Scoliosis Model

Caio Vallio

2025-12-17

Contexto: A escoliose idiopática do adolescente pode apresentar resposta clínica variável a intervenções conservadoras. **Objetivo:** Explorar associações entre características basais e melhora radiográfica em 6 meses. **Métodos:** Estudo observacional com análise exploratória e inferencial. O desfecho contínuo foi Δ (maior curva em 6 meses − baseline, em graus). O desfecho binário (MCID) foi melhora definida como $\Delta \leq -5°$. Ajustamos um modelo linear para Δ e um modelo logístico para o MCID, reportando estimativas com IC95% e verificações de adequação dos modelos.

Como reproduzir

- O arquivo de dados é lido de `data/modelagem final.xlsx` (aba `dados`).
- As análises dependem de pacotes R listados no chunk de setup; caso algum pacote esteja ausente, o documento interrompe a execução com uma mensagem explícita.

Objetivo e delineamento

Este é um relatório **exploratório e inferencial**: o objetivo é **estimar associações ajustadas** entre variáveis basais e a evolução radiográfica em 6 meses.

Definições de desfecho

- **Desfecho contínuo:** Δ = (maior curva em 6 meses − maior curva baseline), em graus. Valores mais negativos indicam maior melhora.

- **Desfecho binário (MCID):** melhora clínica definida como $\Delta \leq -5°$ (redução de pelo menos 5 graus).

Plano de análise estatística

- **Descrição da amostra**: estatísticas descritivas das variáveis basais e do desfecho.
- **Inferência (associações ajustadas)**:
  - **Modelo linear** para Δ (efeitos como betas, IC95%).
  - **Modelo logístico** para MCID (efeitos como odds ratios, IC95%).
- **Adequação dos modelos**: diagnósticos de assunções (colinearidade, resíduos, heteroscedasticidade quando aplicável) e medidas de qualidade de ajuste (por exemplo, R^2).
- **Forma funcional**: para preditores contínuos, assume-se relação aproximadamente linear.

Por se tratar de análise **exploratória**, os resultados devem ser interpretados com cautela, com ênfase em **magnitude/direção** e incerteza (IC95%).

```
## Setup (reprodutibilidade)
pkgs <- c(
    "tidyverse", "gtsummary",
    "readxl", "janitor",
    "performance", "qqplotr", "PupillometryR"
)
missing_pkgs <- pkgs[!vapply(pkgs, requireNamespace, logical(1), quietly = TRUE)]
if (length(missing_pkgs) > 0) {
    stop("Pacotes ausentes: ", paste(missing_pkgs, collapse = ", "))
}
invisible(lapply(pkgs, library, character.only = TRUE))

set.seed(123)
theme_set(theme_minimal())


# read data
df_raw <- readxl::read_excel("data/modelagem final.xlsx", sheet = "dados")

# clean names
df <- df_raw |> janitor::clean_names()

# clean escoliometro
df <- df |>
    mutate(
        escoliometro = pmax(
            escoliomertro_cervical,
            escoliometro_torarica,
```

```
            escoliometro_lombar,
            na.rm = TRUE
        )
    ) |>
    mutate(
        regiao = case_when(
            escoliometro == escoliomertro_cervical ~ "cervical",
            escoliometro == escoliometro_torarica ~ "toracica",
            escoliometro == escoliometro_lombar ~ "lombar",
            TRUE ~ NA_character_
        )
    ) |>
    select(
        -escoliomertro_cervical,
        -escoliometro_torarica,
        -escoliometro_lombar
    )

# definicao do desfecho (ver secao acima)
df <- df |>
    mutate(delta = maior_curva_6_meses - maior_curva_baseline) |>
    mutate(delta_cat = if_else(delta <= -5, 1, 0)) |>
    mutate(
        delta_cat_f = factor(
            delta_cat,
            levels = c(0, 1),
            labels = c("Sem melhora (MCID)", "Melhora (MCID)")
        )
    )

# casting de variaveis (garantir tipos adequados no ajuste)
df <- df |>
    mutate(
        idade = as.double(idade),
        lenke = as.factor(lenke),
        risser = as.factor(risser),
        sexo = as.factor(sexo),
        regiao = as.factor(regiao),
        cifose_categ = as.factor(cifose_categ)
    )
```

Dados faltantes e tamanho amostral

```
missing_tbl <- df |>
    summarise(across(everything(), ~ sum(is.na(.)))) |>
    pivot_longer(everything(), names_to = "variavel", values_to = "n_missing") |>
    mutate(pct_missing = n_missing / nrow(df)) |>
    arrange(desc(pct_missing))

missing_tbl |>
    mutate(pct_missing = scales::percent(pct_missing, accuracy = 0.1)) |>
    print(n = 50)
```

```
# A tibble: 21 x 3
   variavel             n_missing pct_missing
   <chr>                    <int> <chr>
 1 id                           0 0.0%
 2 idade                        0 0.0%
 3 sexo                         0 0.0%
 4 altura                       0 0.0%
 5 peso                         0 0.0%
 6 imc                          0 0.0%
 7 regiao                       0 0.0%
 8 maior_curva_baseline         0 0.0%
 9 maior_curva_6_meses          0 0.0%
10 delta                        0 0.0%
11 lenke                        0 0.0%
12 risser                       0 0.0%
13 cifose_toracica              0 0.0%
14 lordose_lombar               0 0.0%
15 cifose_categ                 0 0.0%
16 inclinacao                   0 0.0%
17 dif_colete                   0 0.0%
18 correcao_colete              0 0.0%
19 escoliometro                 0 0.0%
20 delta_cat                    0 0.0%
21 delta_cat_f                  0 0.0%
```

Observação: os modelos abaixo usam **análise por caso completo** (listwise deletion), que é o comportamento padrão do `glm()`/`lm()` quando há dados faltantes nas variáveis do modelo.

Codificação de variáveis categóricas (referências)

As variáveis categóricas entram como fatores. A **categoria de referência** (baseline) é o **primeiro nível** do fator (conforme `levels()`), a menos que seja explicitamente reordenado.

```
categoricas <- c("sexo", "regiao", "lenke", "risser", "cifose_categ")
map_dfr(categoricas, function(v) {
    tibble(
        variavel = v,
        niveis = paste(levels(df[[v]]), collapse = " | ")
    )
}) |>
    print(n = Inf)
```

```
# A tibble: 5 x 2
  variavel     niveis
  <chr>        <chr>
1 sexo         feminino | masculino
2 regiao       lombar | toracica
3 lenke        1 | 2 | 3 | 4 | 5 | 6
4 risser       0 | 1 | 2 | 3 | 4
5 cifose_categ hipercifose | hipocifose | normal
```

Análise descritiva

Variáveis de entrada do modelo

Variáveis coletadas na linha de base.

```
df |>
    gtsummary::tbl_summary(
        include = c(
            idade,
            altura,
            peso,
            imc,
            cifose_toracica,
            lordose_lombar,
            dif_colete,
            correcao_colete,
```

```
            escoliometro,
            inclinacao,
            cifose_categ,
            sexo,
            regiao,
            lenke,
            risser
        ),
        type = list(
            idade ~ "continuous"
        ),
        statistic = list(
            gtsummary::all_continuous() ~ "{mean} ({sd})",
            gtsummary::all_categorical() ~ "{n} ({p}%)"
        )
    )
```

Desfecho principal

- Maior curva na linha de base e em 6 meses.
- Delta: maior curva 6 meses - maior curva baseline.
- Delta categórica: delta melhor em 5 graus ou mais (MCID).

```
df |>
    gtsummary::tbl_summary(
        include = c(
            maior_curva_baseline,
            maior_curva_6_meses,
            delta,
            delta_cat_f
        ),
        type = list(
            delta_cat_f ~ "categorical"
        ),
        statistic = list(
            gtsummary::all_continuous() ~ "{mean} ({sd})",
            gtsummary::all_categorical() ~ "{n} ({p}%)"
        )
    )
```

Characteristic	**N = 152**[1]
idade	13.07 (1.48)
altura	1.58 (0.08)
peso	48 (9)
imc	19.05 (2.91)
cifose_toracica	22 (12)
lordose_lombar	54 (12)
dif_colete	-17.6 (6.2)
correcao_colete	49 (19)
escoliometro	13.8 (4.3)
inclinacao	
flexivel	96 (63%)
rigido	56 (37%)
cifose_categ	
hipercifose	8 (5.3%)
hipocifose	17 (11%)
normal	127 (84%)
sexo	
feminino	130 (86%)
masculino	22 (14%)
regiao	
lombar	80 (53%)
toracica	72 (47%)
lenke	
1	8 (5.3%)
2	11 (7.2%)
3	32 (21%)
4	13 (8.6%)
5	16 (11%)
6	72 (47%)
risser	
0	32 (21%)
1	34 (22%)
2	45 (30%)
3	24 (16%)
4	17 (11%)

[1]Mean (SD); n (%)

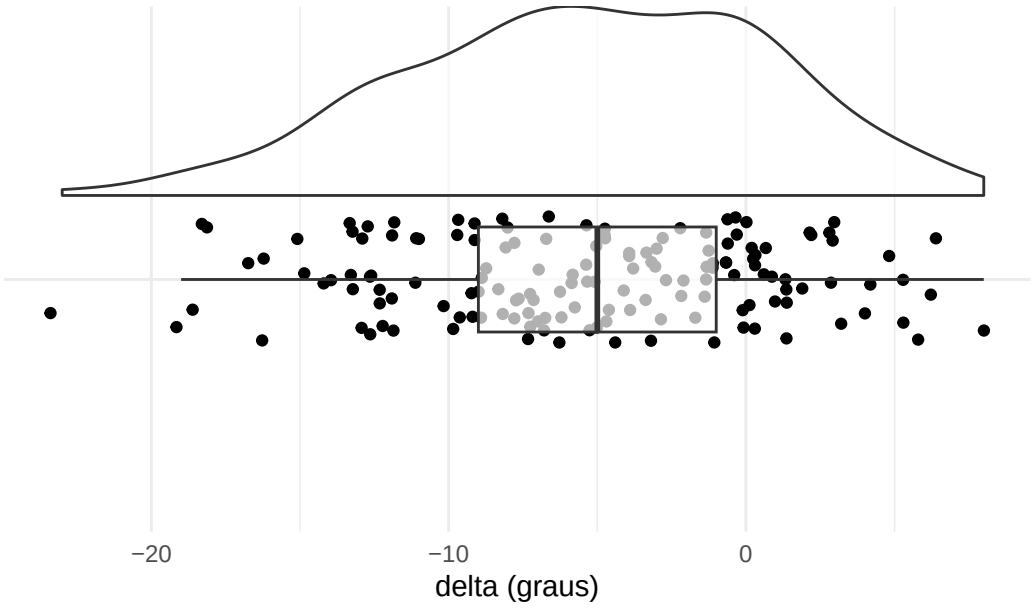
Characteristic	N = 152[1]
maior_curva_baseline	36.6 (5.9)
maior_curva_6_meses	31 (9)
delta	-5.4 (6.1)
delta_cat_f	
Sem melhora (MCID)	68 (45%)
Melhora (MCID)	84 (55%)

[1]Mean (SD); n (%)

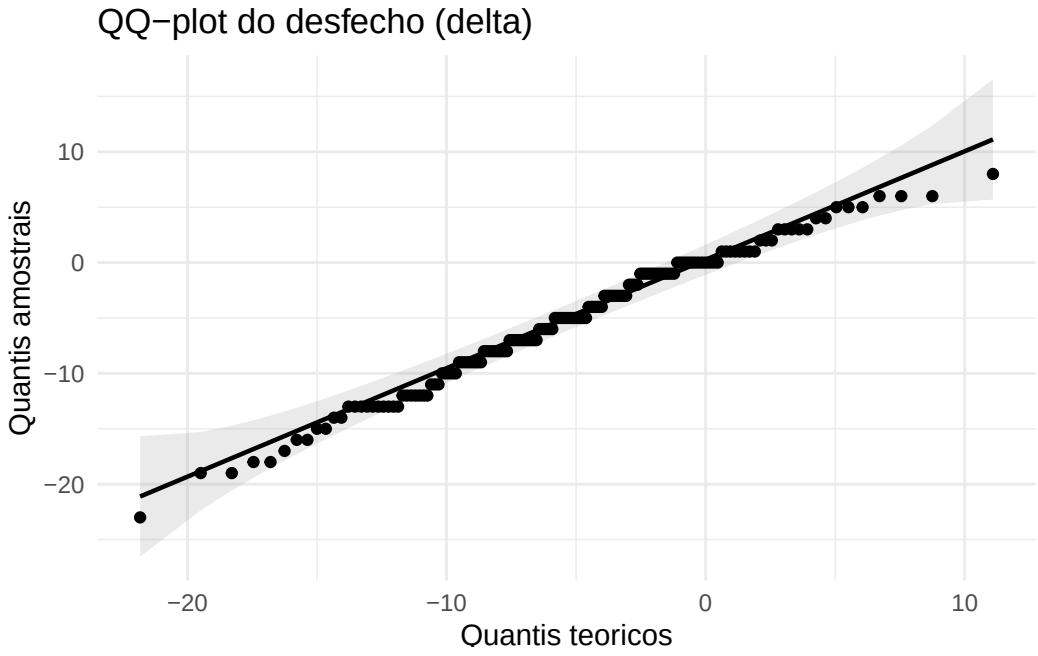
Distribuição do desfecho principal

```
df |>
    ggplot(aes(x = "", y = delta)) +
    PupillometryR::geom_flat_violin(
        position = position_nudge(x = .2),
        # fill = "steelblue",
        alpha = 0.7
    ) +
    labs(
        title = "Distribuicao do desfecho (delta)",
        x = "",
        y = "delta (graus)"
    ) +
    geom_point(position = position_jitter(w = .15)) +
    geom_boxplot(
        width = .25,
        alpha = 0.7,
        outlier.shape = NA
    ) +
    coord_flip() +
    theme(
        axis.title.y = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank()
    )
```

Distribuicao do desfecho (delta)




```
df |>
    ggplot(aes(sample = delta)) +
    qqplotr::stat_qq_band(alpha = 0.2) +
    qqplotr::stat_qq_line() +
    qqplotr::stat_qq_point() +
    labs(
        title = "QQ-plot do desfecho (delta)",
        x = "Quantis teoricos",
        y = "Quantis amostrais"
    )
```



Modelagem

Modelo de regressão logística

Este modelo estima associações ajustadas com a **chance de melhora (MCID)**, definida como $\Delta \leq -5°$.

```
model_bin <- glm(
    delta_cat ~
        idade +
        altura +
        peso +
        # imc +
        cifose_toracica +
        lordose_lombar +
        dif_colete +
        correcao_colete +
        escoliometro +
        inclinacao +
        cifose_categ +
        sexo +
        regiao +
```

```
        lenke +
        risser,
    data = df, family = "binomial"
)
```

Parâmetros do modelo

Parâmetros do modelo de regressão logística. Resposta em *Odds Ratio*.

```
model_bin |>
    gtsummary::tbl_regression(exponentiate = TRUE, conf.level = 0.95) |>
    # gtsummary::add_global_p() |>
    gtsummary::bold_p() |>
    gtsummary::bold_labels() |>
    gtsummary::italicize_levels()
```

Verificações de adequação do modelo

Colinearidade

Verificação de colinearidade entre as variáveis independentes (*Variance Inflation Factor*). Valores maiores que 10 indicam colinearidade entre as variáveis.

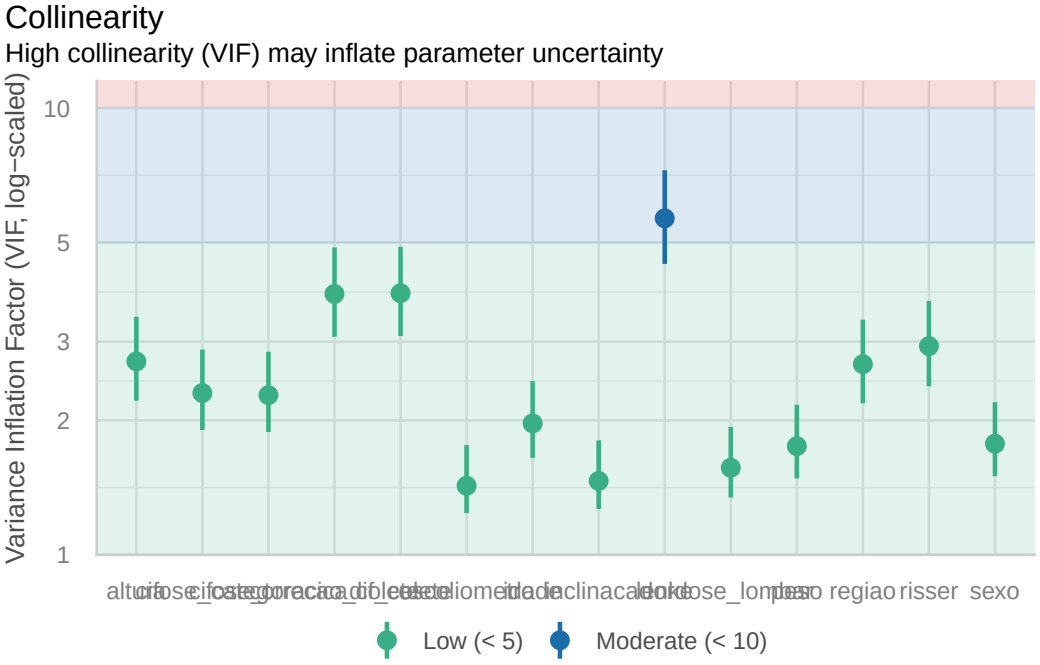
```
performance::check_collinearity(model_bin) |>
    arrange(desc(VIF)) |>
    select(Term, VIF, VIF_CI_low, VIF_CI_high) |>
    performance::print_html()
```

```
performance::check_collinearity(model_bin) |> plot()
```

Characteristic	OR	95% CI	p-value
idade	1.19	0.79, 1.85	0.4
altura	0.65	0.00, 3,725	>0.9
peso	1.01	0.94, 1.07	0.8
cifose_toracica	1.01	0.95, 1.07	0.7
lordose_lombar	0.98	0.93, 1.02	0.3
dif_colete	0.85	0.72, 1.01	0.064
correcao_colete	1.02	0.96, 1.08	0.5
escoliometro	1.12	1.00, 1.28	0.074
inclinacao			
flexivel	—	—	
rigido	0.23	0.07, 0.64	**0.007**
cifose_categ			
hipercifose	—	—	
hipocifose	4.92	0.14, 185	0.4
normal	16.7	1.33, 326	**0.042**
sexo			
feminino	—	—	
masculino	1.46	0.29, 7.93	0.6
regiao			
lombar	—	—	
toracica	1.32	0.33, 5.54	0.7
lenke			
1	—	—	
2	0.42	0.02, 7.63	0.5
3	0.05	0.00, 0.41	**0.009**
4	0.17	0.01, 1.57	0.13
5	2.21	0.15, 30.2	0.6
6	0.36	0.04, 2.60	0.3
risser			
0	—	—	
1	3.94	1.00, 16.7	0.054
2	1.93	0.51, 7.54	0.3
3	0.69	0.13, 3.55	0.7
4	1.60	0.26, 10.5	0.6

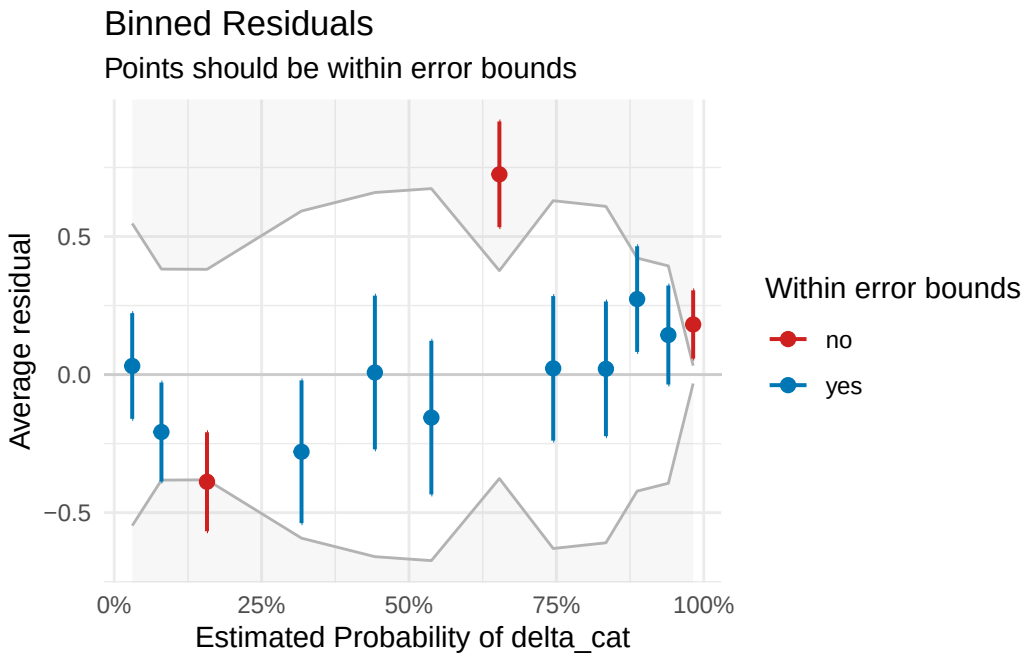
Abbreviations: CI = Confidence Interval, OR = Odds Ratio

Term	VIF	VIF_CI_low	VIF_CI_high
lenke	5.66	4.48	7.26
dif_colete	3.85	3.09	4.89
correcao_colete	3.84	3.08	4.87
risser	2.93	2.38	3.70
altura	2.71	2.21	3.41
regiao	2.67	2.18	3.36
cifose_categ	2.30	1.90	2.88
cifose_toracica	2.28	1.88	2.85
idade	1.97	1.65	2.45
sexo	1.77	1.50	2.20
peso	1.75	1.48	2.17
lordose_lombar	1.57	1.34	1.93
inclinacao	1.46	1.27	1.80
escoliometro	1.43	1.24	1.76



Resíduos binarizados

```
performance::binned_residuals(model_bin) |> plot()
```

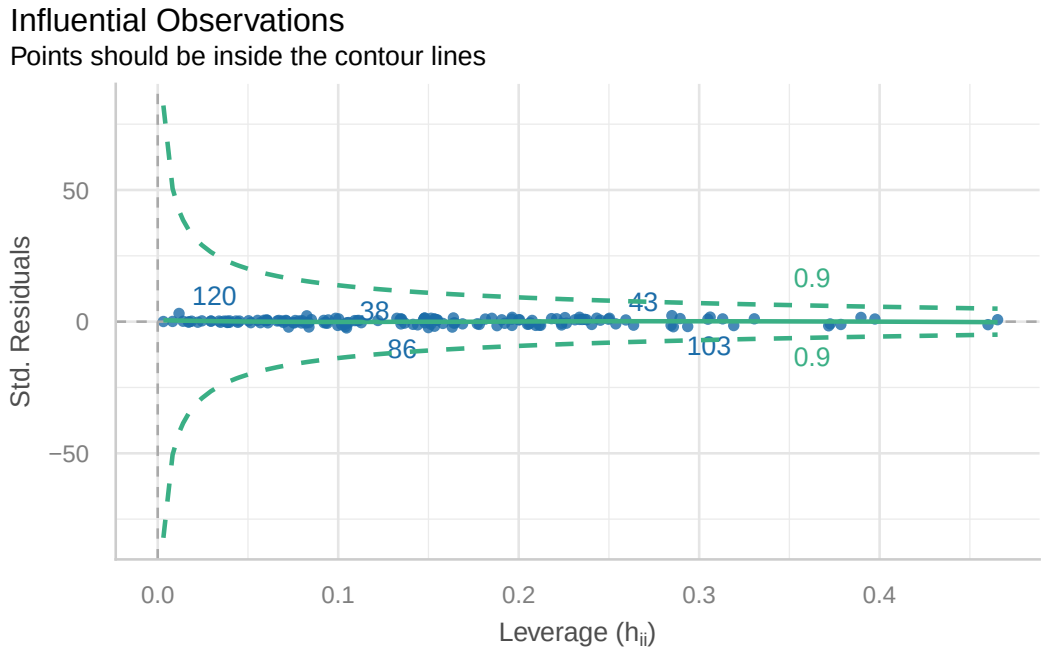


Outliers

```
performance::check_outliers(model_bin)
```

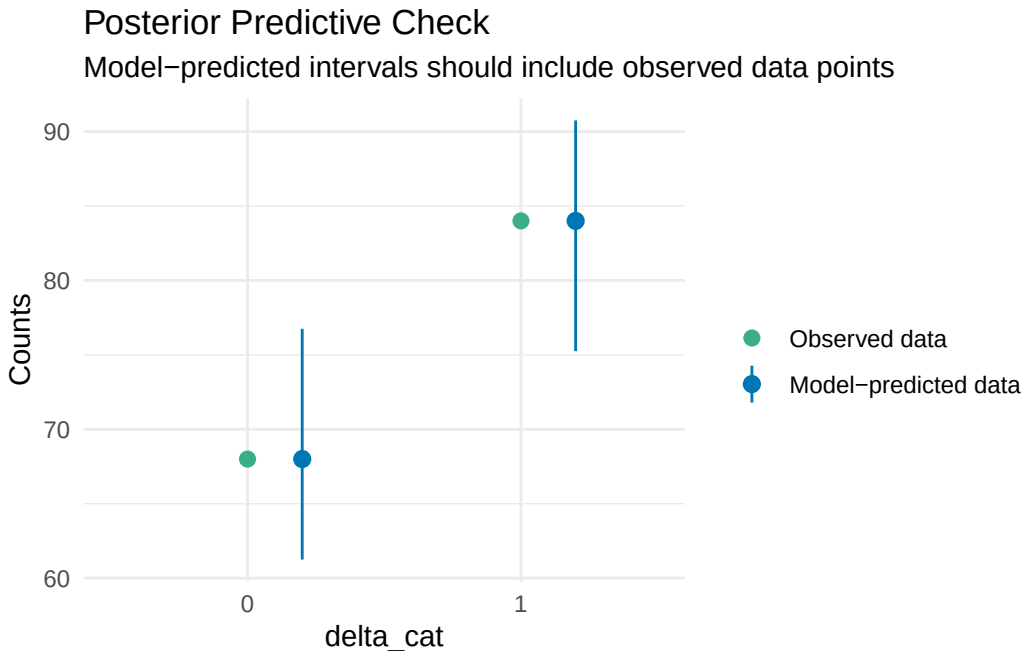
```
OK: No outliers detected.
- Based on the following method and threshold: cook (0.922).
- For variable: (Whole model)
```

```
performance::check_outliers(model_bin) |> plot()
```



Valores ajustados (diagnóstico)

```
performance::check_predictions(model_bin) |> plot()
```

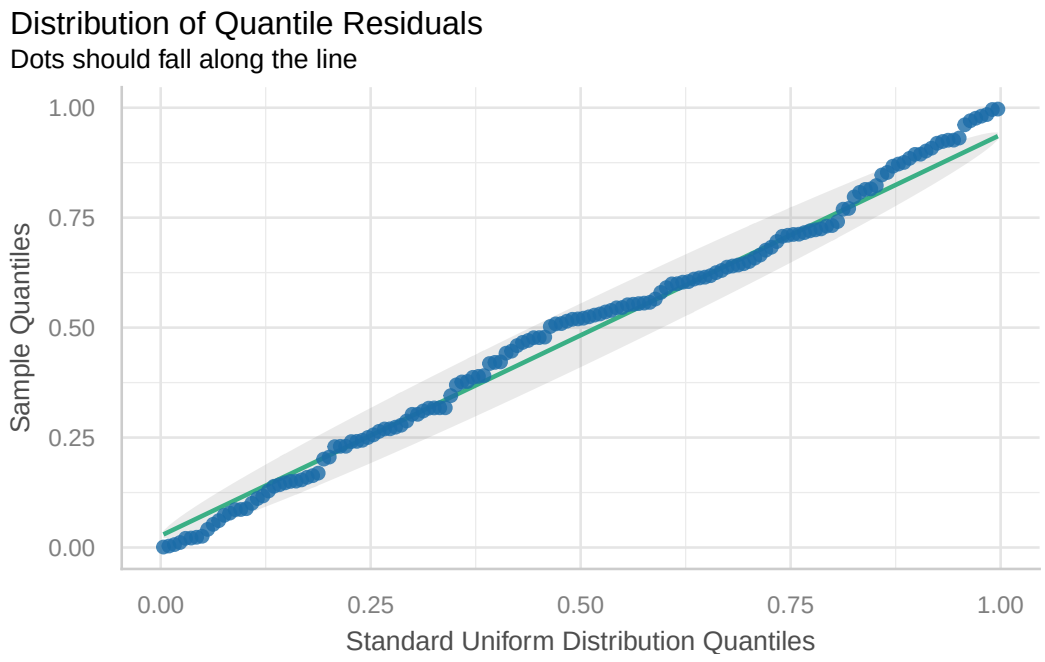


Resíduos simulados

```
performance::check_residuals(model_bin)
```

```
OK: Simulated residuals appear as uniformly distributed (p = 0.435).
```

```
performance::check_residuals(model_bin) |> plot()
```



Qualidade do ajuste do modelo

Teste de Hosmer-Lemeshow para avaliar a qualidade do ajuste de modelos binomiais. P-valor < 0.05 indica que o modelo não ajusta bem os dados.

```
performance::performance_hosmer(model_bin)
```

```
# Hosmer-Lemeshow Goodness-of-Fit Test

  Chi-squared: 7.009
           df: 8
      p-value: 0.536
```

Coeficiente de determinação (Tjur)

Coeficiente de determinação de Tjur (R^2_{Tjur}). Teste específico para modelos binomiais.

```
performance::r2(model_bin)
```

```
# R2 for Logistic Regression
  Tjur's R2: 0.448
```

Tamanho amostral efetivo (casos completos)

```
mf_bin <- model.frame(model_bin)
bin_n <- nrow(mf_bin)
bin_events <- sum(mf_bin[[1]] == 1, na.rm = TRUE)
bin_nonevents <- sum(mf_bin[[1]] == 0, na.rm = TRUE)
tibble(
    n_modelo = bin_n,
    eventos_mcid = bin_events,
    nao_eventos = bin_nonevents
)
```

```
# A tibble: 1 x 3
  n_modelo eventos_mcid nao_eventos
     <int>        <int>       <int>
1      152           84          68
```

Modelo de regressão linear

Este modelo estima associações ajustadas com o desfecho contínuo Δ (maior curva em 6 meses − baseline, em graus).

```
model_lin <- lm(
    delta ~
        idade +
        altura +
        peso +
        # imc +
        cifose_toracica +
        lordose_lombar +
        dif_colete +
```

```
        correcao_colete +
        escoliometro +
        inclinacao +
        cifose_categ +
        sexo +
        regiao +
        lenke +
        risser,
    data = df
)
```

Parâmetros do modelo

```
model_lin |>
    gtsummary::tbl_regression(conf.level = 0.95) |>
    # gtsummary::add_global_p() |>
    gtsummary::bold_p() |>
    gtsummary::bold_labels() |>
    gtsummary::italicize_levels()
```

Verificações de adequação do modelo

Colinearidade

Verificação de colinearidade entre as variáveis independentes (*Variance Inflation Factor*). Valores maiores que 10 indicam colinearidade entre as variáveis.

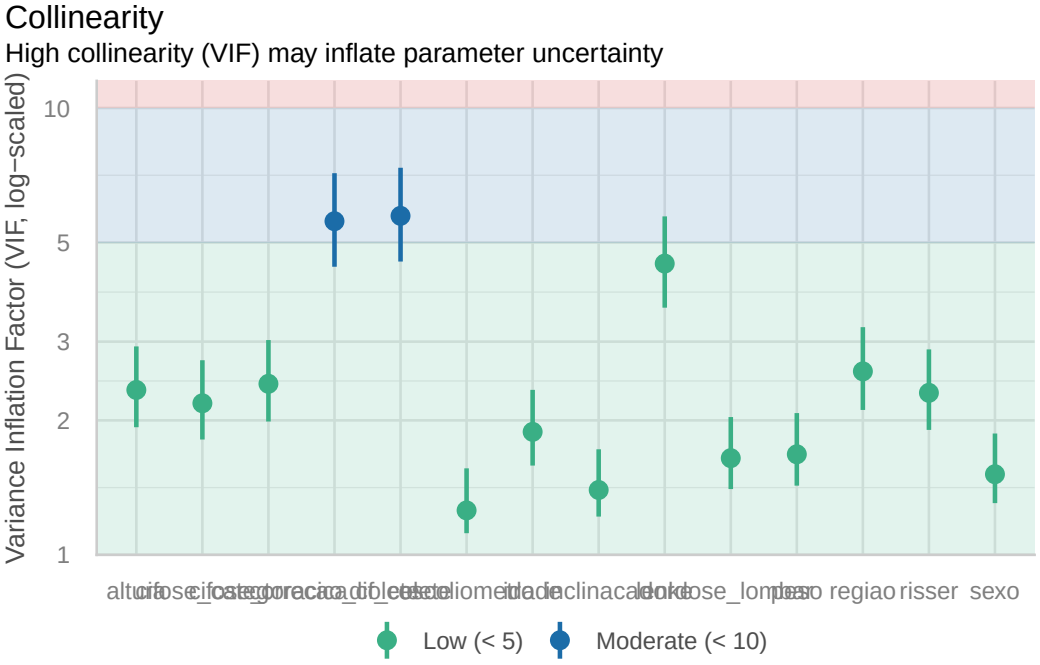
```
performance::check_collinearity(model_lin) |>
    arrange(desc(VIF)) |>
    select(Term, VIF, VIF_CI_low, VIF_CI_high) |>
    performance::print_html()
```

```
performance::check_collinearity(model_lin) |> plot()
```

Characteristic	Beta	95% CI	p-value
idade	-0.54	-1.3, 0.20	0.2
altura	-3.0	-19, 13	0.7
peso	0.02	-0.10, 0.13	0.8
cifose_toracica	-0.03	-0.14, 0.07	0.5
lordose_lombar	0.01	-0.08, 0.09	0.9
dif_colete	0.34	0.03, 0.65	**0.032**
correcao_colete	0.01	-0.09, 0.11	0.8
escoliometro	0.00	-0.21, 0.21	>0.9
inclinacao			
flexivel	—	—	
rigido	2.6	0.63, 4.5	**0.010**
cifose_categ			
hipercifose	—	—	
hipocifose	-3.7	-9.5, 2.1	0.2
normal	-4.9	-9.3, -0.49	**0.030**
sexo			
feminino	—	—	
masculino	0.81	-2.0, 3.6	0.6
regiao			
lombar	—	—	
toracica	0.24	-2.3, 2.8	0.9
lenke			
1	—	—	
2	1.0	-3.8, 5.9	0.7
3	6.3	2.2, 10	**0.003**
4	5.0	0.46, 9.6	**0.031**
5	-1.6	-6.4, 3.2	0.5
6	2.9	-1.1, 6.9	0.2
risser			
0	—	—	
1	-2.0	-4.6, 0.57	0.12
2	-1.3	-3.8, 1.1	0.3
3	0.23	-2.7, 3.1	0.9
4	0.58	-2.9, 4.1	0.7

Abbreviation: CI = Confidence Interval

Term	VIF	VIF_CI_low	VIF_CI_high
dif_colete	5.74	4.53	7.35
correcao_colete	5.58	4.41	7.15
lenke	4.49	3.57	5.72
regiao	2.57	2.11	3.23
cifose_toracica	2.41	1.99	3.02
altura	2.34	1.93	2.93
risser	2.30	1.90	2.88
cifose_categ	2.18	1.81	2.73
idade	1.88	1.58	2.34
peso	1.68	1.43	2.08
lordose_lombar	1.65	1.40	2.03
sexo	1.52	1.31	1.87
inclinacao	1.40	1.22	1.72
escoliometro	1.26	1.12	1.56



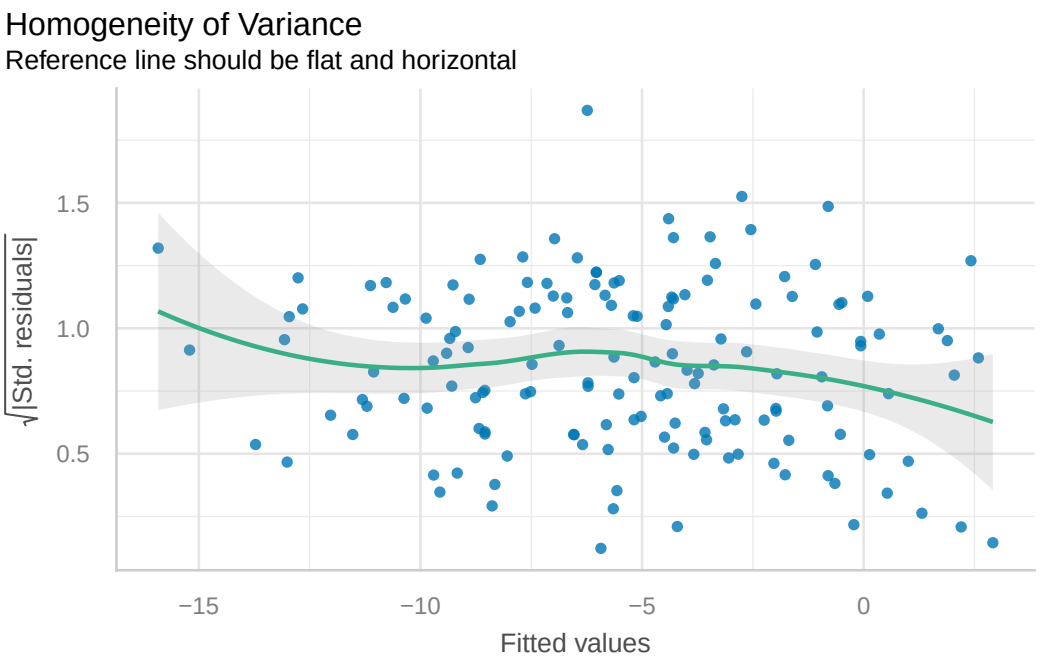
Heteroscedasticidade

Verificação de heteroscedasticidade entre as variáveis independentes.

```
performance::check_heteroscedasticity(model_lin)
```

```
OK: Error variance appears to be homoscedastic (p = 0.695).
```

```
performance::check_heteroscedasticity(model_lin) |> plot()
```

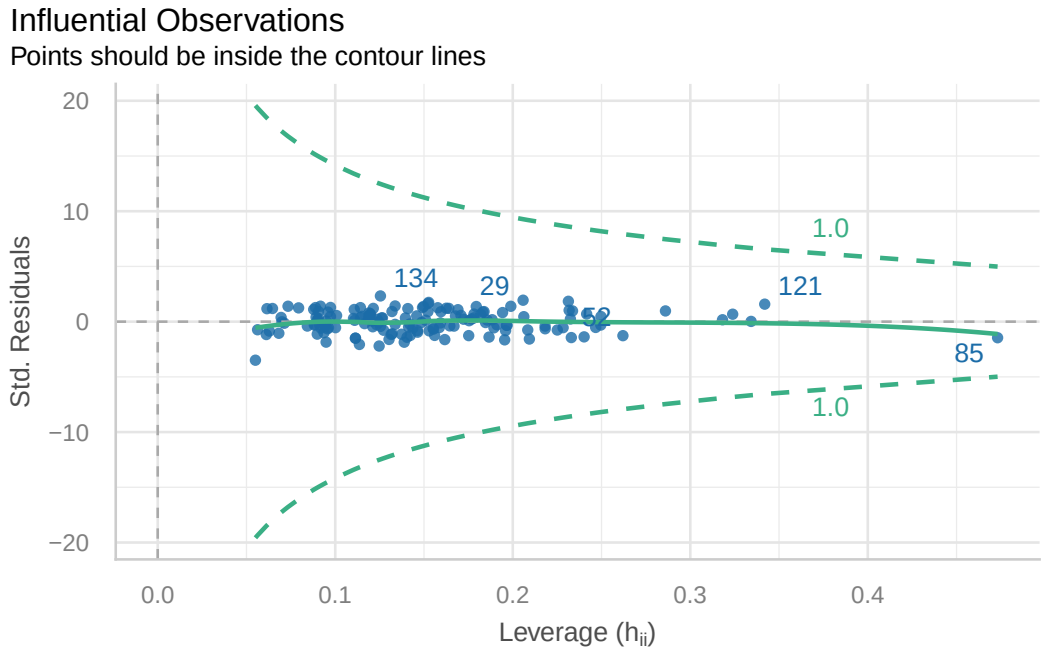


Outliers

```
performance::check_outliers(model_lin)
```

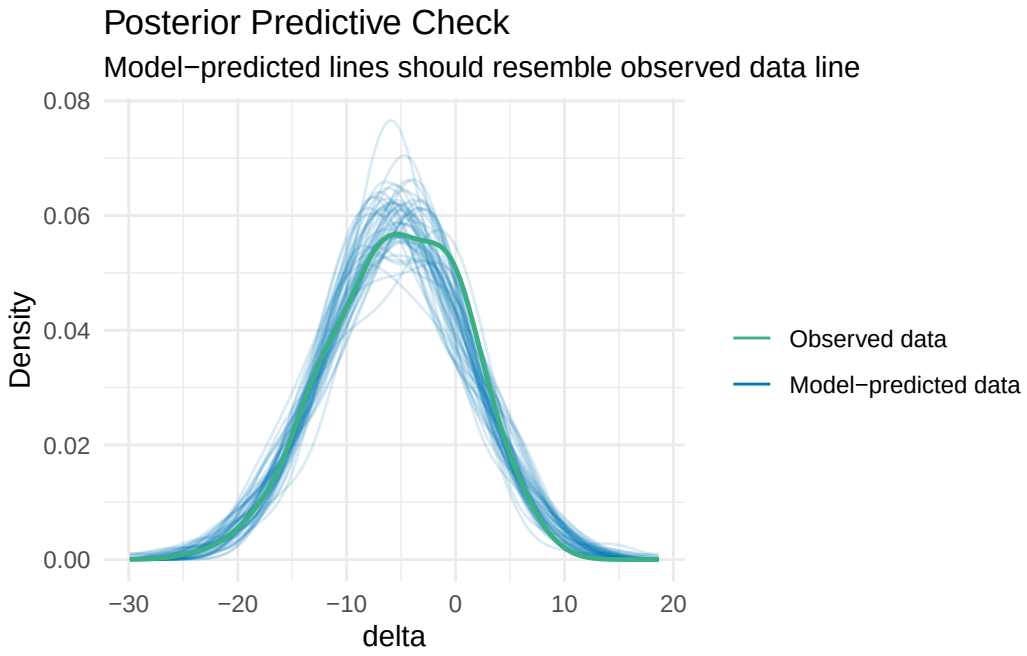
```
OK: No outliers detected.
- Based on the following method and threshold: cook (0.931).
- For variable: (Whole model)
```

```
performance::check_outliers(model_lin) |> plot()
```



Valores ajustados (diagnóstico)

```
performance::check_predictions(model_lin) |> plot()
```

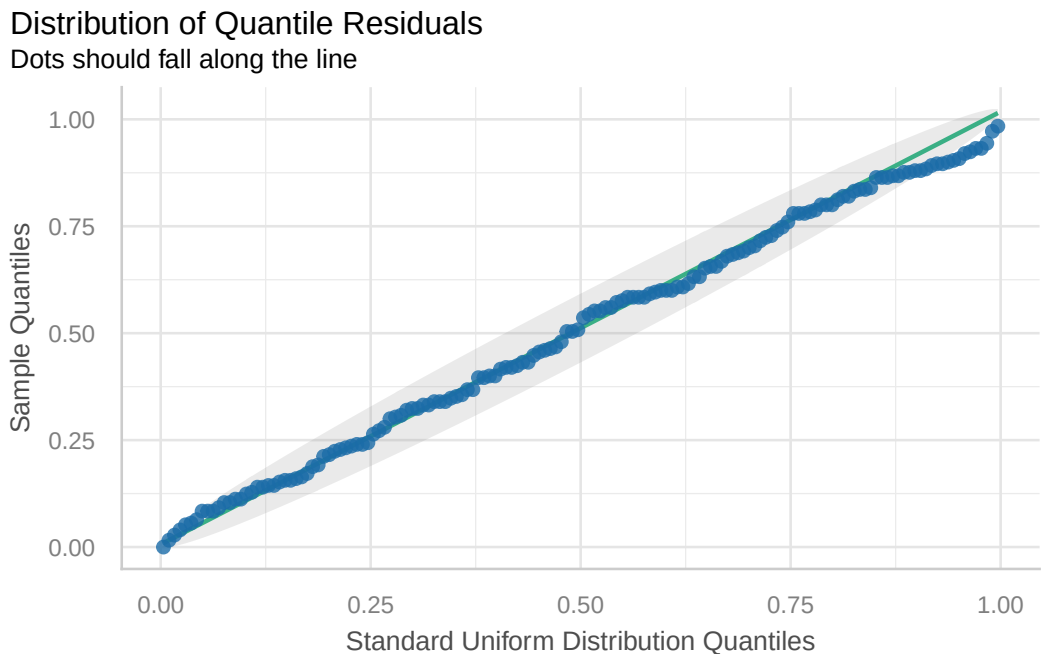


Resíduos simulados

```
performance::check_residuals(model_lin)
```

```
OK: Simulated residuals appear as uniformly distributed (p = 0.871).
```

```
performance::check_residuals(model_lin) |> plot()
```



Qualidade do ajuste do modelo

Coeficiente de determinação

R^2.

```
# Qualidade do ajuste
performance::r2(model_lin)
```

```
# R2 for Linear Regression
       R2: 0.436
  adj. R2: 0.340
```

Tamanho amostral efetivo (casos completos)

```
tibble(n_modelo = nobs(model_lin))
```

```
# A tibble: 1 x 1
  n_modelo
     <int>
1      152
```