Probabilidade e Estatística

Eduardo Horta

Correlação

Coeficiente de correlação

- Se X e Y são duas variáveis aleatórias em uma mesma população (ou seja, definidas em um mesmo espaço amostral), então definimos o **coeficiente de correlação entre X e Y** como sendo o parâmetro

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

onde σ_X é o desvio padrão de X, e σ_Y é o desvio padrão de Y.

Coeficiente de correlação

- Lembrando:
  - $\mathrm{Cov}(X, Y) = \mathbb{E}\{(X - \mathbb{E}X)(Y - \mathbb{E}Y)\} = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$;
  - $\sigma_X^2 = \mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$, etc.
- O *parâmetro* ρ é uma medida do grau de associação *linear* entre as variáveis X e Y.
- Em alguns casos é conveniente explicitar, na notação, as variáveis às quais o coeficiente se refere: escrevemos $\rho(X, Y)$ ou $\rho_{X,Y}$ ao invés de apenas ρ.

Coeficiente de correlação

- O coeficiente de correlação é sempre um número entre -1 e 1: $-1 \leq \rho \leq 1$.
- De fato, como

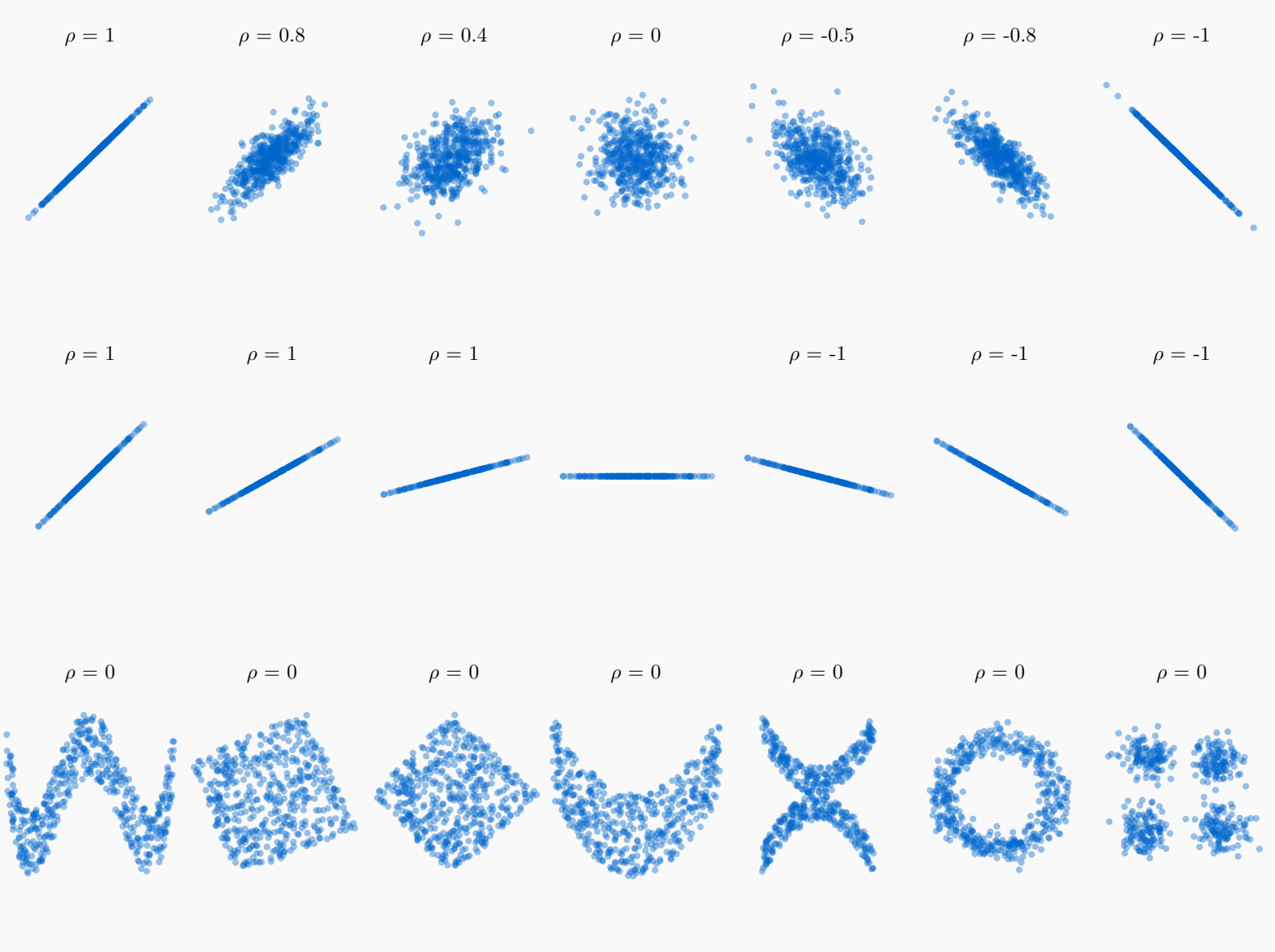
$$\left(\frac{X - \mathbb{E}X}{\sigma_X} \pm \frac{Y - \mathbb{E}Y}{\sigma_Y} \right)^2 \geq 0,$$

tomando o valor esperado em ambos os lados da desigualdade acima e expandindo o quadrado, obtemos a quota desejada $|\rho| \leq 1$.

Coeficiente de correlação

- O valor de ρ nos permite interpretar o tipo de relação entre as variáveis X e Y, da seguinte forma:
  - $\rho = 1$: associação linear *positiva* perfeita entre X e Y.
  - $0 < \rho < 1$: associação linear *positiva* entre X e Y (quanto mais próximo de 1, mais forte é a associação linear entre essas variáveis. Quanto mais próximo de zero, mais fraca é tal associação);
  - $\rho = 0$: ausência de associação *linear* entre X e Y (embora possa haver dependência *não linear* entre elas);
  - $-1 < \rho < 0$: associação linear *negativa* entre X e Y (quanto mais próximo de -1, mais forte é a associação linear entre essas variáveis. Quanto mais próximo de zero, mais fraca é tal associação);
  - $\rho = -1$: associação linear *negativa* perfeita entre X e Y.

- Na figura abaixo temos o gráfico de dispersão de amostras **simuladas** $(X_1, Y_1), \ldots, (X_n, Y_n)$, de acordo com diferentes modelos populacionais.



Estimação pontual de ρ

Estimação pontual de ρ

- Dada uma amostra aleatória $(X_1, Y_1), \ldots, (X_n, Y_n)$ de uma população descrita por uma função de distribuição $F_{X,Y}$ com coeficiente de correlação $\rho = \rho(X, Y)$, um estimador natural para ρ é dado por

$$\widehat{\rho} = \frac{\sum_{i=1}^{n} X_i Y_i \ - \ n\bar{X}\bar{Y}}{\sqrt{(\sum_{i=1}^{n} X_i^2 \ - \ n\bar{X}^2) \cdot (\sum_{i=1}^{n} Y_i^2 \ - \ n\bar{Y}^2)}}$$

- Pode-se mostrar que, nessas condições, o estimador $\widehat{\rho}$ definido acima é consistente para ρ (embora seja viesado).

Testes de hipóteses para ρ

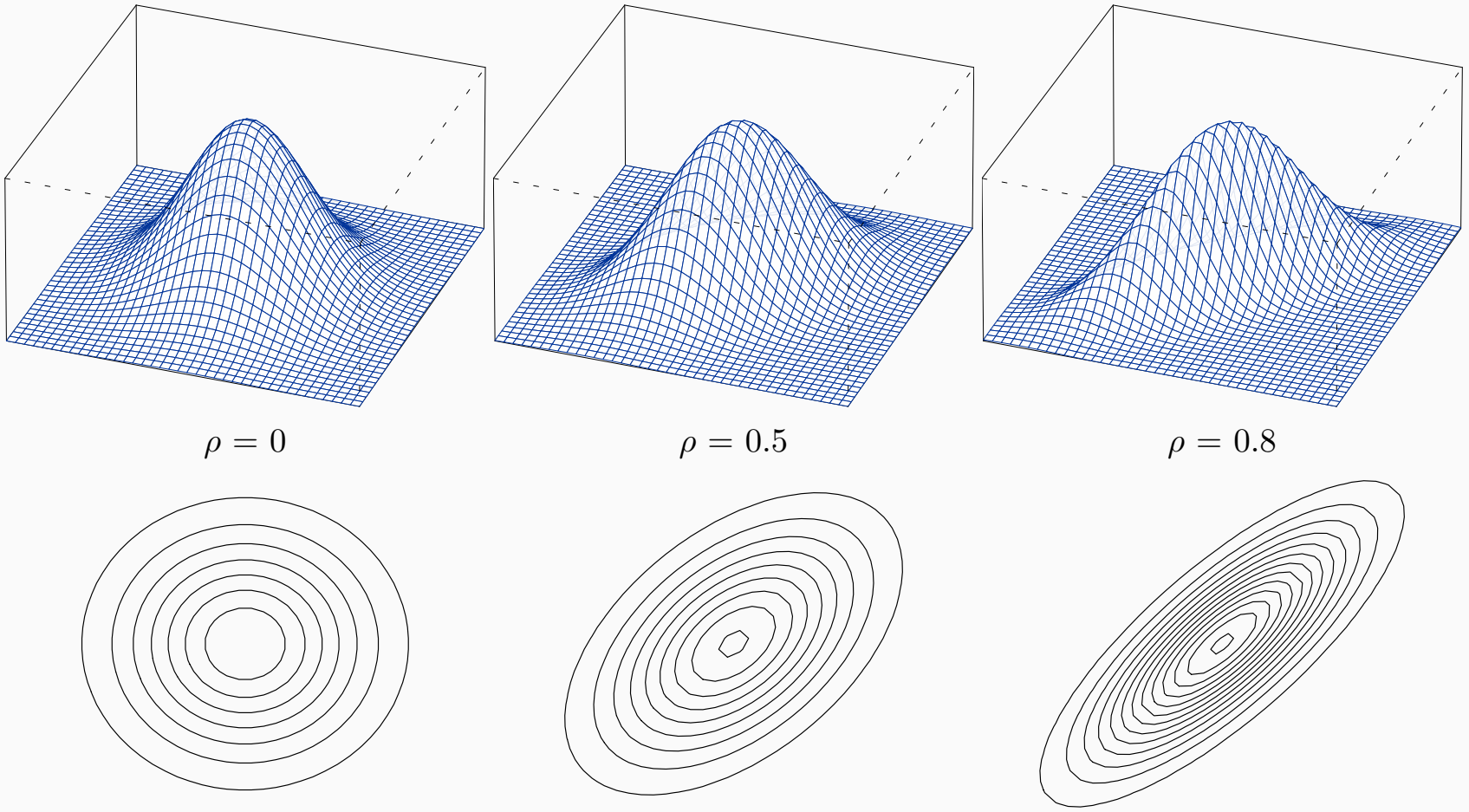
Testes de hipóteses para ρ

- A estatística $\sqrt{n}(\widehat{\rho} - \rho)$ tem distribuição aproximadamente normal (se o tamanho da amostra, n, for suficientemente grande).
- Contudo, para explicitar a variância dessa estatística são necessárias suposições adicionais sobre a distribuição conjunta (modelo teórico) das variáveis X e Y.
- Por exemplo, se (X, Y) tem distribuição normal bivariada, então

$$\sqrt{n}(\widehat{\rho} - \rho) \overset{\text{aprox.}}{\sim} N(0, (1 - \rho^2)^2),$$

e com isso podemos testar hipóteses sobre ρ (ver mais abaixo).

- Lembrando: distribuição *normal bivariada*



Testes de hipóteses para ρ

- A estatística de teste que usaremos para testar hipóteses sobre o coeficiente de correlação vai depender da hipótese nula que propusermos.
- Em qualquer caso, a validade do teste de hipóteses se baseia na suposição inicial de que o vetor aleatório (X, Y) tem distribuição normal bivariada.

Testes de hipóteses para ρ

- **Caso 1**. "$H_0 : \rho = 0$". Nesse caso, a estatística de teste a ser usada é

$$T = \widehat{\rho}\sqrt{\frac{n-2}{1-\widehat{\rho}^2}},$$

a qual tem distribuição t de Student com $n-2$ graus de liberdade (em particular, para valores de n razoavelmente grandes, essa distribuição é aproximadamente normal padrão).

Testes de hipóteses para ρ

- **Caso 2**. "$H_0 : \rho = \rho_0$" (onde ρ_0 é um valor numérico fixo diferente de 0). Nesse caso, a estatística de teste a ser usada é

$$\xi = \frac{\sqrt{n-3}}{2}(\ln(\frac{1+\widehat{\rho}}{1-\widehat{\rho}}) - \ln(\frac{1+\rho_0}{1-\rho_0})),$$

  a qual tem distribuição aproximadamente Normal padrão (isto é, com média igual a zero e variância igual a 1).

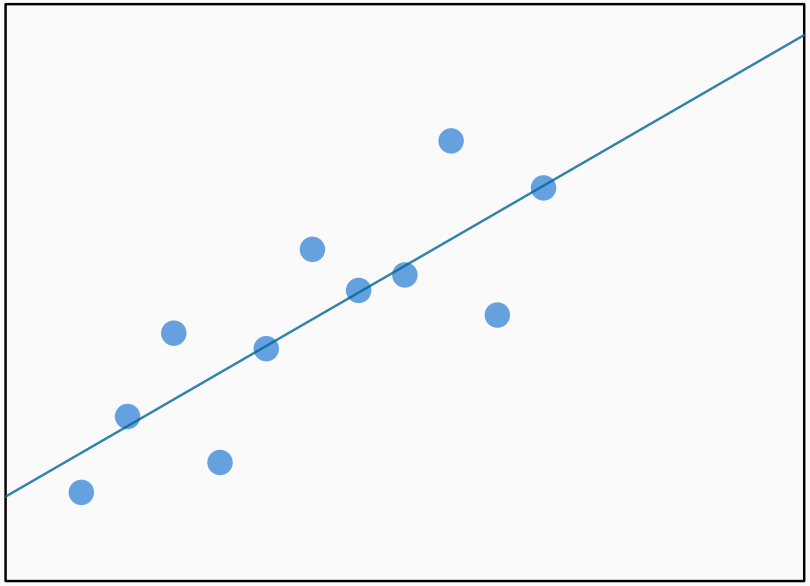
Coeficiente de correlação: alguns comentários

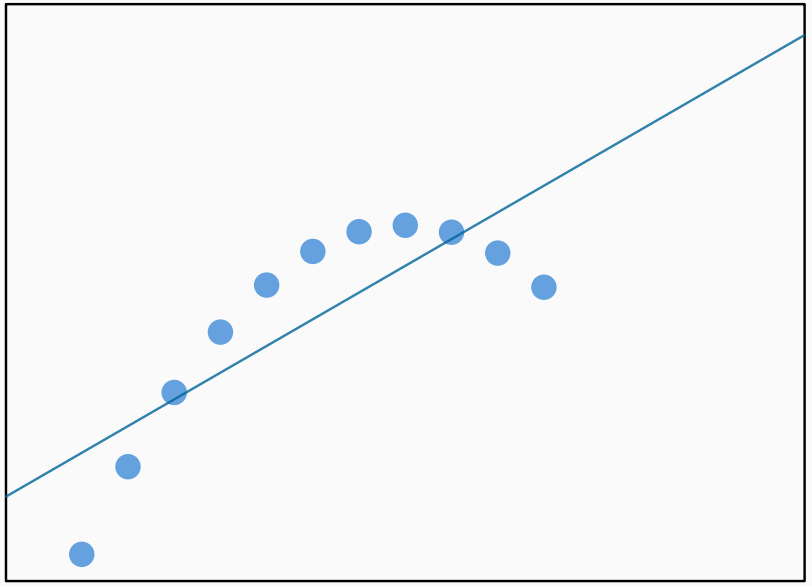
Correlação: alguns comentários

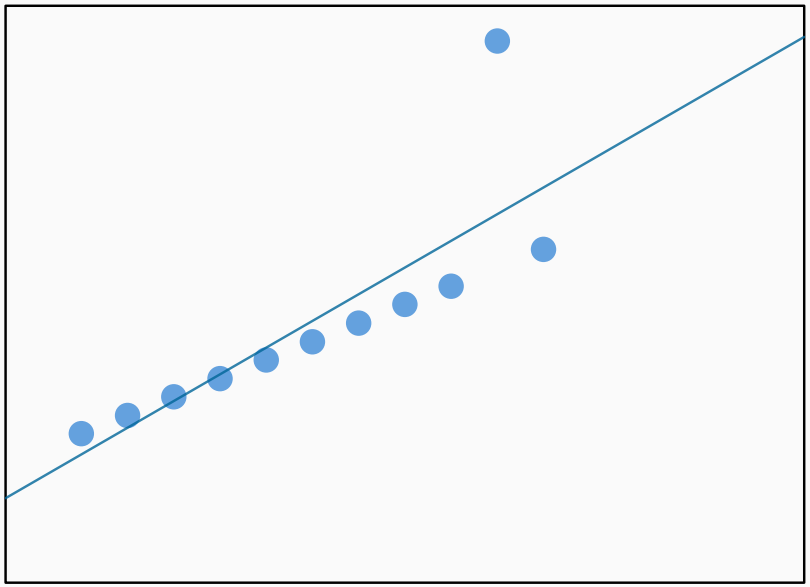
- Conforme mencionado anteriormente, o coeficiente de correlação captura apenas o grau de associação *linear* entre duas variáveis aleatórias.
- Para dependências de tipo não-linear, o coeficiente de correlação deve ser interpretado com cautela.
- Vejamos mais alguns exemplos:

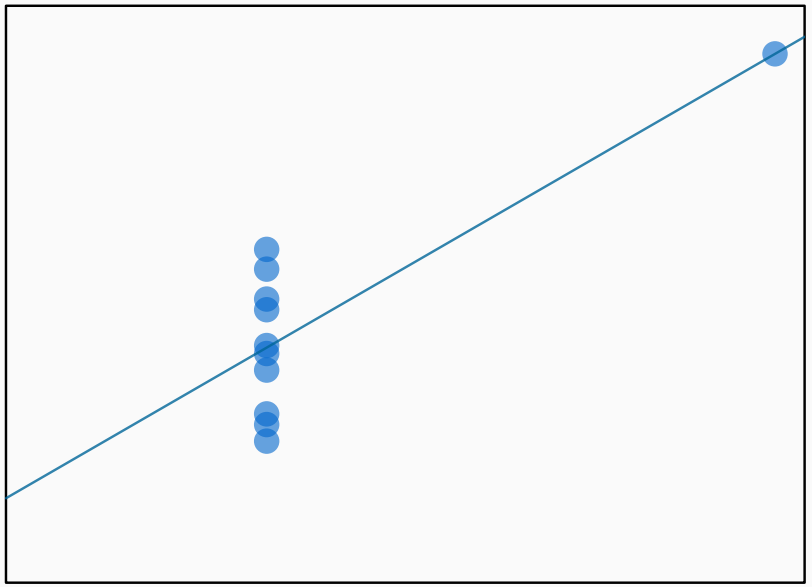
Exemplo

- Os quatro conjuntos de dados na figura abaixo, embora bastante distintos, possuem o mesmo coeficiente de correlação (amostral): $\widehat{\rho} = 0.8165214$









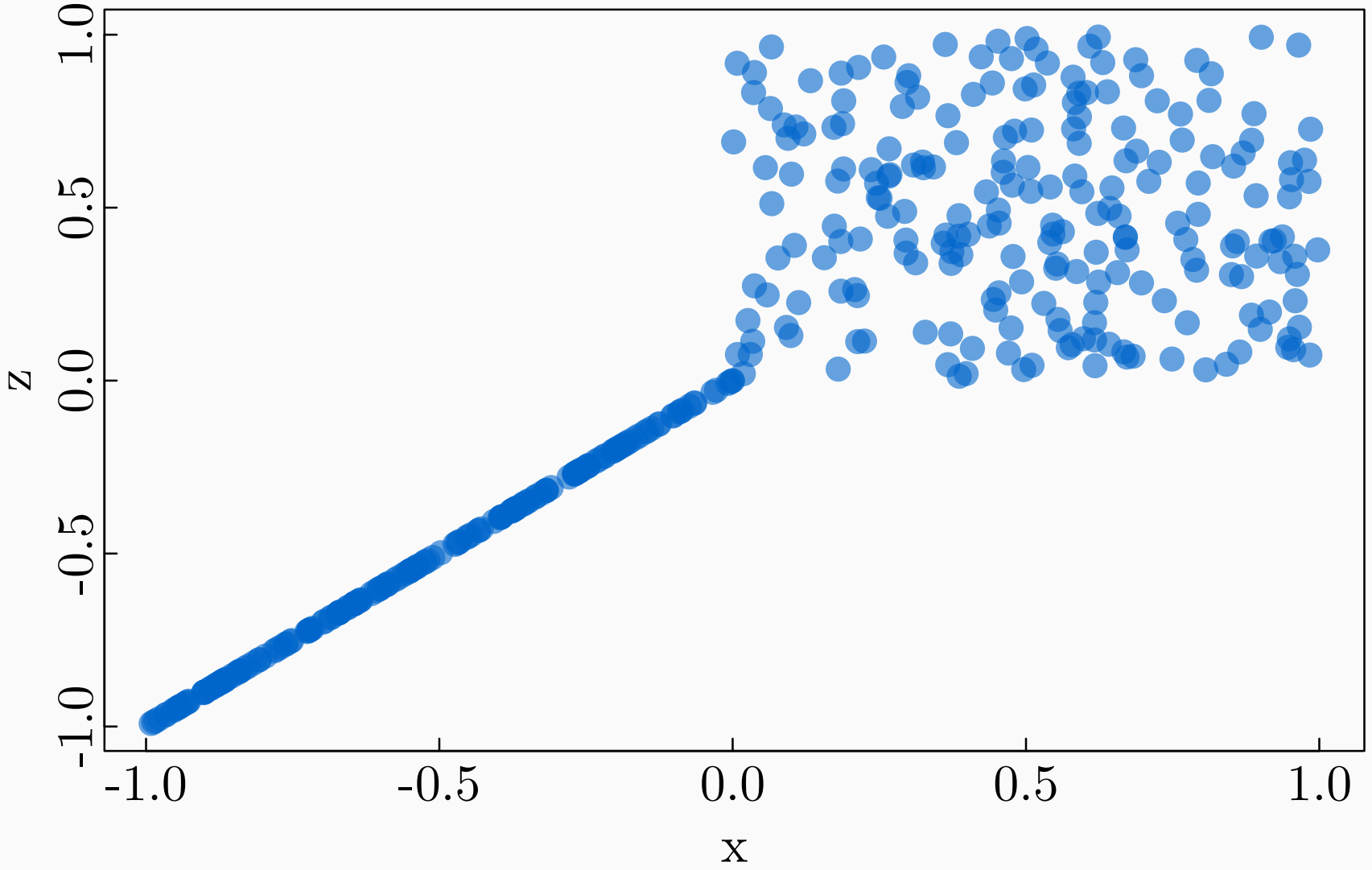
Exemplo

- Considere variáveis aleatórias $X \sim$ Uniforme$[-1, 1]$ e $Y \sim$ Uniforme$[0, 1]$.
- Seja Z a variável aleatória

$$Z = \begin{cases} X, & \text{se } X \leq 0; \\ Y, & \text{se } X > 0. \end{cases}$$

- Pode-se mostrar que o coeficiente de correlação entre X e Z é maior do que 0.87, algo que poderia ser interpretado como uma associação muito forte entre essas variáveis.
- Abaixo está o gráfico de dispersão de uma amostra simulada, de tamanho $n = 400$, dessas variáveis.

Exemplo



Coeficiente de correlação vs causalidade

- **Importante**: *correlação* não significa *causalidade*. Duas variáveis podem ser fortemente correlacionadas sem que uma delas seja causa (parcial) da outra.
- Por exemplo: numerosos estudos epidemiológicos mostraram que mulheres submetidas a terapia de reposição hormonal (TRH) também apresentavam uma incidência abaixo da média de doenças coronárias; isto é, esses estudos apontam para uma correlação negativa entre exposição a TRH e incidência de doenças coronárias.

Coeficiente de correlação vs causalidade

- Levantou-se a hipótese de que a TRH seria preventiva de doenças coronárias; todavia, uma análise mais cuidadosa dos dados mostrou que mulheres submetidas a TRH provinham principalmente de classes econômicas ABC, as quais apresentam dietas e práticas de atividades físicas melhores do que a média populacional.
- Conclui-se que não havia uma relação causal entre exposição a TRH e redução do risco cardíaco, mas sim um terceiro fator causando ambas (no caso, boa alimentação e prática de exercícios, mais prevalentes nas classes ABC que também tendem a ser mais expostas a TRH).

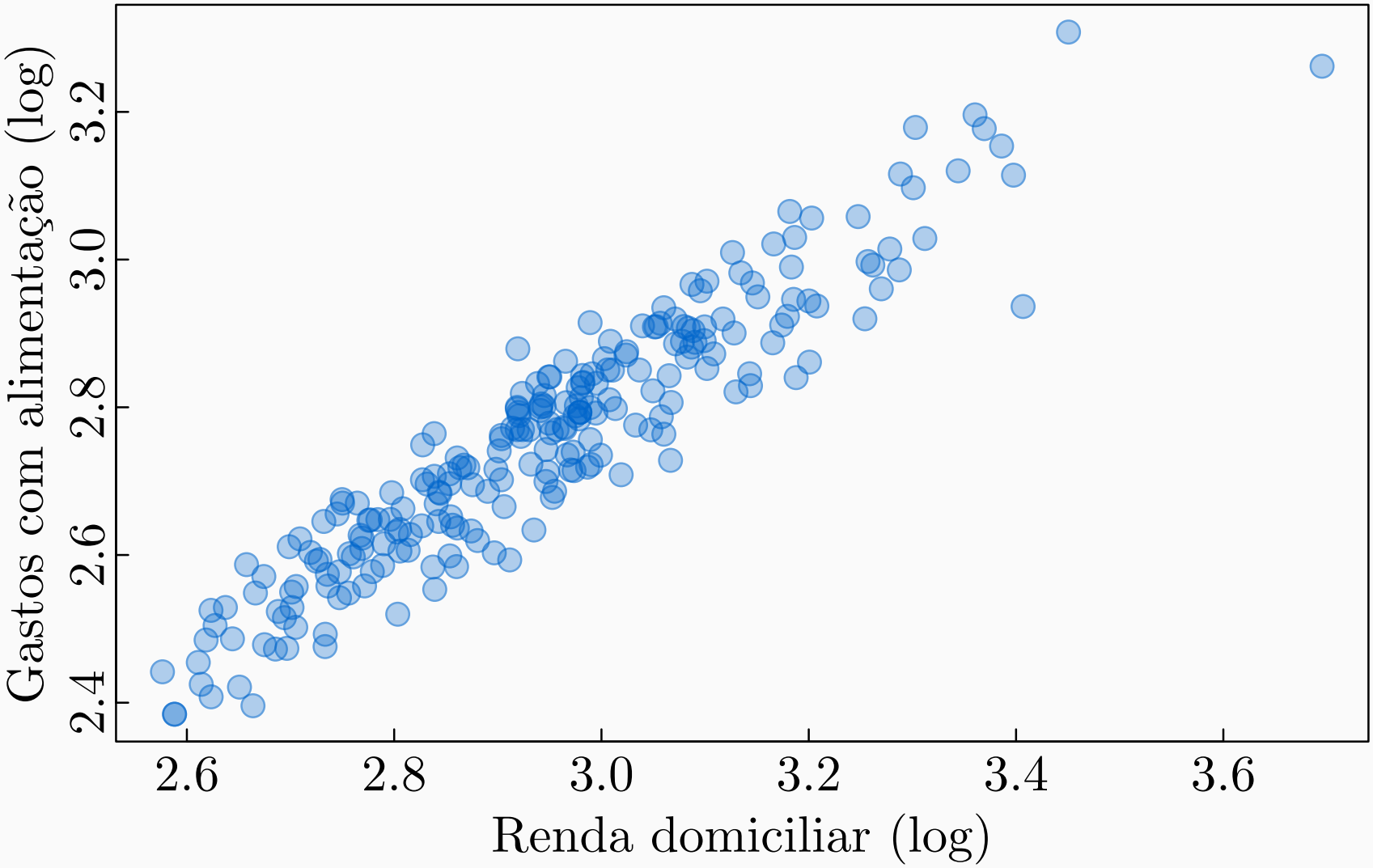
Coeficiente de correlação vs causalidade

- Ver também: https://www.tylervigen.com/spurious-correlations

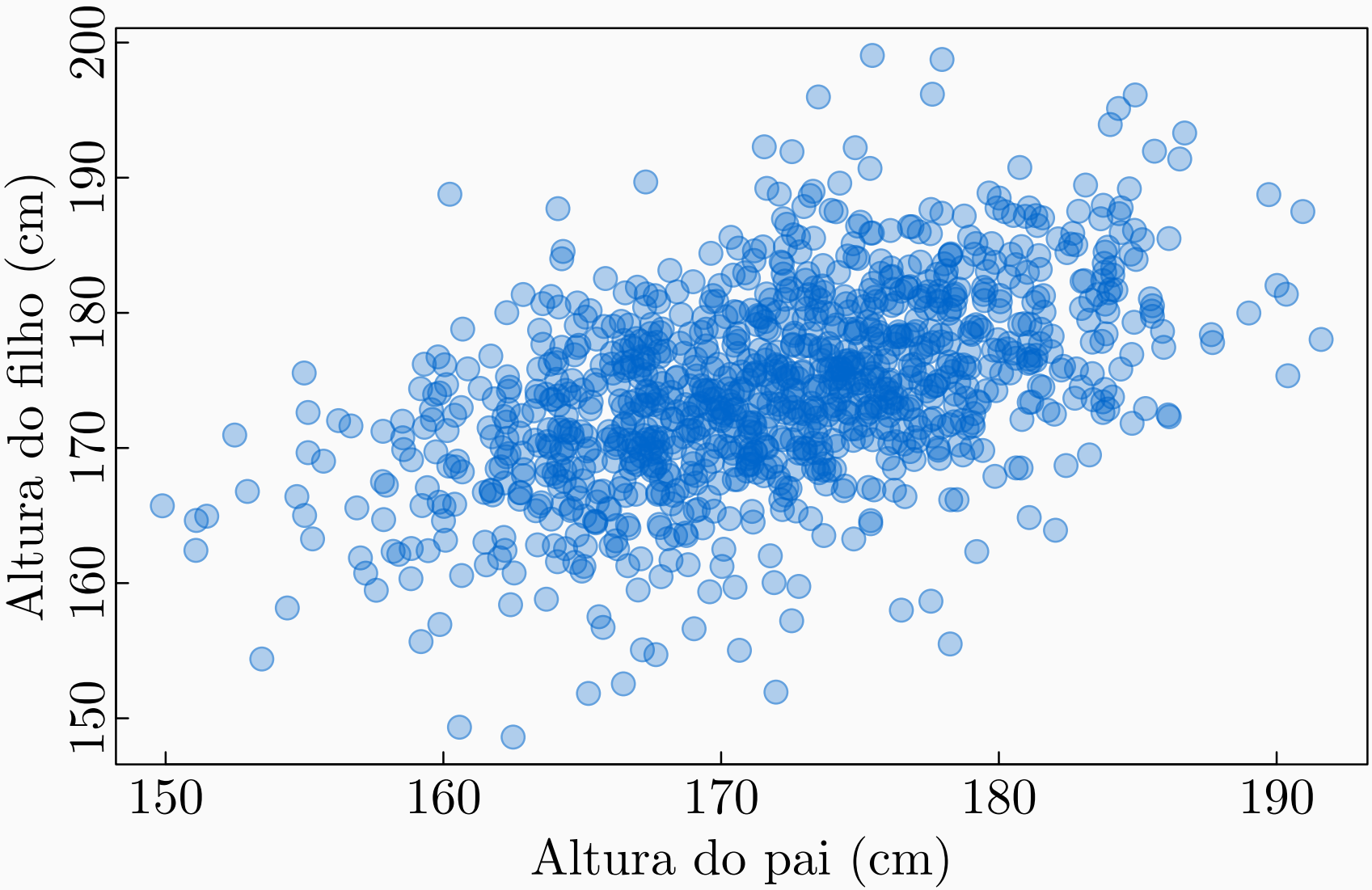
Modelos de regressão linear

Modelos de regressão linear

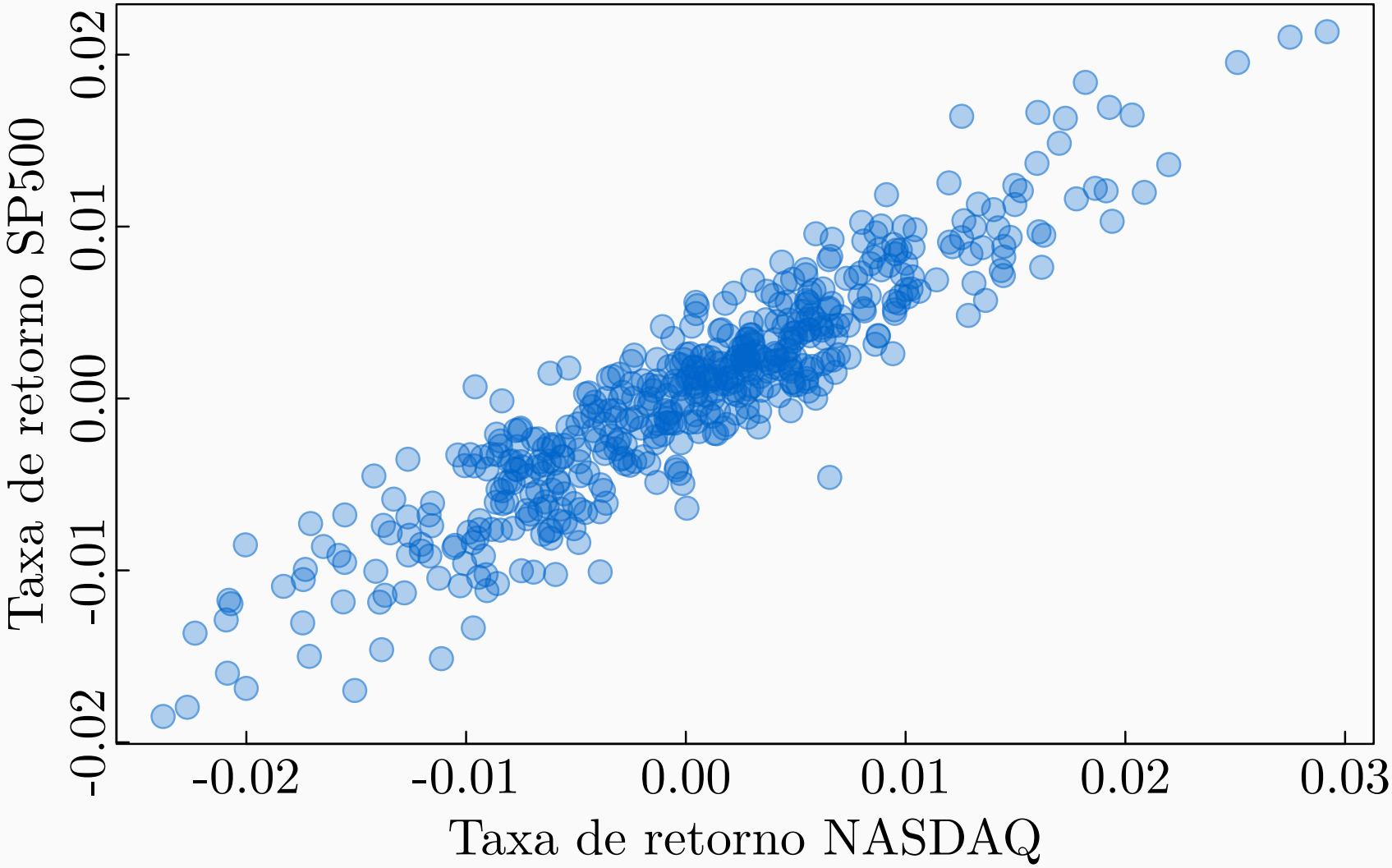
- Vamos começar com alguns exemplos (utilizando dados reais!):



Modelos de regressão linear



Modelos de regressão linear



Modelos de regressão linear

- O modelo de regressão linear estipula que a relação entre duas variáveis aleatórias X e Y é descrita pela equação

$$Y = \alpha + \beta X + \varepsilon,$$

  onde assume-se que $\mathbb{E}(\varepsilon \mid X = x) = 0$.
- Aqui, α e β são números reais fixos (isto é, *parâmetros*), enquanto X, Y e ε são variáveis aleatórias.
- Observação: nesse caso, pode-se mostrar que

$$\rho(X, Y) = \beta \frac{\sigma_X}{\sigma_Y}.$$

  onde $\sigma_X^2 = \mathrm{Var}(X)$, etc.

Modelos de regressão linear: terminologia

- A variável Y é dita o **regressando**, ou a **variável explicada**, ou a **resposta**.
- A variável X é dita o **regressor**, ou **a variável explicativa**, ou a **covariável**, ou o **preditor**.
- A variável ε é dita o **termo de erro**, ou **ruído**.
- O parâmetro α é dito o **intercepto** do modelo.
- O parâmetro β é dito o **coeficiente de inclinação** do modelo.
- A variável aleatória ε é dita o **termo de erro** do modelo.

- Observação: em alguns contextos, é usual chamar Y de variável dependente, e X de variável independente. Essa terminologia pode se sobrepor ao conceito de variáveis aleatórias independentes e, portanto, deve ser evitada.

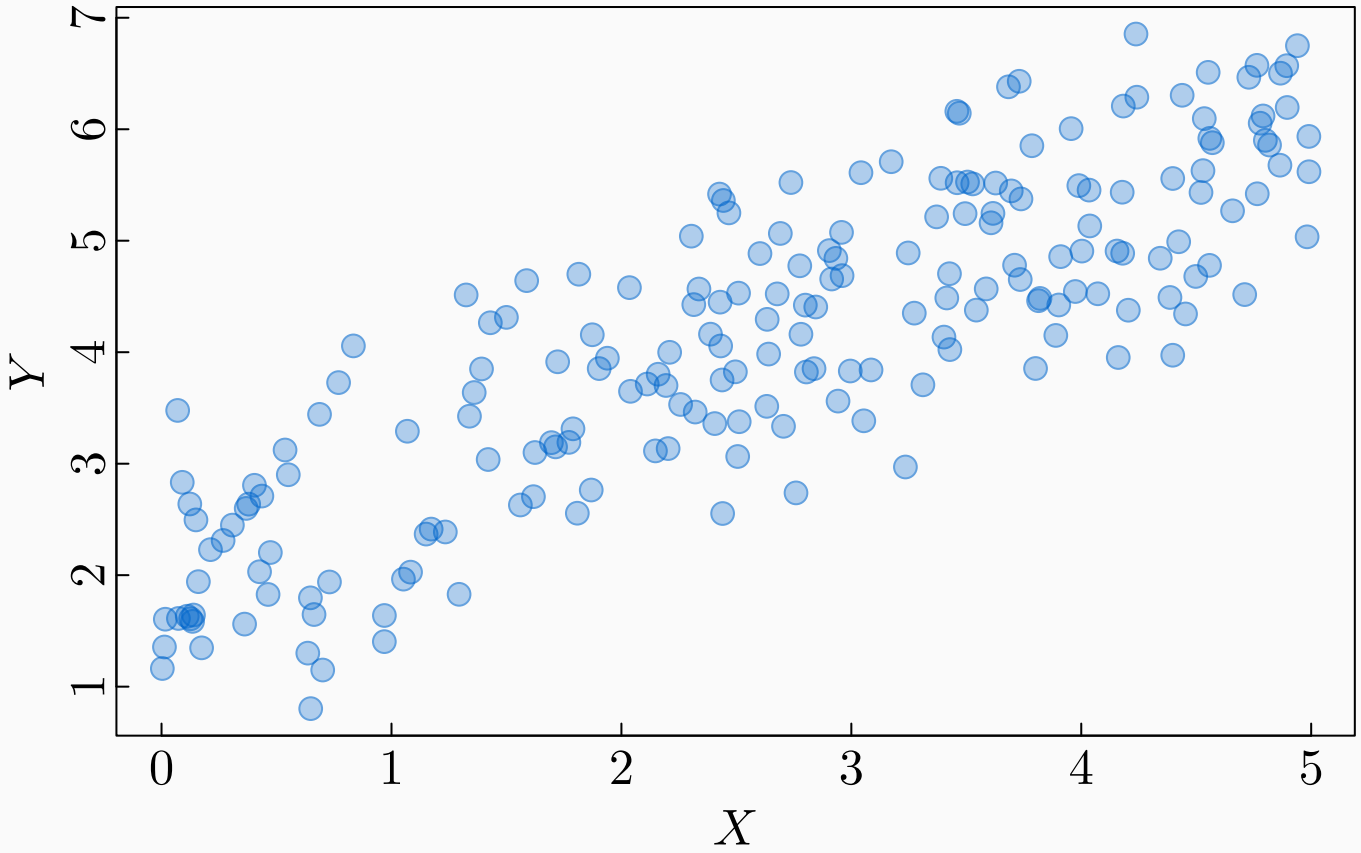
Modelos de regressão linear: comentários

- A ideia do modelo proposto é que somente o preditor X e a resposta Y são observáveis, mas o termo de erro ε não. Além disso, α e β são parâmetros desconhecidos;
- **Em aplicações**, coletamos uma amostra aleatória $(X_1, Y_1), \ldots, (X_n, Y_n)$ de uma população a qual supostamente é descrita por um modelo linear $Y_i = \alpha + \beta X_i + \varepsilon_i$, com os parâmetros α e β desconhecidos, e onde os termos de erro ε_i não são observados.

Modelo de regressão linear: exemplo simulado

A figura a seguir mostra o gráfico de dispersão de uma amostra iid $(X_1, Y_1), \ldots, (X_n, Y_n)$, **simulada** do modelo $Y = 2 + 0.8\,X + \varepsilon$, onde $X \sim$ Uniforme$(0, 5)$ e $\varepsilon \sim$ Normal$(\mu_\varepsilon = 0,\, \sigma^2_\varepsilon = 0.7^2)$, com X e ε mutuamente independentes.

Modelo de regressão linear: exemplo simulado



Modelo de regressão linear: exemplo simulado

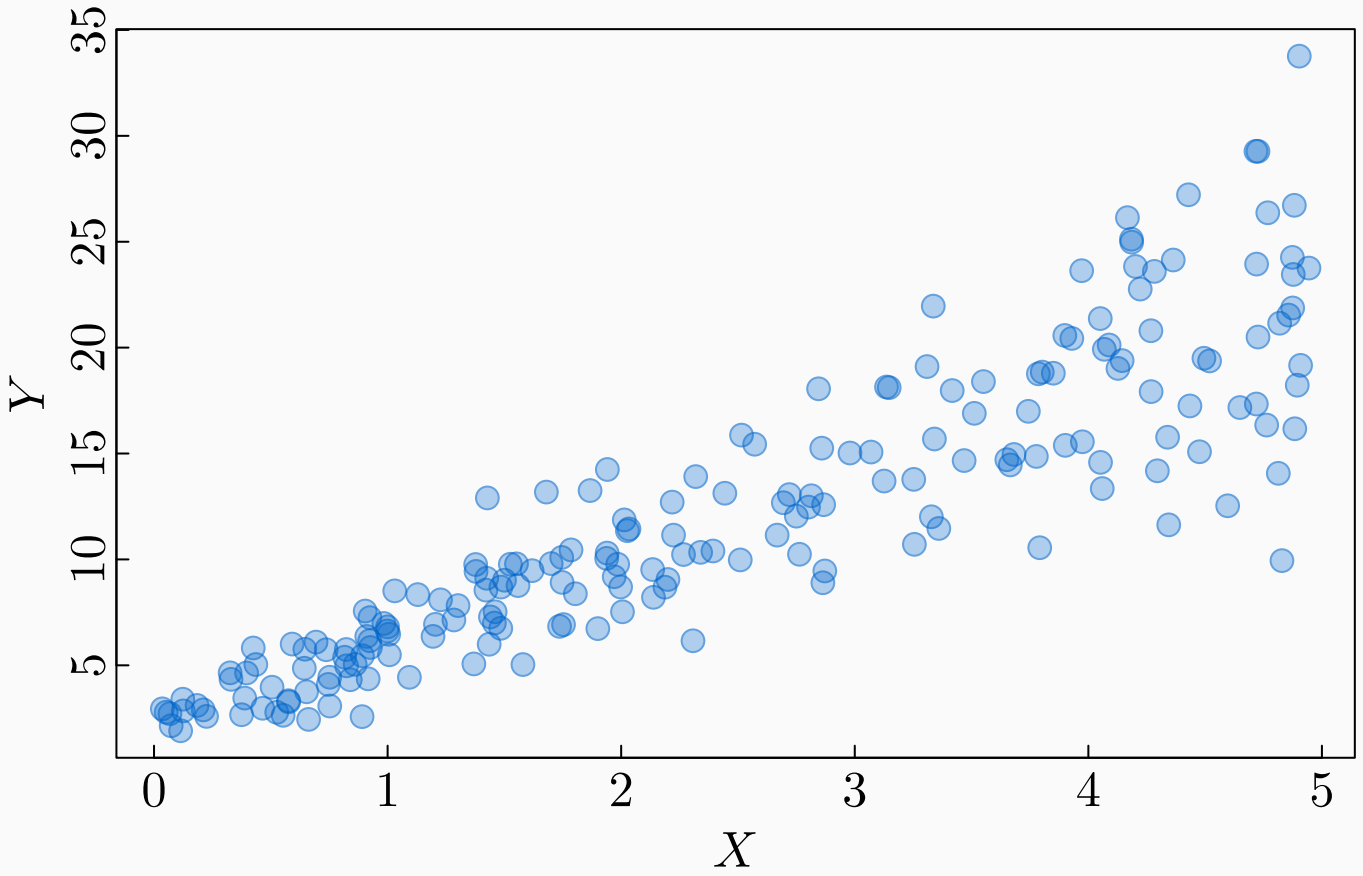
- A sintaxe em R para gerar o gráfico acima é a seguinte:

```
set.seed(2)

alpha = 2
beta = .8

n = 200

epsilon = rnorm(n, sd = .7)
X = runif(n, min = 0, max = 5)

Y = alpha + beta*X + epsilon
```

Modelo de regressão linear: outro exemplo simulado

A figura a seguir mostra o gráfico de dispersão de uma amostra aleatória $(X_1, Y_1), \ldots, (X_n, Y_n)$, **simulada** do modelo $Y = 2 + 4\,X + \varepsilon$, onde $X \sim$ Uniforme$(0, 5)$ e onde ε, condicional em $X = x$, tem distribuição Normal com $\mathbb{E}(\varepsilon|X = x) = 0$ e $\mathrm{Var}(\varepsilon|X = x) = 0.8 + 0.7X$.

Modelo de regressão linear: outro exemplo simulado



Modelos de regressão linear: comentários

- Algumas observações sobre o modelo linear: em muitas circunstâncias, é possível que se tenha algum tipo de controle sobre os valores da covariável X ao se executar um experimento ou coletar uma amostra.

Modelos de regressão linear: comentários

- Isso é particularmente verdadeiro em experimentos laboratoriais, em que o experimentador de fato pode determinar os níveis de X: por exemplo, se Y é a temperatura de ebulição da água e X é o nível de concentração de alguma substância química (digamos, sal), então podemos preestabelecer alguns níveis de concentração $X_1, X_2, \ldots, X_n$ (possivelmente com repetições), e efetuar medições da temperatura de ebulição da água nesses diferentes níveis.
- Em outras situações, tipicamente não se tem esse tipo de controle (por exemplo, dados econômicos, biológicos, meteorológicos). Nos três exemplos acima, a covariável é intrinsicamente aleatória.

Modelos de regressão linear: Estimação

Modelos de regressão linear: Estimação

- Utilizando tão somente os valores amostrais $(X_1, Y_1), \ldots, (X_n, Y_n)$, vamos propor estimadores $\widehat{\alpha}$ e $\widehat{\beta}$ para os parâmetros α e β, dados respectivamente por

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}, \qquad \widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X}$$

Modelos de regressão linear: Estimação

- O par $(\widehat{\alpha},\, \widehat{\beta})$ minimiza a função

$$L(a,\, b) = \sum_{i=1}^{n} (Y_i - a - bX_i)^2.$$

- Dito de outra forma: $(\widehat{\alpha},\, \widehat{\beta})$ é a solução do problema de minimização

$$\min_{a,b} L(a, b);$$

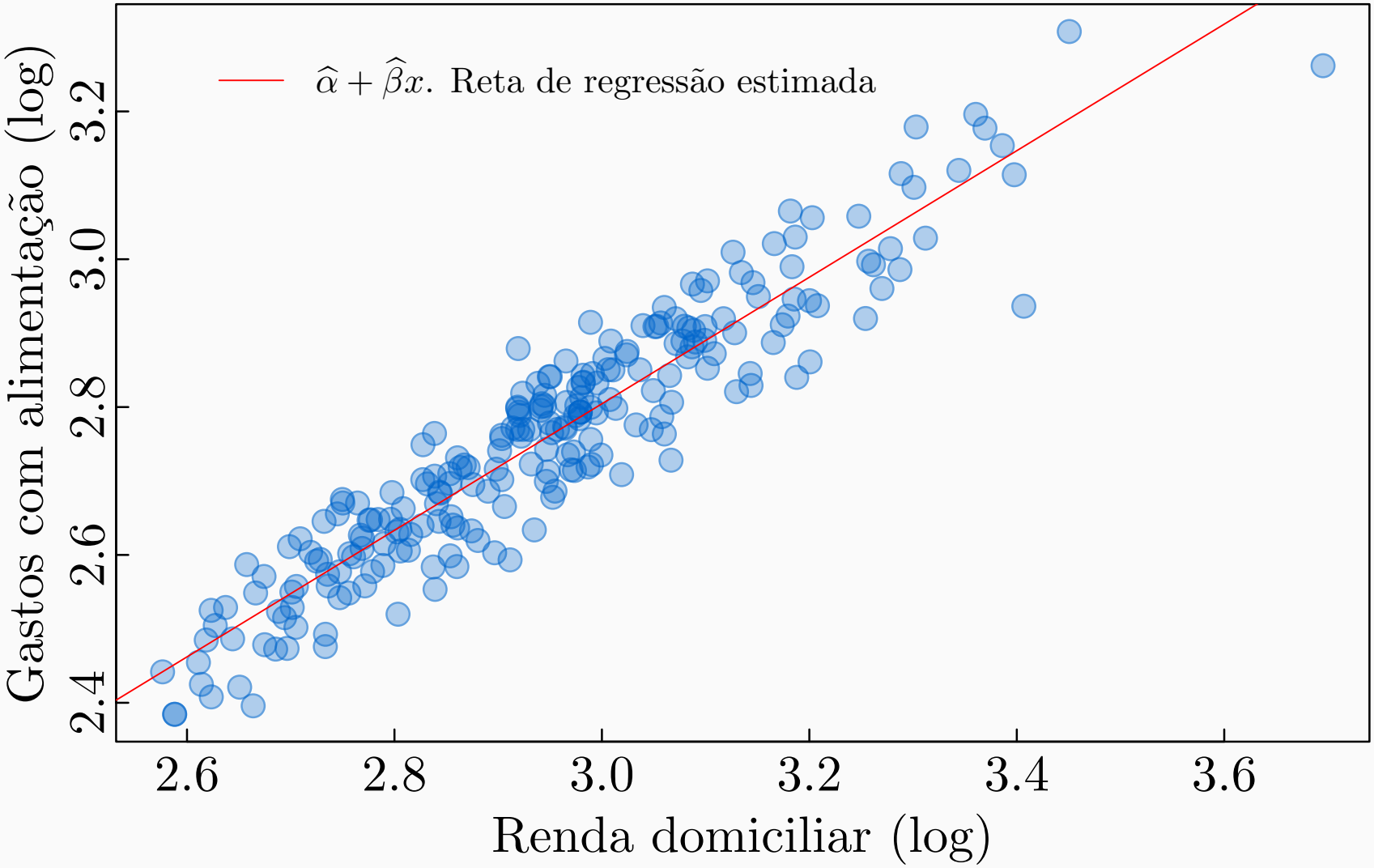
- Por esta razão, $\widehat{\alpha}$ e $\widehat{\beta}$ são chamados de estimadores de mínimos quadrados (ordinários), MQO.

Modelos de regressão linear: exemplo aplicado

- Para o conjunto de dados sobre renda domiciliar e gastos com alimentação ilustrado anteriormente, vamos denotar por Y_i o "logaritmo dos gastos com alimentação da i-ésima unidade amostral" e por X_i o "logaritmo da renda domiciliar da i-ésima unidade amostral".
- Nessa amostra, os valores computados dos estimadores de MQO são

$$\widehat{\beta} \approx 0.856, \qquad \widehat{\alpha} \approx 0.237$$

Modelos de regressão linear: exemplo aplicado



Regressão linear: observações e propriedades

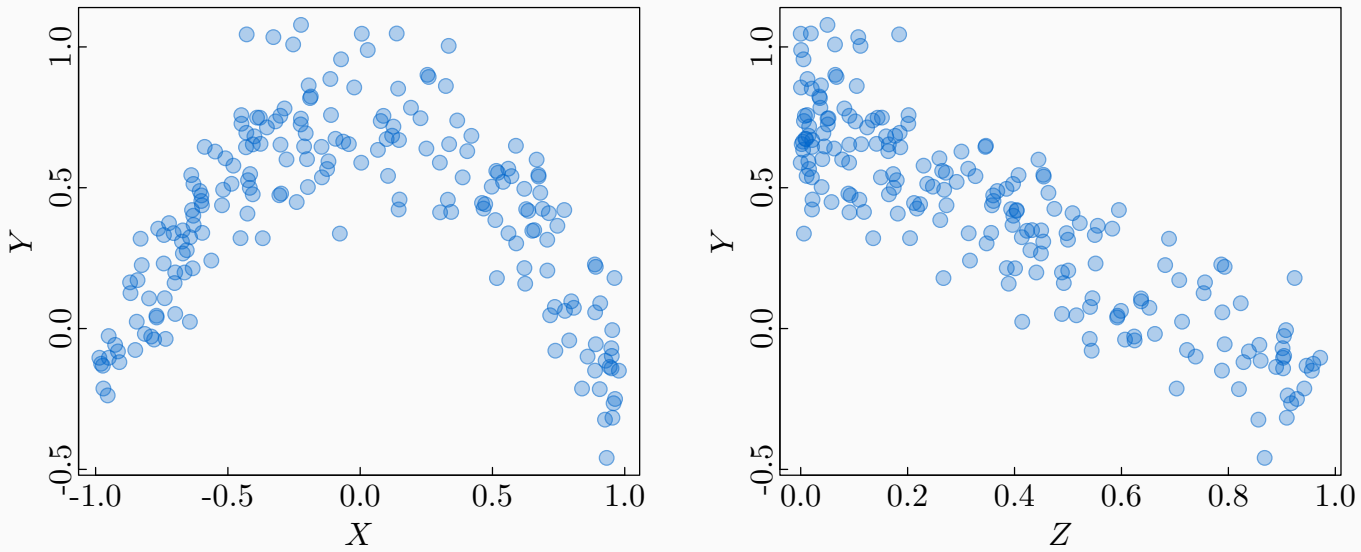
- É importante ressaltar (e o exemplo acima ilustra isso) que o termo *linear* refere-se à relação entre os parâmetros do modelo, e não entre as variáveis envolvidas.
- Por exemplo, digamos que estamos interessados em estudar a relação entre duas variáveis aleatórias Y e X, onde

$$Y = \alpha + \beta X^2 + \varepsilon,$$

  Em particular, queremos estimar os parâmetros α e β.
- Isso pode ser feito escrevendo-se $Z = X^2$: a relação entre Y e Z é linear, logo podemos estimar os parâmetros de interesse através do método de MQO (aplicado aos dados $(Z_1, Y_1) \ldots, (Z_n, Y_n)$).

Regressão linear: observações e propriedades

Gráficos de dispersão de uma amostra aleatória $(X_1, Y_1), \ldots, (X_n, Y_n)$, **simulada** do modelo $Y = 3/4 - X^2 + \varepsilon$, onde $X \sim$ Uniforme$(-1, 1)$ e onde ε, condicional em $X = x$, tem distribuição Normal com $\mathbb{E}(\varepsilon|X = x) = 0$ e $\mathrm{Var}(\varepsilon|X = x) = 0.2 - 0.08|X|$. Aqui, $Z = X^2$.



Regressão linear: ajuste e previsão

Regressão linear: ajuste e previsão

- No modelo de regressão linear

$$Y = \alpha + \beta X + \varepsilon$$

com $\mathbb{E}(\varepsilon \mid X = x)$, vale que a reta $f(x) = \alpha + \beta x$ corresponde à esperança condicional de Y dado que $X = x$:

$$\begin{aligned}
\mathbb{E}(Y \mid X = x) &= \mathbb{E}(\alpha + \beta X + \varepsilon \mid X = x) \\
&= \mathbb{E}(\alpha | X = x) + \beta \mathbb{E}(X | X = x) + \mathbb{E}(\varepsilon | X = x) \\
&= \alpha + \beta x
\end{aligned}$$

Regressão linear: ajuste e previsão

- Por essa razão, definimos o **valor ajustado (ou predito) de Y dado que $X = x$** como sendo a quantidade

$$\widehat{Y}(x) = \widehat{\alpha} + \widehat{\beta}x$$

e escrevemos também

$$\widehat{Y}_i = \widehat{Y}(X_i)$$

quando o valor de x corresponde a um dos valores na amostra.

- Definimos ainda os **resíduos** do ajuste por

$$\widehat{\varepsilon}_i = Y_i - (\widehat{\alpha} + \widehat{\beta}X_i), \qquad i = 1, \ldots, n$$

Regressão linear: ajuste e previsão

- **Importante**: não confundir os resíduos $\widehat{\varepsilon}_i$ com os termos de erro ε_i.
- Os *resíduos* são quantidades estimadas a partir da amostra:

$$\widehat{\varepsilon}_i = Y_i - (\widehat{\alpha} + \widehat{\beta} X_i), \qquad i = 1, \ldots, n$$

- Os *termos de erro* são quantidades não observáveis:

$$\varepsilon_i = Y_i - (\alpha + \beta X_i), \qquad i = 1, \ldots, n$$

Propriedades algébricas dos estimadores de MQO

- **Proposição**: valem as seguintes propriedades:
  1. A média *amostral* dos resíduos $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n$ é zero:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i &= \frac{1}{n}\sum_{i=1}^{n} Y_i - (\widehat{\alpha} + \widehat{\beta} X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{n}{n}\widehat{\alpha} - \frac{\widehat{\beta}}{n}\sum_{i=1}^{n} X_i \\ &= (\bar{Y} - \widehat{\beta}\bar{X}) - \widehat{\alpha} = 0\end{aligned}$$

     Em particular, $\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$.
  2. Os termos de erro são *ortogonais* aos regressores:

$$\sum_{i=1}^{n} X_i \widehat{\varepsilon}_i = 0.$$

Avaliação de ajuste *in-sample*

- Uma vez ajustado o modelo de regressão (isto é, uma vez estimados os parâmetros α e β), podemos avaliar a qualidade desse ajuste respondendo à seguinte pergunta: *quanto da variabilidade observada na variável resposta pode ser explicada pela reta de regressão obtida*?
- Formalmente, procedemos da seguinte maneira: definimos as quantidades

$$\mathrm{SQTot} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2, \quad \mathrm{SQRes} = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2, \quad \mathrm{SQReg} = \sum_{i=1}^{n} \left(\widehat{Y}_i - \bar{Y}\right)^2,$$

  denominados, respectivamente, a **soma dos quadrados totais**, a **soma dos quadrados dos resíduos** e a **soma dos quadrados da regressão**.

Avaliação de ajuste *in-sample*

- Temos a seguinte identidade:

$$\mathrm{SQTot} = \mathrm{SQReg} + \mathrm{SQRes}$$

- De fato, como $Y_i - \bar{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \bar{Y}) = \widehat{\varepsilon}_i + (\widehat{Y}_i - \bar{Y})$, temos que (elevando ambos os lados ao quadrado e somando em i)

$$\mathrm{SQTot} = \mathrm{SQReg} + \mathrm{SQRes} + 2\sum_{i=1}^{n} \widehat{\varepsilon}_i(\widehat{Y}_i - \bar{Y}).$$

  onde o terceiro termo da soma no lado direito da equação acima se anula: de fato

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i(\widehat{Y}_i - \bar{Y}) = \widehat{\alpha}\sum_{i=1}^{n} \widehat{\varepsilon}_i + \widehat{\beta}\sum_{i=1}^{n} \widehat{\varepsilon}_i X_i - \bar{Y}\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$$

Avaliação de ajuste *in-sample*

- A expressão

$$\mathrm{SQTot} = \mathrm{SQReg} + \mathrm{SQRes}$$

  nos diz que a variabilidade dos Y_i's em torno da média $\bar{Y}$ se decompõe em duas componentes: a primeira correspondendo àquela parte dessa variabilidade que pode ser explicada pela reta de regressão e a segunda correspondendo à variabilidade proveniente dos resíduos.
- Sendo assim, a quantidade

$$\mathrm{R}^2 \stackrel{\mathrm{def}}{=} \frac{\mathrm{SQReg}}{\mathrm{SQTot}} = 1 - \frac{\mathrm{SQRes}}{\mathrm{SQTot}}$$

  captura o quanto da variabilidade amostral da resposta pode ser atribuído a variações no preditor.

Regressão linear: predição e ajuste *out-of-sample*

- Na prática, nem sempre estaremos satisfeitos em obter um bom ajuste *in-sample*.
- De fato, uma das principais aplicações do modelo de regressão linear é a possibilidade de serem feitas predições *fora da amostra*
- Digamos que, a partir de uma amostra aleatória $(X_1, Y_1), \ldots, (X_n, Y_n)$ retirada da população (X, Y), obtivemos as estimativas $\widehat{\alpha}$ e $\widehat{\beta}$ para os parâmetros α e β.
- Se obtivermos uma nova unidade para nossa amostra da qual só conhecemos o valor X_{n+1} (mas não o de Y_{n+1}), ainda assim podemos oferecer uma *previsão* para o valor de Y_{n+1}, qual seja

$$\widehat{Y}(X_{n+1}) = \widehat{\alpha} + \widehat{\beta} X_{n+1}.$$

- O valor predito acima é uma estimativa de $\mathbb{E}(Y_{n+1} \mid X_{n+1})$.

Propriedades estatísticas dos estimadores de MQO

Propriedades estatísticas dos estimadores de MQO

- Consideremos uma amostra aleatória $(X_1, Y_1), \ldots, (X_n, Y_n)$ da população descrita por uma função de distribuição $F_{X,Y}$, onde X e Y são relacionados pelo modelo

$$Y = \alpha + \beta X + \varepsilon,$$

com $\mathbb{E}(\varepsilon \mid X = x) = 0$ e $\mathrm{Var}(\varepsilon \mid X = x) = \sigma_\varepsilon^2(x)$.

Propriedades estatísticas dos estimadores de MQO

- **Proposição**: Nas condições acima, valem as seguintes propriedades para o estimador $\widehat{\beta}$:

  (1) $\widehat{\beta}$ é não-viesado para β:

  $$\mathbb{E}\big(\widehat{\beta}\big) = \beta$$

  (2) Escrevendo $\tilde{X} = (X_1, \ldots, X_n)$, a variância de $\widehat{\beta}$ condicional em $\widetilde{X}$ é dada por:

  $$\mathrm{Var}\big(\widehat{\beta} \,|\, \tilde{X}\big) = \frac{\sum_{i=1}^{n} \sigma_\varepsilon^2(X_i)(X_i - \bar{X})^2}{\big(\sum_{i=1}^{n}(X_i - \bar{X})^2\big)^2},$$

  donde a variância incondicional de $\widehat{\beta}$ satisfaz

  $$\mathrm{Var}\big(\widehat{\beta}\big) \approx \frac{\mathbb{E}\big\{\sigma_\varepsilon^2(X)\,(X - \mathbb{E}X)\big\}}{n\sigma_X^4}.$$

  Em particular, $\widehat{\beta}$ é consistente para β.

Propriedades dos estimadores de MQO

- Segue que:
  (3) Condicional em $X_1, \ldots, X_n$, o estimador $\widehat{\beta}$ tem distribuição aproximadamente normal:

$$\widehat{\beta}|\tilde{X} \sim N(\beta,\, \mathrm{Var}(\widehat{\beta}\,|\,\tilde{X}))$$

  (4) Sob a hipótese nula $H_0 : \beta = \beta_0$, a estatística de teste

$$T = \frac{\widehat{\beta} - \beta_0}{\widehat{V}}$$

  tem distribuição aproximadamente Normal padrão. Aqui, $\widehat{V}^2$ é um estimador de $\mathrm{Var}(\widehat{\beta}|\widetilde{X})$ dado por

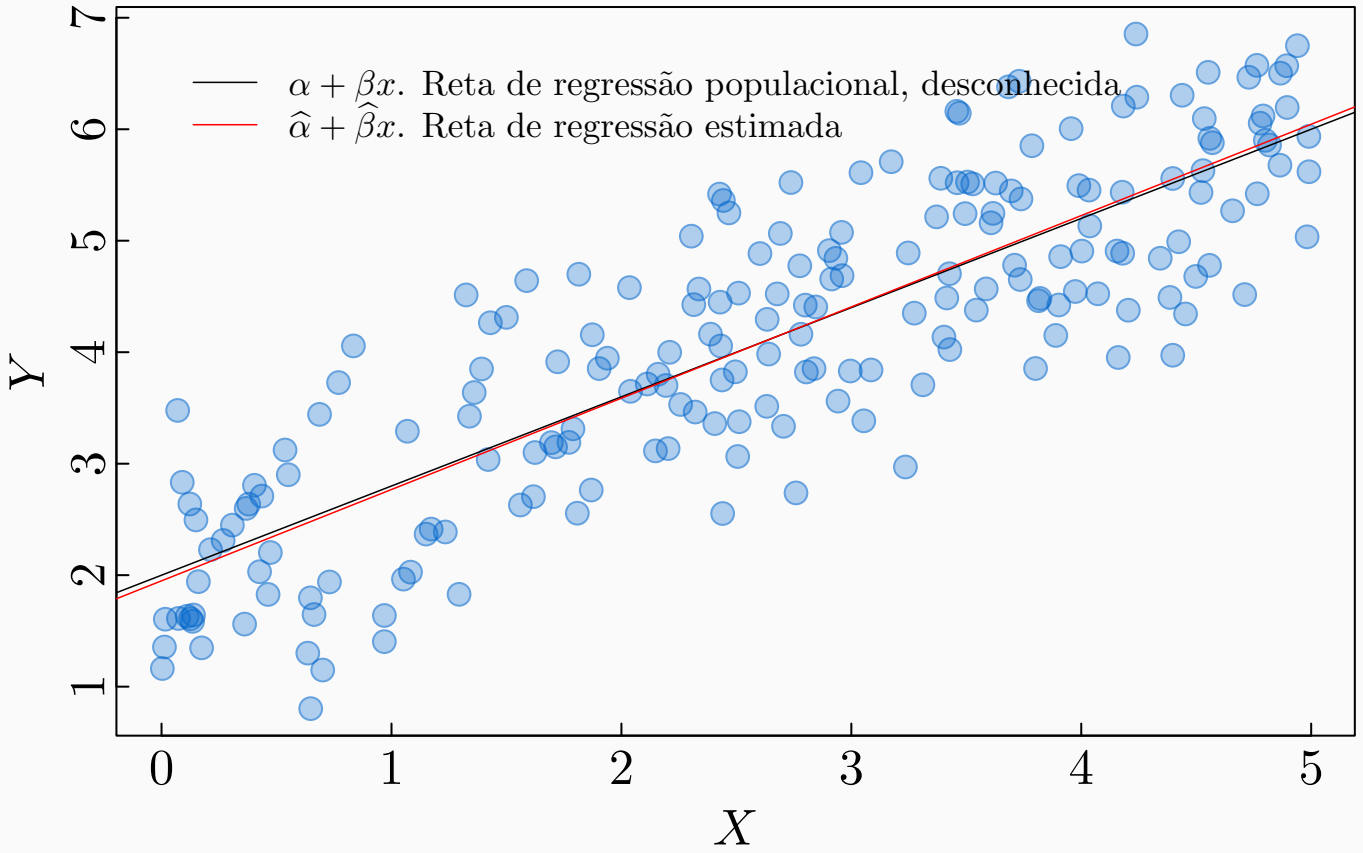
$$\widehat{V}^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2 (X_i - \bar{X})^2}{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)^2}$$

Exemplo

- Voltemos ao nosso **exemplo** simulado: temos uma amostra iid $(X_1, Y_1), \ldots, (X_n, Y_n)$ do modelo $Y = 2 + 0.8\,X + \varepsilon$, onde $X \sim$ Uniforme$(0, 5)$ e $\varepsilon \sim$ Normal$(\mu_\varepsilon = 0,\, \sigma^2_\varepsilon = 0.7^2)$, com X e ε mutuamente independentes. Aqui temos $n = 200$.
- Os valores de $\widehat{\alpha}$ e $\widehat{\beta}$ correspondentes a essa amostra são

$$\widehat{\alpha} = 1.94934 \qquad \text{e} \qquad \widehat{\beta} = 0.81907$$

Exemplo


$\alpha + \beta x$. Reta de regressão populacional, desconhecida
$\widehat{\alpha} + \widehat{\beta} x$. Reta de regressão estimada
Y
X
0
1
2
3
4
5
1
2
3
4
5
6
7

Exemplo: Testes de hipóteses

- Se quisermos testar a hipótese nula "$H_0 : \beta = 0$" contra a alternativa "$H_1 : \beta > 0$", devemos obter o valor da estatística de teste

$$T = \frac{\widehat{\beta}}{\widehat{V}},$$

  a qual segue (sob H_0) uma distribuição (aproximadamente) Normal padrão (conforme vimos acima).
- Nessa particular amostra o valor da estatística de teste é 23.767. Ao nível de significância de 1%, portanto, rejeitamos H_0. (Por quê?)

Importante

Observação: se quiséssemos ser inteiramente rigorosos na notação, deveríamos escrever $X_i(\omega)$, $\bar{X}_n(\omega)$, $\widehat{\alpha}_n(\omega)$, etc, que são os valores numéricos das variáveis aleatórias em questão *na particular amostra* ω *de tamanho* n. Essa notação fica um tanto carregada e, por isso, é costumeiro omitir o "ω" e o n.

Uma aplicação: classificação binária

Uma aplicação: classificação binária

- Pela definição do modelo de regressão linear, temos

$$\mathbb{E}(Y \mid X = x) = \alpha + \beta x.$$

- Quer dizer, a reta dada pela equação $\alpha + \beta x$ descreve o comportamento médio da resposta nos diferentes *níveis* do preditor X.
- Esse modelo tem uma interpretação interessante no caso em que a resposta é uma variável binária (isto é, quando Y tem distribuição Bernoulli).

Uma aplicação: classificação binária

- **Lembrando**: Y tem distribuição Bernoulli com parâmetro u se

$$\mathbb{P}(Y = 1) = u, \qquad \mathbb{P}(Y = 0) = 1 - u.$$

- Nesse caso, temos a identidade

$$\mathbb{E}(Y) = \big(0 \times \mathbb{P}(Y = 0)\big) + \big(1 \times \mathbb{P}(Y = 1)\big) = \mathbb{P}(Y = 1) = u.$$

- Semelhantemente, condicionando em uma covariável X,

$$\mathbb{E}(Y \,|\, X = x) = \mathbb{P}(Y = 1 \,|\, X = x) =: u(x)$$

O modelo de probabilidade linear

- Se Y é uma variável com distribuição Bernoulli e X é uma covariável, o *modelo de probabilidade linear* estipula que

$$\mathbb{P}(Y = 1 \,|\, X = x) = \alpha + \beta x$$

- Evidentemente, o lado esquerdo da equação acima deve ser um número entre 0 e 1, para qualquer nível x da covariável X:

$$0 \leq \alpha + \beta x \leq 1.$$

  Em particular, tal modelo só faz sentido se o preditor for uma variável aleatória limitada.
- Por exemplo, se $\beta > 0$, então deve valer

$$\mathbb{P}\left(-\frac{\alpha}{\beta} \leq X \leq \frac{1-\alpha}{\beta}\right) = 1$$

O modelo de probabilidade linear

- Suponhamos, então, que $\mathbb{P}(Y = 1 \mid X = x) = \alpha + \beta x$.
- Definindo $\varepsilon = Y - (\alpha + \beta X)$, podemos escrever

$$Y = \alpha + \beta X + \varepsilon$$

  e é de fácil verificação que $\mathbb{E}(\varepsilon | X = x) = 0$.
- Quer dizer, de fato o modelo de probabilidade linear nada mais é do que um modelo de regressão linear no caso em que a resposta tem distribuição Bernoulli.

O modelo de probabilidade linear

- Esse modelo ilustra como a suposição (usual em textos introdutórios) de que termo de erro e covariável são independentes é facilmente violada.
- De fato, temos $\mathrm{Var}(\varepsilon|X=x) = \mathbb{E}(\varepsilon^2|X=x)$ já que $\mathbb{E}(\varepsilon|X=x) = 0$. Logo, escrevendo $u(x) = \mathbb{E}(Y|X=x) = \alpha + \beta x$, vemos que

$$\mathrm{Var}(\varepsilon|X=x) = \mathbb{E}\big(Y^2 - 2Yu(x) + u(x)^2 \,|\, X=x\big).$$

- Como Y só pode assumir os valores 0 ou 1, vale que $Y^2 = Y$. Assim,

$$\mathrm{Var}(\varepsilon|X=x) = \mathbb{E}(Y|X=x) - 2u(x)\mathbb{E}(Y|X=x) + u(x)^2.$$

- Segue, da definição de $u(x)$, que

$$\mathrm{Var}(\varepsilon|X=x) = u(x) \times \big(1 - u(x)\big).$$

O modelo de probabilidade linear: simulação

- A figura a seguir mostra o gráfico de dispersão de uma amostra aleatória de tamanho $n = 200$ do modelo de probabilidade linear descrito acima, com $\alpha = {}^{1}/_{8}$ e $\beta = {}^{6}/_{8}$.
- Nessa amostra, o valor das estimativas dos parâmetros de intercepto e inclinação foram, respectivamente $\widehat{\alpha} = 0.1058$ e $\widehat{\beta} = 0.7264$.

O modelo de probabilidade linear: simulação