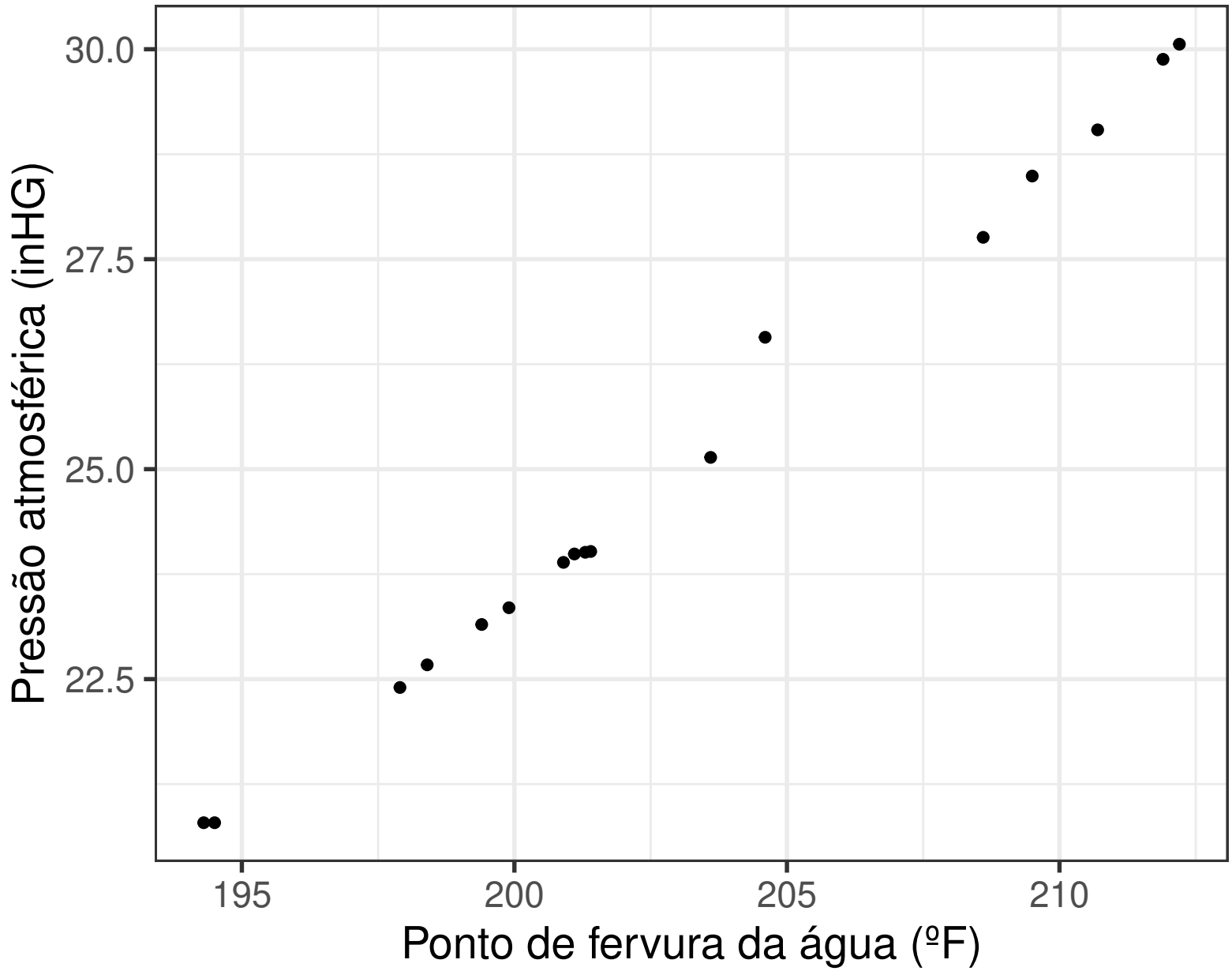
Regressão linear

- Mínimos quadrados;
- Modelo de regressão simples (univariado);
  - ◇ Formulação;
  - ◇ Premissas.
- Distribuição amostral dos estimadores;
- Intervalos de confiança para os coeficientes;
- Testes para os coeficientes;
- Predição: pontual e intervalar.

Regressão linear: motivação



Mínimos quadrados

Suponha que estamos interessados na reta

$$y_i = \beta_0 + \beta_1 x_i. \tag{32}$$

- β_0 é chamado o **intercepto** (*intercept*) da reta;
- β_1 é chamado o **coeficiente angular** (*slope*) da reta.

Teorema 35 (A linha de mínimos quadrados)

Sejam $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ uma coleção de n pontos. Os valores dos coeficientes que minimizam a soma de quadrados são

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

onde $\bar{x} = (1/n)\sum_{i=1}^n x_i$ e $\bar{y} = (1/n)\sum_{i=1}^n y_i$.

Prova: Escrever a equação de estimação, $Q = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$, diferenciar Q com respeito aos coeficientes e igualar a zero. Ver Teorema 11.1.1 em DeGroot.

O modelo linear

Podemos construir um modelo estatístico explícito para a relação entre as variáveis[18] $\boldsymbol{X}$ e Y:

$$E[Y \mid \boldsymbol{X} = x_1, x_2, \ldots, x_P] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_P x_P. \tag{33}$$

Terminologia:

- Y é chamada de desfecho, **variável-resposta** ou variável dependente;
- $\boldsymbol{X}$ são chamados covariáveis, **preditores** ou, ainda, variáveis independentes;
- $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_P\}$ são os **coeficientes de regressão**.

Podemos então idealizar o seguinte modelo

Ideia 6 (Modelo linear)

$$Y_i = \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij} + \epsilon_i, \ \epsilon_i \sim \text{Normal}(0, \sigma^2).$$

[18] Em notação de matrizes, $E[Y] = \boldsymbol{X}^T \boldsymbol{\beta}$.

Premissas (importante!)

Como todo modelo, a regressão linear se apoia em premissas sobre os dados e o seu processo gerador.

P1. O(s) preditor(es) é (são) conhecido(s);

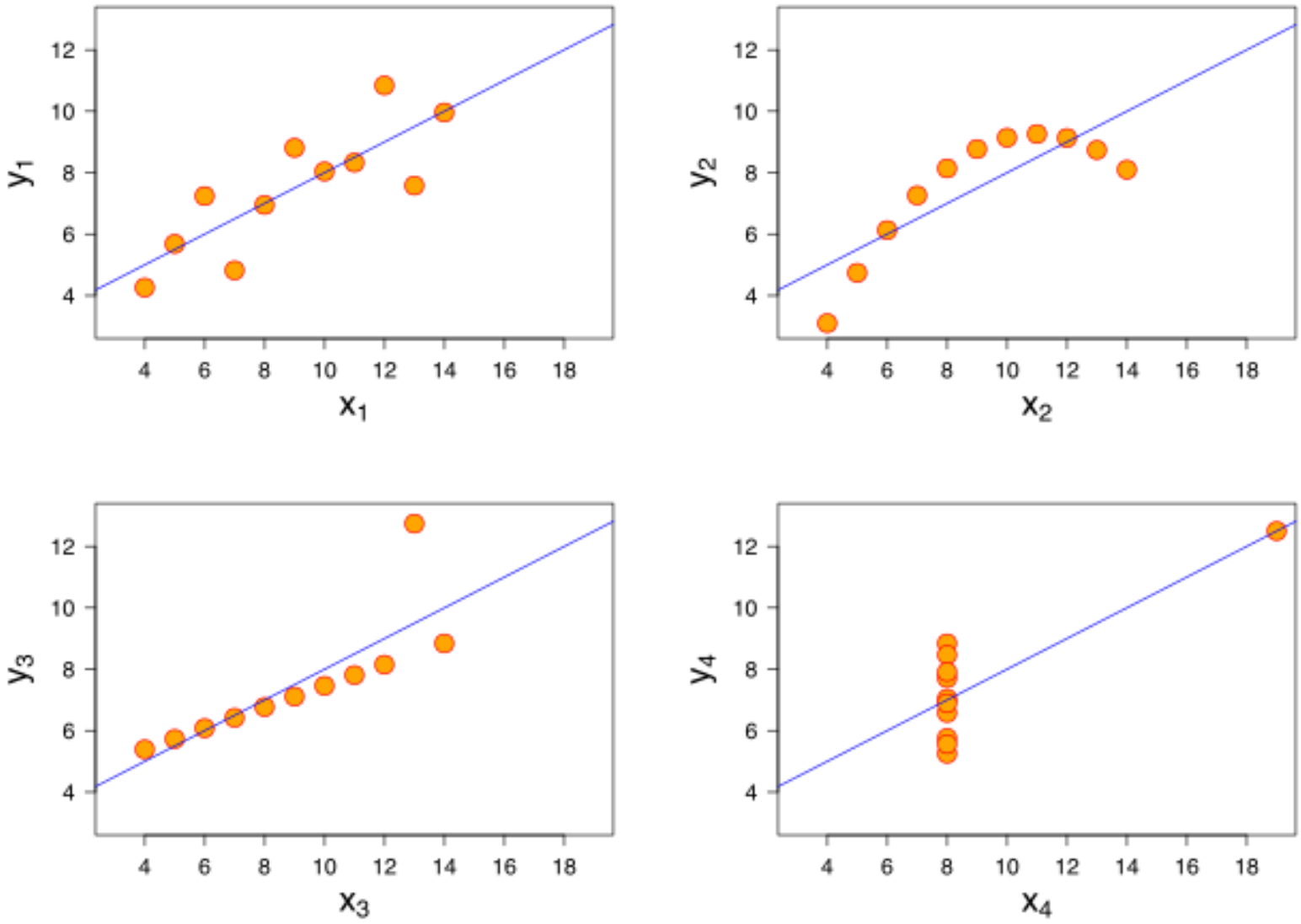
P2. Normalidade: dados os preditores $\boldsymbol{X}$, a resposta Y tem distribuição normal;

P3. Linearidade na média: a esperança condicional de Y é dada por $\beta_0 + \sum_{j=1}^{P} \beta_j x_{ij}$;

P4. Variância comum (**homocedasticidade**): a variância condicional de Y_i é σ^2 para todo $i = 1, 2, \ldots, n$;

P5. Independência (condicional): dados os valores de $\boldsymbol{X}$, os valores de Y são idependentes entre si.

Cuidado! Quarteto de Anscombe[19]



[19] Em homenagem ao estatístico Britânico Francis Anscombe (1918-2001).

Um teorema interessante

No modelo linear, a solução de mínimos quadrados e a de máxima verossimilhança coincidem!

Teorema 36 (EMV para os coeficientes de uma regressão linear (simples))

Sob as premissas já listadas, os estimadores de máxima verossimilhança para $\theta = (\beta_0, \beta_1, \sigma^2)$ *são*

$$\hat{\beta}_{0\,EMV} = \bar{y} - \hat{\beta}_{1\,EMV}\bar{x},$$

$$\hat{\beta}_{1\,EMV} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\sigma^2}_{EMV} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (\hat{\beta}_{0\,EMV} + \hat{\beta}_{1\,EMV}x_i)\right)^2,$$

ou seja, os estimadores de máxima verossimilhança dos coeficientes minimizam a soma de quadrados da reta estimada.

Prova: Ver Teorema 11.2.1 de DeGroot.

Distribuição amostral dos estimadores

Sob as premissas já discutidas, podemos fazer afirmações sobre a distribuição amostral dos estimadores obtidos:

Teorema 37 (Distribuição amostral dos estimadores dos coeficientes)

$$\hat{\beta}_{0\,EMV} \sim \text{Normal}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_x^2}\right)\right),$$

$$\hat{\beta}_{1\,EMV} \sim \text{Normal}\left(\beta_1, \frac{\sigma^2}{s_x^2}\right),$$

$$\text{Cov}\left(\hat{\beta}_{0\,EMV}, \hat{\beta}_{1\,EMV}\right) = -\frac{\bar{x}\sigma^2}{s_x^2},$$

onde $s_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Prova: Usar as leis de esperanças e variâncias. Ver Teorema 11.2.2 de DeGroot.

Intervalos de confiança para os coeficientes

Podemos computar intervalos de confiança para os coeficientes da regressão linear de maneira muito similar ao que já vimos para o caso da média da Normal.

Teorema 38 (Intervalos de confiança para os coeficientes de uma regressão linear)

$$\hat{\beta}_0 \pm \hat{\sigma}' c \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_x^2}} \quad e \quad \hat{\beta}_1 \pm c \frac{\hat{\sigma}'}{s_x},$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_{pred} \pm c \hat{\sigma}' \sqrt{\frac{1}{n} + \frac{(x_{pred} - \bar{x})^2}{s_x^2}},$$

onde $c = T^{-1}(1 - \frac{\alpha_0}{2}; n-2)$ *e*

$$\hat{\sigma}' := \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2}}.$$

Prova: Usar o Teorema 11.3.5 de DeGroot e os valores apropriados de c_0 e c_1.

Testes de hipóteses para o coeficiente angular

Em geral, estamos interessados em testar a hipótese

$$H_0 : \beta_1 = \beta^\star,$$
$$H_1 : \beta_1 \neq \beta^\star.$$

Para tanto, podemos computar a estatística

$$U_1 = s_x \frac{\hat{\beta}_1 - \beta^\star}{\hat{\sigma}'},$$

e computar o p-valor como

$$\Pr(U_1 \geq |u_1|) + \Pr(U_1 \leq -|u_1|).$$

Notando que U_1 tem distribuição t de Student com $n-2$ graus de liberdade sob H_0, podemos computar o p-valor exatamente.
Resultados bem similares valem para testar hipóteses sobre β_0 ou $\hat{Y}$.

Predição pontual

Suponha que queremos prever o valor de Y para um certo x_{pred} que não foi observado no experimento. Podemos compor nossa predição (pountual) como

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{pred}}. \tag{34}$$

Teorema 39 (Erro quadrático médio da predição)

A predição como em (34) tem erro quadrático médio (EQM) igual a

$$E\left[\left(\hat{Y} - Y\right)^2\right] = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{pred} - \bar{x})^2}{s_x^2}\right).$$

Prova: Ver Teorema 11.2.3 de DeGroot.

Observação 26 (EQM fora da amostra)

O EQM aumenta quanto mais longe x_{pred} estiver dos valores de X que foram medidos (observados).

Predição intervalar

Muitas vezes estamos interessados em produzir um *intervalo* para a nossa predição, ao invés de um único valor (predição pontual). Nesta situação, podemos fazer uso do seguinte teorema:

Teorema 40 (Intervalos de predição para $\hat{Y}$)

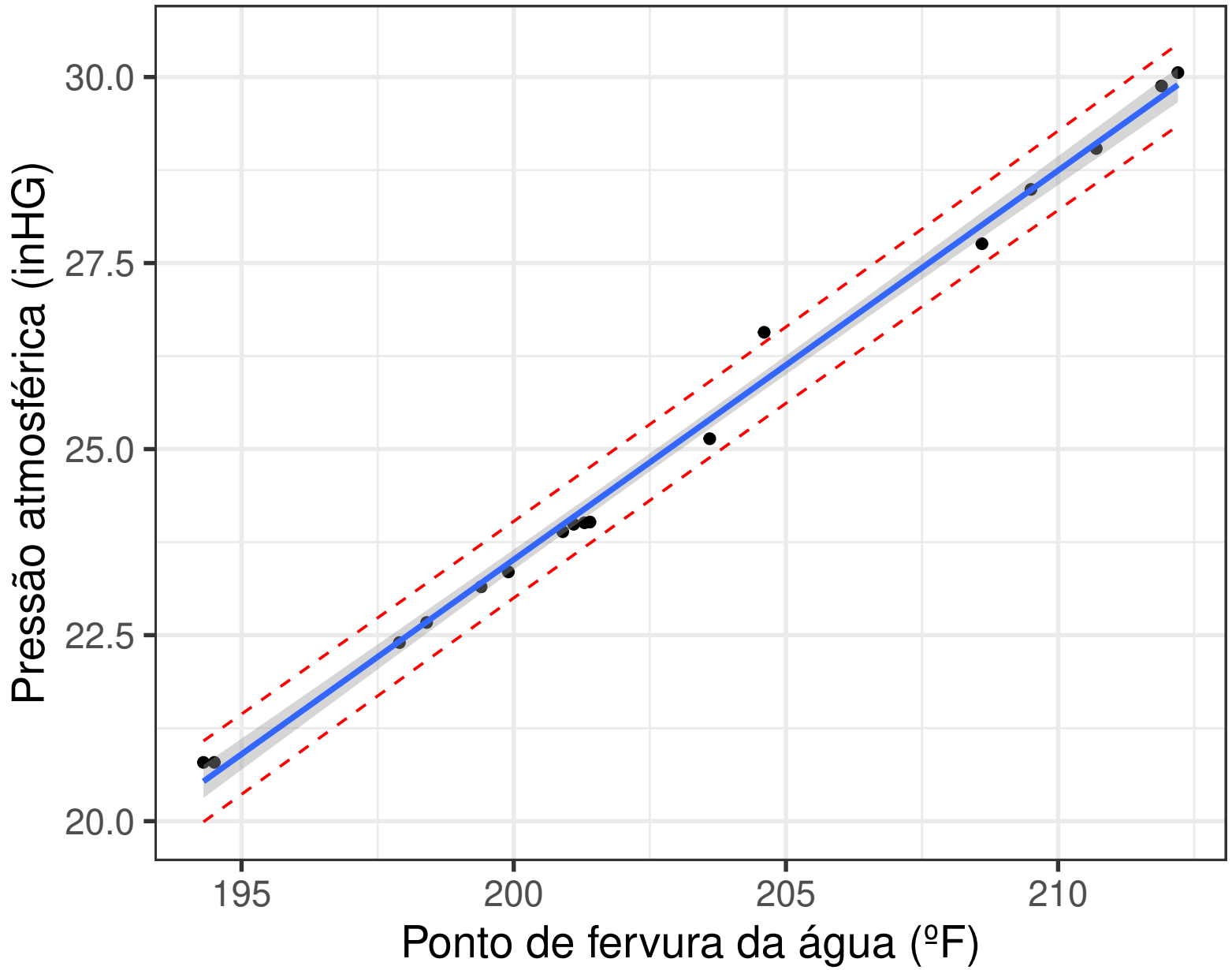
A probabilidade de $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{pred}$ *estar no intervalo*

$$\hat{Y} \pm T^{-1}\left(1 - \frac{\alpha_0}{2}; n-2\right) \hat{\sigma}' \sqrt{\left[1 + \frac{1}{n} + \frac{(x_{pred} - \bar{x})^2}{s_x^2}\right]},$$

é $1 - \alpha_0$.

Prova: Ver Teorema 11.3.6 de DeGroot.

Intervalos de confiança e de predição: ilustração



O que aprendemos?

FGV EMAp

- O modelo linear permite modelar a relação (linear) entre uma (ou mais) variável(is) independente(s) e uma variável dependente;
- A estimação dos coeficientes pode ser feita por mínimos quadrados;
- A solução de mínimos quadrados é também a solução de máxima verossimilhança!
- Podemos aplicar a teoria Normal para testar hipóteses sobre os coeficientes e calcular intervalos de confiança;
- Podemos produzir predições sobre a variável dependente para valores não-observados da(s) variável(is) independente(s).

Leitura recomendada

FGV EMAp

- DeGroot seções 11.1, 11.2 e 11.3;
- * Casella & Berger (2002), seção 11.3.
- Próxima aula: De Groot, seção 9.9;
- **Exercícios recomendados**
  - DeGroot, seção 11.1: exercício 3.
  - DeGroot, seção 11.2: exercícios 2, 3 e 6.
  - * Bônus: DeGroot, seção 11.2: exercício 19 (valendo 0.5 na média).