Mineração de Dados

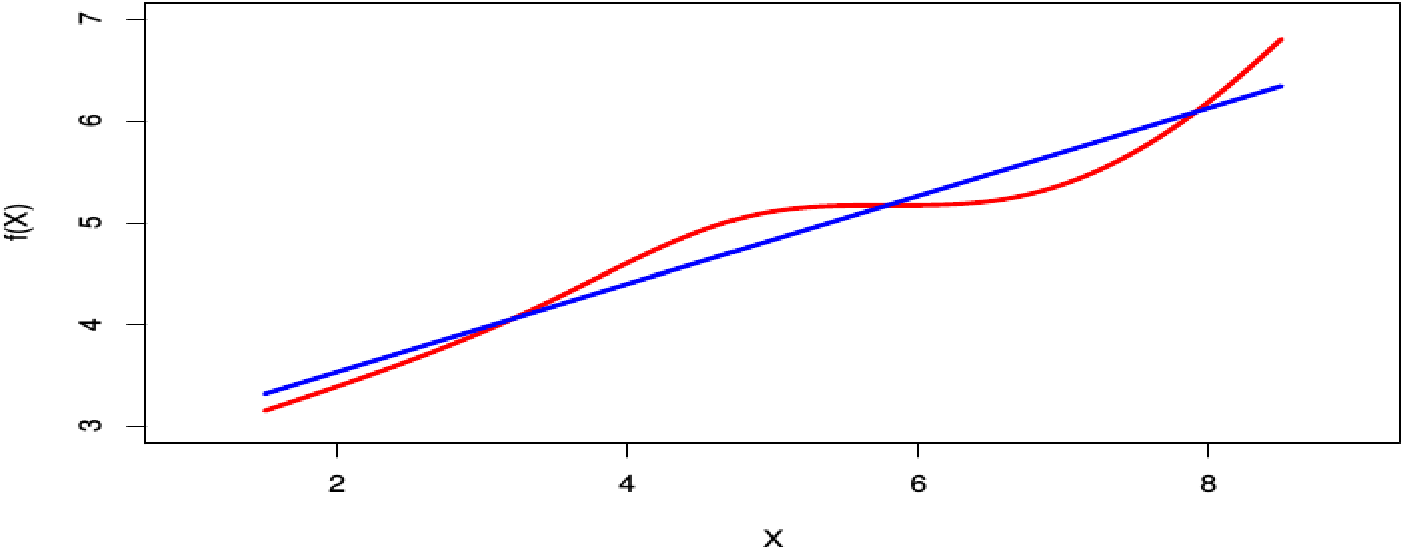
Aula 2 – parte 1

Especialização em Ciência de Dados e suas Aplicações

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Regressão linear

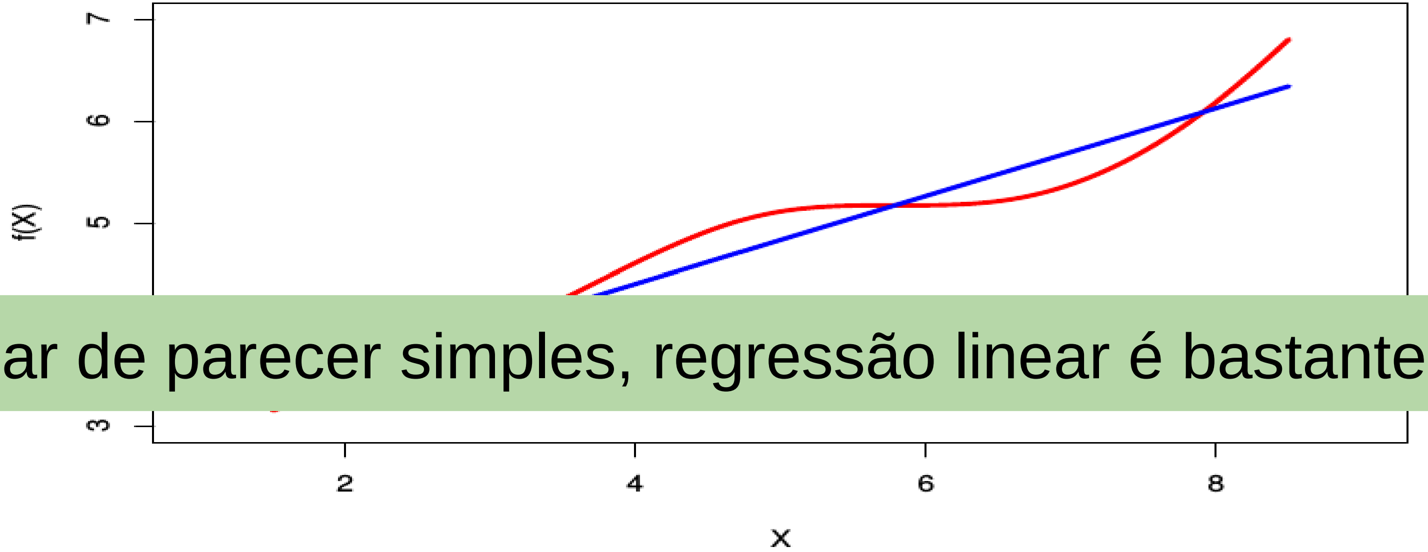
- É uma estratégia simples de aprendizado supervisionado
- Assume que a dependência de Y em $X_1, X_2, \ldots X_p$ é linear



A função real não é linear

Regressão linear

- É uma estratégia simples de aprendizado supervisionado
- Assume que a dependência de Y em $X_1,X_2,...X_p$ é linear



Apesar de parecer simples, regressão linear é bastante poderosa

A função real não é linear

Modelos de RL

- O que é um bom modelo?
- Como estimar os parâmetros do modelo?

Modelos de RL – Exemplo (propaganda)

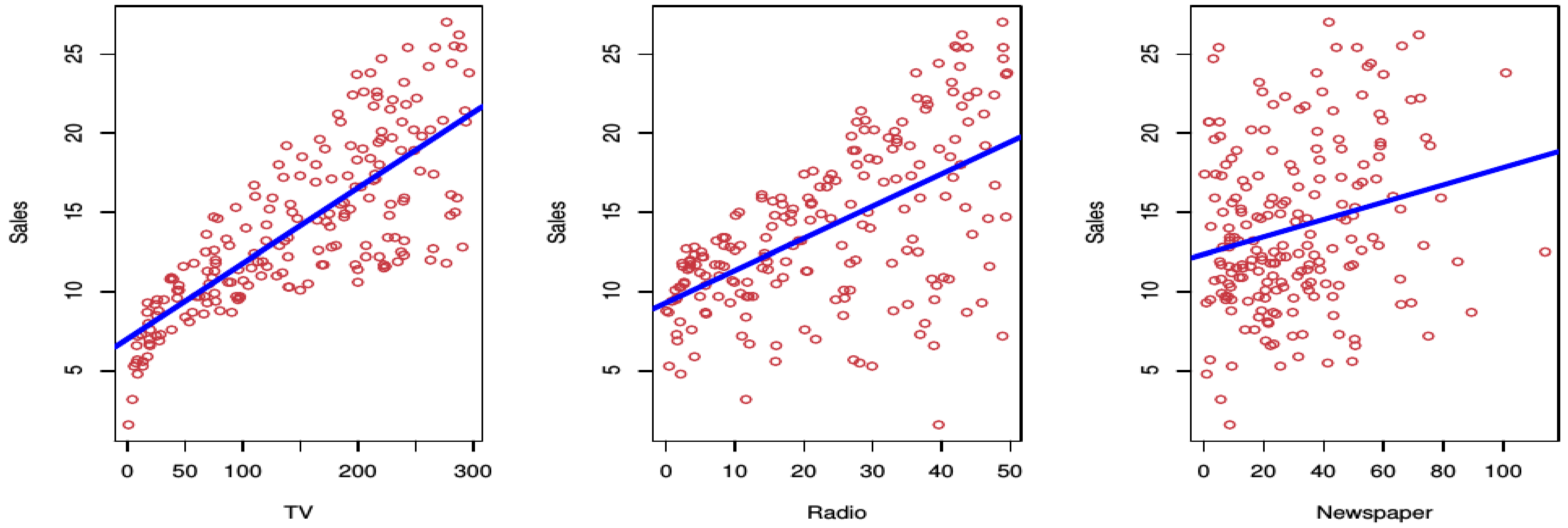
Existe um relacionamento entre propaganda e vendas?

O quão forte é o relacionamento entre o orçamento de propaganda e as vendas?

Qual mídia contribui para aumentar as vendas?

Quão precisamente podemos predizer vendas futuras?

RL simples com um único preditor



RL simples com um único preditor

Assumindo o modelo $Y = \beta_0 + \beta_1 X + \epsilon,$

onde β_0 é o ponto onde a reta cruza o eixo Y e β_1 é a inclinação da reta (coeficientes ou parâmetros) e ϵ é um erro

Dado algumas estimativas b_0 e b_1 para os coeficientes, nós realizamos previsões com

$$\hat{y} = b_0 + b_1 x$$

Previsão de Y

Estimando os parâmetros com mínimos quadrados

Se $\hat{y} = b_0 + b_1 x$ então o erro na estimativa para x_i é:

$$e_i = y_i - \hat{y}_i$$

Definimos a Soma dos Erros ao Quadrado (Sum of Squared Errors **SSE**)

$$SSE = e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2,$$

que é equivalente a:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Estimando os parâmetros com mínimos quadrados

A abordagem dos mínimos quadrados escolhe b_0 e b_1 que minimiza o **SSE**

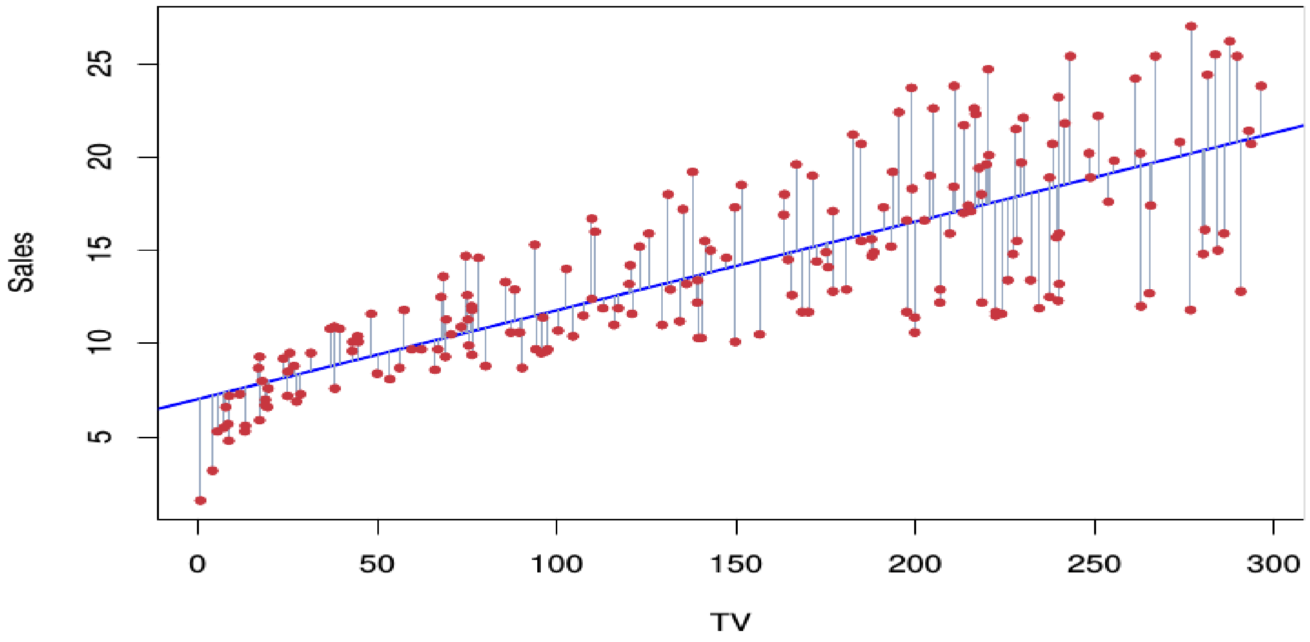
Os melhores parâmetros da regressão (que levam à menor variância dos erros) são:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

Onde $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ e $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

são as médias da amostra

Exemplo - propaganda



Uma aproximação linear captura a essência do relacionamento, apesar da “deficiência” no início

Exemplo – estimando os parâmetros

Tempo de execução de uma *query* para várias palavras:

x	Palavras	3	5	7	9	10
y	Tempo	1.19	1.73	2.53	2.89	3.26

x<-c(3,5,7,9,10)
y<-c(1.19,1.73,2.53,2.89,3.26)

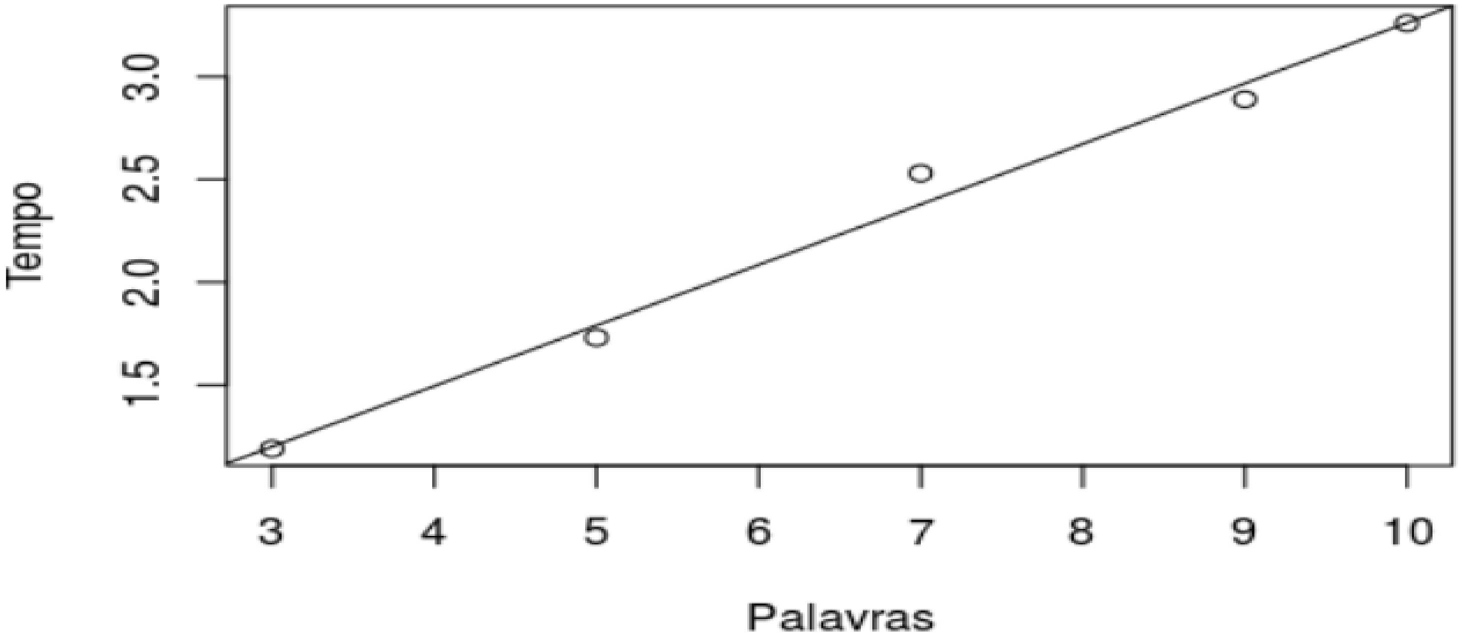
mediaY <- sum(y)/length(y)
mediaX = 2,32

mediaX <- sum(x)/length(x)
mediaX = 6,8

b_1 <- sum((x-mediaX)*(y-mediaY))/sum((x-mediaX)^2)
b_1 = 0,2945122

b_0 <- mediaY - (b1)*(mediaX)
b_0 = 0,3173171

Exemplo – estimando os parâmetros

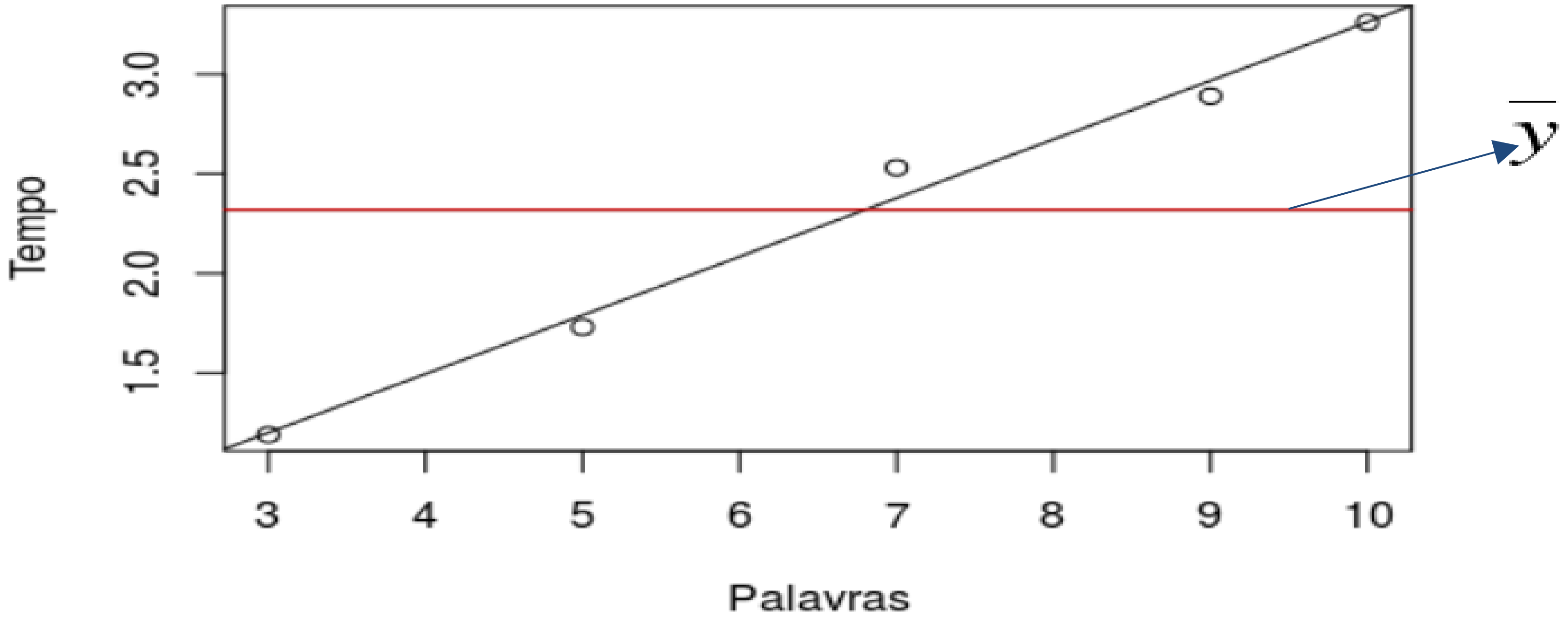



```
plot(x,y,xlab = "Palavras",ylab = "Tempo")
abline(b0, b1)
```

Acurácia geral do modelo

- Sem regressão, a melhor estimativa de y é $\bar{y}$
- Regressão provê uma melhor estimativa, mas ainda existem erros

Acurácia geral do modelo



Acurácia geral do modelo

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Sum of Squared Errors

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Total Sum of Squares

Qualidade da regressão medida pelo coeficiente de determinação:

$$R^2 = \frac{SST - SSE}{SST}$$

Quanto maior o valor de R^2, melhor a regressão (*fitting* dos dados).

Exemplo - Acurácia geral do modelo

Tempo de execução de uma query para várias palavras:

x	Palavras	3	5	7	9	10
y	Tempo	1.19	1.73	2.53	2.89	3.26

b_1 <- 0.2945122
b_0 <- 0.3173171
MediaX <- 2.32

$$R^2 = \frac{SST - SSE}{SST}$$

yPred <- b0+b1*x
1.200854   1.789878   2.378902   2.967927   3.262439

SSE<- sum((y-yPred)^2)
SST <-sum((y- mediaY)^2)
R^2 <- (SST-SSE)/SST   = **0.98**

Regressão linear pode ser “enganadora”

- Regressão despreza alguma informação sobre os dados
  - Para permitir uma sumarização compacta

- Algumas vezes características vitais são perdidas
  - No geral, examinando os gráficos de dados pode-se determinar se há um problema ou não

Exemplos de conjuntos

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

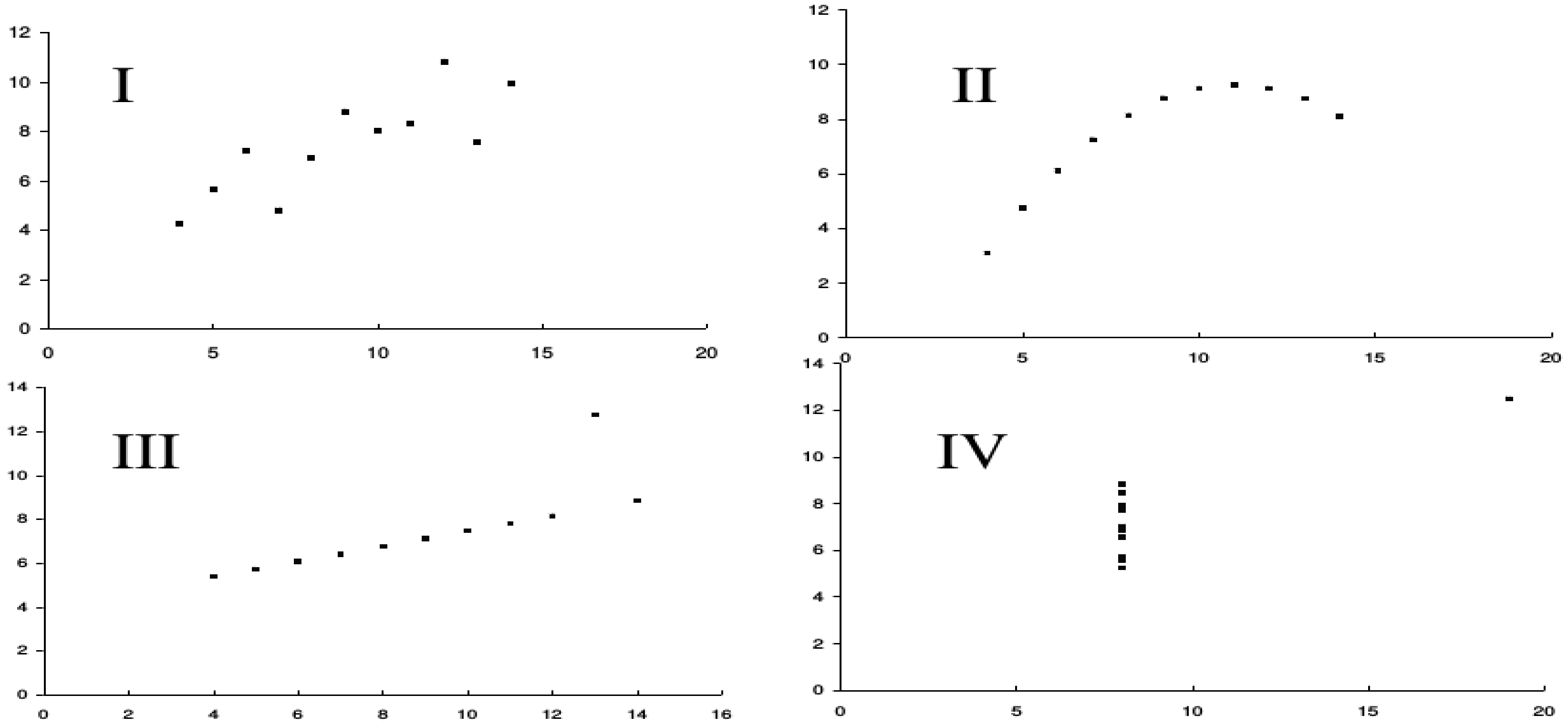
O que a regressão nos diz?

Exatamente a mesma coisa para cada um deles!

N = 11
Média de y = 7.5
Y = 3 + .5 X
Erro padrão da regressão é 0.118
Todas as somas de quadrados são as mesmas

R2 = .67

Visualização de cada conjunto



Regressão linear múltipla

- Modelos com mais de uma variável previsora

- Cada variável previsora tem uma relação linear com a variável de resposta

- Conceitualmente, seria equivalente a fazer um gráfico de uma linha de regressão num espaço n-dimensional, em vez de 2-dimensões

Regressão linear múltipla

Nosso modelo: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$

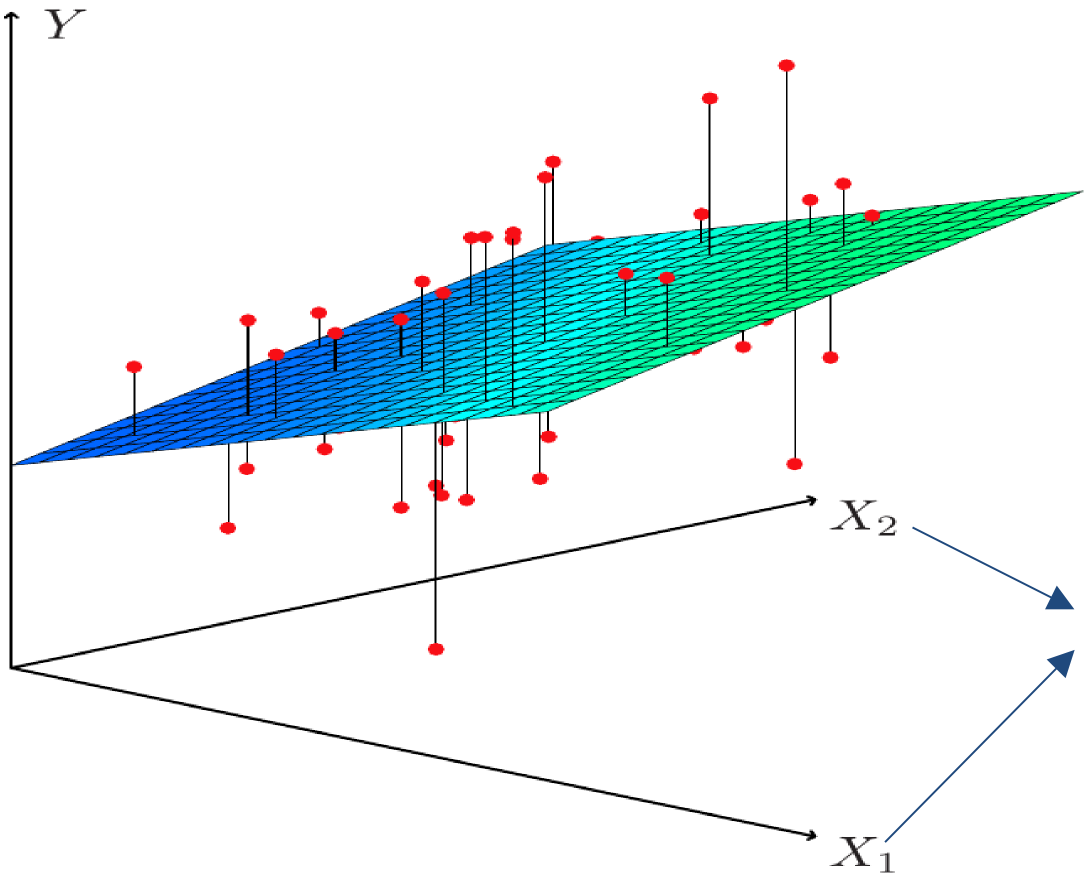
Interpretamos β_j como o efeito médio em Y no aumento de uma unidade em X_j, mantendo todos os outros parâmetros fixos.

No exemplo de propaganda:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

Regressão linear múltipla

Hiperplano para duas dimensões (difícil desenhar para 3+)



Variáveis previsoras

Interpretando os coeficientes

O cenário ideal é quando os previsores **não** são correlacionados

- Correlação entre previsores causa problemas:
  - *Interpretação é difícil: e.g., quando* X_j *muda, os outros previsores também mudam (no exemplo de propaganda, o orçamento de uma empresa pode ter aumentado para todas os tipo de propaganda)*

- Causalidade deve ser evitada para dados observacionais

Estimando e prevendo

Dado as estimativas b_0, b_1, b_2, ..., b_p

$$\hat{y} = b_0 + b_1 x_1 + b_2 x2 + \ldots + b_p x_p$$

Estimamos β_0 , β_1 , ... , β_p como os valores que minimizam a Soma dos Erros ao Quadrado (SSE)

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$
$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_p x_{ip})^2$$

Cálculo realizado com software estatístico. Fórmula *messy*

Procedimento similar à regressão simples

Resultado do exemplo

Valor das vendas quando todos os investimentos são 0

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Não é significativo na presença de TV e rádio no modelo

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

São correlacionados

Indício de que ao colocar rádio no modelo newspaper não é necessário

Avaliação

$$R^2 = \frac{SST - SSE}{SST}$$

Quantity	Value
Residual Standard Error	1.69
R^2	0.897

Melhorou com relação a uma única variável

Regressão Ridge

Os métodos de seleção de subconjunto usam mínimos quadrados (SSE) para ajustar um modelo linear que contém um subconjunto dos preditores.

Como alternativa, podemos ajustar um modelo contendo todos os p preditores usando uma técnica que restringe ou regulariza as estimativas do coeficiente

>> ou seja, que diminuem a estimativas de coeficiente para próximo de zero.

Regressão Ridge

Métodos que se baseiam em mínimos quadrados escolhem B_0,...B_j que minimizam:

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Ridge estima coeficiente que minimizam:

$$\text{SSE} + \lambda \sum_{j=1}^{p} \beta_j^2$$

Regressão Ridge

Métodos que se baseiam em mínimos quadrados escolhem $B_0,...B_j$ que minimizam:

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

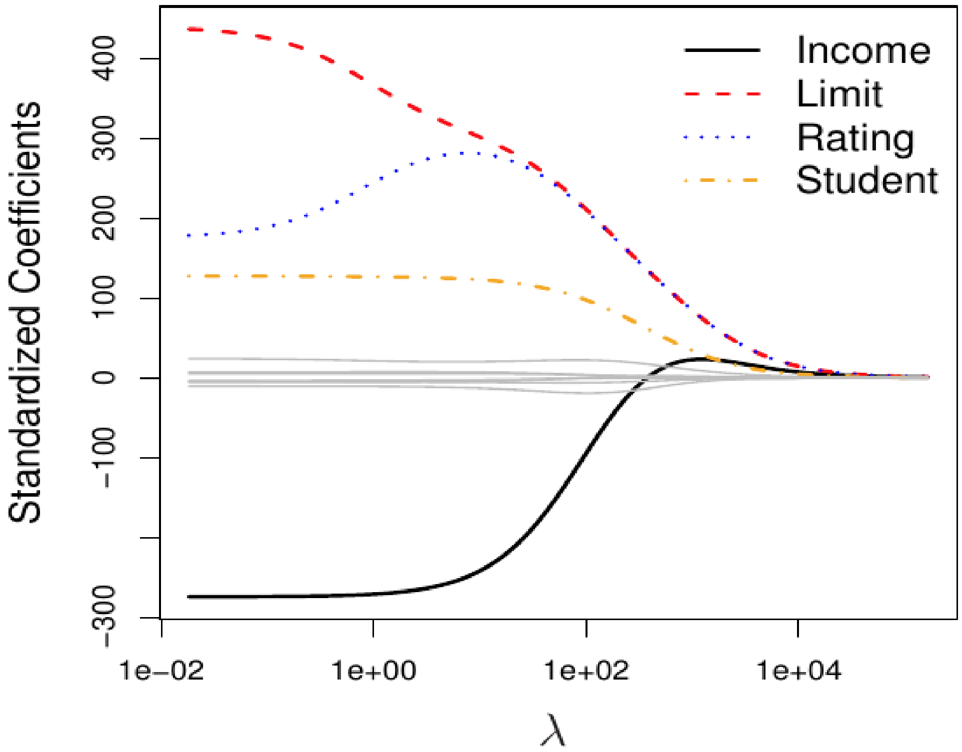
Ridge estima coeficiente que minimizam:

$$\text{SSE} + \boxed{\lambda \sum_{j=1}^{p} \beta_j^2}$$

Penalidade para valores de coeficientes altos

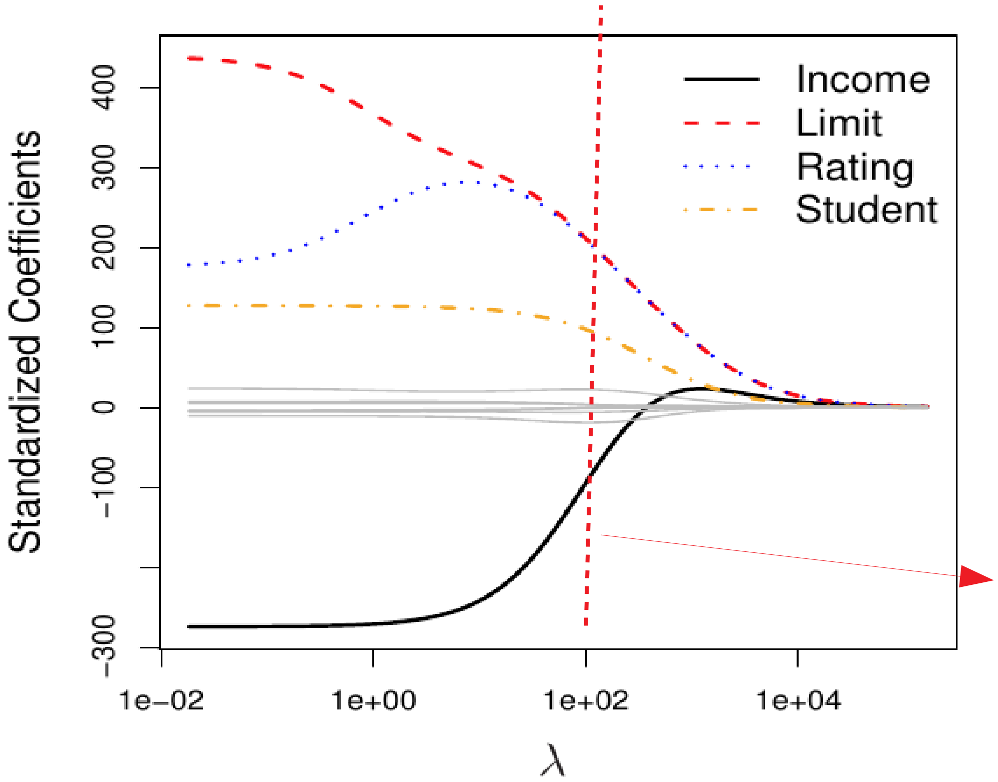
Parâmetro que precisa ser determinado separadamente

Regressão Ridge



Não seleciona variáveis, mas deixa os coeficientes próximo de zero.

Regressão Ridge



λ = 100, e os diferentes parâmetros do modelo estimado

Outras métricas para avaliar a regressão

Mean Absolute Error (MAE) $$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE) $$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Pune erros maiores

Root Mean Squared Error (RMSE) $$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Interpretável nos valores de Y

R^2 não tem unidade associada

Agradecimentos

Slides parcialmente derivados do material de aula de:

Jussara Almeida e Virgilio Almeida – Departamento de Ciência da Computação da UFMG. Curso: Métodos Quantitativos para a Ciência da Computação Experimental

Curso do livro The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Trevor Hastie, Robert Tibshirani e Jerome Friedman

Livro: Raj Jain. The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling.