Análise de regressão

Tiago M. Magalhães

Departamento de Estatística - ICE-UFJF

Juiz de Fora, 11 de março de 2024

Roteiro

Roteiro

Motivação

Os primeiros conceitos de regressão foram propostos por Galton (1889), aplicados na Antropometria, em que o objetivo era equacionar as relações de dependência entre a altura dos pais (**variável explicativa**, variável preditora ou covariável)

Motivação

Os primeiros conceitos de regressão foram propostos por Galton (1889), aplicados na Antropometria, em que o objetivo era equacionar as relações de dependência entre a altura dos pais (**variável explicativa**, variável preditora ou covariável) e a altura dos filhos (variável de interesse, **variável resposta** ou desfecho).

Motivação

Os primeiros conceitos de regressão foram propostos por Galton (1889), aplicados na Antropometria, em que o objetivo era equacionar as relações de dependência entre a altura dos pais (**variável explicativa**, variável preditora ou covariável) e a altura dos filhos (variável de interesse, **variável resposta** ou desfecho).

Motivação

Porém, esses conceitos puderam ser aplicados em qualquer contexto, como:

- na necessidade de uma empresa em analisar os fatores que podem interferir nas vendas;

Motivação

Porém, esses conceitos puderam ser aplicados em qualquer contexto, como:

- na necessidade de uma empresa em analisar os fatores que podem interferir nas vendas;
- na predição de mortalidade de uma pessoa ao contrair uma doença, baseada nas suas características, como idade e doenças pré existentes.

Motivação

Porém, esses conceitos puderam ser aplicados em qualquer contexto, como:

- na necessidade de uma empresa em analisar os fatores que podem interferir nas vendas;
- na predição de mortalidade de uma pessoa ao contrair uma doença, baseada nas suas características, como idade e doenças pré existentes.

Motivação

- caracterização de um cliente em bom ou mau pagador, baseada no máximo de dias de atraso nos últimos 6 meses, se fez ou não saque no cartão e no tempo que o cliente tem o cartão;

Motivação

- caracterização de um cliente em bom ou mau pagador, baseada no máximo de dias de atraso nos últimos 6 meses, se fez ou não saque no cartão e no tempo que o cliente tem o cartão;
- distância de uma galáxia até a Terra, baseada no valor dos tons de cinza de cada pixel de uma foto da galáxia em observação.

Motivação

- caracterização de um cliente em bom ou mau pagador, baseada no máximo de dias de atraso nos últimos 6 meses, se fez ou não saque no cartão e no tempo que o cliente tem o cartão;
- distância de uma galáxia até a Terra, baseada no valor dos tons de cinza de cada pixel de uma foto da galáxia em observação.

Motivação

A **análise de regressão** é uma técnica estatística para investigar e modelar a relação entre as variáveis.

Na verdade, a análise de regressão pode ser a técnica estatística mais utilizada (Montgomery et al., 2021).

Motivação

A **análise de regressão** é uma técnica estatística para investigar e modelar a relação entre as variáveis.

Na verdade, a análise de regressão pode ser a técnica estatística mais utilizada (Montgomery et al., 2021).

Motivação

Os modelos de regressão ganharam ainda mais notoriedade com a popularização do Aprendizado máquina (*machine learning*), em que o foco é preditivo, isto é, a partir de valores de entrada (covariáveis), se deseja prever o resultado esperado (variável resposta).

Motivação

Os modelos de regressão ganharam ainda mais notoriedade com a popularização do Aprendizado máquina (*machine learning*), em que o foco é preditivo, isto é, a partir de valores de entrada (covariáveis), se deseja prever o resultado esperado (variável resposta).

Roteiro

Modelo de regressão linear simples

Como nós vimos, existem fenômenos (variável resposta, Y) que podem ser descritos por uma variável preditora (x), isto é feito da seguinte forma:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

Modelo de regressão linear simples

Como nós vimos, existem fenômenos (variável resposta, Y) que podem ser descritos por uma variável preditora (x), isto é feito da seguinte forma:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

em que $\mu = \beta_1 + \beta_2 x$ e σ^2 são, respectivamente, a média e variância da distribuição de Y, x é um valor conhecido, β_1, β_2 e σ^2 são parâmetros desconhecidos.

Modelo de regressão linear simples

Como nós vimos, existem fenômenos (variável resposta, Y) que podem ser descritos por uma variável preditora (x), isto é feito da seguinte forma:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

em que $\mu = \beta_1 + \beta_2 x$ e σ^2 são, respectivamente, a média e variância da distribuição de Y, x é um valor conhecido, β_1, β_2 e σ^2 são parâmetros desconhecidos. Além disso, β_1 e β_2 são chamados de coeficientes de regressão.

Modelo de regressão linear simples

Como nós vimos, existem fenômenos (variável resposta, Y) que podem ser descritos por uma variável preditora (x), isto é feito da seguinte forma:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

em que $\mu = \beta_1 + \beta_2 x$ e σ^2 são, respectivamente, a média e variância da distribuição de Y, x é um valor conhecido, β_1, β_2 e σ^2 são parâmetros desconhecidos. Além disso, β_1 e β_2 são chamados de coeficientes de regressão.

Modelo de regressão linear simples

É conveniente escrever a relação entre as variáveis resposta e preditora da seguinte forma:

$$Y = \beta_1 + \beta_2 x + \varepsilon,$$

em que $\varepsilon \sim \mathcal{D}(0, \sigma^2)$,

Modelo de regressão linear simples

É conveniente escrever a relação entre as variáveis resposta e preditora da seguinte forma:

$$Y = \beta_1 + \beta_2 x + \varepsilon,$$

em que $\varepsilon \sim \mathcal{D}(0, \sigma^2)$, sendo ε denominado de erro.

Modelo de regressão linear simples

É conveniente escrever a relação entre as variáveis resposta e preditora da seguinte forma:

$$Y = \beta_1 + \beta_2 x + \varepsilon,$$

em que $\varepsilon \sim \mathcal{D}(0, \sigma^2)$, sendo ε denominado de erro.

Interpretação dos parâmetros do modelo

Parâmetro β_1

Ele é o intercepto da reta de regressão.

Interpretação dos parâmetros do modelo

Parâmetro β_1

Ele é o intercepto da reta de regressão. Ele é o valor esperado da variável resposta quando a variável preditora vale zero.

Interpretação dos parâmetros do modelo

Parâmetro β_1

Ele é o intercepto da reta de regressão. Ele é o valor esperado da variável resposta quando a variável preditora vale zero.

Interpretação dos parâmetros do modelo

Parâmetro β_2

Ele é o declive da reta de regressão, o coeficiente angular.

Interpretação dos parâmetros do modelo

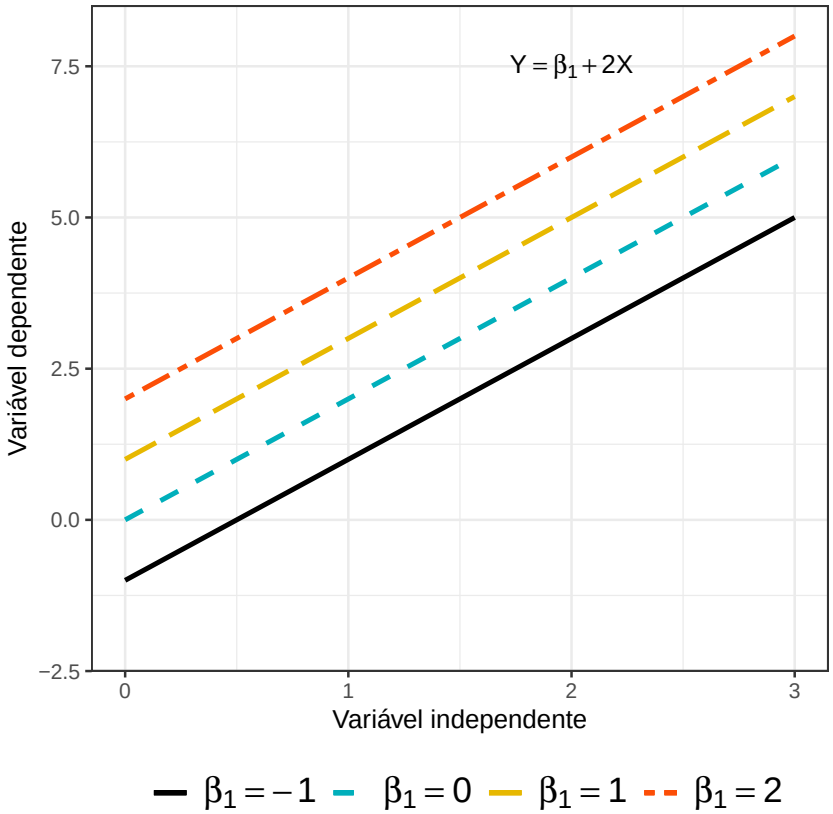
Parâmetro β_2

Ele é o declive da reta de regressão, o coeficiente angular. Ele indica o acréscimo esperado na variável resposta quando aumenta-se uma unidade na variável preditora.

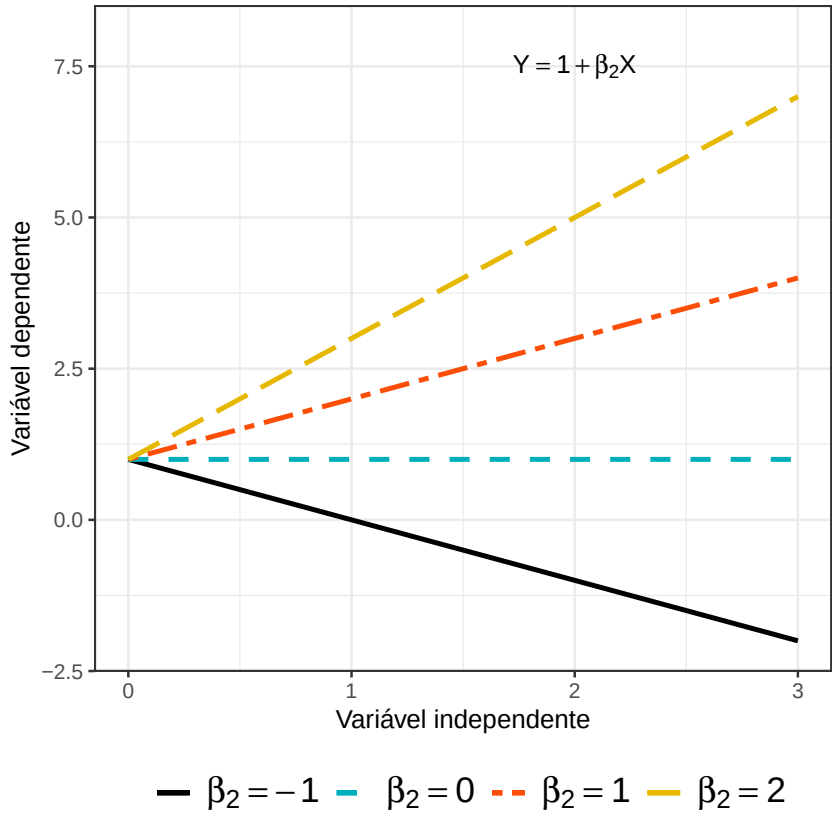
Interpretação dos parâmetros do modelo

Parâmetro β_2

Ele é o declive da reta de regressão, o coeficiente angular. Ele indica o acréscimo esperado na variável resposta quando aumenta-se uma unidade na variável preditora.



(a) β_1 livre e $\beta_2 = 2$.



(b) $\beta_1 = 1$ e β_2 livre.

Figura 1: Retas de regressão em função de β_1 e β_2.

Modelo de regressão linear

A fim de estimar β_1, β_2 e σ^2, e explicitar a relação entre as variáveis, é necessário retirar uma amostra independente de tamanho n do par (Y, x),

i.e., $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$,

Modelo de regressão linear

A fim de estimar β_1, β_2 e σ^2, e explicitar a relação entre as variáveis, é necessário retirar uma amostra independente de tamanho n do par (Y, x), i.e., $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$, com

$$Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2),$$

Modelo de regressão linear

A fim de estimar β_1, β_2 e σ^2, e explicitar a relação entre as variáveis, é necessário retirar uma amostra independente de tamanho n do par (Y, x), i.e., $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$, com

$$Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2),$$

em que $\mu_\ell = \beta_1 + \beta_2 x_\ell$, $\ell = 1, 2, \ldots, n$.

Modelo de regressão linear

A fim de estimar β_1, β_2 e σ^2, e explicitar a relação entre as variáveis, é necessário retirar uma amostra independente de tamanho n do par (Y, x), i.e., $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$, com

$$Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2),$$

em que $\mu_\ell = \beta_1 + \beta_2 x_\ell$, $\ell = 1, 2, \ldots, n$.

Modelo de regressão linear

De forma alternativa,

$$Y_\ell = \beta_1 + \beta_2 x_\ell + \varepsilon_\ell,$$

com $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$. Em outras palavras, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2.

Modelo de regressão linear

De forma alternativa,

$$Y_\ell = \beta_1 + \beta_2 x_\ell + \varepsilon_\ell,$$

com $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$. Em outras palavras, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são variáveis aleatórias independentes e com a mesma variância σ^2.

Suposições

Resumindo, as **suposições** de um modelo de regressão linear simples são de que:

- A relação entre as variáveis resposta e as preditoras é linear;

Suposições

Resumindo, as **suposições** de um modelo de regressão linear simples são de que:

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros,

Suposições

Resumindo, as **suposições** de um modelo de regressão linear simples são de que:

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros,
  - têm média zero;

Suposições

Resumindo, as **suposições** de um modelo de regressão linear simples são de que:

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros,
  - têm média zero;
  - variância constante;

Suposições

Resumindo, as **suposições** de um modelo de regressão linear simples são de que:

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros,
  - têm média zero;
  - variância constante;
  - e são não correlacionados.

Suposições

Resumindo, as **suposições** de um modelo de regressão linear simples são de que:

- A relação entre as variáveis resposta e as preditoras é linear;
- Os erros,
  - têm média zero;
  - variância constante;
  - e são não correlacionados.

Estimação

Pelo Método de Mínimos Quadrados, os estimadores de β_1 e β_2 são dados, respectivamente, por

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \text{ e } \hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \quad (1)$$

Estimação

Pelo Método de Mínimos Quadrados, os estimadores de β_1 e β_2 são dados, respectivamente, por

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \text{ e } \hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \quad (1)$$

Os estimadores de mínimos quadrados são não viesados.

Estimação

Pelo Método de Mínimos Quadrados, os estimadores de β_1 e β_2 são dados, respectivamente, por

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \text{ e } \hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \quad (1)$$

Os estimadores de mínimos quadrados são não viesados.

Estimação da variância

O estimador da variância, σ^2, é dado por

$$\hat{\sigma}^2 = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \hat{\beta}_1 - \hat{\beta}_2 X_\ell \right)^2}{n-2}. \tag{2}$$

Estimação da variância

O estimador da variância, σ^2, é dado por

$$\hat{\sigma}^2 = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \hat{\beta}_1 - \hat{\beta}_2 X_\ell \right)^2}{n-2}. \tag{2}$$

O estimador $\hat{\sigma}^2$, dado em (2), também é denominado de **quadrado médio do resíduo** (QMRes) e ele é um estimador não viciado para σ^2.

Estimação da variância

O estimador da variância, σ^2, é dado por

$$\hat{\sigma}^2 = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \hat{\beta}_1 - \hat{\beta}_2 X_\ell \right)^2}{n-2}. \tag{2}$$

O estimador $\hat{\sigma}^2$, dado em (2), também é denominado de **quadrado médio do resíduo** (QMRes) e ele é um estimador não viciado para σ^2.

Roteiro

Aplicação

O campeonato brasileiro de futebol masculino, o Brasileirão, tem sido disputado no sistema de pontos corridos desde 2003.

Aqui, nosso interesse pode ser em avaliar se existe uma relação linear entre o aproveitamento de pontos (%) e o saldo de gols dos times campeões.

Aplicação

O campeonato brasileiro de futebol masculino, o Brasileirão, tem sido disputado no sistema de pontos corridos desde 2003.

Aqui, nosso interesse pode ser em avaliar se existe uma relação linear entre o aproveitamento de pontos (%) e o saldo de gols dos times campeões.

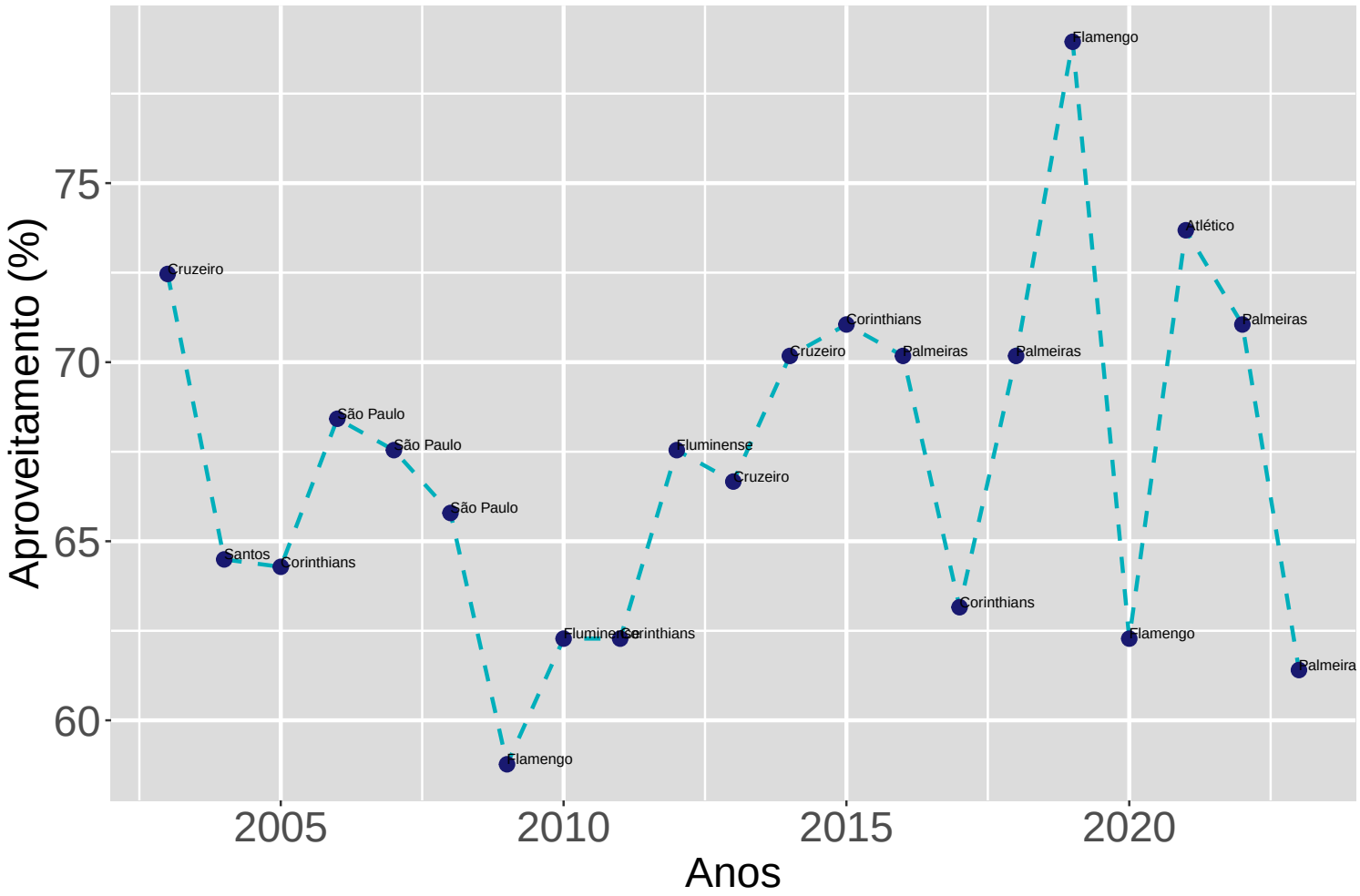


Figura 2: Aproveitamento de pontos (%) dos times campeões ao longo dos anos.

Aplicação

Tabela 1: Estados dos times campeões.

Minas Gerais	Rio de Janeiro	São Paulo
4	5	12
(19%)	(24%)	(57%)
2	2	4

Aplicação

Tabela 2: Aproveitamento de pontos (%) dos times campeões.

Mín.	Q1	Md	Média	Q3	Máx.	DP	CV
58,8	63,2	67,5	67,3	70,2	78,9	4,9	7,3

Aplicação

Tabela 3: Aproveitamento de pontos (%) dos times campeões, por estado.

	Mín.	Q1	Md	Média	Q3	Máx.	DP	CV
Minas Gerais	66,7	69,3	71,3	70,7	72,8	73,7	3,1	4,4
Rio de Janeiro	58,8	62,3	62,3	66,0	67,5	78,9	7,9	12,0
São Paulo	61,4	64,0	66,7	66,7	70,2	71,1	3,5	5,3

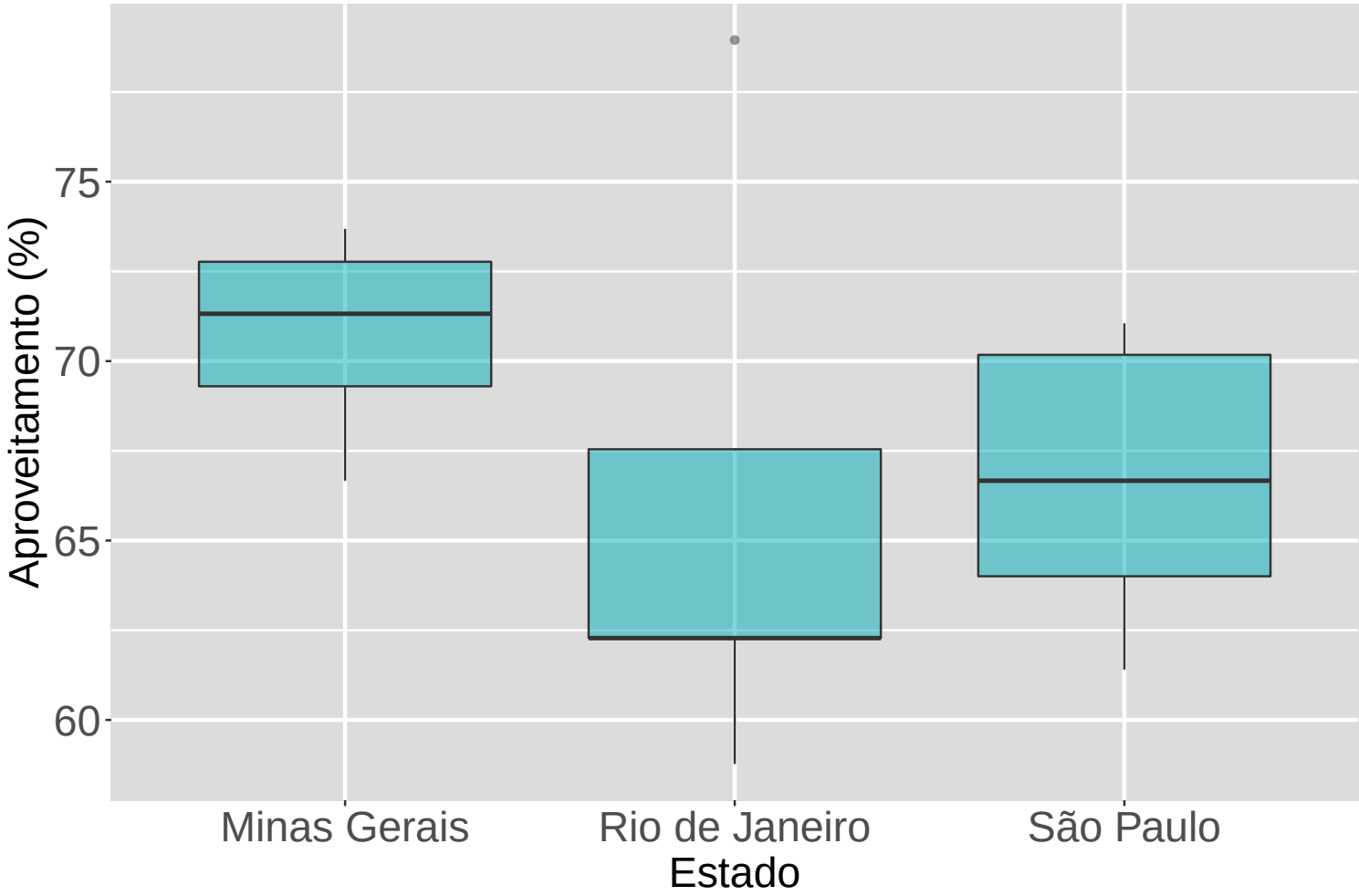


Figura 3: Boxplot do aproveitamento de pontos (%) dos times campeões por UF.

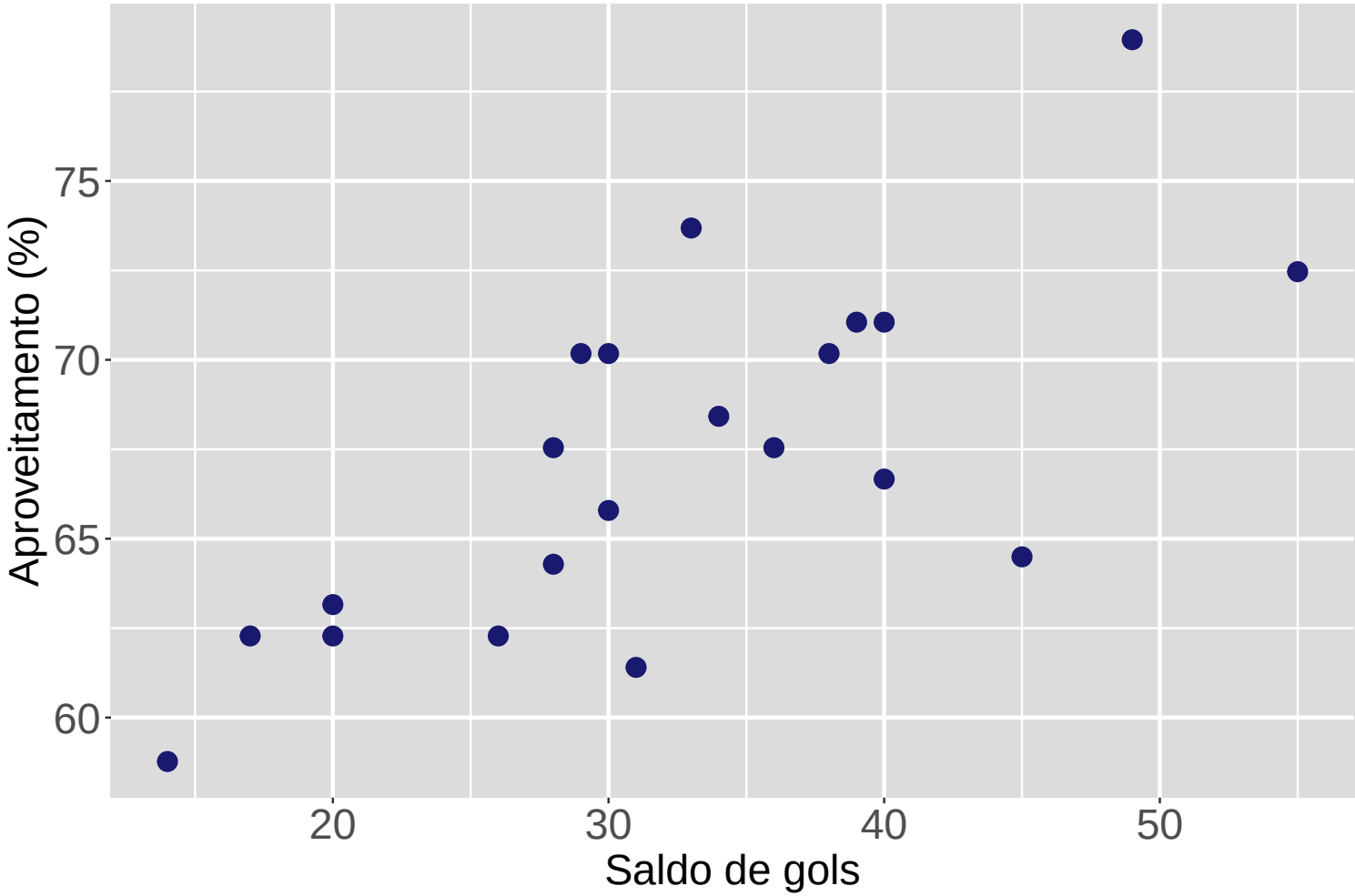


Figura 4: Gráfico de dispersão do aproveitamento de pontos (%) e saldo de gols dos times campeões.

Aplicação

Após o ajuste (estimação dos parâmetros), utilizando (1) e (2), nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = 56{,}1 + 0{,}3x_{\ell 2},$$

Aplicação

Após o ajuste (estimação dos parâmetros), utilizando (1) e (2), nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = 56{,}1 + 0{,}3x_{\ell 2},$$

em que $\hat{Y}_\ell$: aproveitamento de pontos (%) esperado do ℓ-ésimo time campeão, $x_{\ell 2}$: saldo de gols do ℓ-ésimo time campeão, $\ell = 1, 2, \ldots, 21$ e QMRes $= 11{,}9$.

Aplicação

Após o ajuste (estimação dos parâmetros), utilizando (1) e (2), nós temos o seguinte modelo estimado,

$$\hat{Y}_\ell = 56{,}1 + 0{,}3x_{\ell 2},$$

em que $\hat{Y}_\ell$: aproveitamento de pontos (%) esperado do ℓ-ésimo time campeão, $x_{\ell 2}$: saldo de gols do ℓ-ésimo time campeão, $\ell = 1, 2, \ldots, 21$ e QMRes $= 11{,}9$.

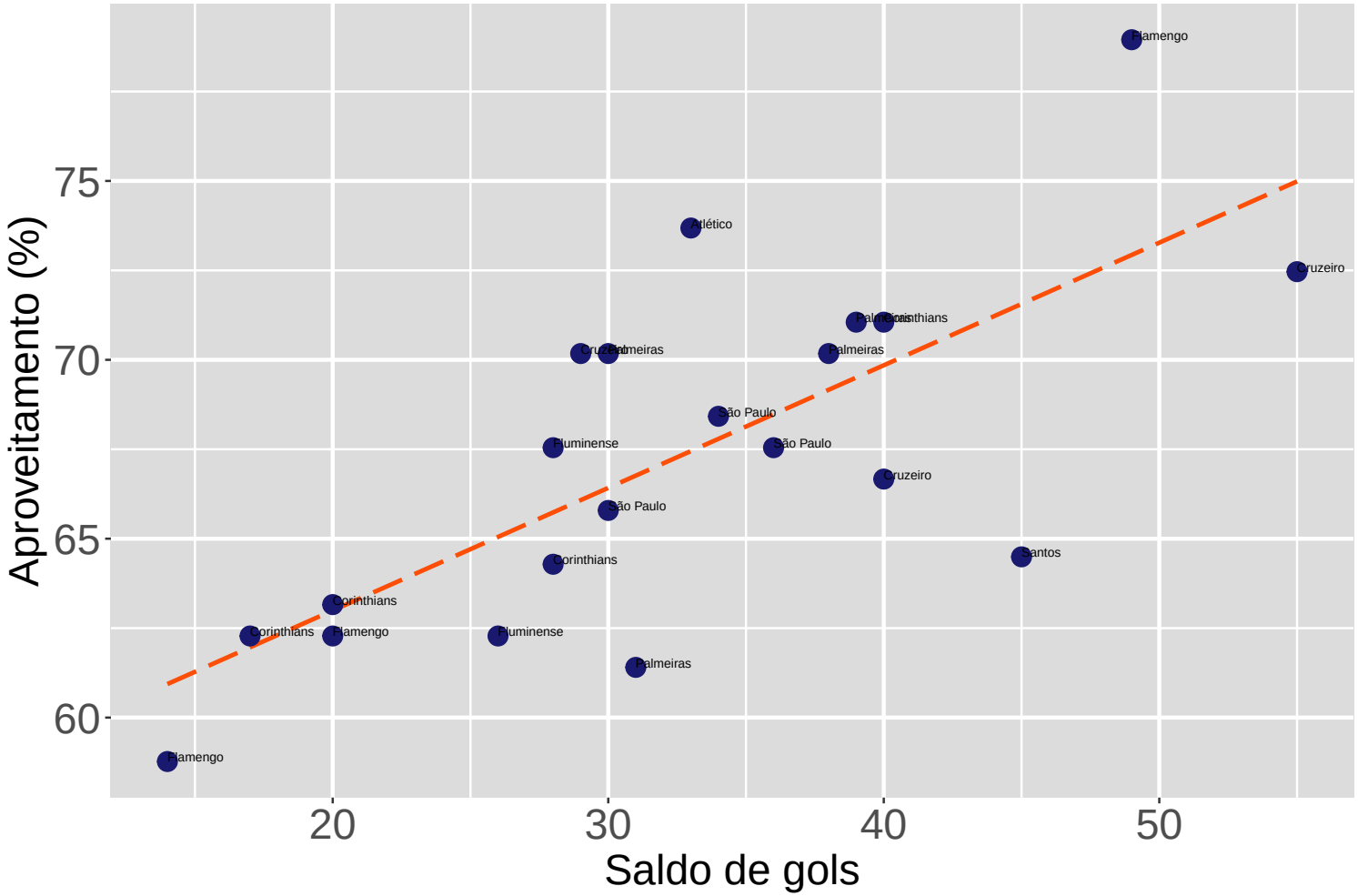


Figura 5: Gráfico de dispersão do aproveitamento de pontos (%) e saldo de gols dos times campeões e a reta de regressão estimada.

Aplicação

Os coeficientes de regressão podem ser interpretados da seguinte forma

- $\hat{\beta}_1 = 56{,}1$:

Aplicação

Os coeficientes de regressão podem ser interpretados da seguinte forma

- $\hat{\beta}_1 = 56{,}1$: se o saldo de gols fosse zero, aproveitamento de pontos do time campeão seria de 56,5%;

Aplicação

Os coeficientes de regressão podem ser interpretados da seguinte forma

- $\hat{\beta}_1 = 56{,}1$: se o saldo de gols fosse zero, aproveitamento de pontos do time campeão seria de 56,5%;
- $\hat{\beta}_2 = 0{,}3$:

Aplicação

Os coeficientes de regressão podem ser interpretados da seguinte forma

- $\hat{\beta}_1 = 56{,}1$: se o saldo de gols fosse zero, aproveitamento de pontos do time campeão seria de 56,5%;
- $\hat{\beta}_2 = 0{,}3$: a cada um gol a mais no saldo, ocorreria um aumento médio de 0,3 pontos percentuais no aproveitamento de pontos.

Aplicação

Os coeficientes de regressão podem ser interpretados da seguinte forma

- $\hat{\beta}_1 = 56{,}1$: se o saldo de gols fosse zero, aproveitamento de pontos do time campeão seria de 56,5%;
- $\hat{\beta}_2 = 0{,}3$: a cada um gol a mais no saldo, ocorreria um aumento médio de 0,3 pontos percentuais no aproveitamento de pontos.

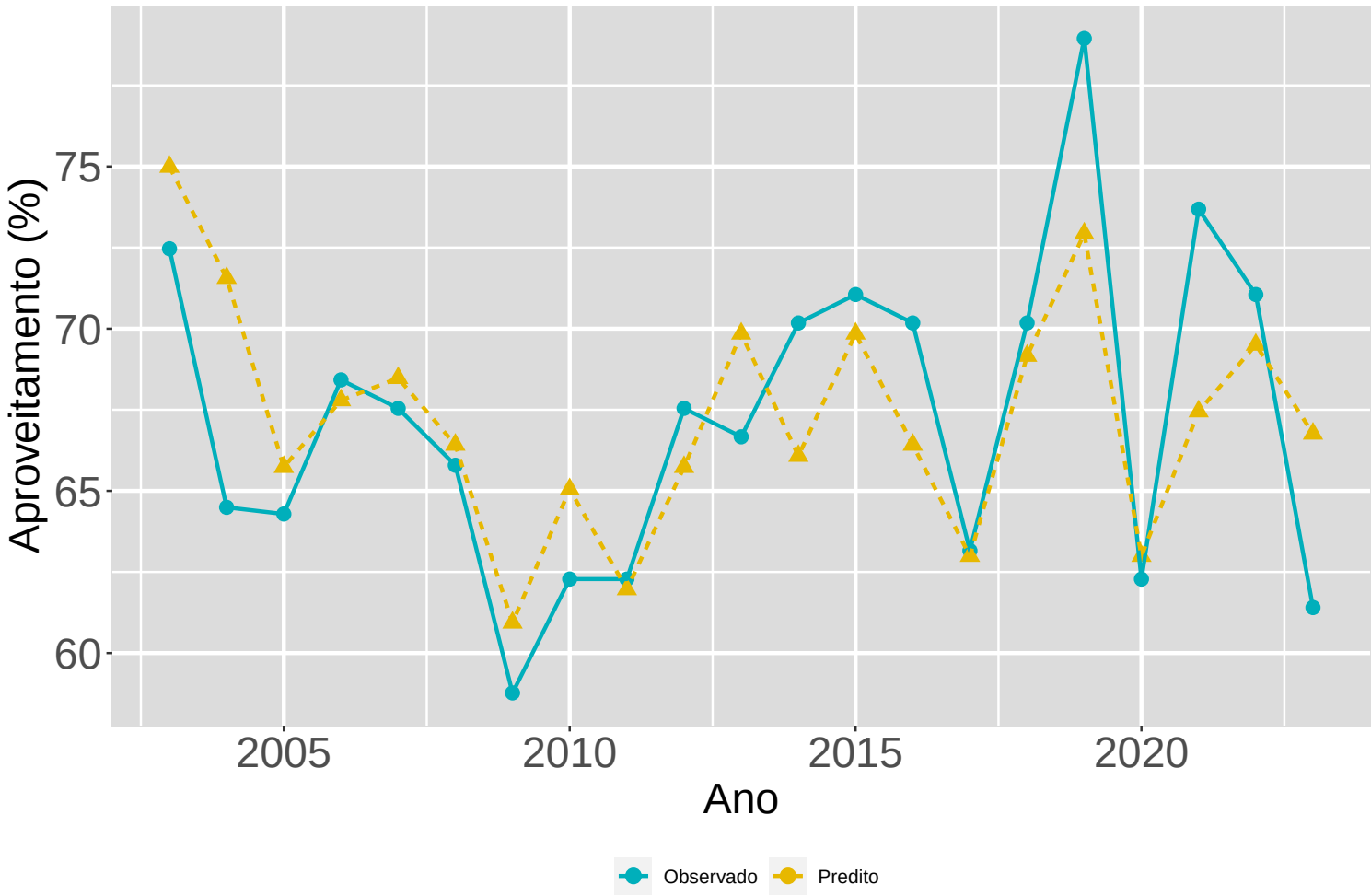


Figura 6: Aproveitamento de pontos (%) observado e predito pelo modelo.

Roteiro

Referências bibliográficas

Galton, F. (1889). *Natural Inheritance*. London: Macmillan.

Montgomery, D. C., E. A. Peck, and G. G. Vining (2021). *Introduction to linear regression analysis* (6th ed.). New York: Wiley.

Obrigado!

✉ tiago.magalhaes@ufjf.br

ufjf.br/tiago_magalhaes

Dept. de Estatística, Sala 319