Regressão linear múltipla: *the madness continues*

Motivação

O modelo de regressão é extremamente útil como ferramenta explanatória e preditiva, mas até agora nos limitamos à situação ao em que temos apenas uma variavel independente. Na vida real é comum estudarmos fenômenos com múltiplas causas. Em aplicações reais, em geral dispomos de muitas covariaveis que podem, em princípio, estar relacionadas à variável dependente (desfecho/resposta).

Nesta lição vamos fazer o nosso modelo de regressão ficar mais realista incluindo múltiplas covariáveis (ou variáveis explanatórias) de uma vez. Vamos entender como estimar quantidades-chaves do modelo e também como diagnosticar o seu ajuste.

O modelo

Vamos escrever

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

com

$$\boldsymbol{\varepsilon} \sim \text{Normal}(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Estimação

É possível mostrar que os estimadores de máxima verossimilhança das quantidades desconhecidas são

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{1}$$

$$\hat{\sigma^2} = \frac{1}{n-2}\left(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}\right). \tag{2}$$

Nos exercícios abaixo, você vai mostrar que estes estimadores têm boas propriedades, como não-viesamento. Em particular,

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}) \tag{3}$$

Além disso, a informação de Fisher para $\boldsymbol{\beta}$ vale

$$\boldsymbol{I}(\boldsymbol{\beta}) = \frac{\boldsymbol{X}^T \boldsymbol{X}}{\sigma^2}. \tag{4}$$

Diagnósticos

Parte integral de qualquer análise é diagnosticar o ajuste do modelo aos dados. Em uma análise de regressão, um diagnóstico importante é a *análise de resíduos*. Defina

$$\hat{e}_i = y_i - \hat{\theta}_i.$$

Como o resíduo da i-ésima observação em relação ao seu valor ajustado. Podemos então escrever

$$E[\hat{\boldsymbol{e}}\hat{\boldsymbol{e}}^T] = \sigma^2 \left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right],$$

para a matriz de covariância dos resíduos. Um passo importante é *padronizar* os resíduos:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.$$

Uma boa medida para a *influência* de uma observação é a distância de Cook:

$$D_i := \frac{1}{p}\left(\frac{h_{ii}}{1 - h_{ii}}\right) r_i^2. \tag{5}$$

Análise

Agora vamos empregar o modelo sob estudo e os resultados listados para analisar dados reais.

Os dados

Os dados que vamos analisar aqui são as notas (*scores*) em um teste padronizado de $n =$ crianças, para os quais foram medidos o quociente de inteligência (QI) da mãe e também se a mãe completou o ensino médio (*high school*, `hs`). Os dados estão aqui e são discutidos no capítulo 10 de Gelman, Hill & Vehtari (2020).

Nosso objetivo é entender como a habilidade inata da criança (predita presumivelmente pelo QI da mãe) e sua condição socioeconômica (sinalizado pelo `hs` da mãe) influenciam no desempenho.

Vamos olhar as principais estatísticas descritivas

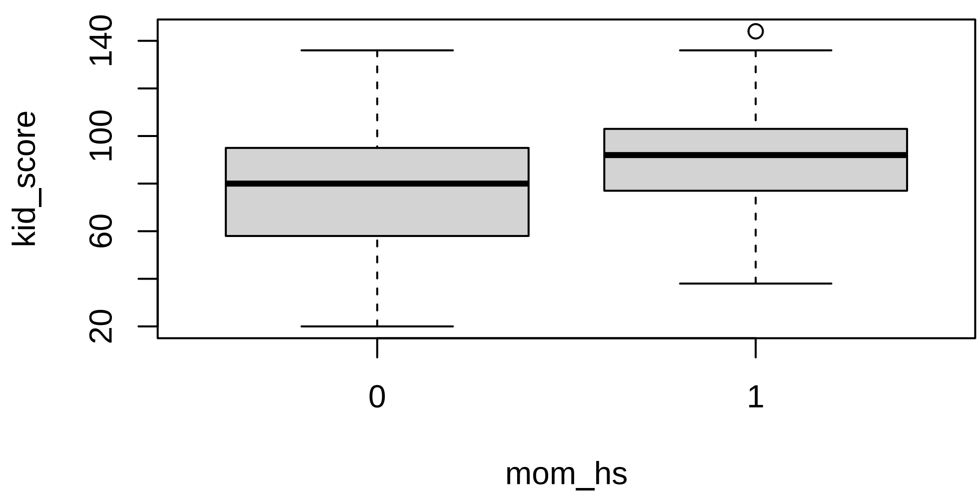
```r
head(kidiq)
```

```
# A tibble: 6 x 5
  kid_score mom_hs mom_iq mom_work mom_age
      <dbl> <fct>   <dbl>    <dbl>   <dbl>
1        65 1       121.         4      27
2        98 1        89.4        4      25
3        85 1       115.         4      27
4        83 1        99.4        3      25
5       115 1        92.7        4      27
6        98 0       108.         1      18
```

```r
summary(kidiq)
```

```
   kid_score      mom_hs       mom_iq            mom_work         mom_age
 Min.   : 20.0   0: 93   Min.   : 71.04   Min.   :1.000   Min.   :17.00
 1st Qu.: 74.0   1:341   1st Qu.: 88.66   1st Qu.:2.000   1st Qu.:21.00
 Median : 90.0           Median : 97.92   Median :3.000   Median :23.00
 Mean   : 86.8           Mean   :100.00   Mean   :2.896   Mean   :22.79
 3rd Qu.:102.0           3rd Qu.:110.27   3rd Qu.:4.000   3rd Qu.:25.00
 Max.   :144.0           Max.   :138.89   Max.   :4.000   Max.   :29.00
```

```r
boxplot(kid_score ~ mom_hs, kidiq)
```



As perguntas

De posse desses dados, podemos nos fazer várias perguntas sobre *associações* nos dados. Por exemplo, queremos saber se o QI da mãe tem alguma associação (leia-se: capacidade preditiva) com as notas (*scores*) da criança. Além disso, essa associação é mediada pelo nível educacional da mãe? Como o fato de que a mãe trabalha fora impacta a variável resposta (na presença das outras covariáveis relevantes)?

Ajustando e investigando modelos

Em primeiro lugar, vamos ajustar o modelo

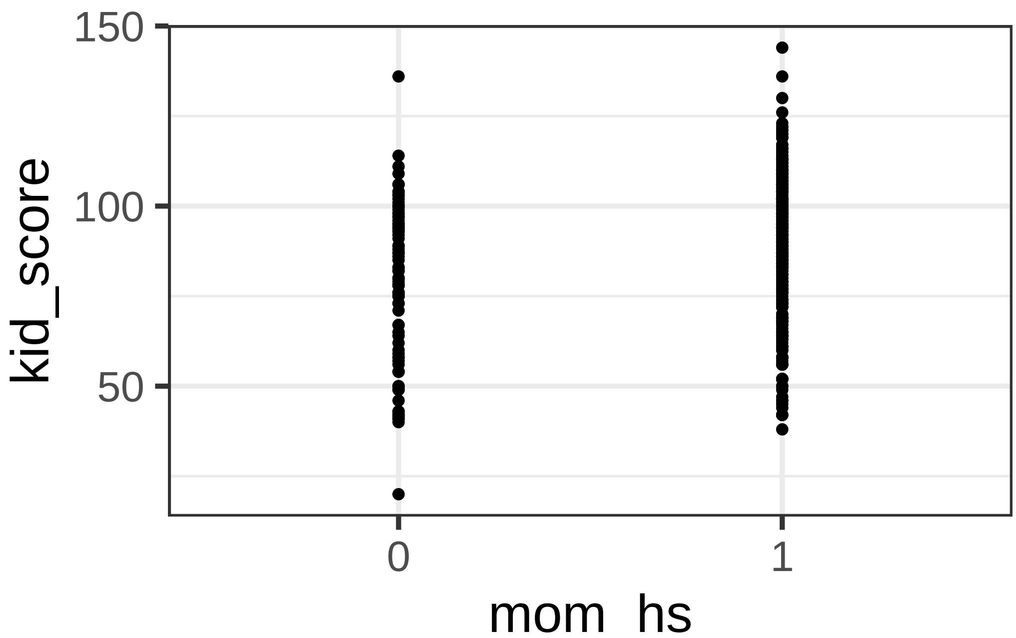
$$\text{kid\_score} = \beta_0 + \beta_{\text{hs}}\text{mom\_hs} + \varepsilon.$$

```r
modelo1 <- lm(kid_score ~ mom_hs, kidiq)
summary(modelo1)
```

```
Call:
lm(formula = kid_score ~ mom_hs, data = kidiq)

Residuals:
   Min     1Q Median     3Q    Max 
-57.55 -13.32   2.68  14.68  58.45 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   77.548      2.059  37.670  < 2e-16 ***
mom_hs1       11.771      2.322   5.069 5.96e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.85 on 432 degrees of freedom
Multiple R-squared:  0.05613,	Adjusted R-squared:  0.05394 
F-statistic: 25.69 on 1 and 432 DF,  p-value: 5.957e-07
```

Deste ajuste fica claro que que a média dos scores é ≈ 77.5 e que condicionado em `mom_hs=1`, em média a criança tem pontuação ≈ 11.8 maior. Isso não é exatamente surpresa, já que se compararmos as médias nos boxplots acima, vemos uma diferença mais ou menos igual a essa.

Vamos visualizar a reta ajustada:



Vamos agora olhar o modelo

$$\text{kid\_score} = \beta_0 + \beta_{\text{iq}}\text{mom\_iq} + \varepsilon.$$

```r
modelo2 <- lm(kid_score ~ mom_iq, kidiq)
summary(modelo2)
```

```
Call:
lm(formula = kid_score ~ mom_iq, data = kidiq)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.753 -12.074   2.217  11.710  47.691 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.79978    5.91741    4.36 1.63e-05 ***
mom_iq       0.60997    0.05852   10.42  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

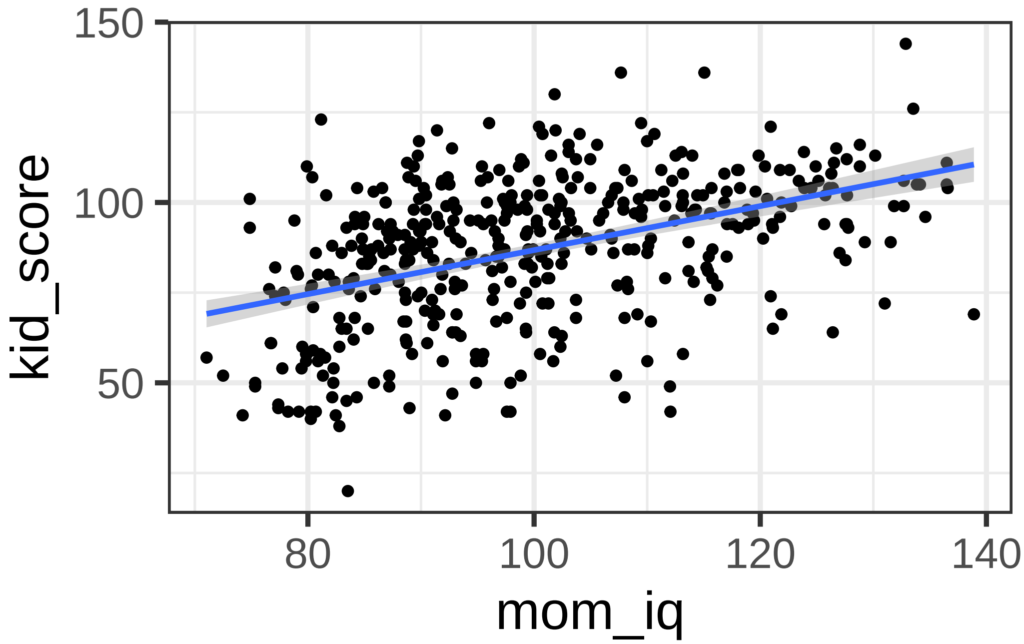
```
Residual standard error: 18.27 on 432 degrees of freedom
Multiple R-squared:  0.201, Adjusted R-squared:  0.1991
F-statistic: 108.6 on 1 and 432 DF,  p-value: < 2.2e-16
```

Vemos que a estimativa do intercepto muda (porquê?) e vemos também que para cada ponto de QI da mãe, a pontuação da criança aumenta ≈ 0.61 pontos. Nessa situação fica claro que é difícil intepretar $\hat{\beta}_0$ (porquê?). Para resolver isso, vamos centrar a variável contínua e reajustar o modelo.

```r
kidiq$c_mom_iq <- kidiq$mom_iq - mean(kidiq$mom_iq)
modelo3 <- lm(kid_score ~ c_mom_iq, kidiq)
summary(modelo3)
```

```
Call:
lm(formula = kid_score ~ c_mom_iq, data = kidiq)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.753 -12.074   2.217  11.710  47.691 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 86.79724    0.87680   98.99   <2e-16 ***
c_mom_iq     0.60997    0.05852   10.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.27 on 432 degrees of freedom
Multiple R-squared:  0.201, Adjusted R-squared:  0.1991
F-statistic: 108.6 on 1 and 432 DF,  p-value: < 2.2e-16
```

Agora fica mais fácil interpretar $\hat{\beta}_0 \approx 86.8$ como o valor esperado da nota quando a mãe tem um QI médio. Agora vamos olhar a reta ajustada junto com um intervalo de confiança para o preditor linear.



Agora vamos finalmente ajustar o modelo com as duas covariáveis:

$$\text{kid\_score} = \beta_0 + \beta_{\text{hs}}\text{mom\_hs} + \beta_{\text{iq}}\text{mom\_iq} + \varepsilon.$$

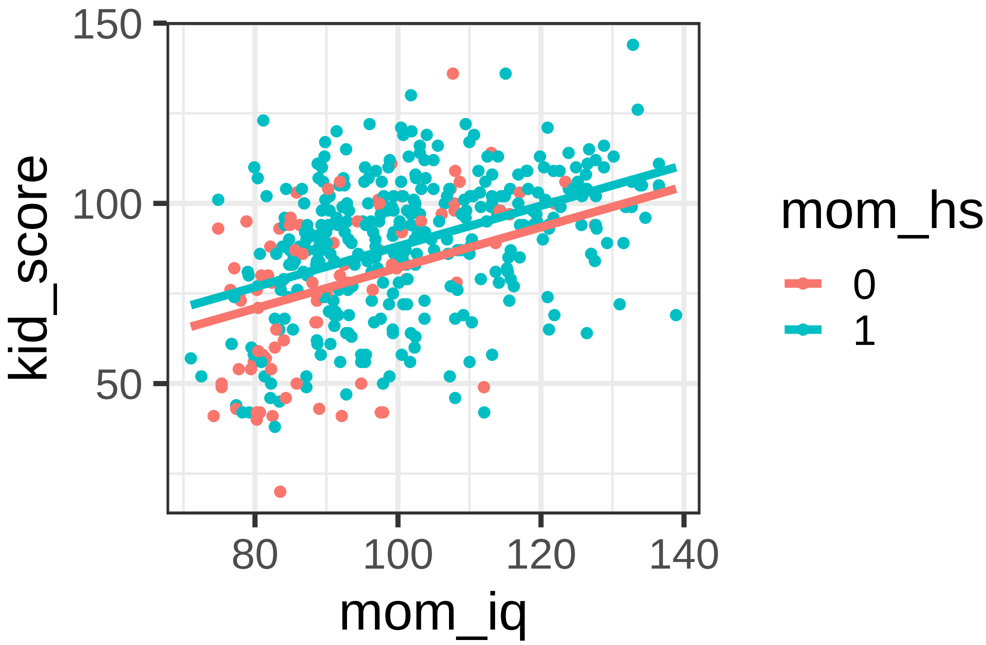
```
modelo4 <- lm(kid_score ~ c_mom_iq + mom_hs, kidiq)
summary(modelo4)
```

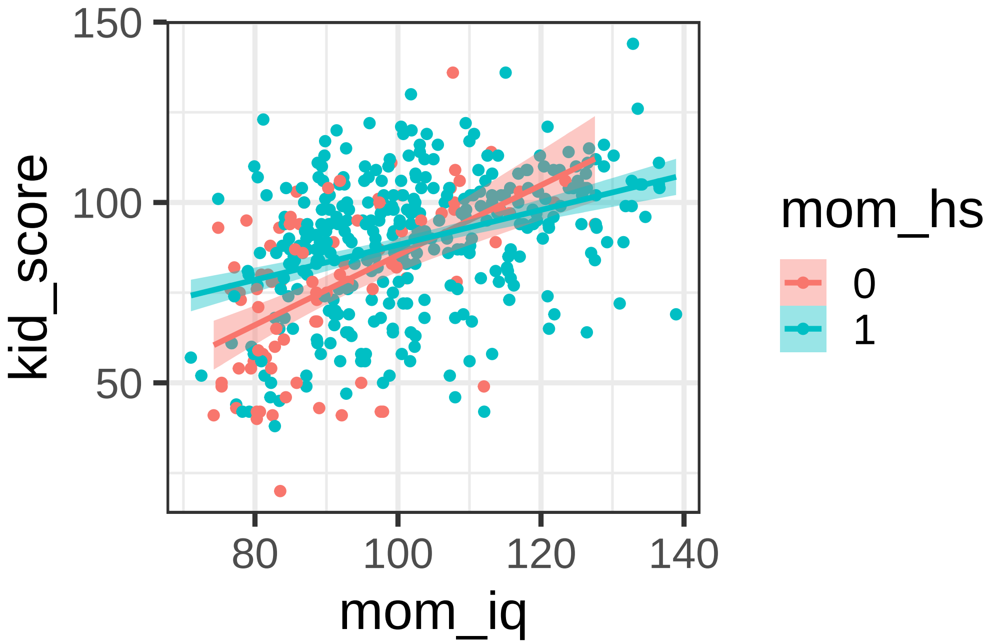
```

Call:
lm(formula = kid_score ~ c_mom_iq + mom_hs, data = kidiq)

Residuals:
    Min      1Q  Median      3Q     Max 
-52.873 -12.663   2.404  11.356  49.545 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 82.12214    1.94370  42.250  < 2e-16 ***
c_mom_iq     0.56391    0.06057   9.309  < 2e-16 ***
mom_hs1      5.95012    2.21181   2.690  0.00742 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 18.14 on 431 degrees of freedom
Multiple R-squared:  0.2141,	Adjusted R-squared:  0.2105
F-statistic: 58.72 on 2 and 431 DF,  p-value: < 2.2e-16
```

Como `mom_iq` está centrada, conseguimos interpretar o intercepto estimado como a média da nota quando a mãe tem um QI médio e não completou o ensino médio.

Vamos dar uma olhada em duas coisas: (i) as retas induzidas pelo modelo que inclui as duas covariáveis e (ii) como ficariam modelos ajustados separadamente para os grupos `mom_hs =0` e `mom_hs=1`.





A primeira figura não traz muitas surpresas; o modelo que ajustamos força o coeficiente angular a ser o mesmo, enquanto os interceptos diferem por β_{hs}. A segunda figura sugere que o coeficiente angular dos dois grupos pode ser bem diferente. Para avaliar essa possiblidade formalmente sem dividir os dados (isto é, ajustando um modelo só), vamos incluir um termo de *interação*:

$$\text{kid\_score} = \beta_0 + \beta_{\text{hs}}\text{mom\_hs} + \beta_{\text{iq}}\text{mom\_iq} + \beta_{\text{hs:iq}}\text{mom\_hs} \times \text{mom\_iq} + \varepsilon,$$

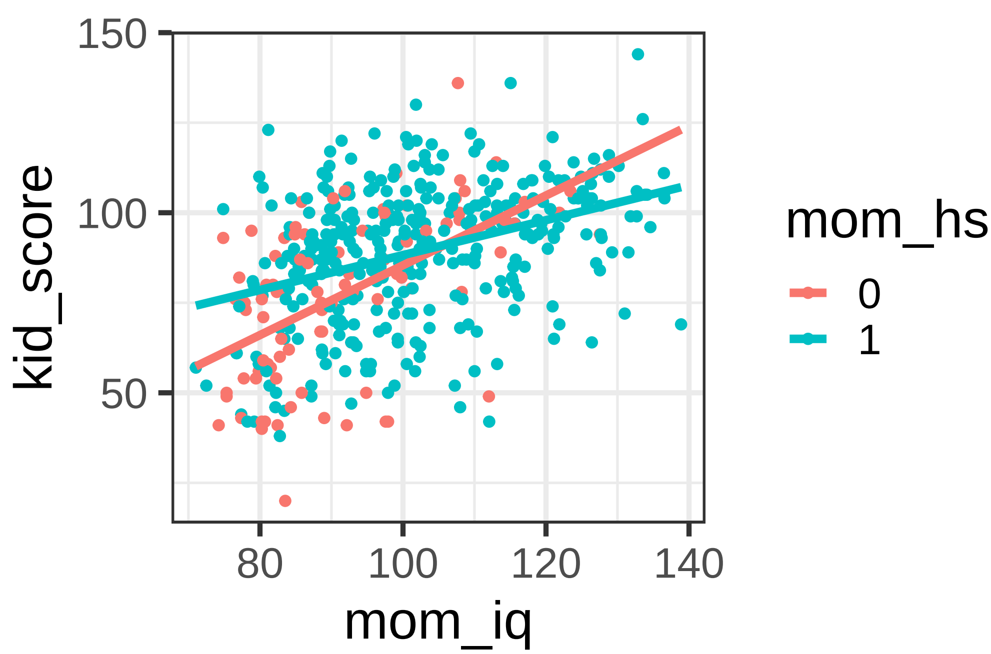
onde vamos entender $\beta_{\text{hs:iq}}$ como a *diferença* entre os coeficientes angulares dos dois grupos.

```
modelo5 <- lm(kid_score ~ c_mom_iq + mom_hs + mom_hs:c_mom_iq, kidiq)
summary(modelo5)
```

```
Call:
lm(formula = kid_score ~ c_mom_iq + mom_hs + mom_hs:c_mom_iq,
    data = kidiq)

Residuals:
    Min      1Q  Median      3Q     Max
-52.092 -11.332   2.066  11.663  43.880

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)       85.4069     2.2182  38.502  < 2e-16 ***
c_mom_iq           0.9689     0.1483   6.531 1.84e-10 ***
mom_hs1            2.8408     2.4267   1.171  0.24239
c_mom_iq:mom_hs1  -0.4843     0.1622  -2.985  0.00299 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.97 on 430 degrees of freedom
Multiple R-squared:  0.2301,    Adjusted R-squared:  0.2247
F-statistic: 42.84 on 3 and 430 DF,  p-value: < 2.2e-16
```

```
conf.line.m5 <- predict(modelo5, pred.grid)
preds.m5 <- data.frame(pred.grid, linpred = conf.line.m5)
preds.m5$mom_iq <- preds.m5$c_mom_iq + mean(kidiq$mom_iq)
ggplot(kidiq,
       aes(x = mom_iq, y = kid_score,
           colour = mom_hs, fill = mom_hs)) +
  geom_point() +
  geom_line(data = preds.m5,
            aes(x = mom_iq, y = linpred,
                colour = mom_hs),
            size = 1.5) +
  theme_bw(base_size = 20)
```



Exercícios de fixação

Tome $\boldsymbol{X}$ uma matriz real $n \times P$ e $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\}^T \in \mathbb{R}^n$ um vetor contendo os valores da variável dependente.

Nosso modelo (um pouco menos geral que o dado acima) é

$$E[Y_i] =: \mu_i(\boldsymbol{\beta}) = \tilde{\boldsymbol{X}}_i^T \boldsymbol{\beta},$$

onde $\boldsymbol{\beta} \in \mathbb{R}^{P+1}$ é o vetor de coeficientes e parâmetro de interesse e $\tilde{\boldsymbol{X}}$ é uma matriz obtida adicionando uma coluna de uns, $\boldsymbol{X_0} = \{1, \ldots, 1\}^T$, a $\boldsymbol{X}$. Para completar a especificação do modelo, vamos assumir que os erros em torno do preditor linear são normalmente distribuídos com variância comum:

$$\begin{aligned} Y_i &= \mu_i(\boldsymbol{\beta}) + \epsilon_i \\ \epsilon_i &\overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma^2), \end{aligned}$$

com $\sigma^2 \in \mathbb{R}_+$ desconhecida.

1. Escreva a log-verossimilhança e deduza seu gradiente e a sua derivada segunda (hessiana);
2. Com base nos cálculos do item anterior, mostre a forma do estimador de máxima verossimilhança para $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$;
3. Mostre que $\hat{\boldsymbol{\beta}}$ é não-viesado;
4. Considere um outro estimador <u>não-viesado</u> de $\boldsymbol{\beta}$: $\tilde{\boldsymbol{\beta}} = \boldsymbol{M}\boldsymbol{y}$, onde

   $$\boldsymbol{M} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{D},$$

   e $\boldsymbol{D}$ é uma matriz $P \times n$ cujas entradas são não-zero. Mostre que $\boldsymbol{R} := \text{Var}(\tilde{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}})$ é positiva-definida.

   Dica: Compute $E[\tilde{\boldsymbol{\beta}}]$ e considere o que deve valer para $\boldsymbol{D}$ sob a premissa de que $\tilde{\boldsymbol{\beta}}$ é não-viesado.

Comentário: Ao resolver o último item, você terá mostrado que o estimador de máxima verossimilhança (e também o estimador de mínimos quadrados) é o melhor estimador linear não-viesado (*best linear unbiased linear estimator, BLUE*). Em particular esta é a versão de Gauss[1] do famoso teorema de Gauss-Markov.

5. (*Desafio*) Considere o seguinte modelo alternativo:

   $$\boldsymbol{\varepsilon} \sim \text{Normal}(\boldsymbol{0}, \sigma^2\boldsymbol{V}),$$

   onde $\boldsymbol{V}$ é uma matriz positiva semi-definida **conhecida**.

[1] Carl Friedrich Gauss (1777-1855), matemático alemão conhecido como o Príncipe dos Matemáticos.

Deduza $\hat{\boldsymbol{\beta}}_{\text{EMV}}$ e sua distribuição amostral, além de $I(\boldsymbol{\beta})$. Discuta como este modelo viola as premissas de Gauss-Markov e quais os efeitos desta violação sobre as estimativas (são viesadas? De que ordem é o viés?). *Dica*: Ver seção 10.8 de ROS.

Referências

- Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. CRC press. (Cap 6)
- Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and other stories. Cambridge University Press.