CE071 - Análise de Regressão Linear

Cesar Augusto Taconeli

27 de fevereiro, 2018

Aula 1 - Introdução

Introdução

Análise de regressão

Conjunto de técnicas estatísticas aplicadas na investigação e modelagem da relação entre variáveis

Na análise de regressão estudamos a distribuição de uma variável aleatória Y condicional a um conjunto de variáveis explicativas $X_1, X_2, ..., X_p$.

Introdução

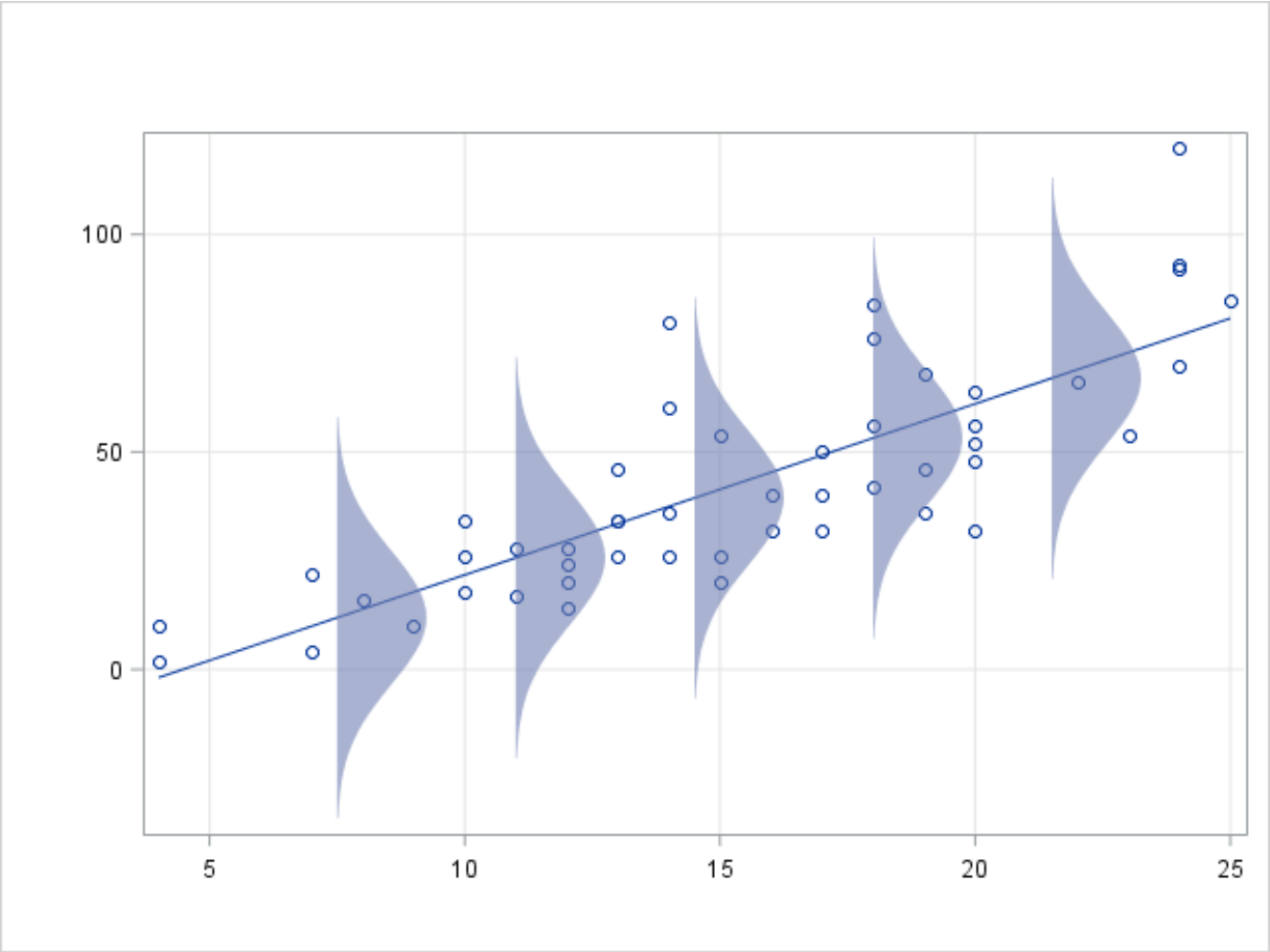

Figura 1: Regressão - Distribuição Normal (1)

Introdução

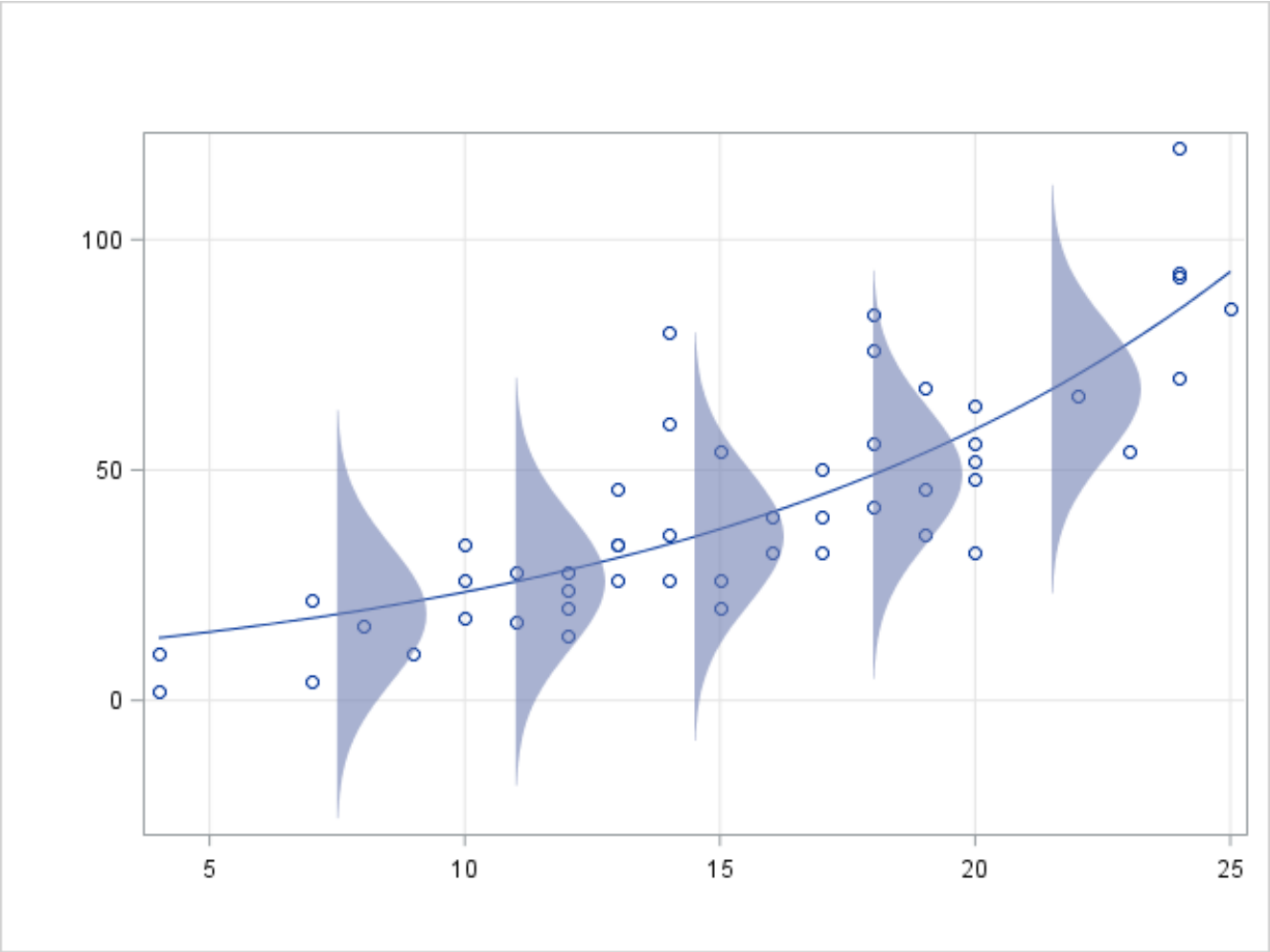

Figura 2: Regressão - Distribuição Normal (2)

Introdução

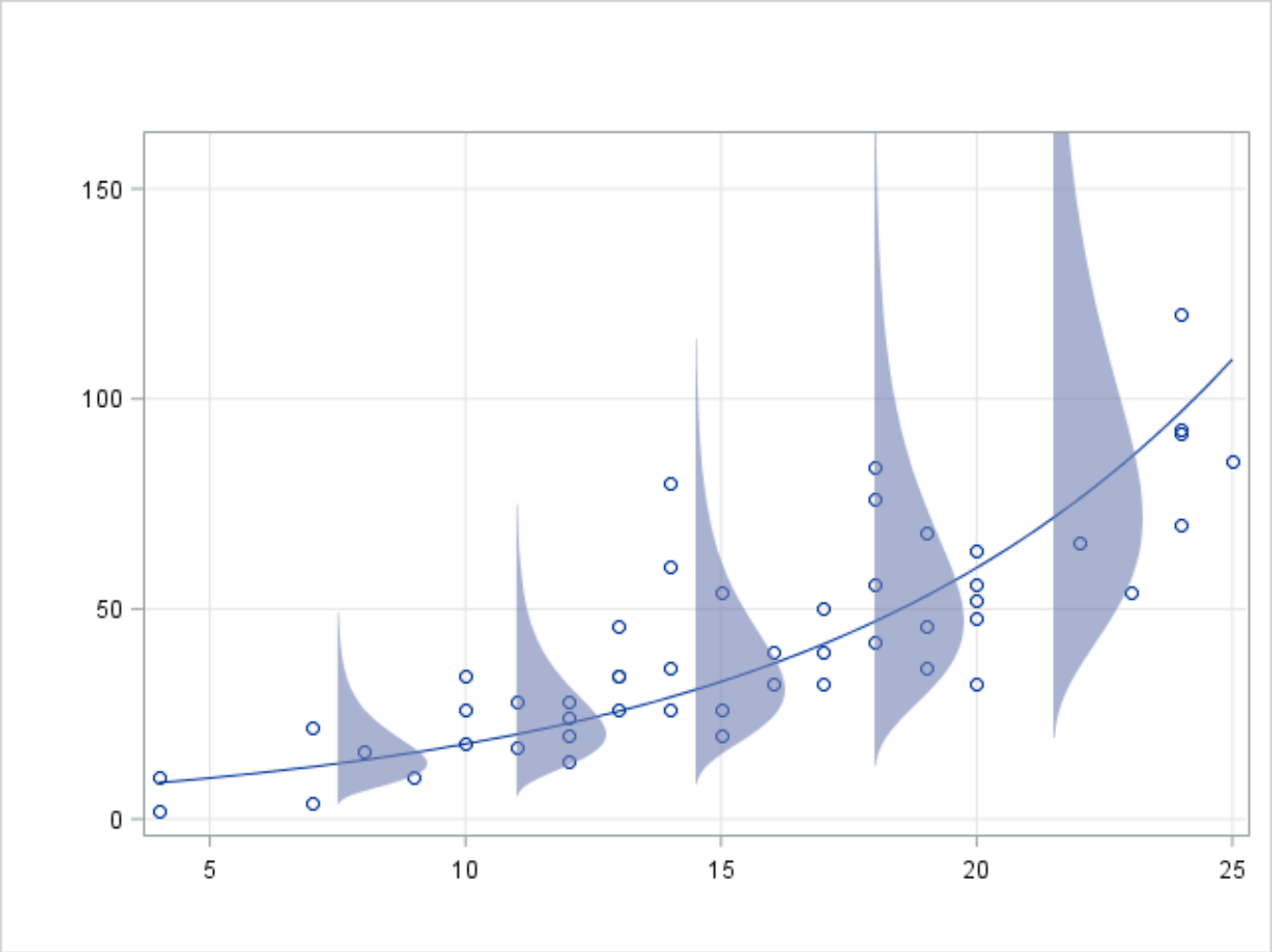

Figura 3: Regressão - Distribuição assimétrica

Introdução

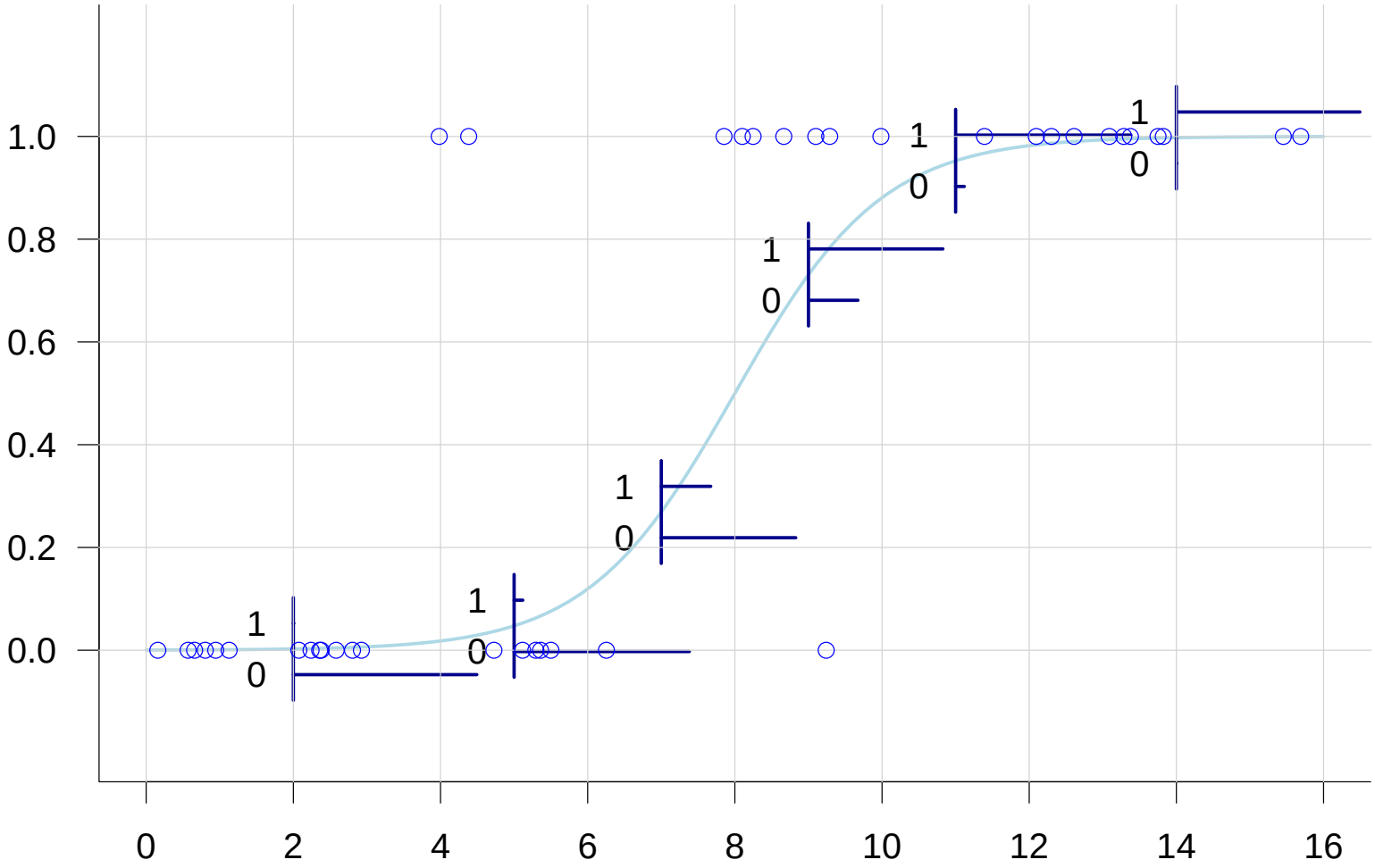

Figura 4: Regressão - dados binários

Introdução

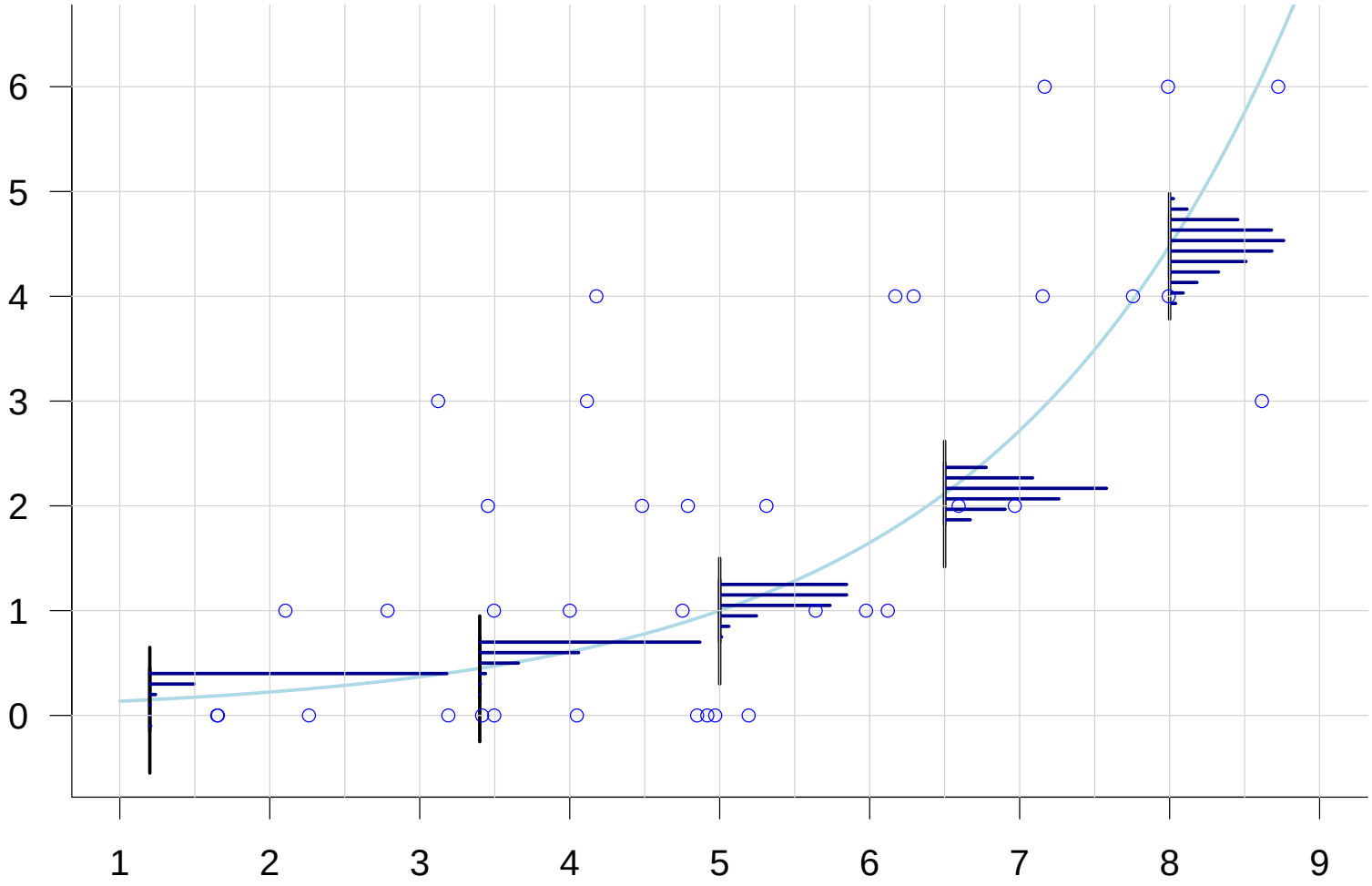

Figura 5: Regressão - dados de contagens

Introdução

All models are wrong but some are useful

George Box

No matter how beautiful your theory, no matter how clever you are or what your name is, if it disagrees with experiment, it's wrong.

Richard Feynman

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.

John W. Tukey

Um breve histórico

- **Século 19:** Desenvolvimento da teoria de mínimos quadrados, que fundamenta o ajuste de modelos lineares;
- A teoria de mínimos quadrados teve origem na física motivada, dentre outros, por problemas envolvendo navegação (século 18);
- Ao longo do século 19, modelos lineares e o método de mínimos quadrados passaram a ser utilizados em outras ciências, baseados não mais em modelos pré-concebidos, mas apenas na evidência empírica.
- O termo regressão foi introduzido por Francis Galton em 1875, baseado no princípio da "regressão à média".

Exemplo 1

- Relação entre altura de pais e filhos (Galton, 1886).
- Os dados estão disponíveis no pacote `Histdata` (base `GaltonFamilies`).
- Os códigos são apresentados em Faraway(2014) e estão disponíveis na página da disciplina.

Introdução

- Os dois objetivos principais de uma análise de regressão são os seguintes:
  - Avaliar o efeito (ou relação) entre as variáveis explicativas e a resposta;
  - Predizer valores não observados da resposta para um conjunto de valores especificados das variáveis explicativas.

Exemplo - reposição de máquinas de refrigerantes

- Dados sobre reposição de máquinas de venda de refrigerantes (Montgomery et al., 2006).
- Os dados estão disponíveis no pacote `MPV` (base `p8.3`)
- Os códigos estão disponíveis na página da disciplina.

Exemplo - reposição de máquinas de refrigerantes

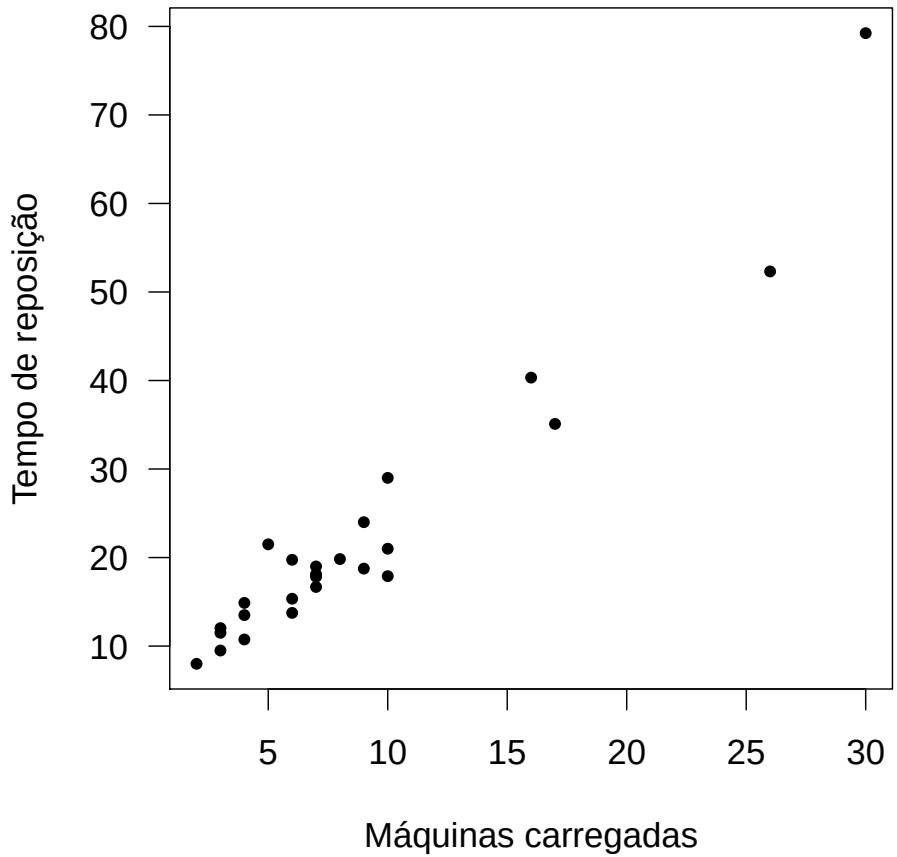

Figura 6: Gráfico de dispersão para o número de máquinas e o tempo de carregamento

Exemplo - reposição de máquinas de refrigerantes

- Há relação entre o tempo de reposição e o número de máquinas? De que tipo? (linear, quadrática...)
- Qual a variação no tempo de reposição para um particular incremento no número de máquinas?
- Qual o tempo médio de reposição para carregamentos de 15 máquinas?
- Quanto tempo previsto para carregar 15 máquinas?
- Para um tempo de reposição de 10 minutos, qual a previsão para o número de máquinas que podem ser carregadas?

Introdução

- Observe que as questões propostas não podem ser respondidas exatamente, uma vez que, fixado o número de máquinas, existe variação no tempo de carregamento.

- Essa variação pode ser explicada por outras variáveis não consideradas na análise, medições incorretas,. . .

- Na análise de regressão, chamamos de *erro* a diferença entre a resposta observada e a resposta esperada segundo o modelo.

Introdução

- Caso não houvesse variação, assumindo-se relação linear entre o tempo de reposição (y) e o número de máquinas carregadas (x), o seguinte modelo explicaria a relação entre as variáveis:

$$y = \beta_0 + \beta_1 x. \tag{1}$$

- Na presença dos erros, uma formulação mais adequada para o modelo seria a seguinte:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2}$$

- A equação (2) configura um **modelo de regressão linear** (simples).

Introdução

- Os termos β_0 e β_1 constituem os **parâmetros** do modelo, e configuram a relação entre as variáveis.

- O componente $\beta_0 + \beta_1 x$ é o preditor (parte fixa) do modelo, enquanto ϵ representa o erro, que é o componente aleatório.

- Denominamos **ajuste** de um modelo de regressão o processo de estimação dos parâmetros do modelo com base nos dados disponíveis, obtendo a equação da regressão ajustada (no caso, a reta ajustada).

- A Figura 6 apresenta o diagrama de dispersão para os dados de carregamento de máquinas de refrigerante com a reta de regressão ajustada.

Exemplo 2 - reposição de máquinas de refrigerantes

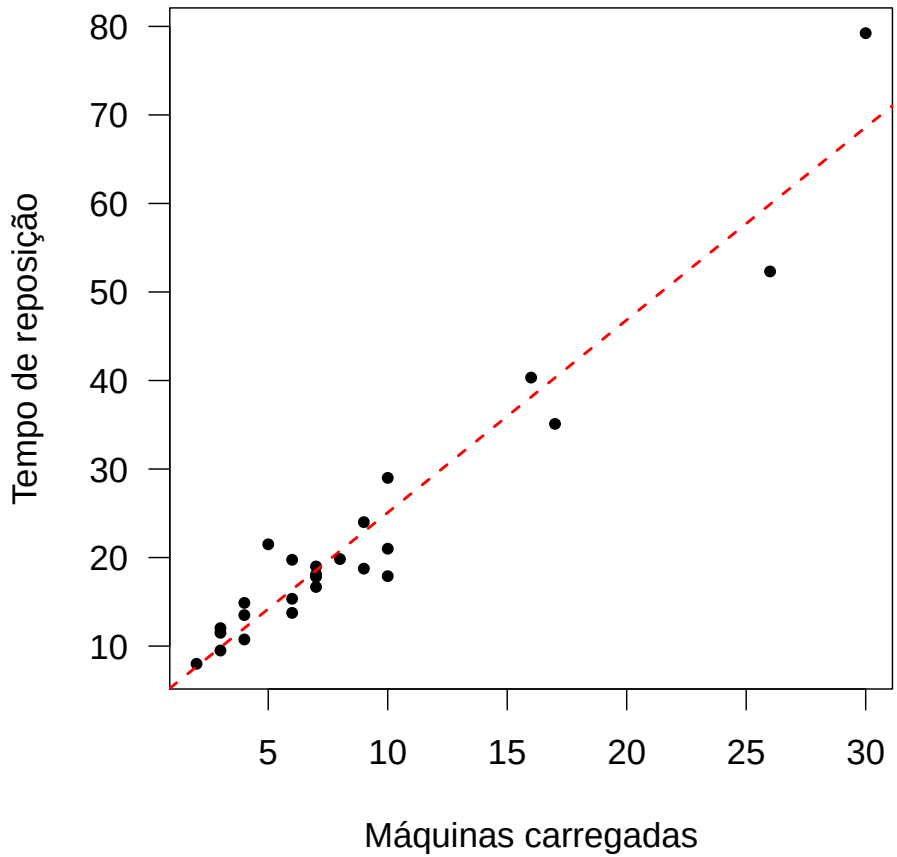

Figura 7: Gráfico de dispersão para o volume e o tempo de carregamento com reta de regressão ajustada

Introdução

- O modelo de regressão linear simples (2) pode ser expresso numa forma mais geral como:

$$y = f(\boldsymbol{x}; \boldsymbol{\beta}) + \epsilon, \tag{3}$$

em que $\boldsymbol{\beta}' = (\beta_0, \beta_1, ..., \beta_k)$ é o vetor de parâmetros e $\boldsymbol{x}' = (x_1, x_2, ..., x_q)$ é o vetor de variáveis explicativas.

- Neste caso, $f(\boldsymbol{x}; \boldsymbol{\beta})$ representa a parte fixa e ϵ a parte aleatória (erro) do modelo.

Introdução

- A formulação apresentada em (3) pode representar tanto modelos de regressão lineares quanto não lineares.

- Um modelo de regressão linear se caracteriza por uma combinação linear de variáveis.

- O que determina um modelo de regressão linear é a linearidade com relação aos parâmetros (e não com relação aos preditores).

Introdução

- Mais formalmente, um modelo de regressão é linear se

$$\frac{\partial f(\boldsymbol{x};\boldsymbol{\beta})}{\partial \beta_j} \tag{4}$$

não depender de β_j, $j = 0, 1, ..., k$.

- Ou seja, num modelo de regressão linear, ao se derivar parcialmente em relação a um parâmetro β_j, o resultado não depende de β_j.

Introdução

Considere os modelos de regressão descritos na sequência. Verifique se cada um deles corresponde a um modelo de regressão linear.

1. $y = \beta_0 + \beta_1 x + \epsilon;$
2. $y = \beta_0 + \beta_1 \ln(x) + \epsilon;$
3. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon;$
4. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon;$
5. $y = \beta_0 + \beta_1 \ln(x_1) + \frac{\beta_2}{x_2} + \beta_3 x_1 x_2 + \epsilon;$
6. $y = \beta_0 + \beta_1 e^{\beta_2 x_2} + \epsilon;$
7. $y = \frac{\beta_1}{1+e^{\beta_2 x}} + \epsilon;$
8. $y = \beta_0 sen(\beta_1 + \beta_2 x) + \epsilon.$

Exemplo - Relação entre tempo de aquecimento e desempenho de ginastas

- Dados de 19 atletas de uma equipe de ginástica artística referentes ao tempo de aquecimento (t, em minutos) e o desempenho (y, um escore numérico) na realização de uma específica apresentação.

Tabela 1: Tempo de aquecimento e desempenho dos atletas

Atleta	x	y	Atleta	x	y
1	1	6.3	11	7	42.0
2	1.5	11.1	12	8	46.1
3	2	20.0	13	9	48.2
4	3	24.0	14	10	45.0
5	4	26.1	15	11	42.0
6	4.5	30.0	16	12	40.0
7	5	33.8	17	13	38.5
8	5.5	34.0	18	14	35
9	6	38.1	19	15	36.8
10	6.5	39.9			

Exemplo - Relação entre tempo de aquecimento e desempenho de ginastas

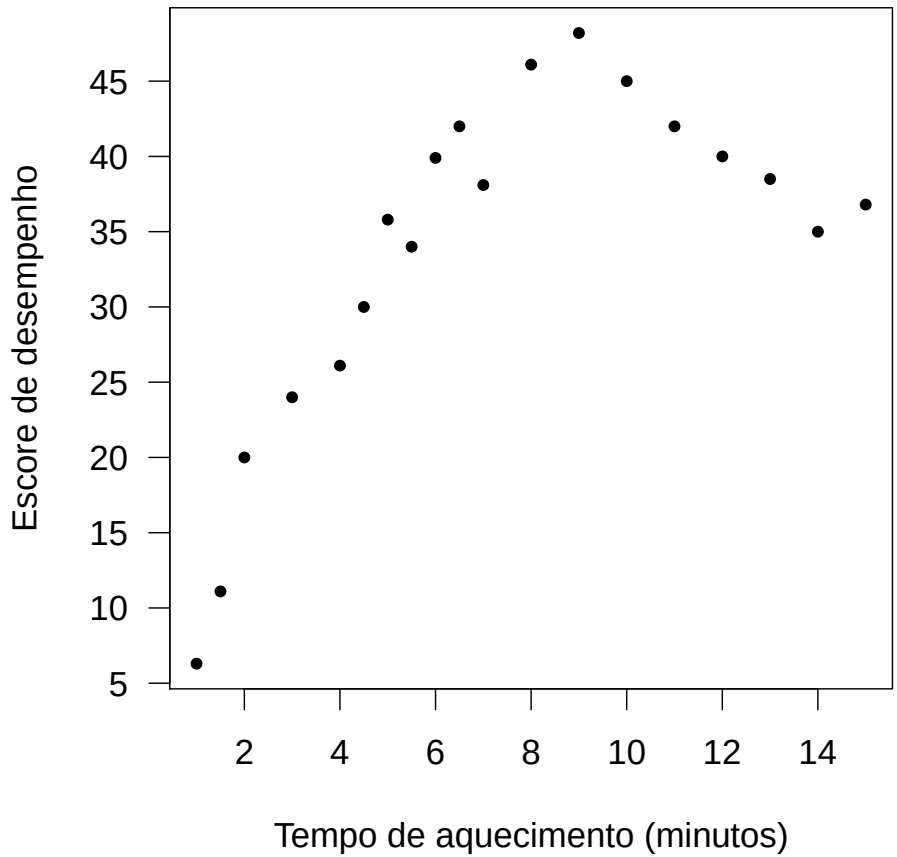

Figura 8: Gráfico de dispersão para os tempos de aquecimento e desempenho dos ginastas

Exemplo - Relação entre tempo de aquecimento e desempenho de ginastas

- A Figura 8 claramente não sugere relação linear entre as variáveis.
- Um polinômio de grau 2, aparentemente, pode explicar bem a relação entre o tempo de aquecimento e o desempenho.
- Nesse caso, vamos considerar o modelo

$$y = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon, \tag{5}$$

que, como pode ser verificado, corresponde a um modelo de regressão linear (é um modelo polinomial).

- O modelo ajustado está representado na Figura 9. Aparentemente, a relação entre as variáveis é bem explicada por um polinômio de grau 2.

Exemplo - Relação entre tempo de aquecimento e desempenho de ginastas

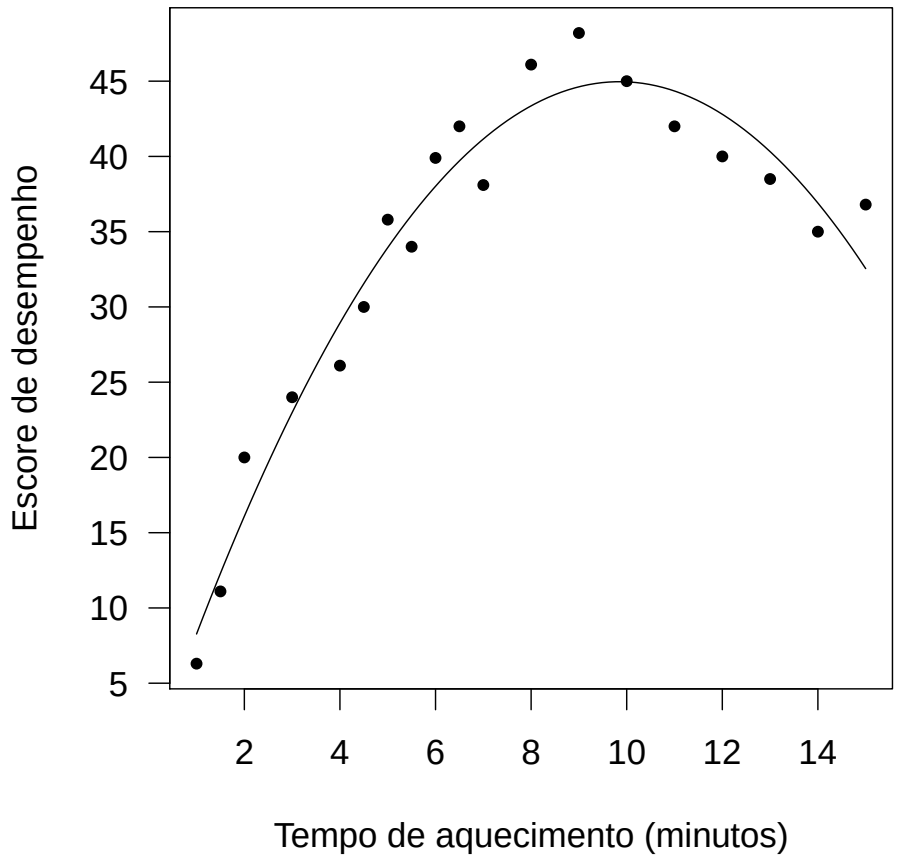

Figura 9: Gráfico de dispersão para os dados dos ginastas com modelo de regressão ajustado

Exemplo - Reação química

- Dados referentes à velocidade de uma reação enzimática (y, expressa em $contagem/min^2$) e à concentração de certo substrato (x, em partes por milhão) em 12 repetições de um experimento.

Tabela 2: Dados - reação química

Repetição	x	y	Repetição	x	y
1	0.02	47	7	0.02	76
2	0.06	97	8	0.06	107
3	0.11	123	9	0.11	139
4	0.22	152	10	0.22	159
5	0.56	191	11	0.56	201
6	1.10	200	12	1.10	207

Exemplo - Reação química

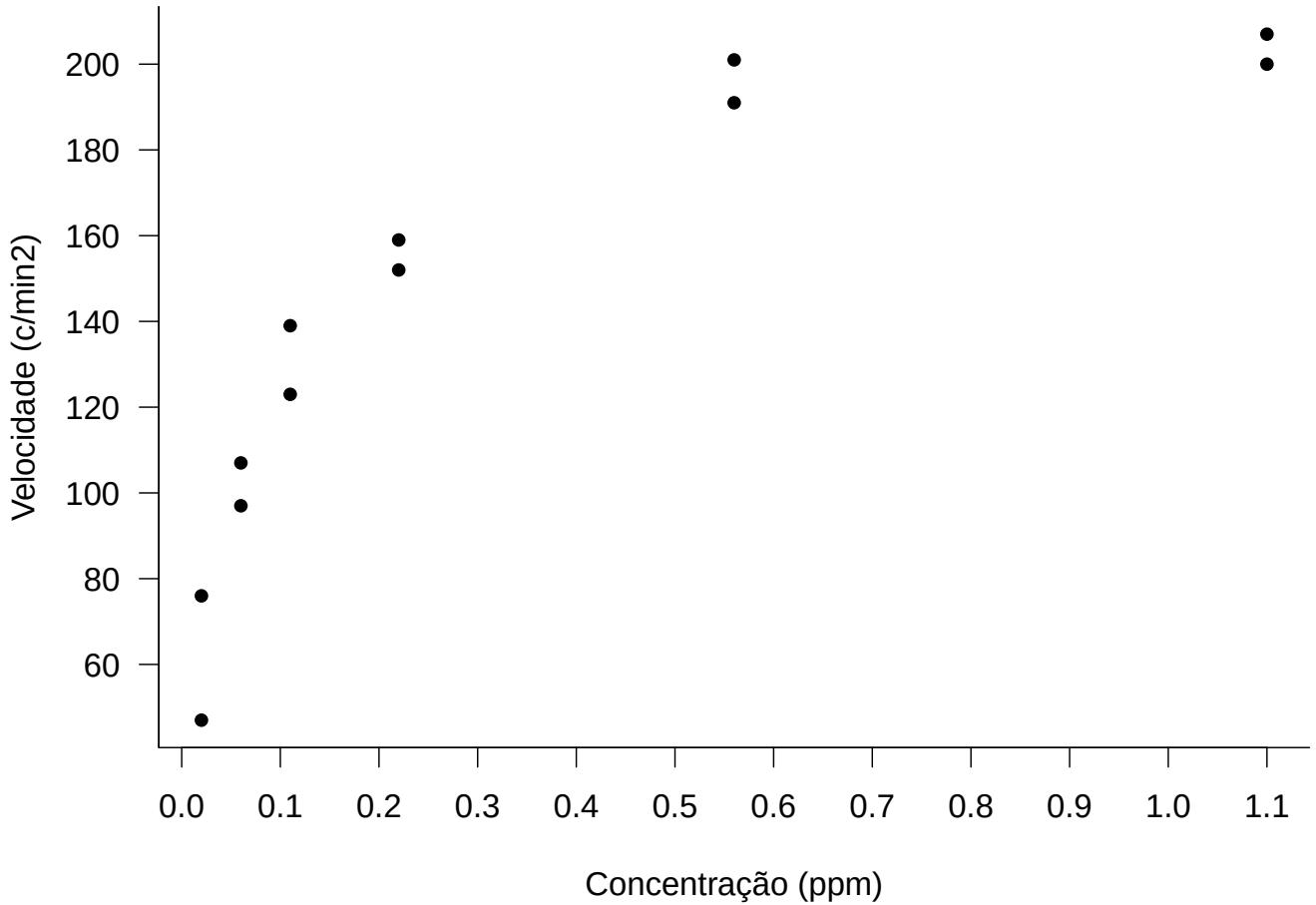

Figura 10: Gráfico de dispersão para os dados da reação química

Exemplo - Reação química

- Novamente observamos uma relação não linear entre as variáveis.

- Ao invés de considerar um modelo polinomial, podemos recorrer à teoria da Química e utilizar o modelo de Michaelis-Menten, que descreve adequadamente problemas de cinética enzimática.

- Dessa forma, vamos considerar $f(x, \boldsymbol{\beta}) = \frac{\beta_1 x}{x+\beta_2}$, ajustando o seguinte modelo de regressão não linear:

$$y = \frac{\beta_1 x}{x + \beta_2} + \epsilon. \tag{6}$$

Exemplo - Reação química

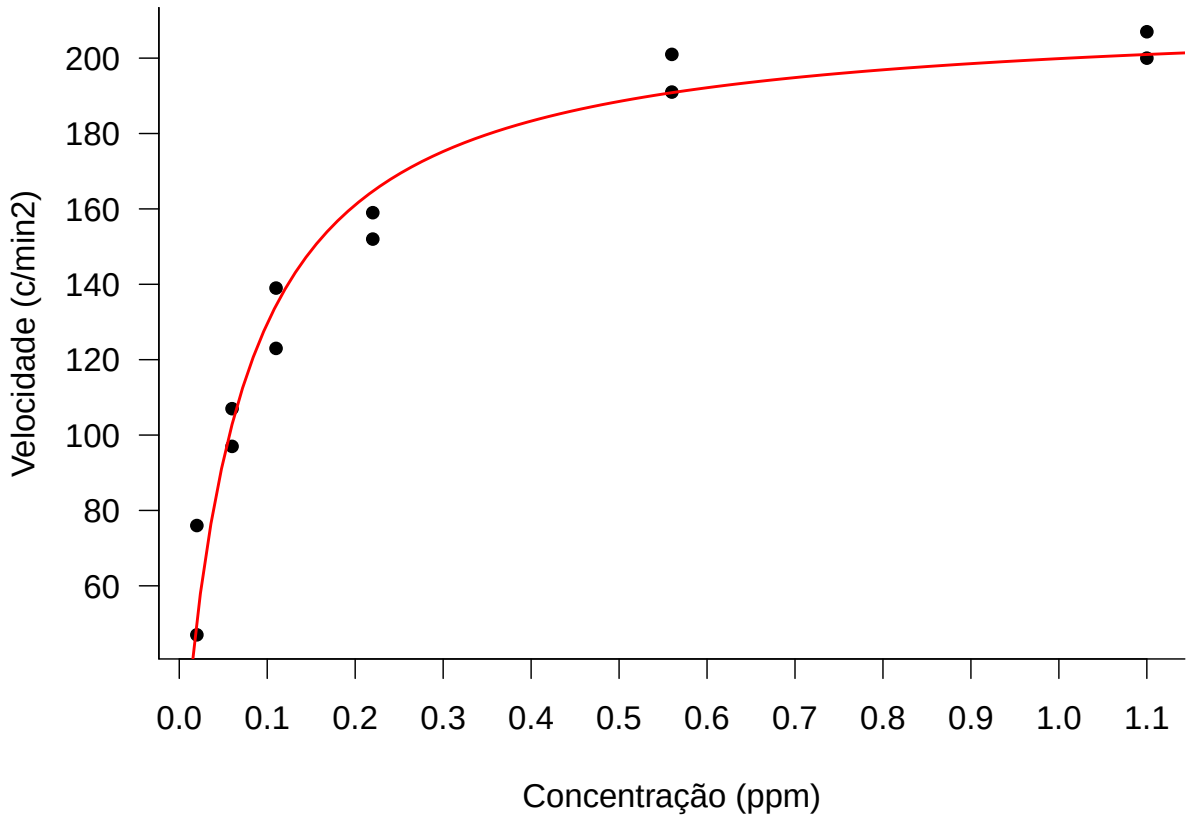

Figura 11: Gráfico de dispersão para os dados da reação química com o modelo de regressão não linear ajustado

Exemplo - Renda e tempo de serviço de profissionais

- Dados referentes ao tempo de serviço (x_1), renda (y, em reais) de profissionais de certo segmento. Além disso, outra variável a ser considerada é se os indivíduos possuem ou não curso superior:

$$x_2 = \begin{cases} 1, & \text{se possui curso superior} \\ 0, & \text{caso contrário} \end{cases}.$$

- Este exemplo envolve uma variável quantitativa e outra categórica. Para problemas como esse podemos usar **análise de covariância**, conforme estudaremos mais adiante.

Exemplo - Dados sobre rendimentos de profissionais de certo segmento

Tabela 3: Dados sobre renda e tempo de serviço de profissionais

Profissional	x_1	x_2	y	Profissional	x_1	x_2	y
1	21	0	3414	11	3	1	3414
2	6	0	3195	12	20	1	6928
3	10	0	3539	13	11	1	4651
4	11	0	3742	14	5	1	3836
5	24	0	4707	15	15	1	5595
6	24	0	5034	16	12	1	5172
7	15	0	4331	17	12	1	4732
8	10	0	3748	18	24	1	7612
9	17	0	3887	19	18	1	6478
10	17	0	4436	20	20	1	6779

Exemplo - Renda e tempo de serviço de profissionais

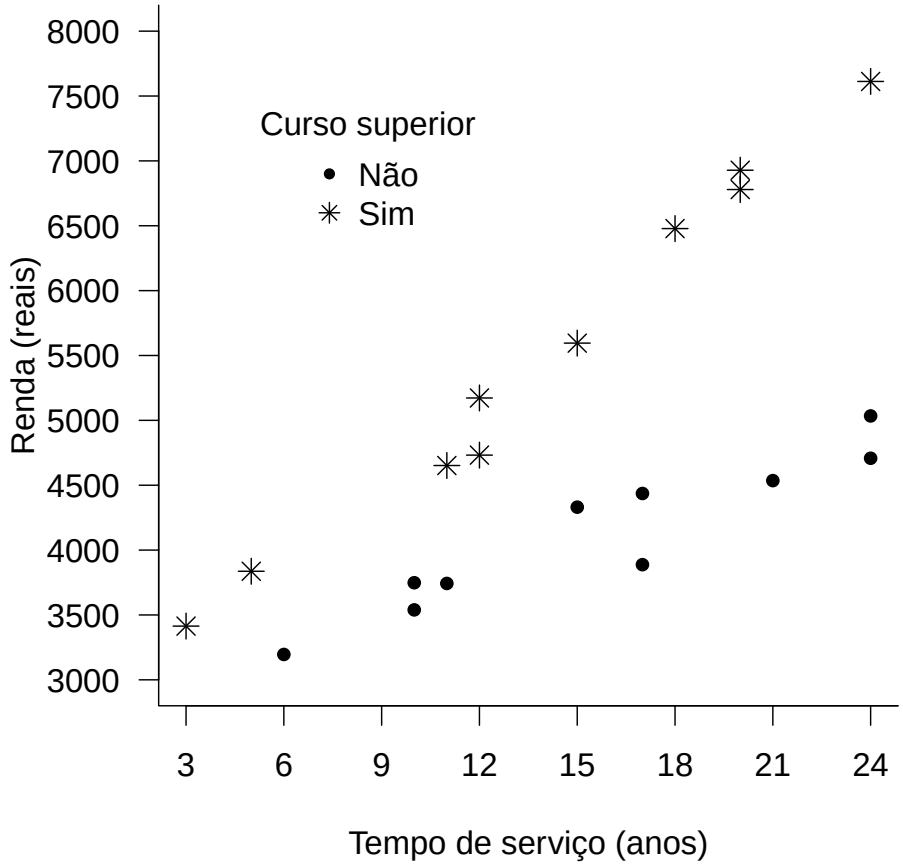

Figura 12: Relação entre renda e tempo de serviço.

Exemplo - Renda e tempo de serviço de profissionais

- Analisando o gráfico, parece haver uma relação (linear) crescente entre tempo de serviço e renda.
- Aparentemente, a relação não é a mesma para profissionais com e sem curso superior.
- Na sequência estão descritos diferentes modelos de regressão que poderíamos considerar na análise desses dados.

Exemplo - Renda e tempo de serviço de profissionais

- **Modelo 1:** Uma única reta é capaz de explicar a relação entre renda e tempo de serviço para os dois grupos (com e sem ensino superior):

$$y = \beta_0 + \beta_1 x_1 + \epsilon; \tag{7}$$

- **Modelo 2:** Cada grupo tem sua própria reta, mas as taxas de variação são as mesmas (apenas os interceptos são diferentes):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon. \tag{8}$$

- As retas de regressão implicadas pelo modelo 2 são as seguintes:

$$\begin{aligned} &y = \beta_0 + \beta_1 x_1 + \epsilon \text{ (para os que não possuem curso superior);} \\ &y = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon \text{ (para os que possuem curso superior).} \end{aligned} \tag{9}$$

Exemplo - Renda e tempo de serviço de profissionais

- **Modelo 3:** Cada grupo tem sua própria reta, com interceptos e taxas de variação distintos:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon. \tag{10}$$

- As retas de regressão implicadas pelo modelo 3 são as seguintes:

$$y = \beta_0 + \beta_1 x_1 + \epsilon \text{ (para os que não possuem curso superior);}$$
$$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon \text{ (para os que possuem curso superior).} \tag{11}$$

Exemplo - Renda e tempo de serviço de profissionais

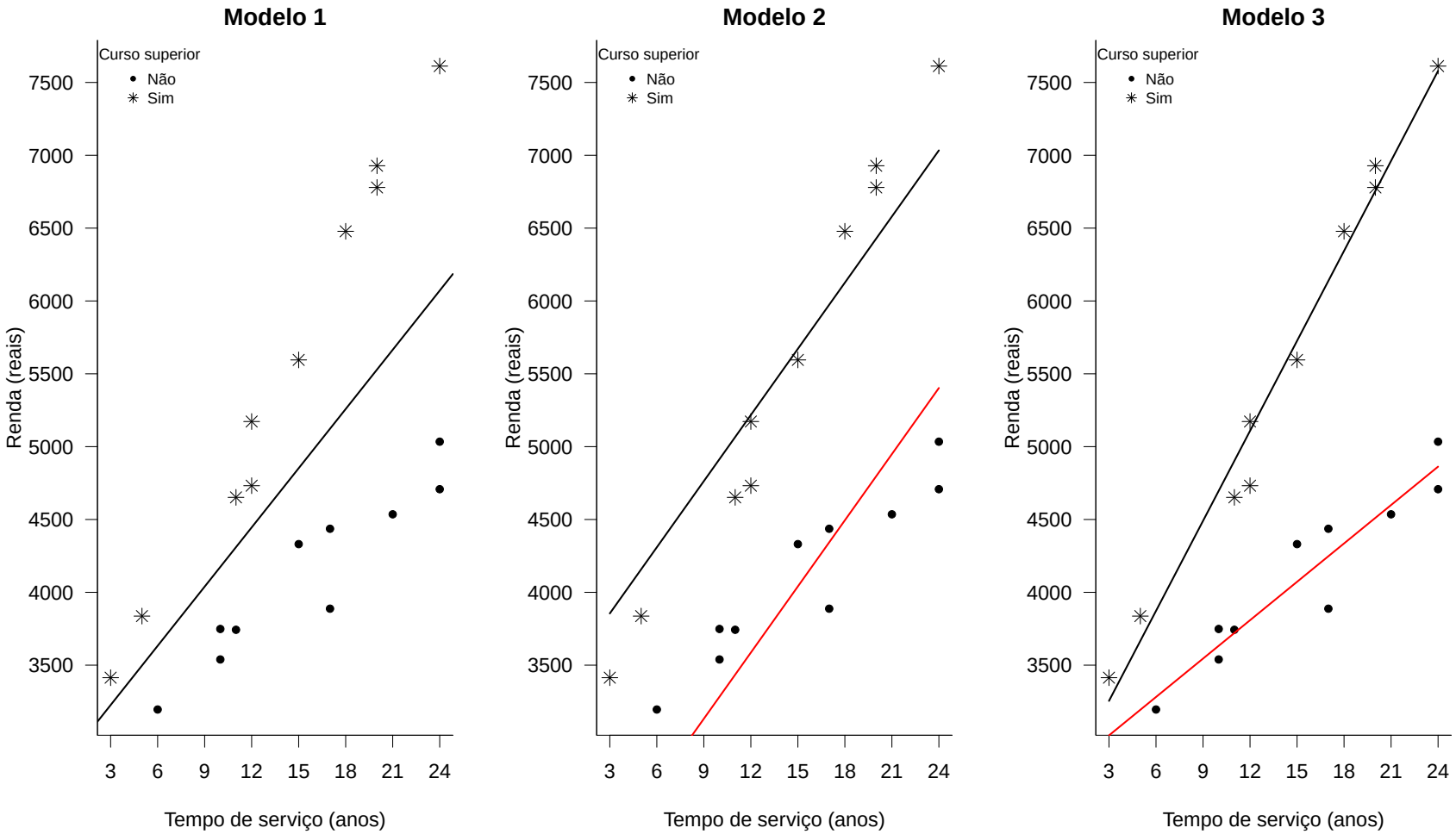

Figura 13: Modelos de regressão ajustados para os dados de renda e tempo de serviço.

Exemplo - Predição de massa gorda

- Base de dados `fat` do pacote `faraway`. O objetivo é ajustar um modelo de regressão para predizer a porcentagem de massa gorda de um indivíduo (`brozek`) com base nas seguintes variáveis:
- `neck`: Circunferência do pescoço;
- `chest`: Circunferência do peito;
- `abdom`: Circunferência do abdomen;
- `hip`: Circunferência do quadril;
- `thigh`: Circunferência da coxa;
- `knee`: Circunferência do joelho;
- `ankle`: Circunferência do tornozelo;
- `biceps`: Circunferência do bíceps estendido;
- `forearm`: Circunferência do antebraço;
- `wrist`: Circunferência do pulso.

Exemplo - Predição de massa gorda

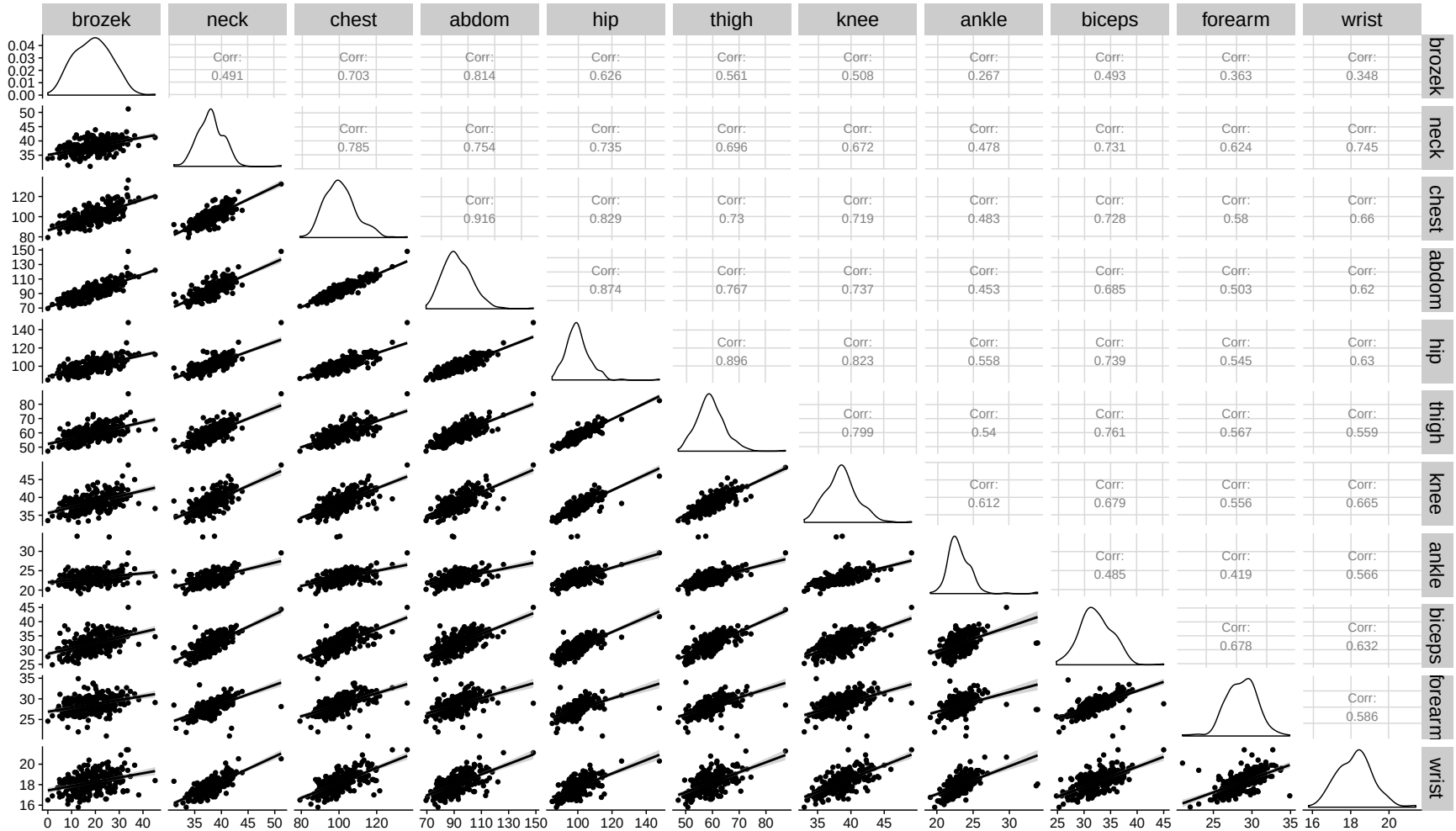

Figura 14: Modelos de regressão ajustados para os dados de renda e tempo de serviço.

Exemplo - Predição de massa gorda

- De maneira geral, as medidas realizadas estão (positivamente) correlacionadas com o percentual de massa gorda;
- Além disso, podemos notar, conforme esperado, que as medidas de circunferências estão correlacionadas entre si;
- O ajuste de um modelo de regressão linear múltipla permitiria predizer o percentual de massa gorda usando as medidas realizadas;
- Com base nos resultados obtidos, poderíamos avaliar a associação de cada medida com o percentual de massa gorda (ajustada pelos valores das demais medidas), e, eventualmente, selecionar as mais importantes para predição.

Mas afinal, de onde vem os modelos de regressão?

- A teoria (física) pode sugerir o modelo. Por exemplo, segundo a lei de Ohm a voltagem aplicada nos terminais de um condutor é proporcional à corrente elétrica que o percorre. Logo, a relação entre as variáveis é linear, induzindo o modelo;

- Experiência com dados de estudos anteriores. Se um particular modelo proporcionou um bom ajuste aos dados em um estudo similar, possivelmente ele vai se ajustar bem aos dados do presente estudo;

- Não há qualquer modelo indicado a priori. A escolha de um modelo pode resultar da exploração dos dados e comparação de diferentes especificações.