Universidade Estadual de Campinas

Instituto de Matemática, Estatística e Computação Cientifica
Departamento de Estatística

Relatório
Trabalho de ME610

Eliane Ramos de Siqueira RA:155233
Guilherme Pazian RA:160323

Professor: Maurício Zevallos

Campinas-SP, 28 de Abril de 2017

1. Introdução

As consequências do hábito de fumar são um dos assuntos mais importantes para pesquisa médica hoje em dia. Em particular, pesquisas anteriores estão apontando uma possível relação decrescente entre este hábito em mulheres durante a gravidez e o peso da criança ao nascer, ou seja, com a presença do hábito, o peso da criança é menor.

Como relatado no Surgeon General's Report, 1989:

"O tabagismo parece ser um determinante mais significativo do peso ao nascer do que a altura, o peso, o numero de fetos, a renda anual da mãe, o histórico de resultados de gestações anteriores ou o sexo do bebê. A redução do peso ao nascer associada ao hábito de fumar parece ser um efeito direto do tabagismo sobre o crescimento fetal."

O objetivo deste trabalho é verificar o quão válida é esta opinião através da análise do peso da criança ao nascer.

Os dados utilizados foram coletados através de entrevistas com as mães durante a gravidez e correspondem a um ano de estudo. Eles incluem 1236 nascimentos onde a criança sobreviveu pelo menos 28 dias. Juntamente com o peso das crianças, foram coletadas informações de outras variáveis que influenciam no peso das crianças, como sexo do bebê, numero de gestações anteriores, hábito de fumar, numero de cigarros que fuma, se parou de fumar, quanto tempo faz, além de raça, altura, peso, nivel educacional e idade do pai e da mãe, estado civil da mãe e rendimento anual da familia.

descrição/metodologia

3. Metodologia

Para este estudo as variáveis disponíveis no banco de dados são:

- Peso
- Sexo;
- Data_nasc
- Vivo
- Qtd_feto
- Tempo_gestacao
- Nmero_gestacoes
- Estado_civil
- Rendimento_anual
- Numero_cigarros
- Tempo_sem_fumar
- Fuma
- Educacao_mae
- Altura_mae
- Peso_mae
- Cor_mae

- Idade_mae
- Educacao_pai
- Altura_pai
- Peso_pai
- Cor_pai
- Idade_pai

foi considerado o peso da criança ao nascer como variável resposta e as demais variáveis como explicativas, afim de avaliar a influencia das demais variáveis sobre o peso do bebê. Em uma primeira análise do banco de dados disponível, notou-se a presença de crianças apenas do sexo masculino que sobreviveram pelo menos 28 dias e eram feto único**, tal caracerística levou a desconsideração das variáveis correspondentes, Sexo,Vivo e Qtd_feto.

Na variavel Cor_mae, temos uma observação com valor desconhecido (99)
Na variavel Idade_mae temos duas observações desconhecidas (99)
Na variavel Educacao_mae temos uma observação desconhecida (9)
Na variavel Altura_mae temos 22 observações desconhecidas (99)
Na variavel Peso_mae temos 36 observações desconhecidas (999)
Na variavel Cor_pai, temos 5 observações com valor desconhecido (99)
Na variavel Idade_pai, temos 7 observações com valor desconhecido (99)
Na variavel Educacao_pai, temos 13 observações com valor desconhecido (9)
Na variavel Altura_pai temos 492 observações desconhecidas (99)
Na variavel Peso_pai temos 499 observações desconhecidas (999)
Na variavel Rendimento_anual temos 124 observações desconhecidas (98)
Na variavel Fuma temos 10 observações desconhecidas (9)
Na variavel Tempo_sem_fumar temos 9 observações desconhecidas (98) e 1 não perguntado(99)
Na variavel Numero_cigarros temos 10 observações desconhecidas (98)

Foram desconsideradas as variaveis Peso_pai e Altura_pai devido a quantidade elevada de observações desconhecidas.

Escrever aqui sobre as variaveis eliminadas e porque

fazendo as devidas conversões para kg, cm

2. Objetivo

4. Resultados

Escreva os resultados aqui

5. Discussão

Escreva a discussão aqui

5. Anexos

   Coloque os anexos aqui

7. Referências

   Escreva as referencias aqui

Limpando os dados

Excluímos as váriaveis: Sexo (apenas masculino), Qtd_feto (apenas 5) e Vivo (apenas 1).

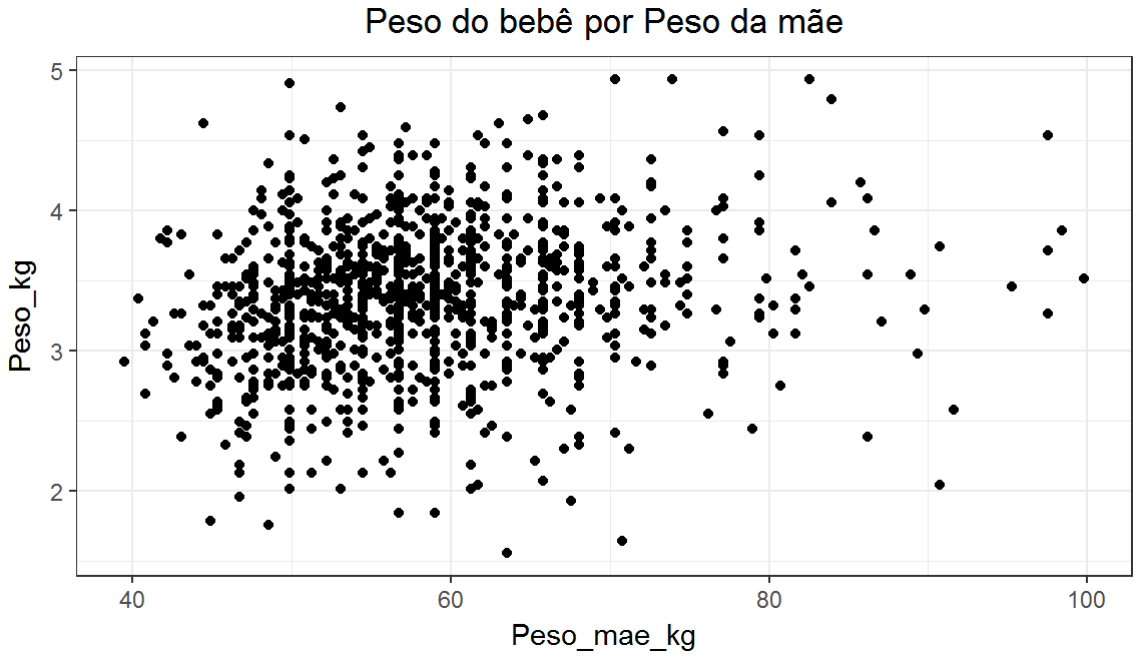
Na variavel Cor_mae, temos uma observação com valor desconhecido (99)
Na variavel Idade_mae temos duas observações desconhecidas (99)
Na variavel Educacao_mae temos uma observação desconhecida (9)
Na variavel Altura_mae temos 22 observações desconhecidas (99)
Na variavel Peso_mae temos 36 observações desconhecidas (999)
Na variavel Cor_pai, temos 5 observações com valor desconhecido (99)
Na variavel Idade_pai, temos 7 observações com valor desconhecido (99)
Na variavel Educacao_pai, temos 13 observações com valor desconhecido (9)
Na variavel Altura_pai temos 492 observações desconhecidas (99)
Na variavel Peso_pai temos 499 observações desconhecidas (999)
Na variavel Rendimento_anual temos 124 observações desconhecidas (98)
Na variavel Fuma temos 10 observações desconhecidas (9)
Na variavel Tempo_sem_fumar temos 9 observações desconhecidas (98) e 1 não perguntado(99)
Na variavel Numero_cigarros temos 10 observações desconhecidas (98)

Tiramos as variaveis Peso_pai e Altura_pai devido a quantidade elevada de observações desconhecidas.

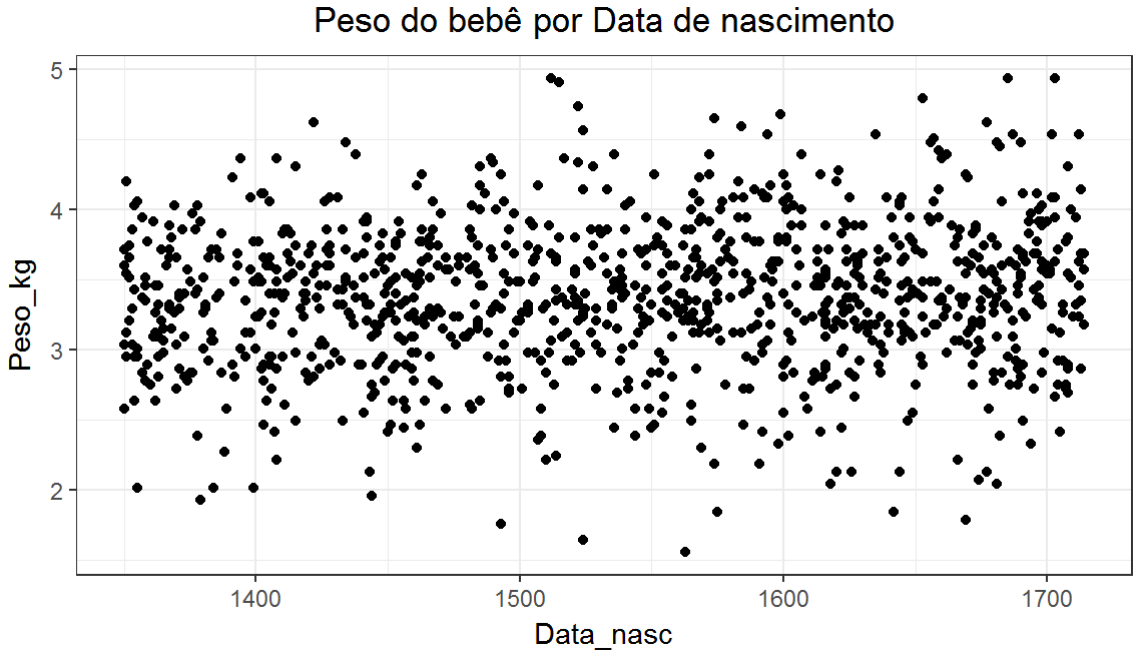
Análise Descritiva

Com base na vizualização de análises descritivas (gráficos de dispersão e boxplos), observou-se as relações entre as variáveis explicativas com a variável resposta (peso do bebê ao nascer).Notou-se que existe pouca relação entre as variáveis explicativas “Altura_mae_cm”,“Idade_pai”, “Idade_mae” e “Data_nasc” em relação ao peso do bebê, ou seja, o peso do bebê parece não estar correlacionado com estas variáveis. Portanto optou-se em não incluir estas variáveis explicativas mesmo num modelo inicial. Em relação às demais variáveis explicativas, todas apresentaram indícios de uma possível corelação com o peso dos bebês, de maneira que a distribuição dos dados referêntes à variável resposta muda conforme os valores observados nas variáveis explicativas. Nenhuma variável contínua/discreta aparentou ter uma relação não-linear com o peso dos bebês.

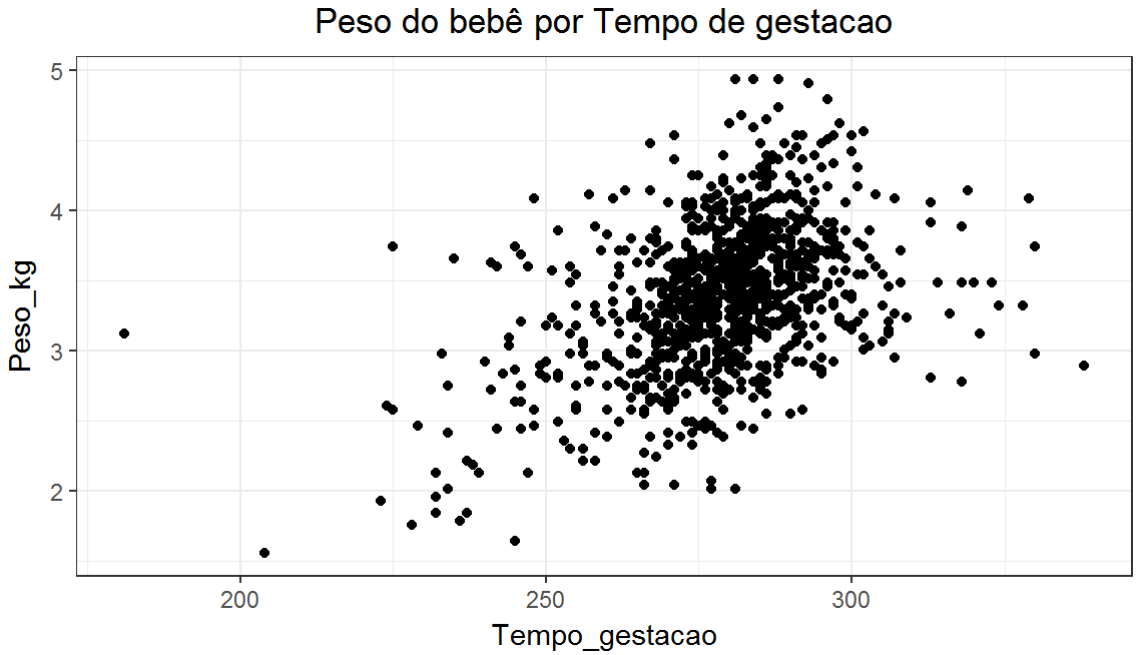
Grafico de Dispersão



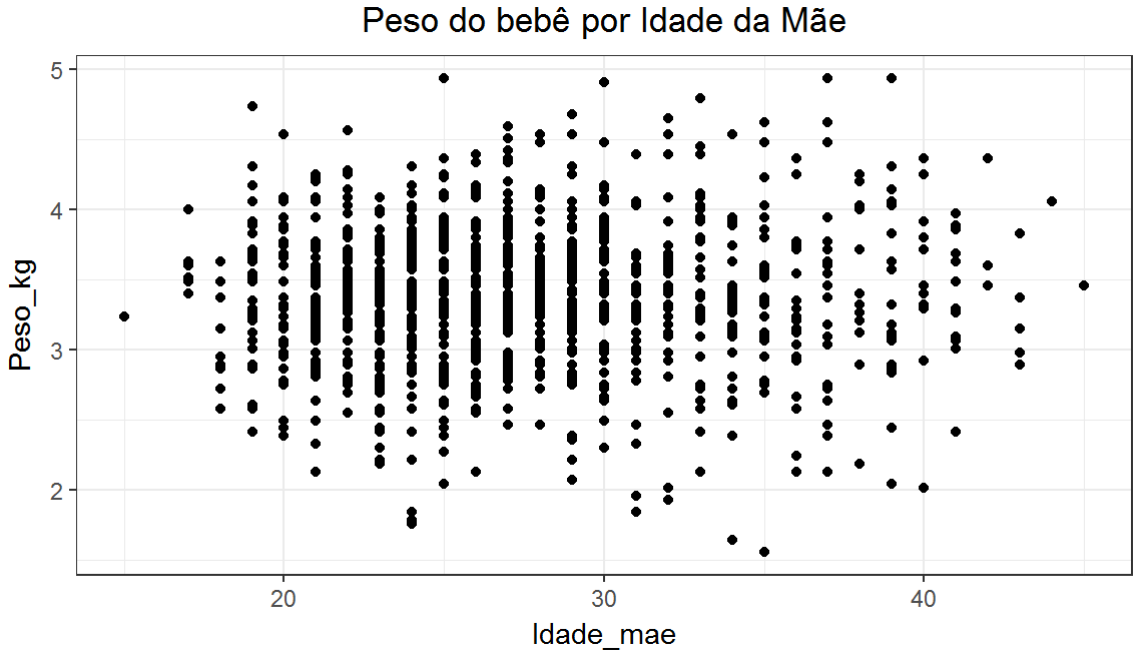
MOderado



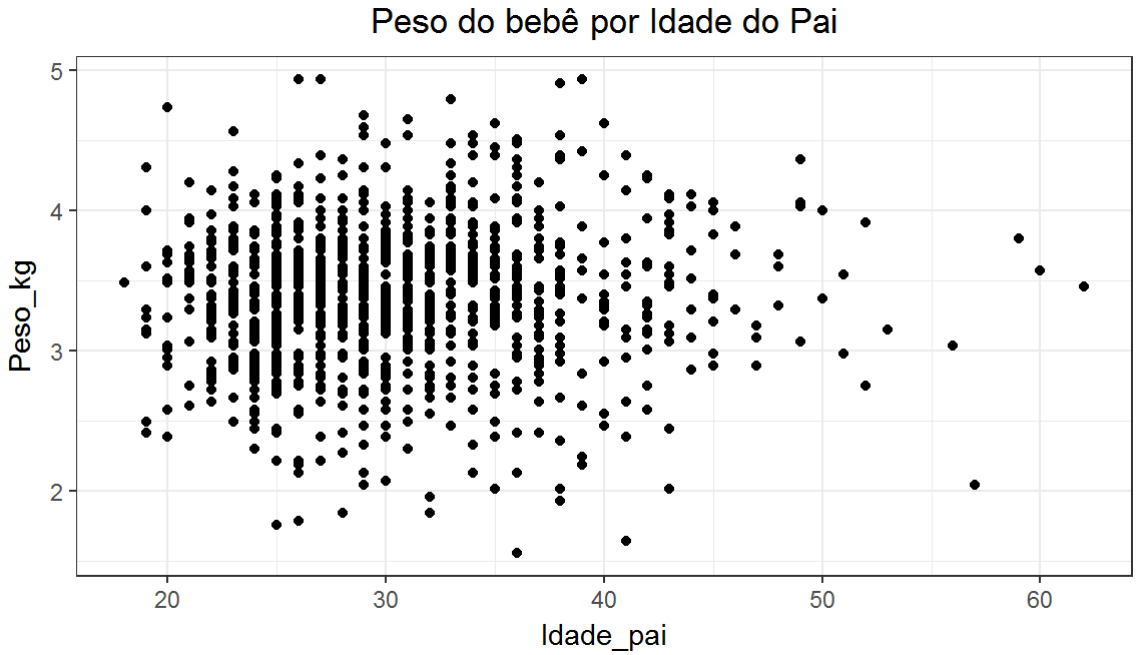
FRACA relaLinear



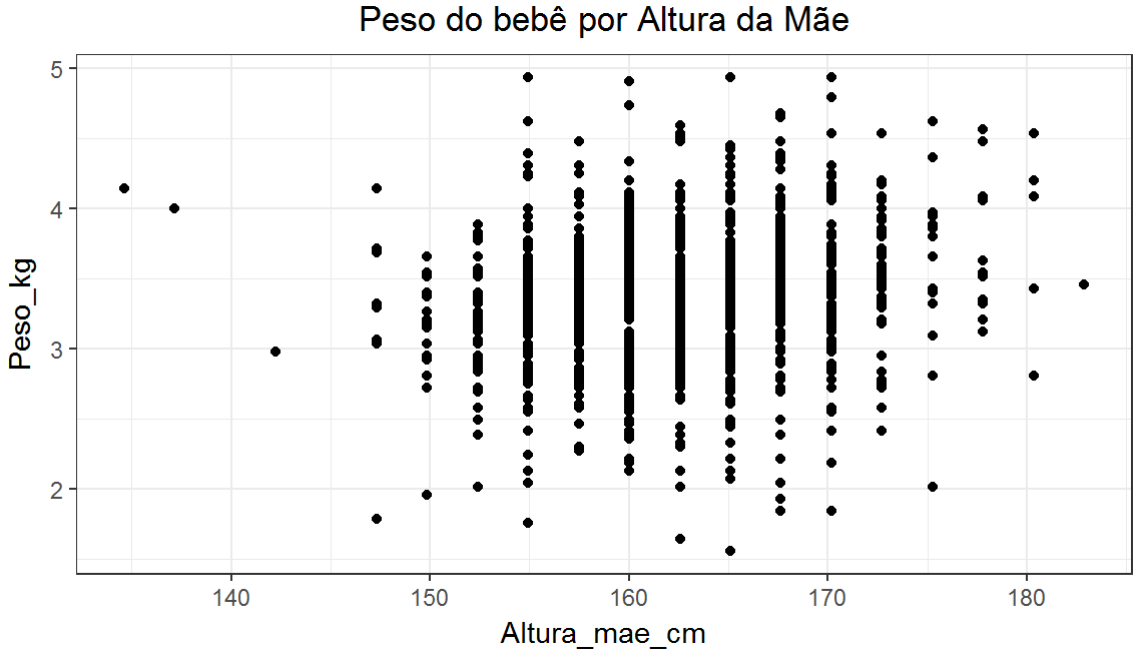
FORTE



Pouca relação linear

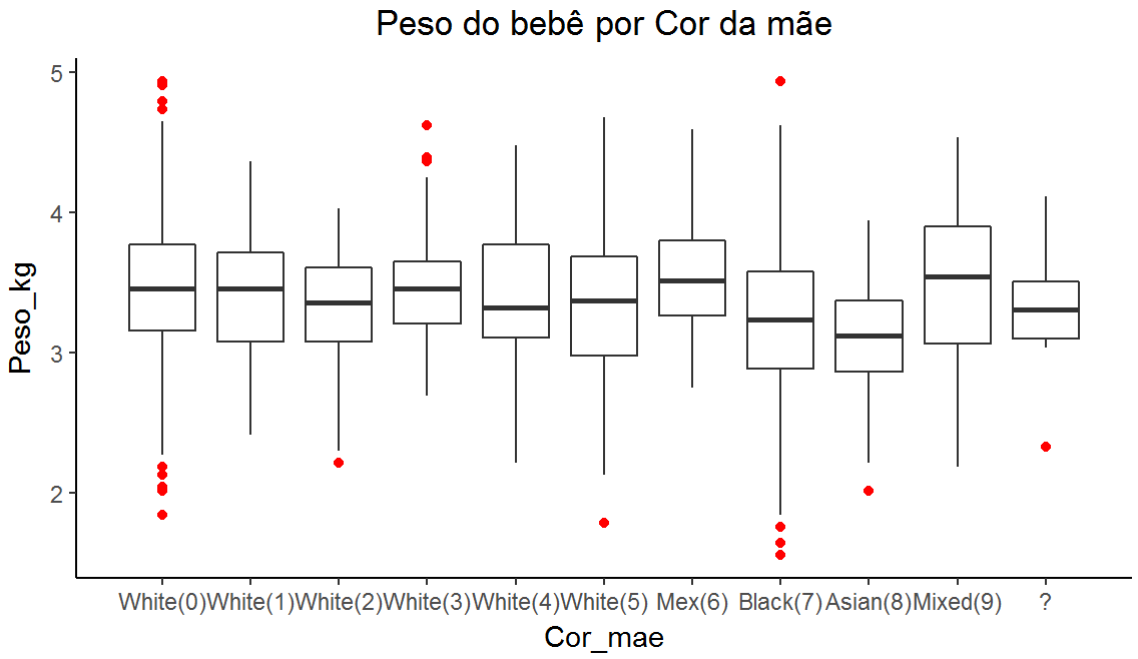


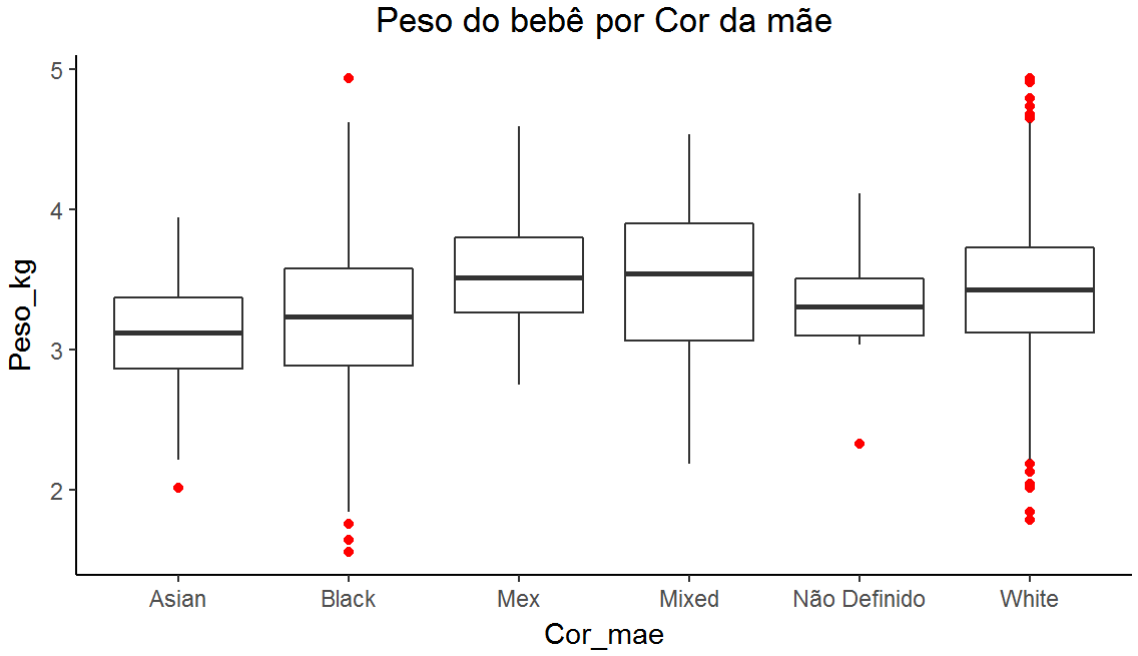
Pouca relação linear

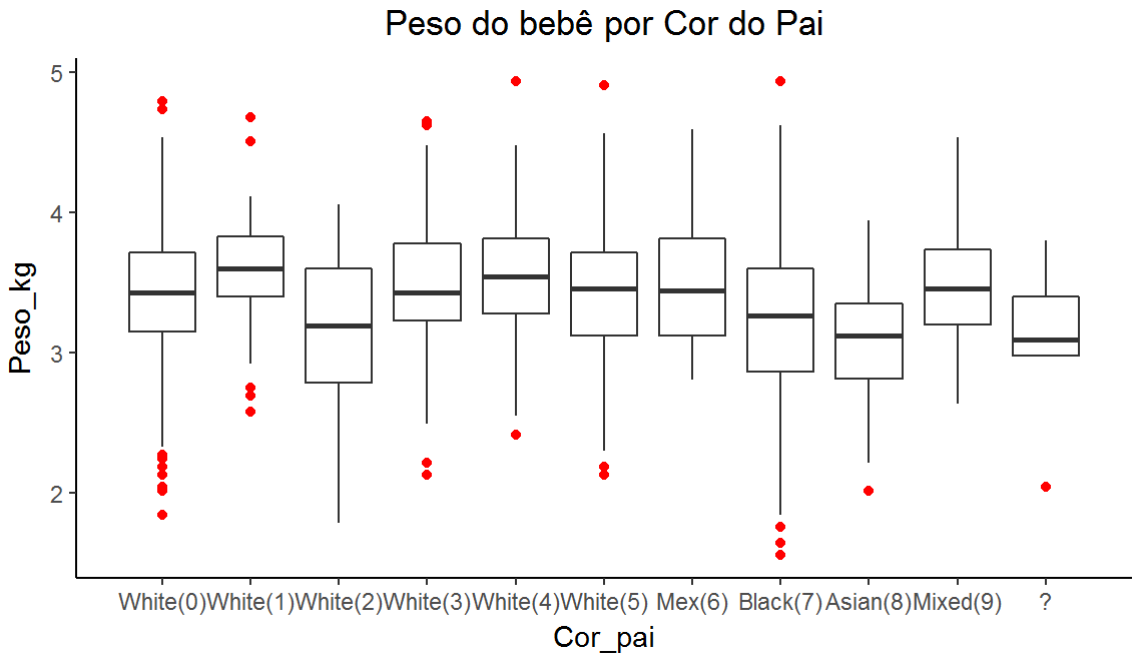


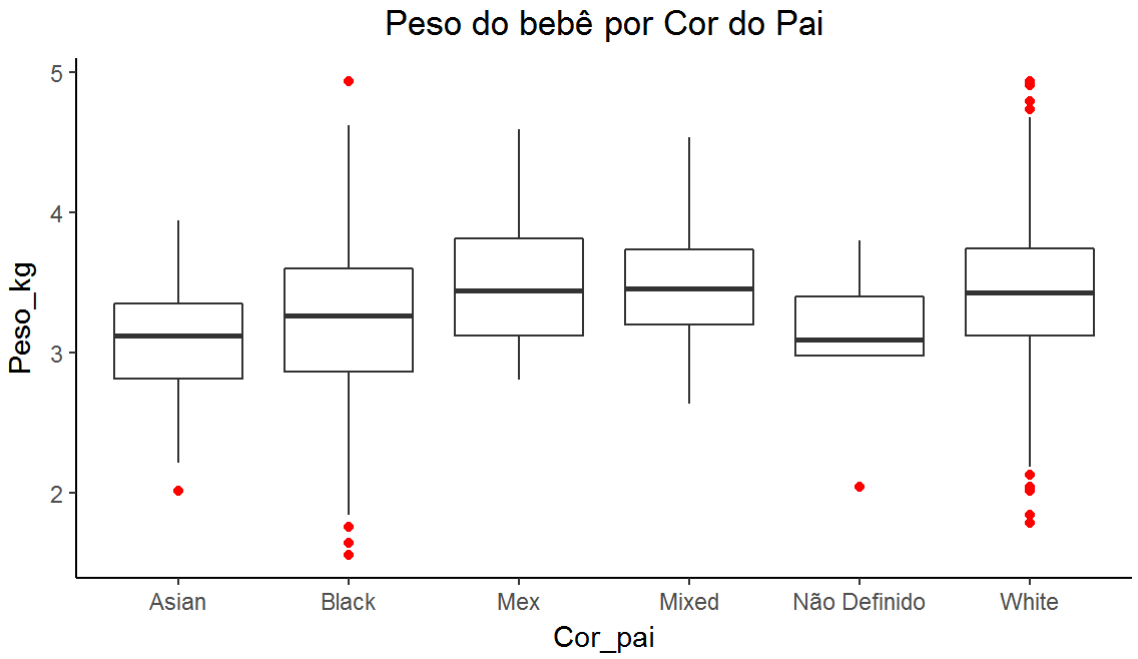
Pouca relação linear

Boxplot para as variaveis categóricas e Dispersão para continua

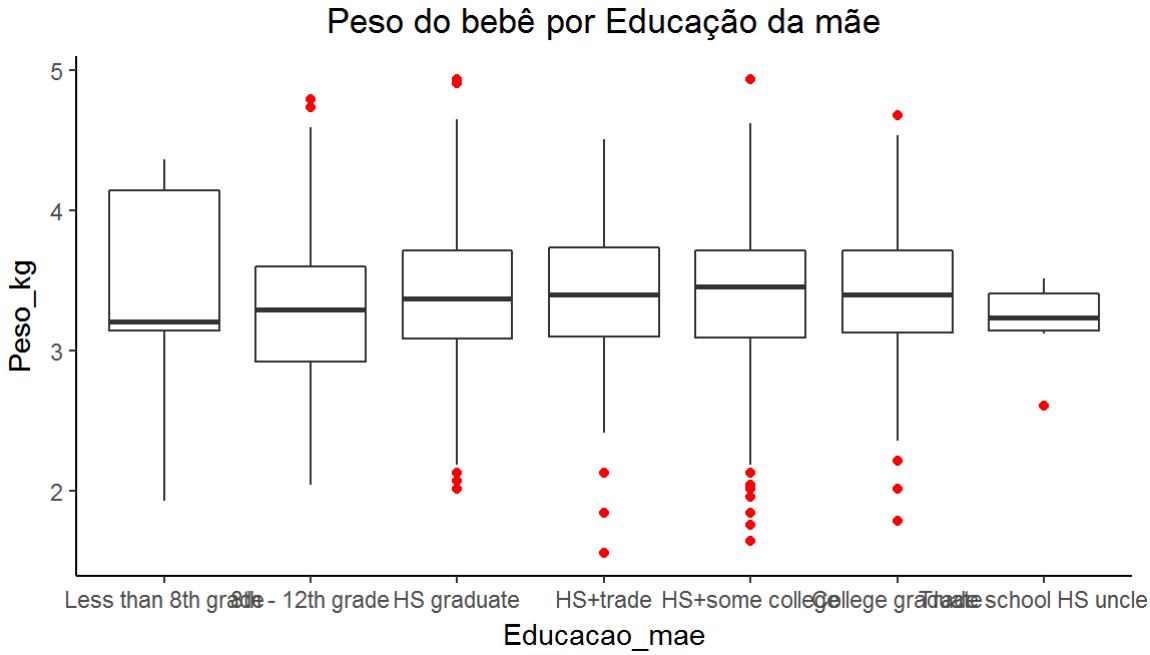


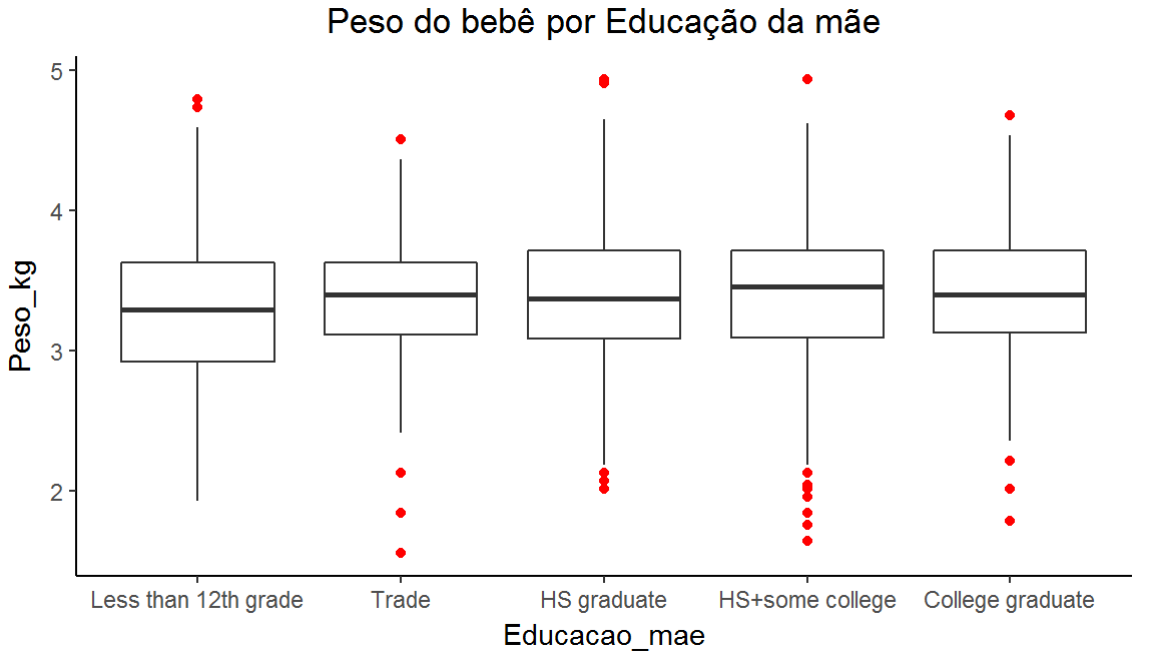



Peso do bebê por Cor do Pai
Peso_kg
5
4
3
2
White(0)
White(1)
White(2)
White(3)
White(4)
White(5)
Mex(6)
Black(7)
Asian(8)
Mixed(9)
?
Cor_pai

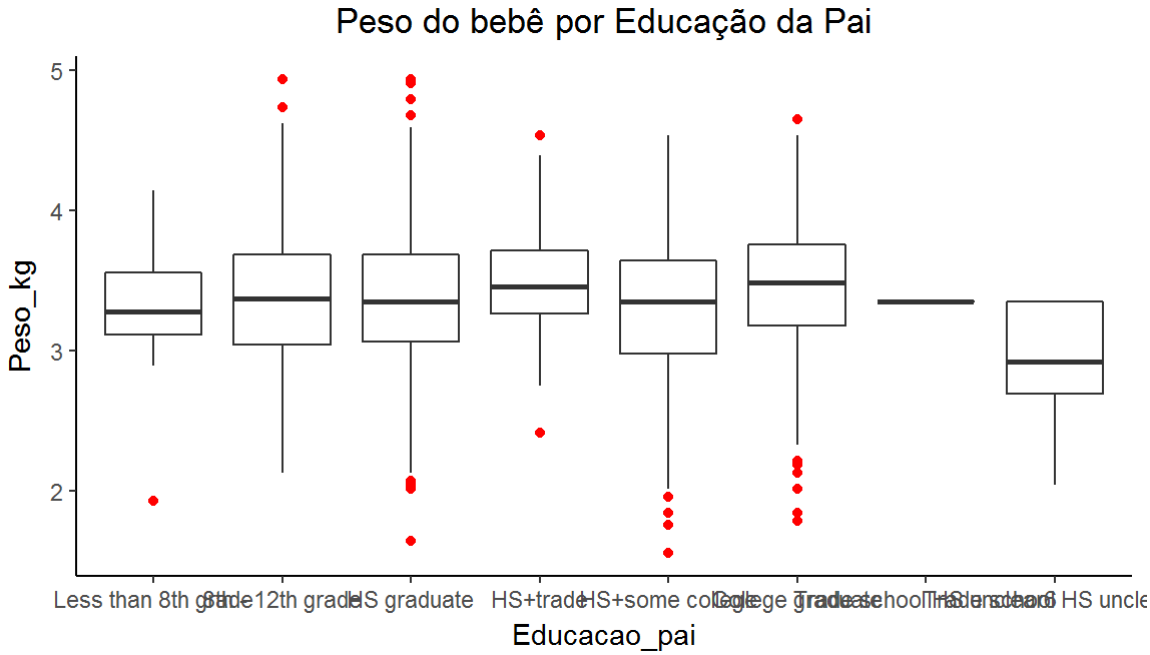

Peso do bebê por Cor do Pai
Peso_kg
5
4
3
2
Asian
Black
Mex
Mixed
Não Definido
White
Cor_pai

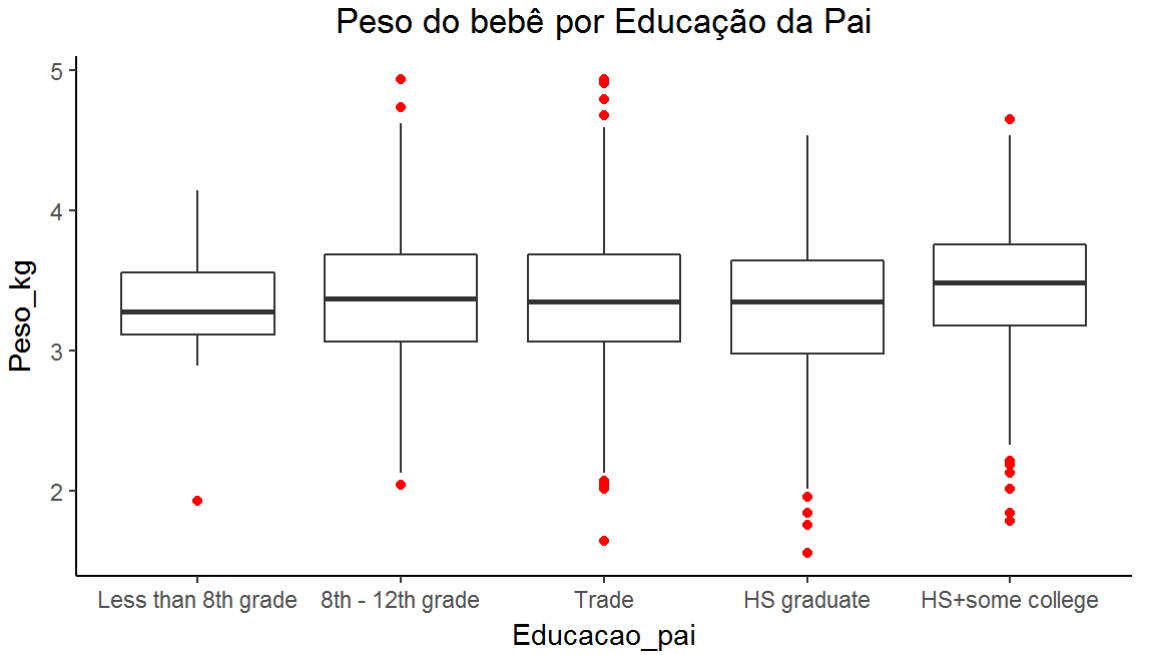
INFLUENCIA


Peso do bebê por Educação da mãe
Peso_kg
5
4
3
2
Less than 8th grade
8th - 12th grade
HS graduate
HS+trade
HS+some college
College graduate
Trade school
HS uncle
Educacao_mae

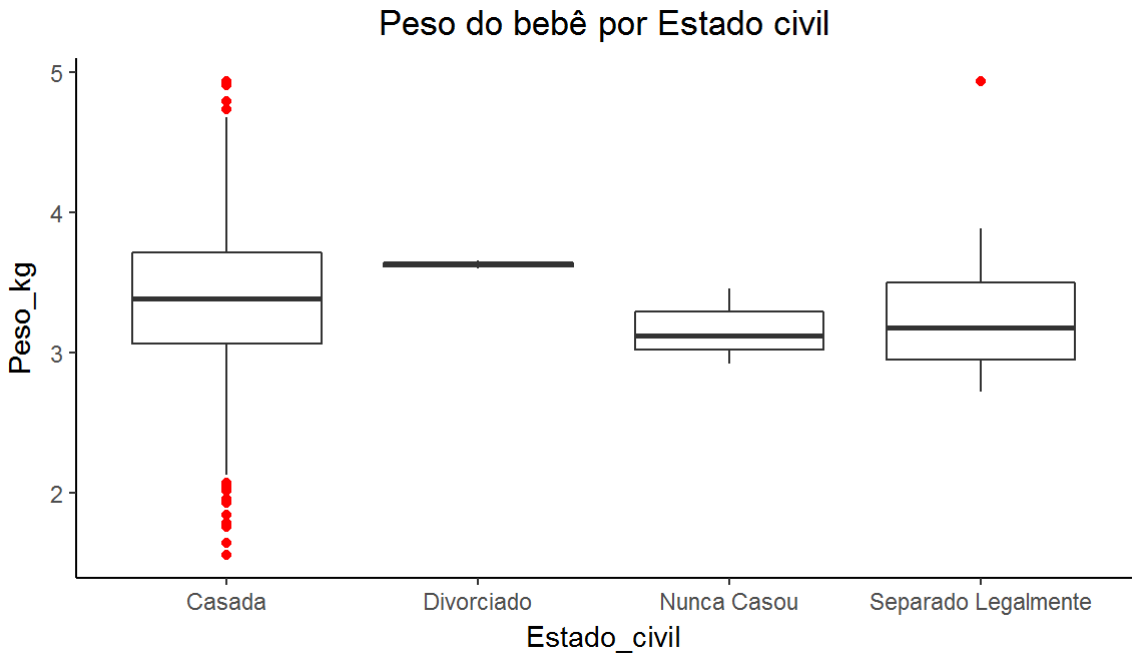

Peso do bebê por Educação da mãe
Peso_kg
5
4
3
2
Less than 12th grade
Trade
HS graduate
HS+some college
College graduate
Educacao_mae

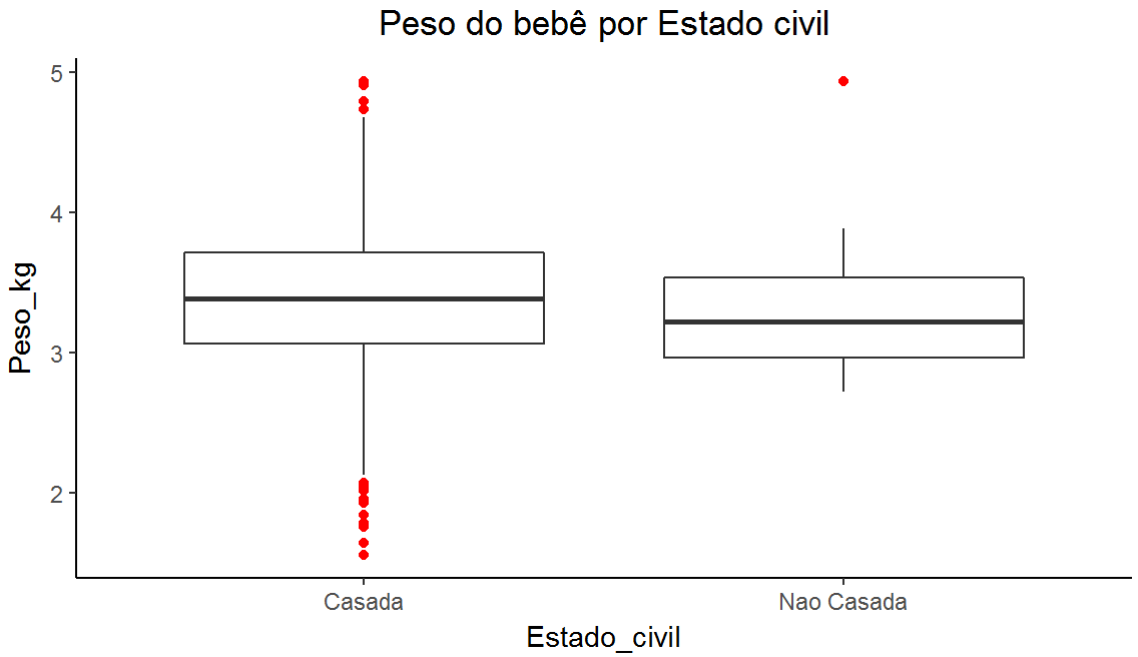
INFLUENCIA


Peso do bebê por Educação da Pai
Peso_kg
5
4
3
2
Less than 8th grade
8th - 12th grade
HS graduate
HS+trade
HS+some college
HS uncle
Educacao_pai

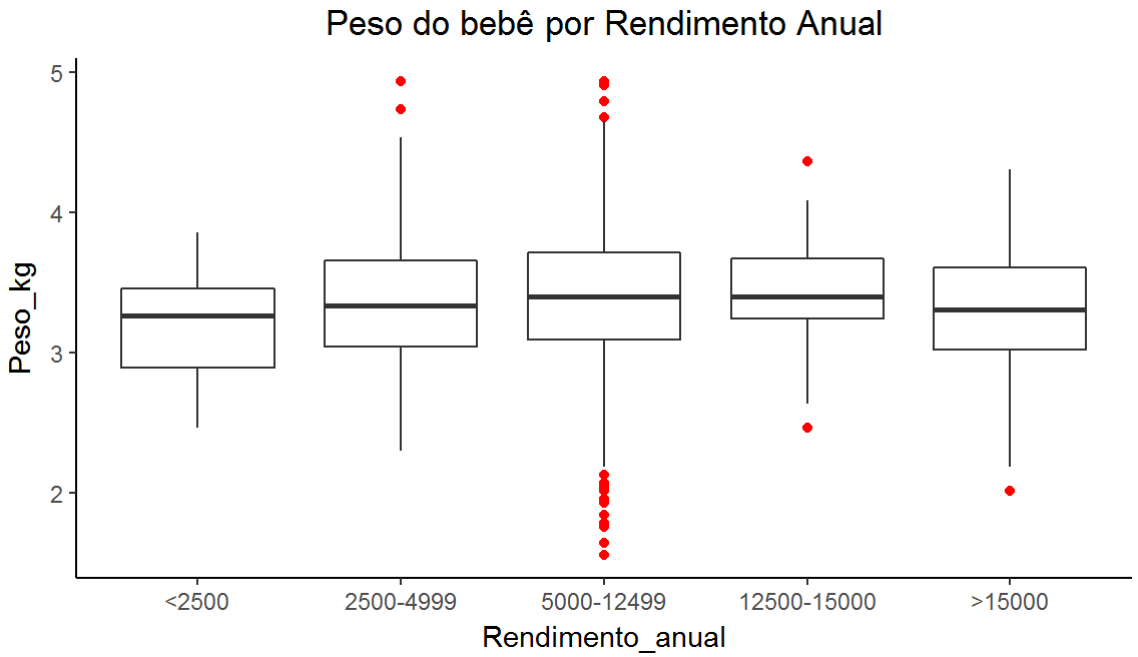

Peso do bebê por Educação da Pai
Peso_kg
5
4
3
2
Less than 8th grade
8th - 12th grade
Trade
HS graduate
HS+some college
Educacao_pai

INFLUENCIA

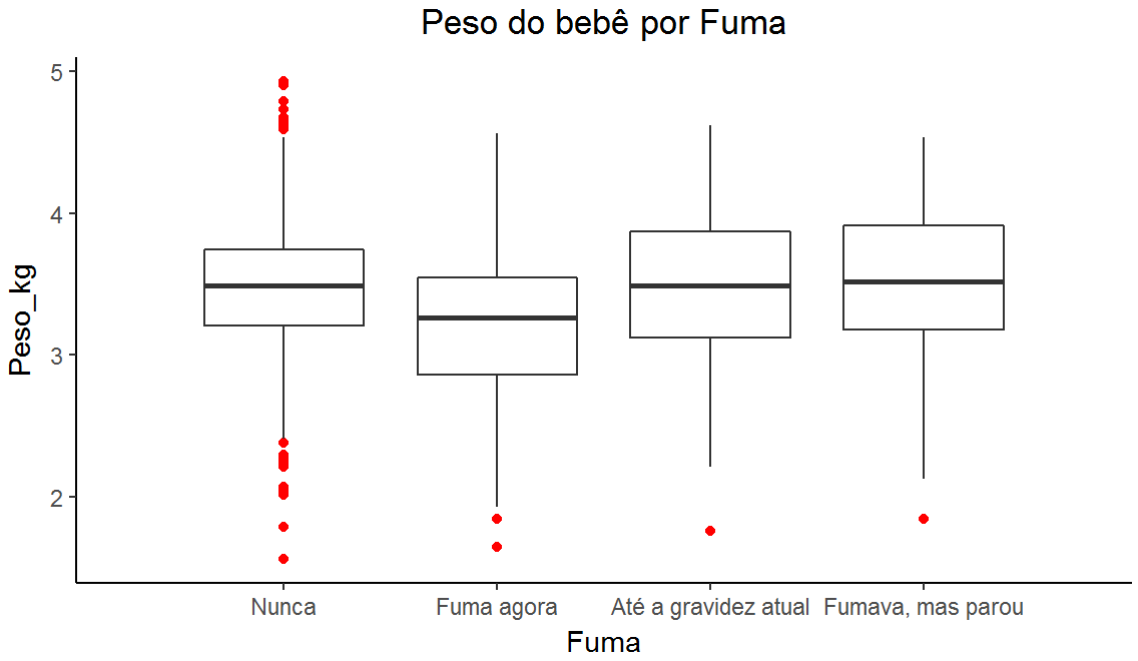

Peso do bebê por Estado civil
Peso_kg
5
4
3
2
Casada
Divorciado
Nunca Casou
Separado Legalmente
Estado_civil

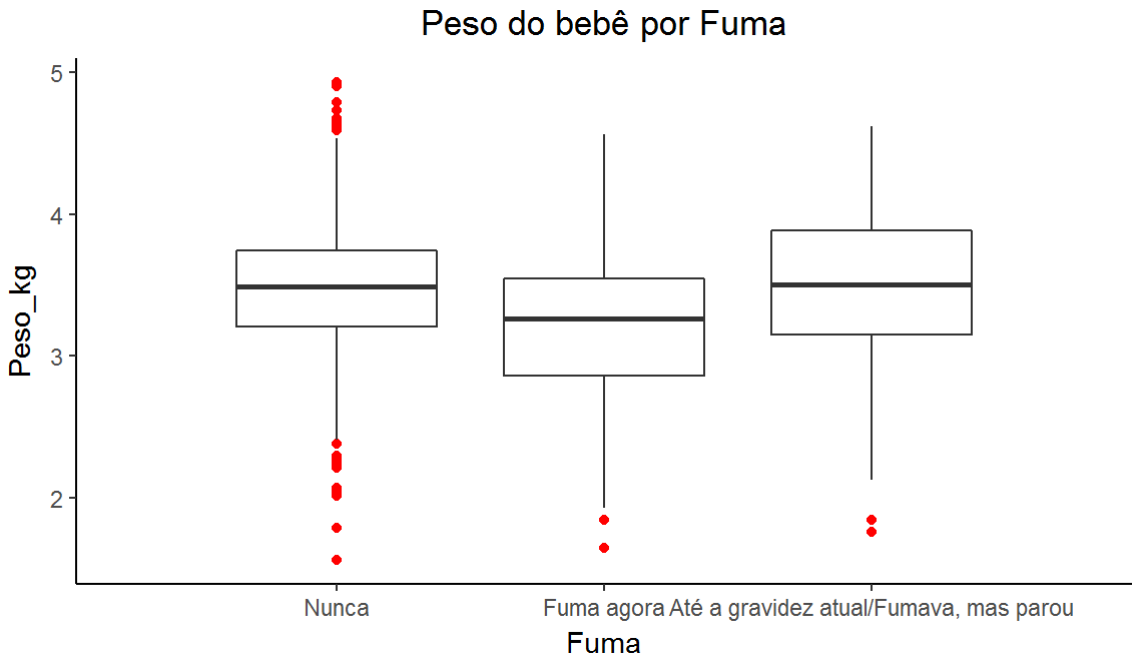

Peso do bebê por Estado civil
Peso_kg
5
4
3
2
Casada
Nao Casada
Estado_civil

INFLUENCIA

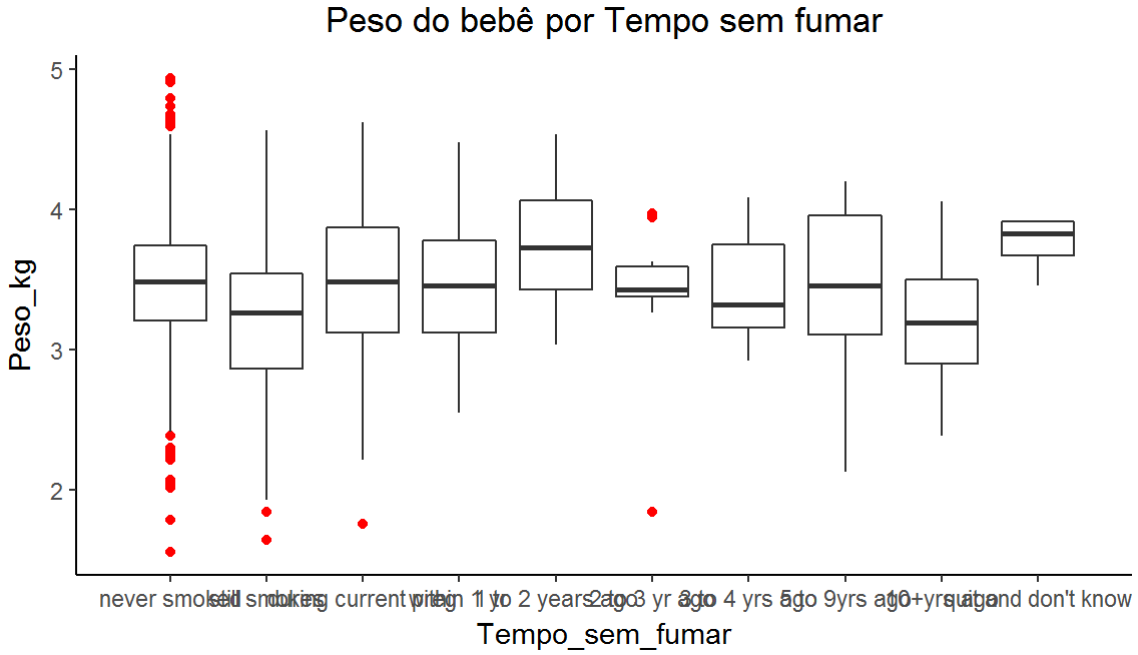

Peso do bebê por Rendimento Anual
Peso_kg
5
4
3
2
<2500
2500-4999
5000-12499
12500-15000
>15000
Rendimento_anual

INFLUÊNCIA****

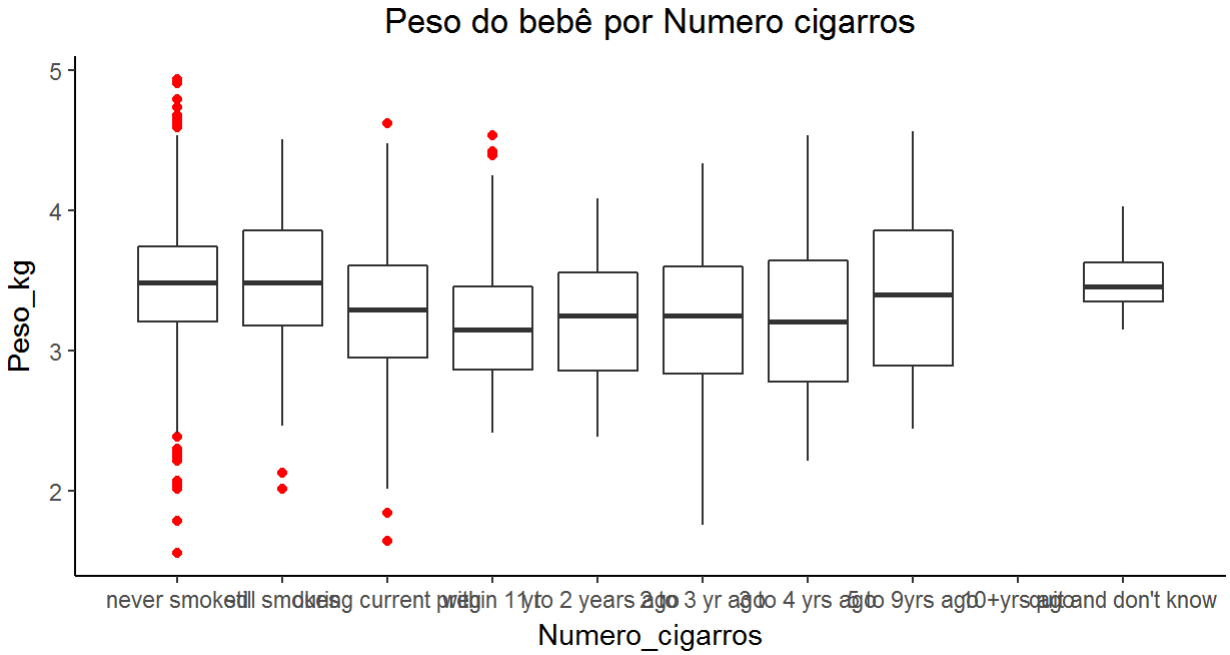

Peso do bebê por Fuma
Peso_kg
5
4
3
2
Nunca
Fuma agora
Até a gravidez atual
Fumava, mas parou
Fuma



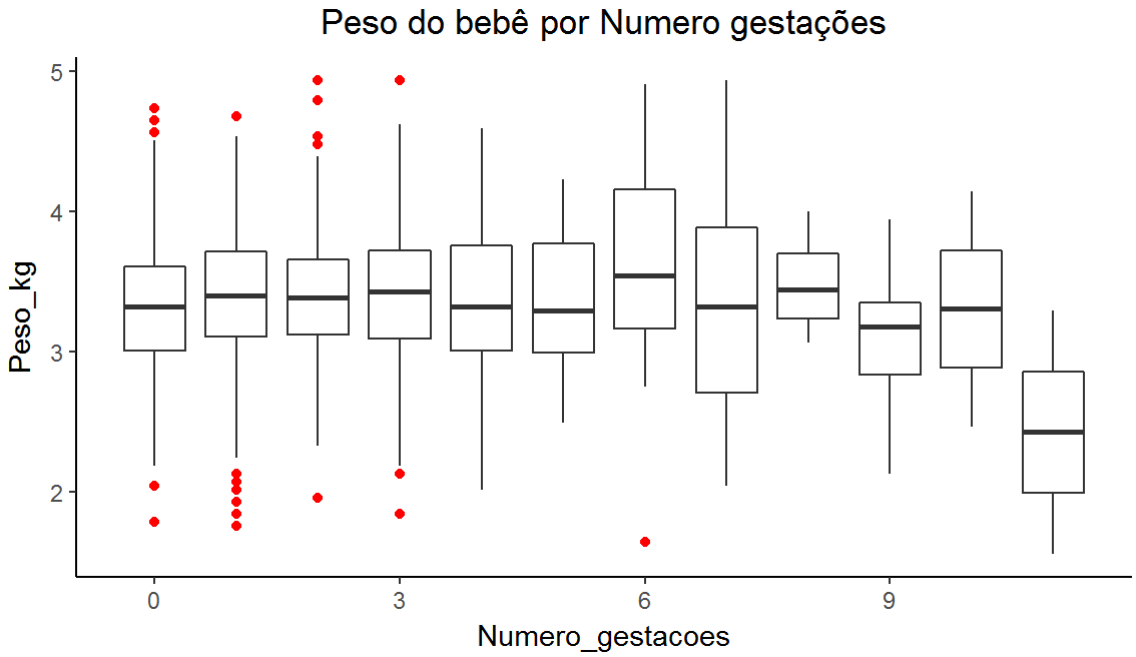
INFLUENCIA



INFLUENCIA



INFLUENCIA



INFLUENCIA

```
##
## Call:
```

```
## lm(formula = Peso_kg ~ Fuma + Rendimento_anual + Estado_civil +
##     Educacao_pai + Educacao_mae + Cor_mae + Cor_pai + Peso_mae_kg +
##     Tempo_gestacao + Data_nasc + Numero_gestacoes + Idade_mae +
##     Altura_mae_cm + Idade_pai + Tempo_sem_fumar, data = dados2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37097 -0.26803 -0.01409  0.26917  1.28249
##
## Coefficients: (2 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.088214   0.158596  19.472  < 2e-16 ***
## FumaFumava                   0.427027   0.220198   1.939 0.052750 .
## FumaNunca                    0.239774   0.030979   7.740 2.44e-14 ***
## Rendimento_anual>15000      -0.072914   0.132423  -0.551 0.582021
## Rendimento_anual12500-15000  0.063823   0.128517   0.497 0.619571
## Rendimento_anual2500-4999    0.103973   0.088327   1.177 0.239421
## Rendimento_anual5000-12499   0.100190   0.084922   1.180 0.238367
## Estado_civil.L               0.066989   0.079584   0.842 0.400140
## Educacao_pai.L               0.003636   0.066574   0.055 0.956457
## Educacao_pai.Q               0.001518   0.053383   0.028 0.977322
## Educacao_pai.C               0.033249   0.040026   0.831 0.406355
## Educacao_pai^4               0.020453   0.030775   0.665 0.506469
## Educacao_mae.L               0.007156   0.045611   0.157 0.875354
## Educacao_mae.Q              -0.029314   0.034077  -0.860 0.389867
## Educacao_mae.C               0.033755   0.043987   0.767 0.443039
## Educacao_mae^4               0.005833   0.036103   0.162 0.871675
## Cor_maeBlack                -0.208267   0.194447  -1.071 0.284397
## Cor_maeMex                   0.009334   0.198765   0.047 0.962554
## Cor_maeMixed                -0.102547   0.194673  -0.527 0.598473
## Cor_maeNão Definido         -0.017288   0.222380  -0.078 0.938051
## Cor_maeWhite                -0.118413   0.150575  -0.786 0.431818
## Cor_paiBlack                 0.207002   0.197627   1.047 0.295152
## Cor_paiMex                   0.362601   0.204732   1.771 0.076851 .
## Cor_paiMixed                 0.313645   0.188852   1.661 0.097070 .
## Cor_paiNão Definido          0.051384   0.235343   0.218 0.827213
## Cor_paiWhite                 0.303496   0.154915   1.959 0.050379 .
## Peso_mae_kg                  0.043345   0.016554   2.618 0.008971 **
## Tempo_gestacao               0.195172   0.014045  13.896  < 2e-16 ***
## Data_nasc                    0.027623   0.013950   1.980 0.047967 *
```

```
## Numero_gestacoes             0.061144   0.017965   3.403 0.000692 ***
## Idade_mae                   -0.004652   0.004607  -1.010 0.312853
## Altura_mae_cm                0.073364   0.016325   4.494 7.82e-06 ***
## Idade_pai                    0.002184   0.003789   0.576 0.564500
## Tempo_sem_fumar1                   NA         NA      NA       NA
## Tempo_sem_fumar2            -0.177057   0.224797  -0.788 0.431102
## Tempo_sem_fumar3            -0.203849   0.236880  -0.861 0.389689
## Tempo_sem_fumar4            -0.014113   0.237222  -0.059 0.952572
## Tempo_sem_fumar5            -0.355631   0.259466  -1.371 0.170802
## Tempo_sem_fumar6            -0.268065   0.267703  -1.001 0.316900
## Tempo_sem_fumar7            -0.294191   0.245949  -1.196 0.231926
## Tempo_sem_fumar8            -0.430564   0.283361  -1.519 0.128958
## Tempo_sem_fumar9                   NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4345 on 992 degrees of freedom
## Multiple R-squared:  0.3157, Adjusted R-squared:  0.2888
## F-statistic: 11.73 on 39 and 992 DF,  p-value: < 2.2e-16
```

```
## [1] 0.3156834
```

```
## [1] 0.2887799
```

Nenhuma estimativa relacionada com as Variáveis “Tempo sem Fumar”,“Estado Civil”, “Educação Pai”, “Educação Mãe”, “Cor da Mãe”,“idade da mãe” e “idade do pai” foram significativos, portanto, será ajustado um novo modelo sem estas variáveis:

```
##
## Call:
## lm(formula = Peso_kg ~ Fuma + Rendimento_anual + Cor_pai + Peso_mae_kg +
##     Tempo_gestacao + Data_nasc + Numero_gestacoes + Altura_mae_cm,
##     data = dados2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37511 -0.27435 -0.00161  0.26382  1.24990
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.96008    0.11325  26.137  < 2e-16 ***
## FumaFumava                   0.23214    0.03990   5.818 7.97e-09 ***
## FumaNunca                    0.24243    0.03005   8.068 2.01e-15 ***
```

```
## Rendimento_anual>15000      -0.08928    0.12778  -0.699  0.48492
## Rendimento_anual12500-15000  0.07040    0.12483   0.564  0.57289
## Rendimento_anual2500-4999    0.11634    0.08743   1.331  0.18362
## Rendimento_anual5000-12499   0.10186    0.08305   1.227  0.22029
## Cor_paiBlack                 0.02669    0.08582   0.311  0.75589
## Cor_paiMex                   0.35963    0.11316   3.178  0.00153 **
## Cor_paiMixed                 0.20926    0.11749   1.781  0.07519 .
## Cor_paiNão Definido          0.02330    0.16460   0.142  0.88745
## Cor_paiWhite                 0.21069    0.08038   2.621  0.00889 **
## Peso_mae_kg                  0.04412    0.01595   2.766  0.00578 **
## Tempo_gestacao               0.19670    0.01387  14.181  < 2e-16 ***
## Data_nasc                    0.02919    0.01373   2.127  0.03367 *
## Numero_gestacoes             0.04790    0.01436   3.336  0.00088 ***
## Altura_mae_cm                0.06958    0.01575   4.419 1.10e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4333 on 1015 degrees of freedom
## Multiple R-squared:  0.3037, Adjusted R-squared:  0.2928
## F-statistic: 27.68 on 16 and 1015 DF,  p-value: < 2.2e-16
```

```
## [1] 0.3037474
```

```
## [1] 0.292772
```

No novo modelo, Nenhuma estimativa relacionada com a Variável “Rendimento Anual” foi estatísticamente significativa, portanto, será ajustado um novo modelo sem estas variáveis:

```
##
## Call:
## lm(formula = Peso_kg ~ Fuma + Cor_pai + Peso_mae_kg + Tempo_gestacao +
##     Data_nasc + Numero_gestacoes + Altura_mae_cm, data = dados2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56007 -0.27306  0.00352  0.26701  1.26376
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.05058    0.08158  37.394  < 2e-16 ***
## FumaFumava           0.23472    0.03986   5.888 5.30e-09 ***
## FumaNunca            0.24166    0.03006   8.040 2.47e-15 ***
```

```
## Cor_paiBlack         0.03698    0.08550   0.432 0.665477
## Cor_paiMex           0.37509    0.11286   3.324 0.000920 ***
## Cor_paiMixed         0.22151    0.11729   1.889 0.059233 .
## Cor_paiNão Definido  0.02207    0.16454   0.134 0.893327
## Cor_paiWhite         0.21606    0.08036   2.689 0.007292 **
## Peso_mae_kg          0.04174    0.01591   2.623 0.008841 **
## Tempo_gestacao       0.19715    0.01387  14.214  < 2e-16 ***
## Data_nasc            0.02796    0.01363   2.051 0.040513 *
## Numero_gestacoes     0.04889    0.01426   3.429 0.000631 ***
## Altura_mae_cm        0.06972    0.01576   4.424 1.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4336 on 1019 degrees of freedom
## Multiple R-squared:  0.2999, Adjusted R-squared:  0.2917
## F-statistic: 36.38 on 12 and 1019 DF,  p-value: < 2.2e-16
```

```
## [1] 0.2999035
```

```
## [1] 0.291659
```

Note que os coeficientes referentes as cores do pai “Não definido” e “Black” são estatísticamente não significativos, portanto, estes são bem similires ao grupo de cor “Asian”, vamos então ajustar um novo modelo com as cores “Asian”, “Não definido” e “Black” pertencendo a um mesmo grupo de cor.

```
##
## Call:
## lm(formula = Peso_kg ~ Fuma + Cor_pai + Peso_mae_kg + Tempo_gestacao +
##     Data_nasc + Numero_gestacoes + Altura_mae_cm, data = dados2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56037 -0.27383  0.00361  0.26626  1.26628
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.08221    0.03459  89.115  < 2e-16 ***
## FumaFumava        0.23481    0.03982   5.896 5.05e-09 ***
## FumaNunca         0.24062    0.02992   8.042 2.44e-15 ***
## Cor_paiMex        0.34476    0.08735   3.947 8.46e-05 ***
## Cor_paiMixed      0.18937    0.09008   2.102 0.035780 *
## Cor_paiWhite      0.18498    0.03430   5.393 8.60e-08 ***
```

```
## Peso_mae_kg       0.04282    0.01571   2.726 0.006519 **
## Tempo_gestacao    0.19703    0.01385  14.222  < 2e-16 ***
## Data_nasc         0.02813    0.01361   2.067 0.039011 *
## Numero_gestacoes  0.04950    0.01416   3.494 0.000496 ***
## Altura_mae_cm     0.07021    0.01563   4.492 7.85e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4332 on 1021 degrees of freedom
## Multiple R-squared:  0.2998, Adjusted R-squared:  0.2929
## F-statistic: 43.71 on 10 and 1021 DF,  p-value: < 2.2e-16
```

```
## [1] 0.2997729
```

```
## [1] 0.2929147
```

9 - Educacao_pai1 36 - Cor_pai10 43 - Idade_pai