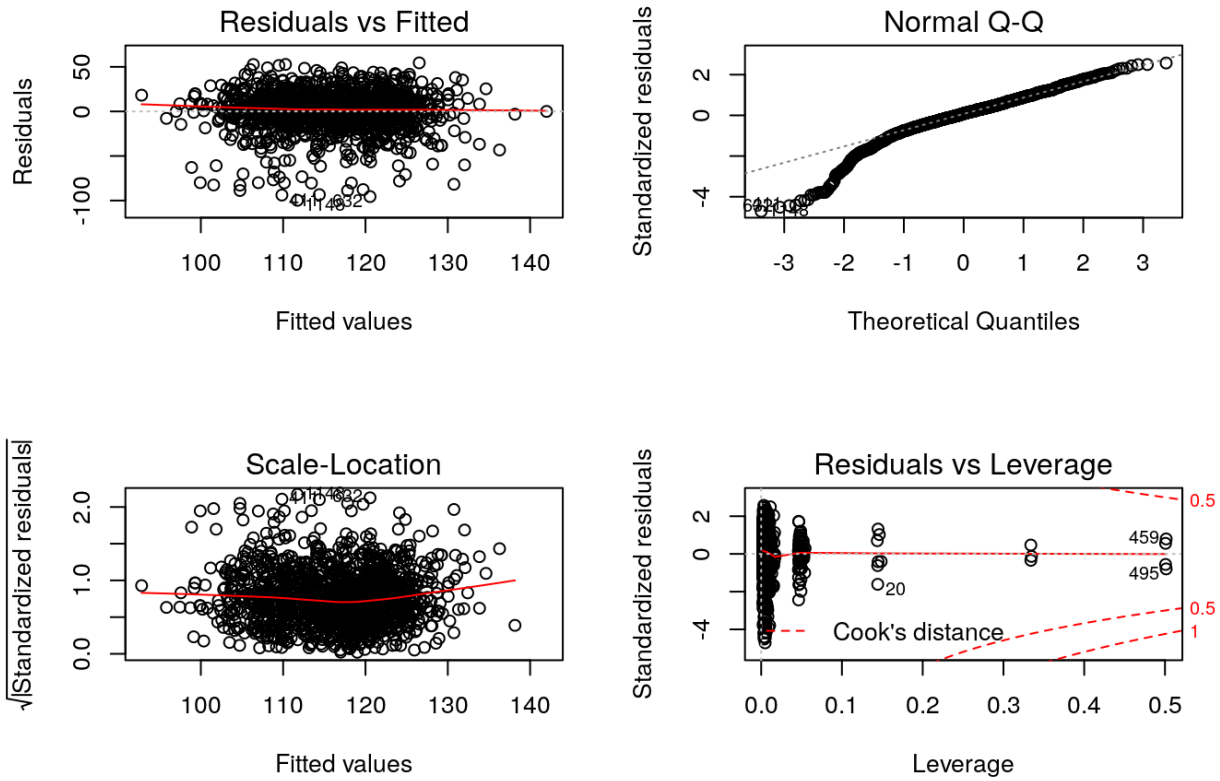


Modelo 7

Sem **Plural**

```
par(mfrow = c(2, 2))
plot(fit_weight_7)
```

```
## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383, 1387

## Warning: not plotting observations with leverage one:
##   1200, 1306, 1383, 1387
```



O gráfico de resíduos x valores ajustados apresenta uma forte heterodascidade, com uma grande concentração em volta de 123 Oz.

Modelo X: Duração da gravidez

NADA POR AQUI

Propostas

d. Proposta de modelo

Com base nas análises, proponha um ou mais modelos lineares multivariados. Explique a sua escolha.

```
# bloco de código - item d
```

Predições

e. Utilize o(s) modelo(s) proposto(s) para fazer pelo menos uma predição.

```
# bloco de código - item e


weeks = c(10,  # Poucas chances de sobrevivência
          11,
          23,  # 17% de sobrevivência
          25,
          39,  # termo
          39.5,
          40.2,
          44,
          45,
          46,  # risco
          47)

gained = c(0.5, 11.3, 43.8, 100, 110)

age = c(4, 6, 8, 10, 12, 14,  # Abaixo
        21.5, 29.3,           # no meio
        44, 45, 46, 60)       # acima



underage_premie = list(Plural="Single", MomAge=10, Sex="M", Weeks=20,
Gained=5, Smoke="N")
underage_tardie = list(Plural="Single", MomAge=10, Sex="M", Weeks=45,
Gained=5, Smoke="N")

up1 = predict(fit_weight_2, underage_premie)
up2 = predict(fit_weight_2, underage_tardie)
```

Peso do bebê

Quantas onças pesaria um bebê nascido nas circunstâncias:

Idade | Tempo de gestação |
<12 | prematuro |
<12 | termo |
<12 | tardio |
>40 | prematuro |
>40 | termo |
>40 | tardio |

Modelo	underage/premie	underage/overtime	overage/premie	overage/tardie
1	26.6069872	130.6760226		

Conclusões e comentários

Comentários

Em retrospecto, penso que teria perdido menos tempo se tivesse feito uma simples análise do R^2 ajustado de todas as variáveis numéricas (já que são poucas). Minha primeira escolha foi justamente a que tem menor potencial de explicação, com os dados disponíveis.

Esses valores sugerem um limite superior para o poder de explicação de um modelo linear que explique as obervações de cara uma dessas variáveis.

Os dados da tabela abaixo foram obtidos da execução do script varios_modelos.R (varios_modelos.R), ideia que só tive perto do final do prazo para entregar o exercício.

R^2 ajustado de modelos naïve para variáveis quantitativas

Coluna	**R^2 ajustado**
MomAge	0.239690016396153
Weeks	0.588115108607321
Gained	0.0818566109518561
BirthWeightOz	0.568165398604105

Suspeitas

Até agora, parece não haver evidências suficientes para compor um modelo linear do peso ganho em função de algumas das outras variáveis. O máximo possível de explicação é de 8% com todos os campos, e vários deles geram coeficientes sem significância estatística ($p > 5\%$).

Quando passei a considerar o peso do bebê ao nascer (BirthWeightOz) como variável dependente, obtive modelos com maior poder de predição ($ {R_{ajust}}^2 > 40%$)

ENTREGA EM 31/10/2018, ÀS 23h59.