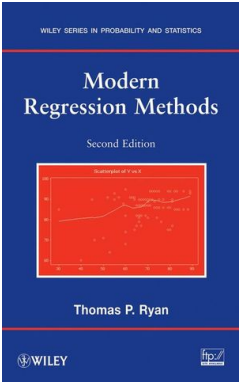
Modelos de Regressão Linear e Não Linear

Dr. Walmes Marques Zeviani
walmes@ufpr.br
Departamento de Estatística - UFPR
Laboratório de Estatística e Geoinformação (LEG)

O que é regressão?



História

- Francis Galton (1822 – 1911) era primo de Charles Darwin e tinha competências em medicina e matemática.
- Galton era fascinado pela biometria humana e herdabilidade.
- Inventou a indentificação pela impressão digital.
- Estudou dados de altura dos pais e filhos adultos (1886).
- "Law of universal regression".
- Altura do filhos *regrediam* para a média.
- Com ajuda de Karl Pearson, ajustou a reta.
- A técnica recebeu o nome **regressão**.

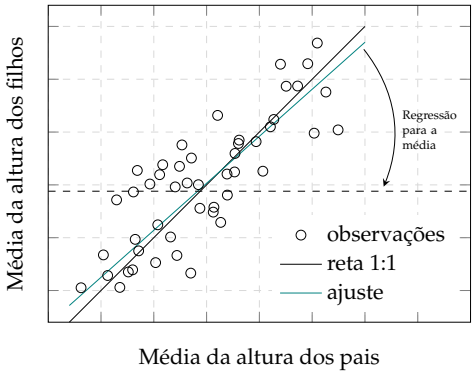


Figura 1: A regressão para a média de Galton.

Motivação

- Estudar a associação entre uma variável y e um conjunto de variáveis x_i, $i \geqslant 1$.
- Fazer a previsão de y.

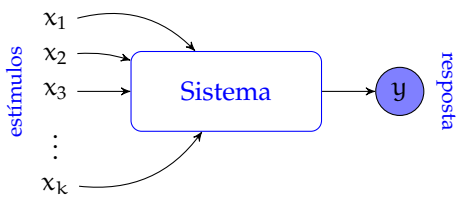


Figura 2: Representação das variáveis relacionadas com o valor de venda de um imóvel.

Figura 3: Representação das variáveis relacionadas com o rendimento escolar de uma criança.

Mais exemplos

- Tempo de uma substância no sangue:
  - Concentração aplicada;
  - Paciente (idade, sexo, pressão, hábitos).
- Produtividade de um pomar:
  - Fertilizante (dose, tipo, frequência);
  - Preparo do solo;
  - Manejo de pragas e doenças;
  - Poda e tratos culturais;
  - Irrigação;
  - Clima.
- Desempenho de time de futebol:
  - Quantidade, intensidade e qualidade de treino;
  - Nutrição e preparo físico dos jogadores;
  - Entrozamento entre jogadores;
  - Estratégia de jogo;
  - Experiência dos jogadores e equipe;
  - Condição do gramado;
  - Apoio da torcida.

Forma genérica

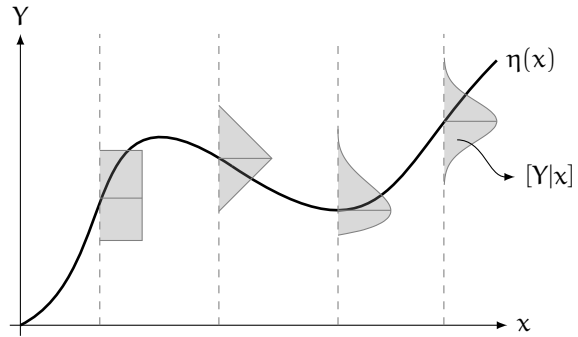


Figura 4: Representação esquemática genérica de um modelo de regressão.

Organização do modelo de regressão

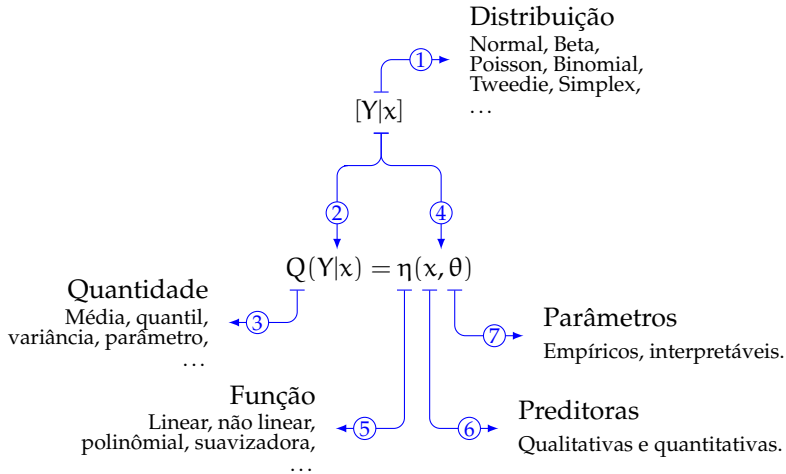
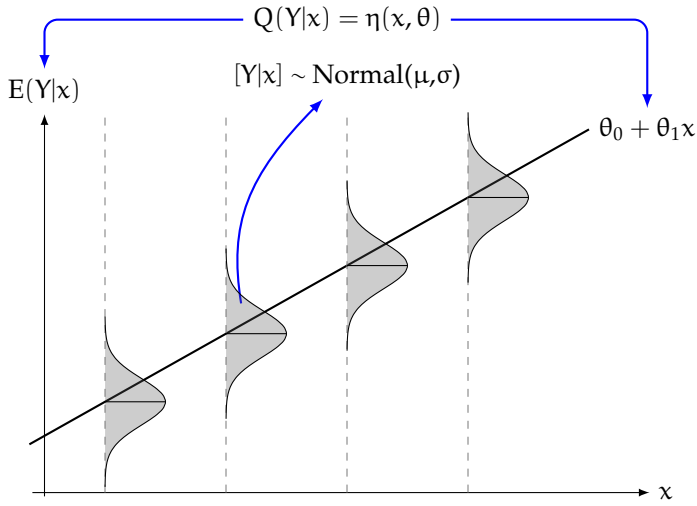
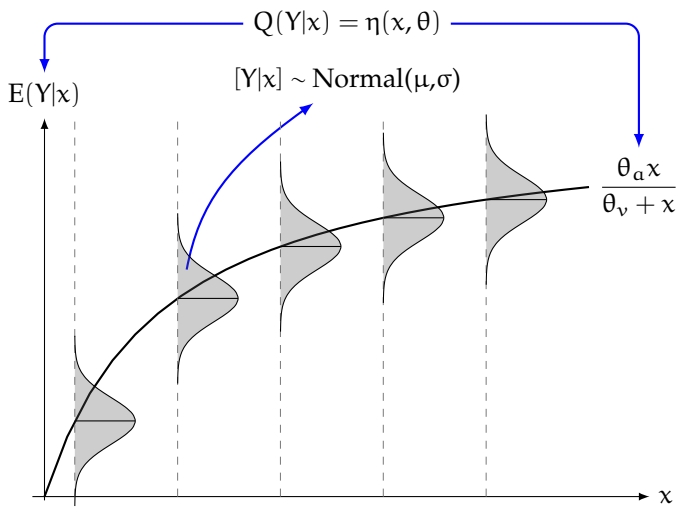


Figura 5: Representação esquemática da construção de um modelo de regressão.

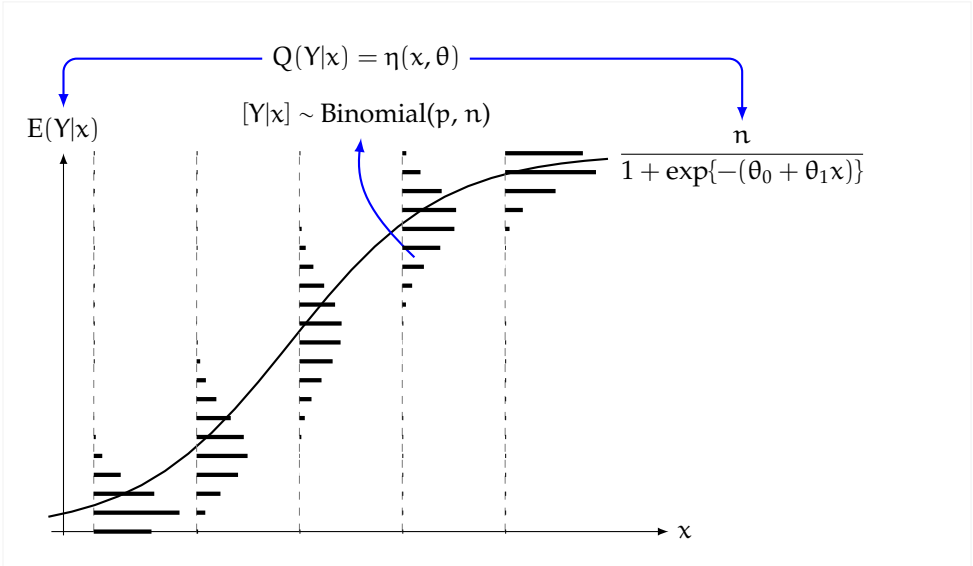
Regressão Linear Simples



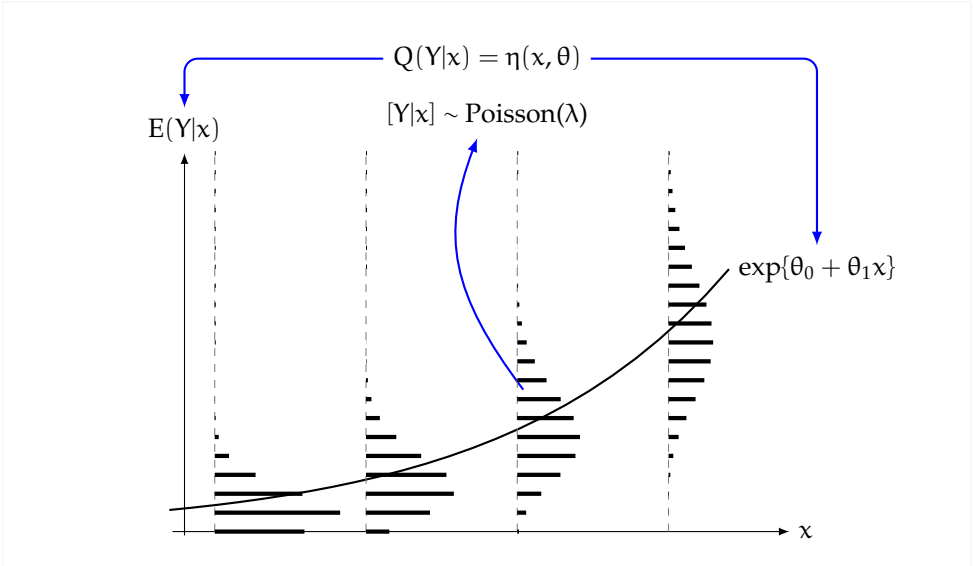
Regressão Não Linear



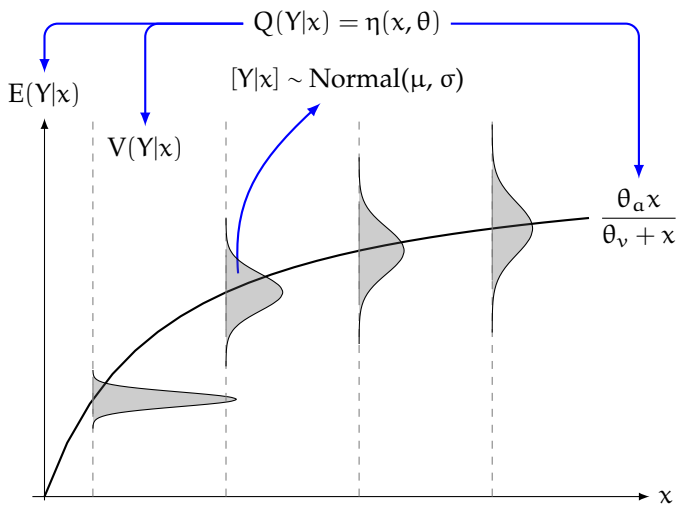
Regressão Binomial



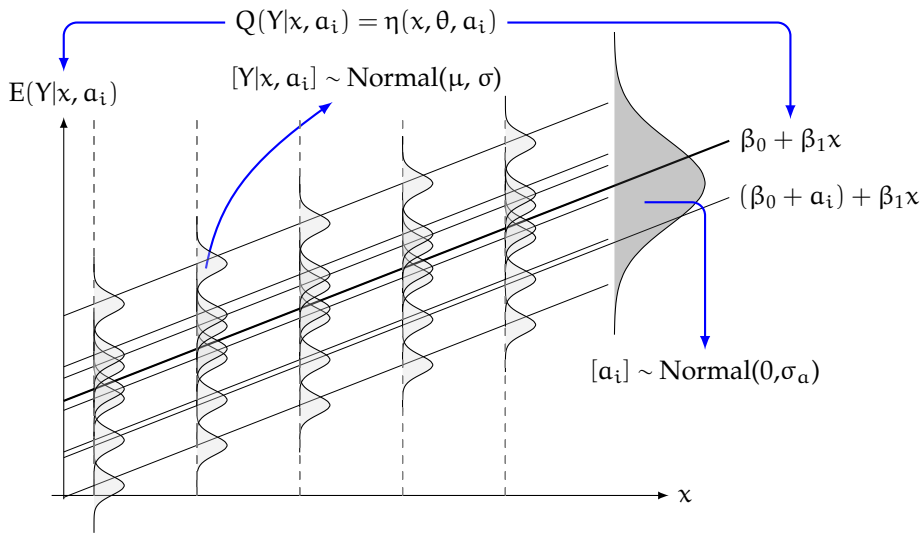
Regressão Poisson



Regressão Não Linear com Variância Não Constante

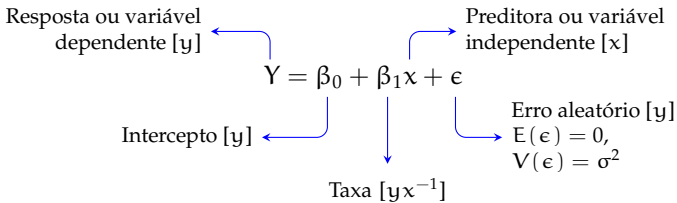


Regressão Linear com Efeito Aleatório

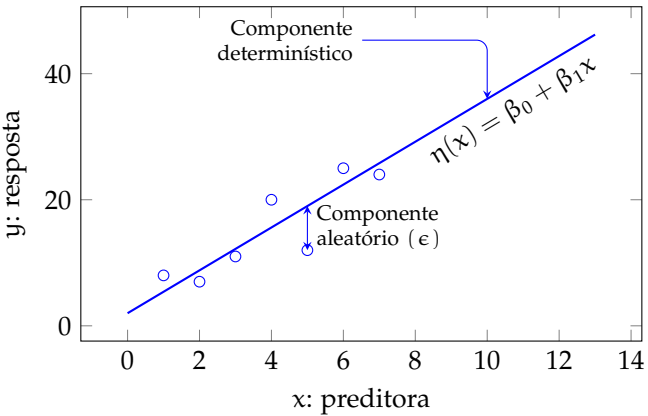


Especificação

- Regressão linear simples

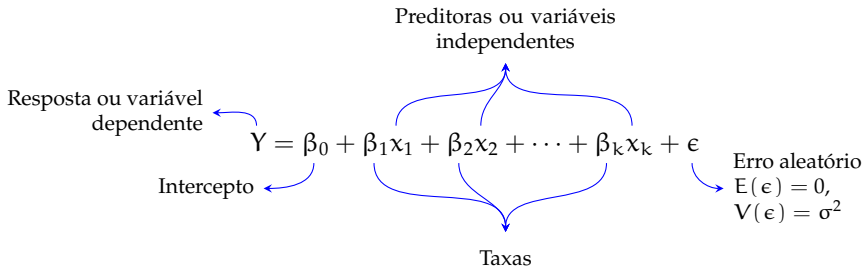


- $E(Y|x) = \beta_0 + \beta_1 x$.
- $V(Y|x) = V(\epsilon) = \sigma^2$.


Componente
determinístico
$\eta(x) = \beta_0 + \beta_1 x$
Componente
aleatório (ϵ)
y: resposta
x: preditora
0
20
40
0
2
4
6
8
10
12
14

Especificação

- Regressão linear múltipla



- $E(Y|x_i, i = 1, \ldots, k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.
- $V(Y|x_i, i = 1, \ldots, k) = V(\epsilon) = \sigma^2$.

Representação matricial

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

$$Y = X\beta + \epsilon$$

$$\underset{n \times 1}{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}} = \underset{\substack{n \times p \\ (p = k+1)}}{\begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}} \underset{p \times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}} + \underset{n \times 1}{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}$$

Estimação

- Critério de mínimos quadrados (ordinários)

$$\begin{aligned} SSE(\beta) &= \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 \\ &= (y - X\beta)^\top (y - X\beta) = \|y - X\beta\| \end{aligned}$$

- Estimador

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta \in \mathbb{R}^p} SSE(\beta) \\ &= (X^\top X)^{-1} X^\top y \end{aligned}$$

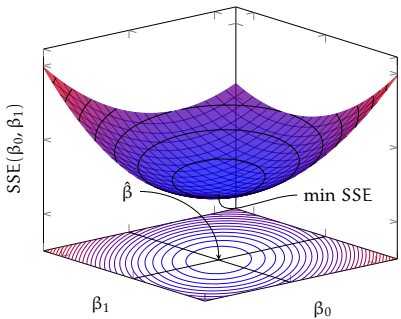


Figura 6: A superfície e mínimos quadrados.

Geometria dos mínimos quadrados

- Otimizar:

$$\begin{aligned} \text{SSE}(\beta) &= (y - X\beta)^\top (y - X\beta) \\ &= y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta. \end{aligned}$$

- Resolver o sistema:

$$\begin{aligned} \frac{\partial \text{SSE}}{\partial \beta^\top} &= 0 - 2X^\top y + 2X^\top X\hat{\beta} = 0 \\ X^\top X\hat{\beta} &= X^\top y \\ \hat{\beta} &= (X^\top X)^{-1} X^\top y. \end{aligned}$$

- $\hat{\beta}$ está no mínimo de SSE pois

$$\frac{\partial^2 \text{SSE}}{\partial \beta \partial \beta^\top} = X^\top X$$

é uma matriz positiva definida.

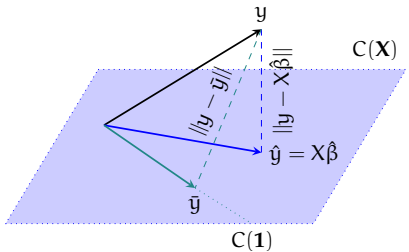
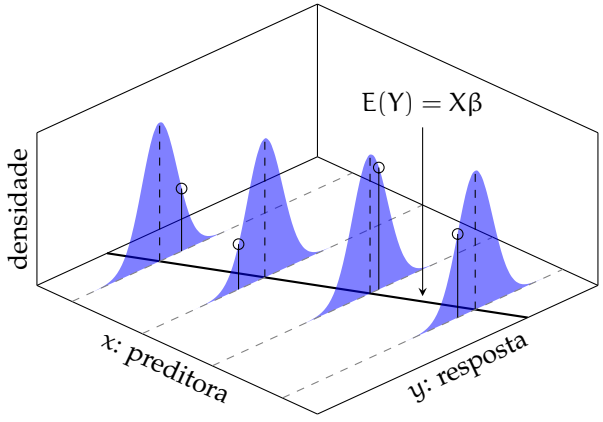


Figura 7: A interpretação geométrica do problema de mínimos quadrados.

Estimação baseada na verossimilhança



- Se $Y \sim \text{Normal}(\mu = X\beta, \sigma^2 = \sigma^2)$, então a log-verossimilhança é

$$ll(\beta, \sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{(y - X\beta)^\top (y - X\beta)}{2\sigma^2}. \quad (1)$$

- Os estimadores correspondem ao máximo da $ll(\theta), \theta = (\beta, \sigma^2)^\top$,

$$\frac{\partial ll(\theta)}{\partial \theta^\top} = 0, \qquad \hat{\beta} = (X^\top X)^{-1} X^\top y, \qquad \hat{\sigma}^2 = \frac{SSE}{n}. \quad (2)$$

Medidas de ajuste

- $R^2 = 1 - \dfrac{SSE(\beta)}{SSE(\beta_0)} = 1 - \dfrac{\|y - \hat{y}\|}{\|y - \bar{y}\|}.$
- $R^2_{adj} = 1 - \dfrac{n-1}{n-p}(1 - R^2).$
- $PRESS = \sum_{i-1}^{n} (y - \hat{y}_{i(-i)})^2$, menor é melhor;
- Log-verossimilhança (maior é melhor)

$$\begin{aligned} ll &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2) - \frac{\|y - X\hat{\beta}\|}{2\hat{\sigma}^2} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(SSE/n) - \frac{n}{2}, \qquad \hat{\sigma}^2 = SSE/n = \|y - X\hat{\beta}\|/n. \end{aligned}$$

- $AIC = 2(p+1) - 2ll$, menor é melhor;
- $BIC = \log(n)(p+1) - 2ll$, menor é melhor;

Medidas de diagnóstico e influência

- Matriz de projeção

$$\hat{y} = Hy, \qquad H = X(X^\top X)^{-1}X^\top,$$

  H é simétrica e indepotente. O posto de H é $\mathrm{tr}(H) = p$.

- Alavancagem (*leverage*)

$$h_i = H_{ii}$$
$$h = \mathrm{diag}(H).$$

- Resíduos ordinários, $V(\hat{e}) = \sigma^2(I - H)$,

$$\hat{e}_i = y_i - \hat{y}_i$$
$$\hat{e} = y - \hat{y}$$
$$\hat{e} = y - X\hat{\beta}.$$

- Resíduos padronizados (ou internamente studentizados),

$$r_i = \frac{\hat{e}_i}{s(\hat{e}_i)} = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_i}}.$$

- Resíduos studentizados (ou externamente studentizados),

$$t_i = \frac{\hat{e}_i}{s(\hat{e}_{i(-i)})} = \frac{\hat{e}_i}{\hat{\sigma}_{-i}\sqrt{1-h_i}}$$

$$\hat{\sigma}^2_{-i} = \frac{(n-p)\hat{\sigma}^2 - \dfrac{\hat{e}_i^2}{1-h_i}}{(n-1)-p}.$$

- Distância de Cook

$$D_i = \frac{(\hat{y} - \hat{y}_{i(-i)})^\top (\hat{y} - \hat{y}_{i(-i)})}{p\hat{\sigma}^2} = \frac{1}{p} \cdot \frac{h_i}{(1-h_i)} \cdot \frac{\hat{e}_i^2}{\hat{\sigma}^2(1-h_i)}.$$

- DFfits

$$dffits_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{\hat{\sigma}_{-i}\sqrt{h_i}} = t_i \left(\frac{h_i}{1-h_i}\right)^{1/2}.$$

- DFbetas

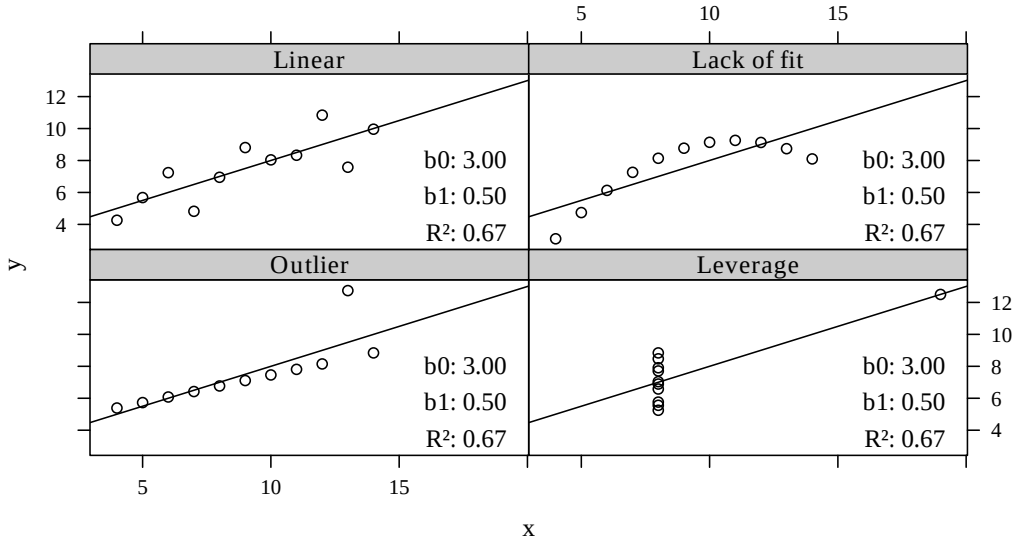
$$dbetas_i = \frac{\hat{\beta} - \hat{\beta}_{-i}}{\hat{\sigma}_{-i}\sqrt{diag((X^\top X)^{-1})}}$$

$$\hat{\beta}_{-i} = \hat{\beta} - \frac{\hat{e}_i}{1-h_i} \cdot (X^\top X)^{-1} x_i.$$

O quarteto de Anscombe

```
data(anscombe)
anscombe
```

```
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```


Linear
Lack of fit
Outlier
Leverage
b0: 3.00
b1: 0.50
R²: 0.67
y
x
4
6
8
10
12
5
10
15

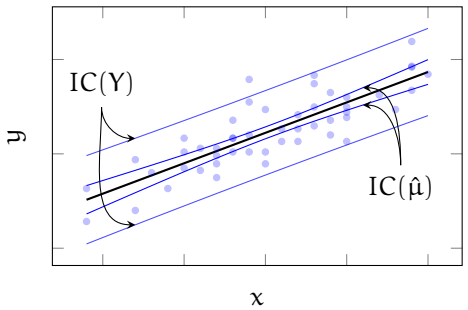


Figura 8: Bandas de confiança para $\hat{\mu}$ e de predição para Y.

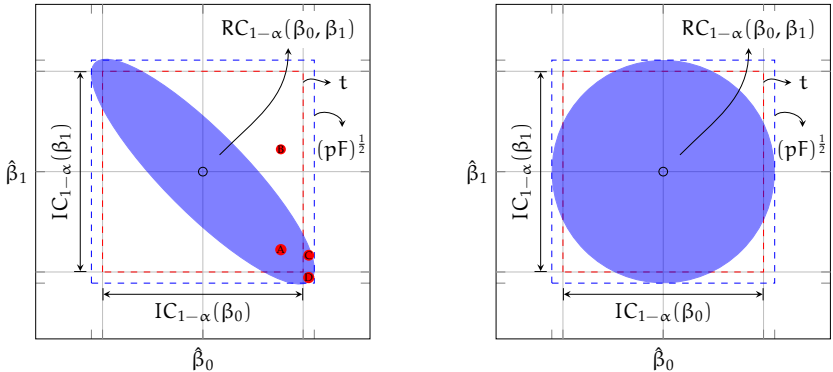


Figura 9: Região e intervalos de confiança para β.

Modelos de regressão não lineares

Benefícios

- Baseados em teoria ou princípios que dão uma relação funcional mais específica entre y e x;
- Parâmetros são interpretáveis;
- São parsimoniosos;
- Podem ser feitas predições fora do domínio observado de x;

Custos

- Requerem procedimentos iterativos de estimação;
- Métodos de inferência são aproximados;

Definição

Linear nos parâmetros

$$\eta(x, \theta) = \theta_0 + \theta_1 x + \theta_2 x^2.$$

$$\frac{\partial \eta}{\partial \theta_0} = 1,$$
$$\frac{\partial \eta}{\partial \theta_1} = x,$$
$$\frac{\partial \eta}{\partial \theta_2} = x^2.$$

Não linear nos parâmetros

$$\eta(x, \theta) = \theta_a(1 - \exp\{-\theta_e(x - \theta_c)\}).$$

$$\frac{\partial \eta}{\partial \theta_a} = 1 - \exp\{-\theta_e(x - \theta_c)\}$$
$$\frac{\partial \eta}{\partial \theta_e} = -\theta_a(\theta_c - x)\exp\{-\theta_b(x - \theta_c)\}$$
$$\frac{\partial \eta}{\partial \theta_c} = -\theta_a\theta_b \exp\{-\theta_b(x - \theta_c)\}.$$

Pontos característicos e formas

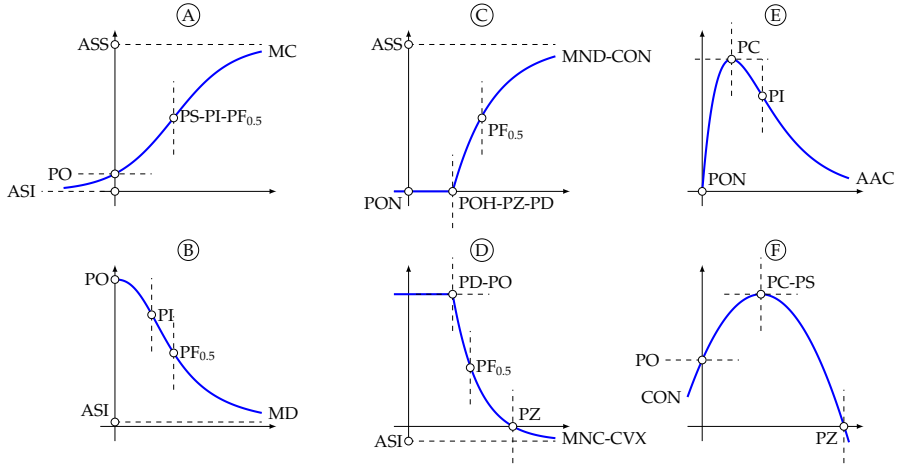
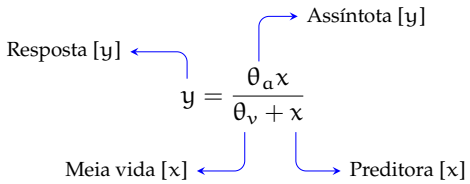
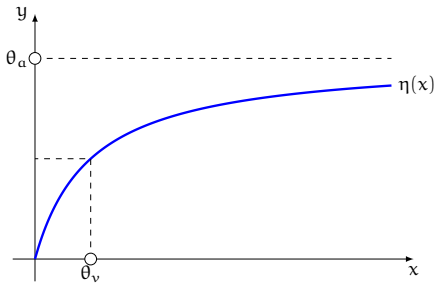


Figura 10: Funções não lineares com destaque para os pontos característicos e formas.

Determinação das unidades de medida (dimensionalidade)

- Modelo Michaelis-Menten





- $\lim_{x \to \infty} \eta(x) = \theta_a$.
- $\eta(x = \theta_v) = \theta_a/2$.

Regressão Local

- LOESS: LOcal regrESSion.
- Ajuste em subconjuntos localizados no domínio de x.
- Para cada ponto x_i no domínio de x, um polinômio de grau baixo é ajustado.
- Observações no entorno de x_i tem pesos que decrescem com a distância.
- A função peso mais usada é a tri-cubo: $(1 - |x|^3)^3 I(|x| < 1)$.
- Se polinômio de grau 0, corresponde à médias móveis.
- Sujeito à *outliers*.
- Requer dados densos para evitar ajustes locais.
- Computacionalmente intensivo.
- Difícil transferir resultados escritos do ajuste.

Splines

- Polinômios ajustados em subconjuntos disjuntos do domínio.
- Um spline é uma função polinomial por partes definida sobre os nós (knots):

$$\xi_1 < \xi_2 < \ldots < \xi_k.$$

- As funções se unem sobre os nós.
- Os nós são colocados nos quantis = mesmo n em cada parte.
- Nós podem ser distribuidos de acordo com a forma da função.

- Um spline de grau D com K é definido por

$$S(x) = \beta_0 + \sum_{d=1}^{D} + \sum_{k=1}^{K} \gamma_k((x - \xi_k)^D I(x \geqslant \xi_k)).$$

- Com um D-polinômio, tem-se:
  - A função $S(x)$ é contínua.
  - A $S(x)$ possui $D - 1$ derivadas.
  - A D-ésima derivada é contínua sobre os nós.
- Natural splines
  - Nas caudas usa-se polinômio de grau 1 (reta).
  - Melhor para previsão fora do domínio observado.
  - Isso as restrições de contínuidade para os nós do interior.

Smooth Splines

- São splines com n nós sobre o domínio de x.
- Normalmente usa-se 3-polinômio.
- Controle da suavidade por meio de penalização

$$PSS = \sum_{i=1}^{n} (y_i - S(x_i))^2 + \lambda \int (S''(x))^2 \, dx.$$

- Com λ fixo, otimiza-se a posição dos nós.
- A escolha do melhor λ pode ser por validação cruzada.

Resumo

- Modelos ”suaves”são bastante flexíveis.
- São mais sensíveis aos outliers.
- Não possuem equação.
- Podem depender de calibração do usuário.

Agradecimentos

- Comissão organizadora do Mgest.
- Participantes do Mgest.
- Colegas do LEG pelo incentivo e colaboração.