Métodos estatísticos

Regressão linear simples

Prof. Wagner Hugo Bonat

Departamento de Estatística
Universidade Federal do Paraná

Estrutura da unidade

- Análise de variância (ANOVA).
- **Regressão linear simples**.

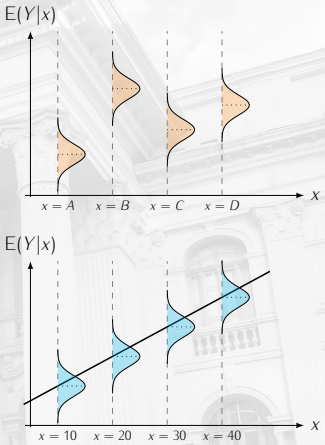

Figura 1. Representações esquemáticas dos modelos de ANOVA e regressão linear simples.

Motivação

- **Problema**: Estudar a relação entre duas variáveis quantitativas x e y.
- Procura-se uma função de x que aproxime y em algum sentido,

$$y \approx f(x; \boldsymbol{\beta}).$$

- Em geral, a aproximação não será perfeita:
  - Variação acidental ou aleatória.
  - Falta de informação sobre y em x.

Figura 2. Foto de George Becker no Pexels.

Motivação

- Cada observação é decomposta em duas partes: a **previsível** e a **aleatória**

  (observação) = (previsível)+(aleatória).

- **Previsível** → reflete o conhecimento sobre o fenômeno (uma função matemática).
- **Aleatória** → deve seguir algum **modelo de probabilidade**.
- Em notação matemática, temos

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

  onde ϵ segue alguma distribuição de probabilidade.

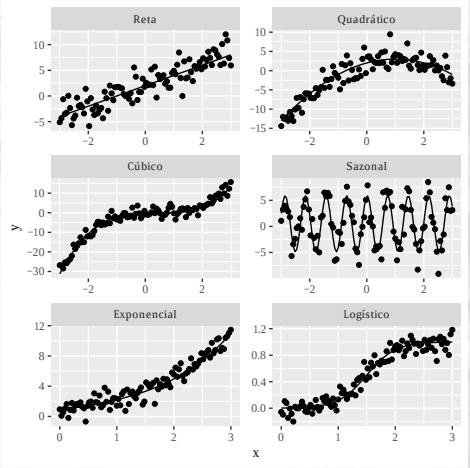

Figura 3. Exemplos de modelos matemáticos.

Exemplo 1

É de interesse entender a relação entre as variáveis altura (x) e peso (y). Temos a seguinte amostra

Altura	Peso
165	55
168	57
170	65
175	68
178	71

- Modelo constante: $y = \mu + \epsilon$.
- Modelo linear: $y = \beta_0 + \beta_1 \text{Altura} + \epsilon$.

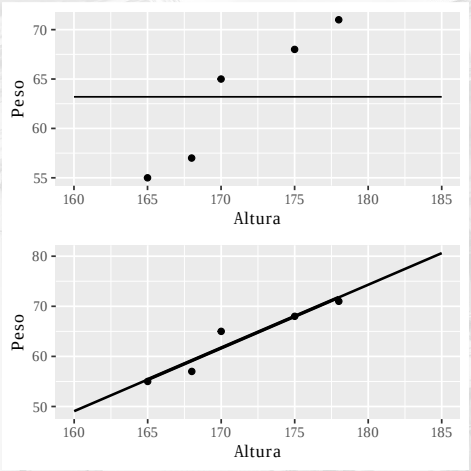

Figura 4. Diagramas de dispersão e exemplos de modelos para descrever a relação entre y e x.

Comentários

- *Peso* é a variável **dependente** ou **resposta** que depende linearmente da *Altura*.
- A variável *Altura* é chamada de **explicativa** ou **covariável** ou **independente**.
- Em ambos os modelos temos **parâmetros** que precisam ser estimados.
- O erro nos modelos é dado por

$$\epsilon = y - \mu$$
$$\epsilon = y - (\beta_0 + \beta_1 \text{Altura}).$$

- Note que, o erro no segundo modelo deve diminuir, pois agora

$$e = f(\cancel{\text{altura}}, \text{sexo}, \text{idade}, \text{país}, \ldots),$$

ou seja, incorporamos uma informação (extra) para explicar o peso.

Modelos de Regressão

- **Análise de regressão**: Técnica estatística que analisa as relações entre uma única variável **dependente** e uma ou mais variáveis **independentes**.
- **Objetivo**: Estudar as relações entre as variáveis, a partir de um **modelo matemático**, permitindo **estimar** o valor de uma variável a partir da outra.
- **Problema**: Definir a **forma** da relação existente entre as variáveis.

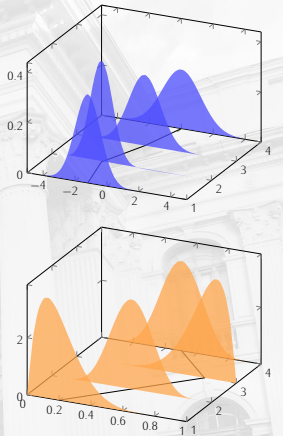

Figura 5. Ilustração dos modelos de regressão Normal e não Normal.

Modelos de Regressão: Descrevendo a variação da média

- Algumas possíveis relações:
  - **Reta** $y = \beta_0 + \beta_1 x$.
  - Quadrático $y = \beta_0 + \beta_1 x + \beta_2 x^2$.
  - Cúbico $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.
  - Sazonal
    $y = \beta_0 + \beta_1 \cos(2\pi x/s) + \beta_2 \sin(2\pi x/s)$.
  - Exponencial $y = \beta_0 \exp(\beta_1 x)$.
  - Logístico
    $y = \exp(\beta_0 + \beta_1 x)/(1 + exp(\beta_0 + \beta_1 x))$.
- Variável **dependente** (y) será **predita** a partir da relação com x.

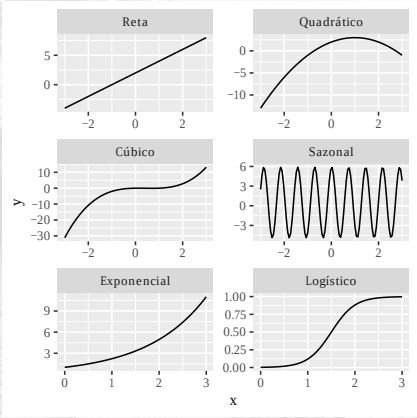

Figura 6. Exemplos de modelos matemáticos.

Regressão linear

- No modelo de **regressão linear** assumimos que a variável resposta possui **relação linear** com as variáveis **explicativas**.
- **Modelo de regressão linear simples** $\rightarrow$ associa uma única variável explicativa (x) à variável dependente (y),

$$y = \beta_0 + \beta_1 x + e.$$

- **Modelo de regressão linear múltipla** $\rightarrow$ associa p variáveis explicativas para descrever o comportamento da **resposta**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon.$$

Regressão linear simples (formulação matemática)

- Sejam y_i e x_i observações da variável resposta e explicativa, para $i = 1, \ldots, n$.
- O interesse é descrever y como uma função linear de x,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- Note que o erro ϵ é uma função dos parâmetros β_0 e β_1.
- Precisamos de um critério para encontrar β_0 e β_1 "ótimos" em algum sentido.
- **Critério de mínimos quadrados**: encontrar β_0 e β_1 tal que

$$\text{SQ}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{1}$$

  seja a menor possível.
- Equação (1) é chamada de **soma de quadrados** (SQ).

Regressão linear simples (formulação matemática)

- Como obter β_0 e β_1 que minimizam a SQ?
- Abordagem padrão é usar cálculo diferencial.
  - Obter o vetor **gradiente** em relação a β_0 e β_1.
  - Resolver o sistema linear resultante (igualar a zero e isolar β_0 e β_1).
- **Vou mostrar os passos básicos apenas por curiosidade**.
- Em termos práticos usamos as equações resultantes ou *softwares* estatísticos.

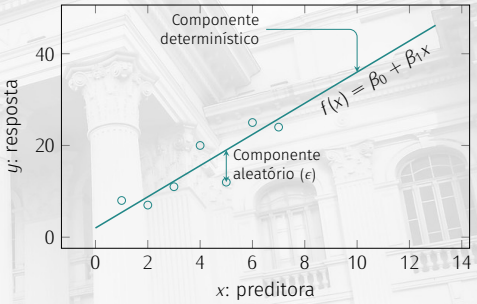

Figura 7. Componentes do modelo de regressão linear simples.

Vetor gradiente

- Chame $y_i - (\beta_0 + \beta_1 x_i) = \epsilon_i$.
- Chame $\beta_0 + \beta_1 x_i = \mu_i$.

$$\nabla f(\beta_0,\beta_1) = \left(\frac{\partial f(\beta_0,\beta_1)}{\partial \epsilon_i} \frac{\partial \epsilon_i}{\mu_i} \frac{\partial \mu_i}{\partial \beta_0}, \frac{\partial f(\beta_0,\beta_1)}{\partial \epsilon_i} \frac{\partial \epsilon_i}{\mu_i} \frac{\partial \mu_i}{\partial \beta_1} \right),$$

em que

$$\frac{\partial f(\beta_0,\beta_1)}{\partial \epsilon_i} = \frac{\partial}{\partial \epsilon_i} \sum_{i=1}^{n} \epsilon_i^2 = 2 \sum_{i=1}^{2} \epsilon_i.$$

$$\frac{\partial \epsilon_i}{\partial \mu_i} = \frac{\partial}{\partial \mu_i}(y_i - \mu_i) = -1.$$

$$\frac{\partial \mu_i}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \beta_0 + \beta_1 x_i = 1.$$

$$\frac{\partial \mu_i}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \beta_0 + \beta_1 x_i = x_i.$$

Vetor gradiente

- Vetor gradiente,

$$\nabla f(\beta_0,\beta_1) = \left(-2\sum_{i=1}^{n} \epsilon_i(1); -2\sum_{i=1}^{n} \epsilon_i x_i \right)$$

$$= \left(-2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i); -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \right).$$

- Resolver o sistema de equações:

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad (2)$$

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \qquad (3)$$

- Pela Eq. (1), temos

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

- Substituindo Eq.(3) na Eq. (2) e resolvendo em $\hat{\beta}_1$ **(após muita manipulação)**, temos

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}.$$

Ilustração do procedimento de mínimos quadrados

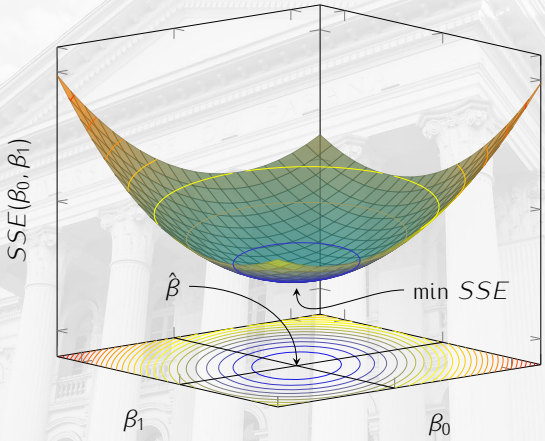

Figura 8. Superfície da soma de quadrados como função dos parâmetros.

Exemplo 1 (cont.)

Temos as seguintes estatísticas amostrais

$$\sum_{i=1}^{n} x_i y_i = 5423, \quad \sum_{i=1}^{n} x_i^2 = 146658,$$

$$\overline{x} = 171.20 \quad \text{e} \quad \overline{y} = 63.20.$$

Assim, $\hat{\beta}_0$ e $\hat{\beta}_1$ são

$$\hat{\beta}_1 = \frac{54239 - 5 \cdot 171.20 \cdot 63.20}{146658 - 5 \cdot (171.20)^2} = \frac{139.80}{110.80} = 1.261733,$$

$$\hat{\beta}_0 = 63.20 - 1.261733 \cdot 171.20 = -152.8087.$$

Exemplo 1 (cont.)

- Valores preditos

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

- Qual o peso esperado de uma pessoa de 1.70?

$$-152.809 + 1.262 \cdot 170 = 61.731.$$

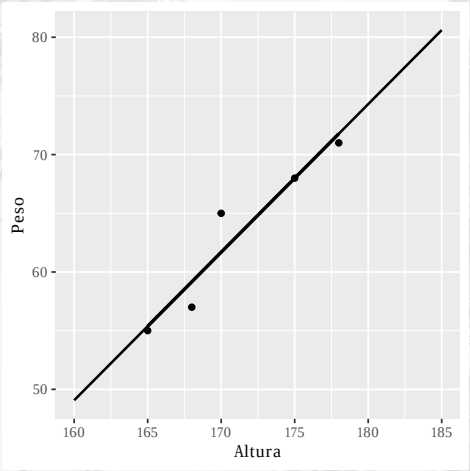

Figura 9. Diagrama de dispersão e reta ajustada.

Regressão linear simples (formulação estatística)

- Sejam Y_i v.a. independentes e x_i uma variável explicatica conhecida.
- **Regressão linear simples** $\rightarrow$ modelo estatístico

$$Y_i \sim \mathrm{N}(\mu_i, \sigma^2).$$

- Note que $\mathrm{E}(Y_i) = \mu_i = \beta_0 + \beta_1 x_i$.
- De forma equivalente, temos

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{em que} \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2).$$

- Estimação $\rightarrow$ mínimos quadrados $=$ máxima verossimilhança.

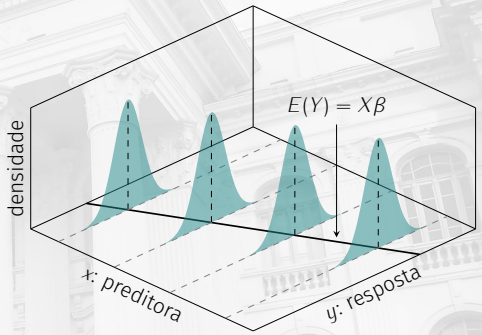

Figura 10. Modelo de regressão linear simples.

Interpretação dos parâmetros

- β_0 (intercepto) representa o ponto onde a reta corta o eixo y (na maioria das vezes não possui interpretação prática).
- β_1 (coeficiente angular) representa a variabilidade em y causada pelo aumento de uma unidade em x. Além disso,
  - $\beta_1 > 0$ mostra que com o aumento de x, também há um aumento em y.
  - $\beta_1 = 0$ mostra que **não há efeito** de x sobre y.
  - $\beta_1 < 0$ mostra que com a aumento de x, há um decréscimo em y.

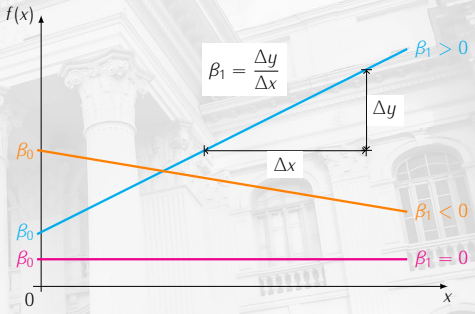

Figura 11. Interpretação dos coeficientes de regressão.

Avaliação do modelo de regressão

Quadro de ANOVA para a regressão linear simples

- Note que na formulação estatística obtemos $\hat{\beta}_0$ e $\hat{\beta}_1$.
- São estimativas/estimadores dos verdadeiros parâmetros β_0 e β_1.
- O principal teste de interesse é verificar se a covariável **influencia** na resposta.
- Em termos de teste de hipóteses,

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_a : \beta_1 \neq 0.$$

- Caso H_0 não seja rejeitada, adotamos o modelo

$$\text{Modelo 0} : Y_i = \beta_0 + \epsilon_i = \mu + \epsilon_i.$$

- Caso H_0 seja rejeitada, o modelo é

$$\text{Modelo 1} : Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Quadro de ANOVA para a regressão linear simples

- Baseado no Modelo 0 obtemos a soma de quadrados total

$$\text{SQT} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

- Baseado no Modelo 1 obtemos a soma de quadrados residual

$$\text{SQRes} = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

- A diferença SQT - SQRes é o que chamamos de **soma de quadrados da regressão**

$$\text{SQReg} = \text{SQT} - \text{SQRes} = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Decomposição da soma de quadrados

- Note que

$$\text{SQT} = \text{SQReg} + \text{SQRes}.$$

- Relação importante

$$y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}) = \hat{\epsilon}_i + (\hat{y}_i - \overline{y}).$$

- Desvio de uma observação em relação à média = desvio da observação em relação à reta de regressão + desvio do valor ajustado em relação à média.

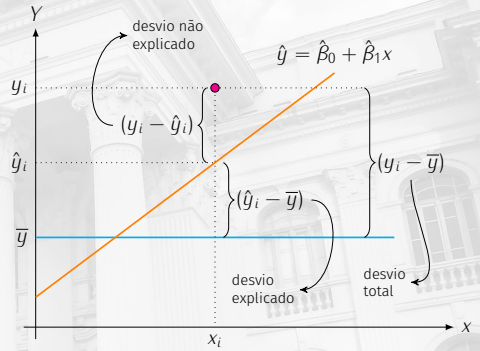

Figura 12. Decomposição da variação dos dados na regressão linear simples.

Quadro de ANOVA para a regressão linear simples

- Importante notar que temos somas de quadrados $\rightarrow$ Qui-quadrado.
- Na SQT temos $n-1$ graus de liberdade (estimamos apenas $\overline{Y}$).
- Na SQRes temos $n-2$ graus de liberdade (estimamos $\hat{\beta}_0$ e $\hat{\beta}_1$).
- Assim, chegamos aos chamados **quadrados médios**

$$\text{QMT} = \frac{\text{SQT}}{n-1} = S^2, \quad \text{QMRes} = \frac{\text{SQRes}}{n-2} \quad \text{e} \quad \text{QMReg} = \frac{\text{SQReg}}{1}.$$

- Chegamos na estatística $F = \frac{\text{QMReg}}{\text{QMRes}} \sim F(1, n-2)$.

Quadro de ANOVA para a regressão linear simples

- Quadro de ANOVA

Fonte de variação	GL	SQ	QM	F
Regressão	1	SQReg	QMReg	QMReg/QMRes
Residual	n-2	SQRes	QMRes	
Total	n-1	SQT		

- **Coeficiente de determinação** é o percentual da variabilidade da resposta explicada pelo modelo, ou seja,

$$R^2 = \frac{\text{SQReg}}{\text{SQT}}.$$

Exemplo 1 (cont.)

Vamos montar o quadro da ANOVA.

- Soma de quadrados da regressão $\text{SQReg} = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2$

$$\text{SQReg} = 1.261733^2((165 - 171.2)^2 + \ldots + (178 - 171.2)^2) = 176.3903.$$

- Soma de quadrados dos resíduos $\text{SQres} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.

$$\text{SQRes} = (55 - (-152.809 + 1.262 \cdot 165))^2 + \ldots + (71 - (-152.809 + 1.262 \cdot 178))^2 = 16.41.$$

OBS: Na prática, pode-se calcular a SQT e SQReg e obter SQRes pela diferença SQT - SQReg.

Quadro da ANOVA

- Interpretação é feita como qualquer teste de hipótese.
- O F crítico para 95% de confiança neste caso é 10.12.

Fonte de variação	GL	SQ	QM	F
Regressão	1	176.39	176.39	32.24
Residual	5-2	16.41	5.47	
Total	5-1	192.80		

- Concluímos que a covariável *Altura* influencia a resposta *Peso*.
- O coeficiente de determinação é $R^2 = \frac{176.39}{192.80} = 0.9148$.

Exemplo 2

A tabela a seguir relaciona as distâncias percorridas por carros (km) e seu consumo de combustível (litros), em uma amostra de 10 carros novos.

Dist	Cons	Dist.	Cons.
20	1.33	80	6.15
60	5.45	70	4.11
15	1.66	73	5.00
45	3.46	28	2.95
35	2.92	85	6.54

- Faça um diagrama de dispersão.
- Trace um modelo linear aproximado.
- Estime os parâmetros $\hat{\beta}_0$ e $\hat{\beta}_1$.
- Monte o quadro de ANOVA.
- Interprete o resultado. Pode-se concluir que para percursos mais longos há maior consumo de combustível?
- Faça uma *predição* do consumo de combustível para uma distância de 50 km.

Exemplo 2 (cont.)

Diagrama de dispersão e reta ajustada

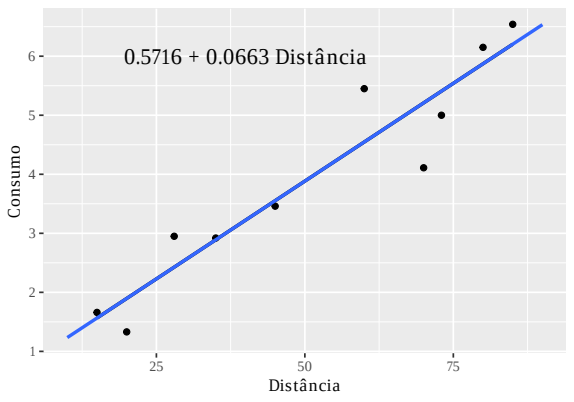

Figura 13. Diagrama de dispersão e reta ajustada.

Exemplo 2 (cont.)

Estatísticas amostrais

$$\sum_{i=1}^{n} x_i y_i = 2419.60, \quad \sum_{i=1}^{n} x_i^2 = 32113$$

$$\overline{x} = 51.10 \quad \text{e} \quad \overline{y} = 3.957$$

Assim, $\hat{\beta}_0$ e $\hat{\beta}_1$ são

$$\hat{\beta}_1 = \frac{2419.60 - 10 \cdot 51.10 \cdot 3.957}{32113 - 10 \cdot (51.10)^2} = \frac{397.573}{6000.9} = 0.0663.$$

$$\hat{\beta}_0 = 3.957 - 0.0663 \cdot 51.10 = 0.5716.$$

Consumo para distância de $50km \rightarrow 0.5716 + 0.0663 \cdot 50 = 3.8841$.

Exemplo 2 (cont.)

Vamos montar o quadro da ANOVA.

- Soma de quadrados da regressão $\text{SQReg} = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2$

$$\text{SQReg} = 0.0663^2((20 - 51.1)^2 + \ldots + (85 - 51.2)^2) = 26.3383.$$

- Soma de quadrados dos resíduos $\text{SQres} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

$$\text{SQRes} = (1.33 - (0.5715 + 0.0663 \cdot 20))^2 + \ldots + (6.54 - (0.5715 + 0.0663 \cdot 85))^2 = 2.9951.$$

Exemplo 2 (cont.)

- Interpretação é feito como qualquer teste de hipótese.
- O F crítico para 95% de confiança neste caso é 5.31.

Fonte de variação	GL	SQ	QM	F
Regressão	1	26.34	26.34	70.35
Residual	8	2.99	0.37	
Total	9	29.33		

- Concluímos que a covariável *Distância* influencia a resposta *Consumo*.
- O coeficiente de determinação é $R^2 = \frac{26.34}{29.33} = 0.8980$.

Inferência completa

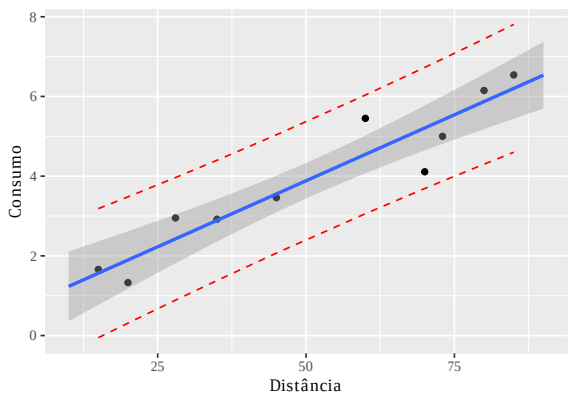

Figura 14. Diagrama de dispersão, reta ajustada e intervalo de confiança de predição

Resumo

- Modelos de regressão são as ferramentas estatísticas mais populares.
- Aplicações aparecem em praticamente todas as áreas da ciência.
- Existem diversas variações dos modelos de regressão:
  - Linear múltiplo.
  - Modelos lineares generalizados.
  - Modelos aditivos generalizados.
  - Modelos para múltiplas respostas.
  - Muito mais!!

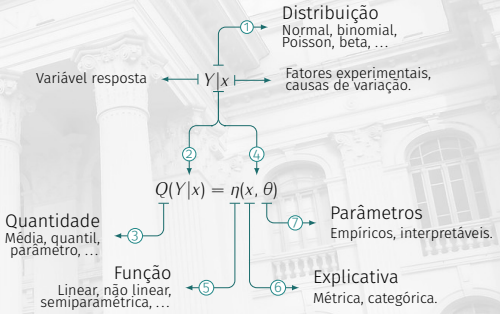

Figura 15. Resumo modelos de regressão.

Estrutura da unidade

- Análise de variância (ANOVA).
- **Regressão linear simples**.

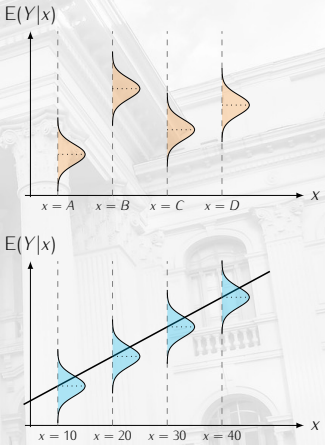

Figura 16. Representações esquemáticas dos modelos de ANOVA e regressão linear simples.