Resumo da Aula

Regressão Linear:

- Simples;
- Múltipla;
- Minimização;
- Diagnóstico.

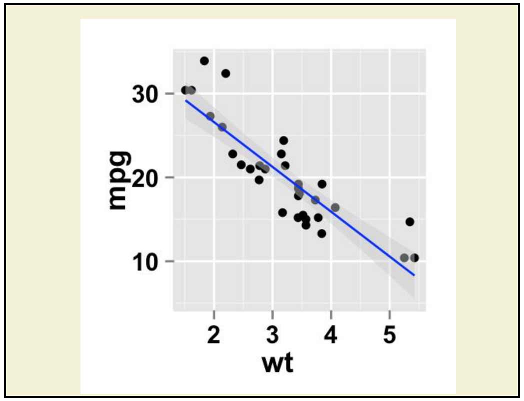
Regressão Logística:

- Arcabouço probabilístico;
- Exemplo.

Regressão Penalizada:

- Ridge;
- LASSO;

Regressão Linear



$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p.$$

Regressão Simples

Suponha que você tem um conjunto de dados com duas ou mais variáveis.

Queremos criar um modelo da forma

$$Y = f(X) + \epsilon$$

onde X representa as variáveis de entrada e ϵ

o erro.

Regressão Simples

Para que serve f(X)??

1 - Fazer predições de Y quando X=x

2 - Entender que componentes de

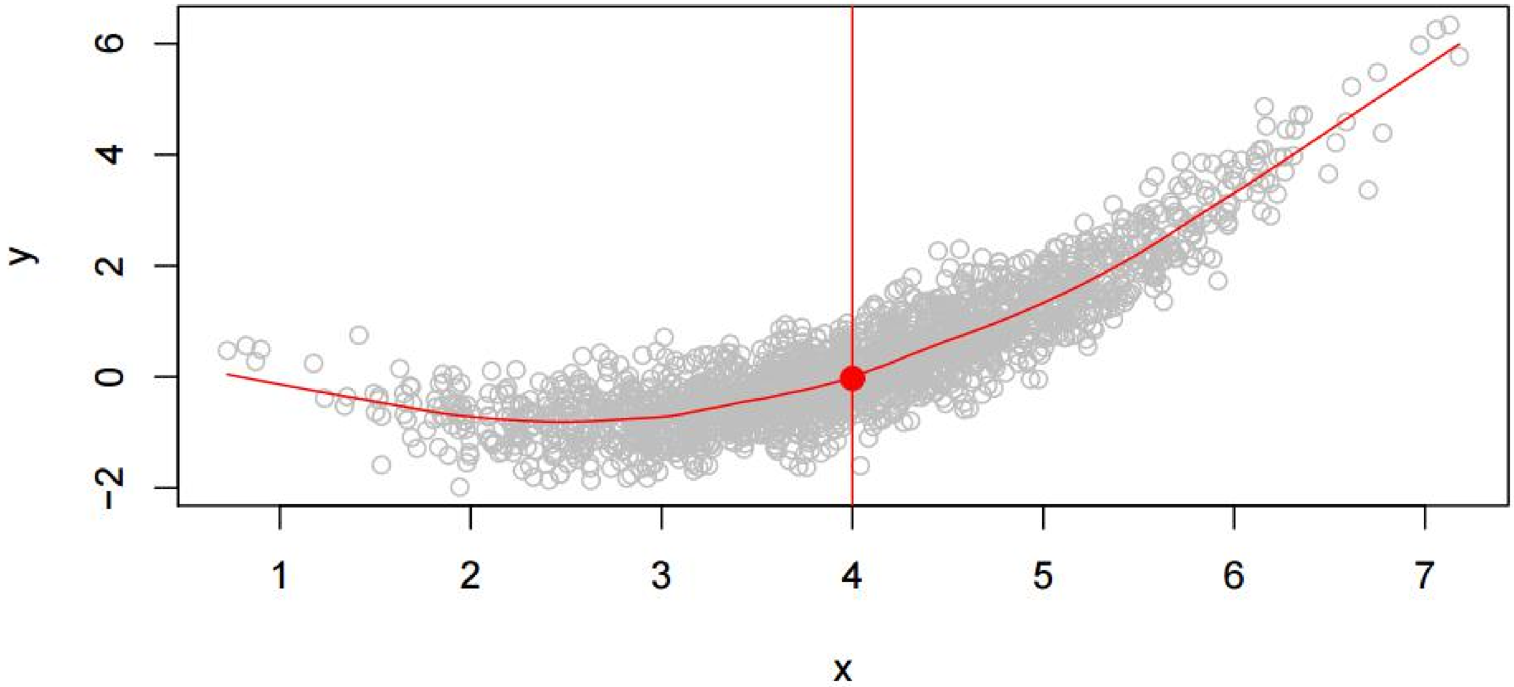
$$X = (X_1, X_2, \ldots, X_p)$$

ϵ

são importantes para explicar Y.

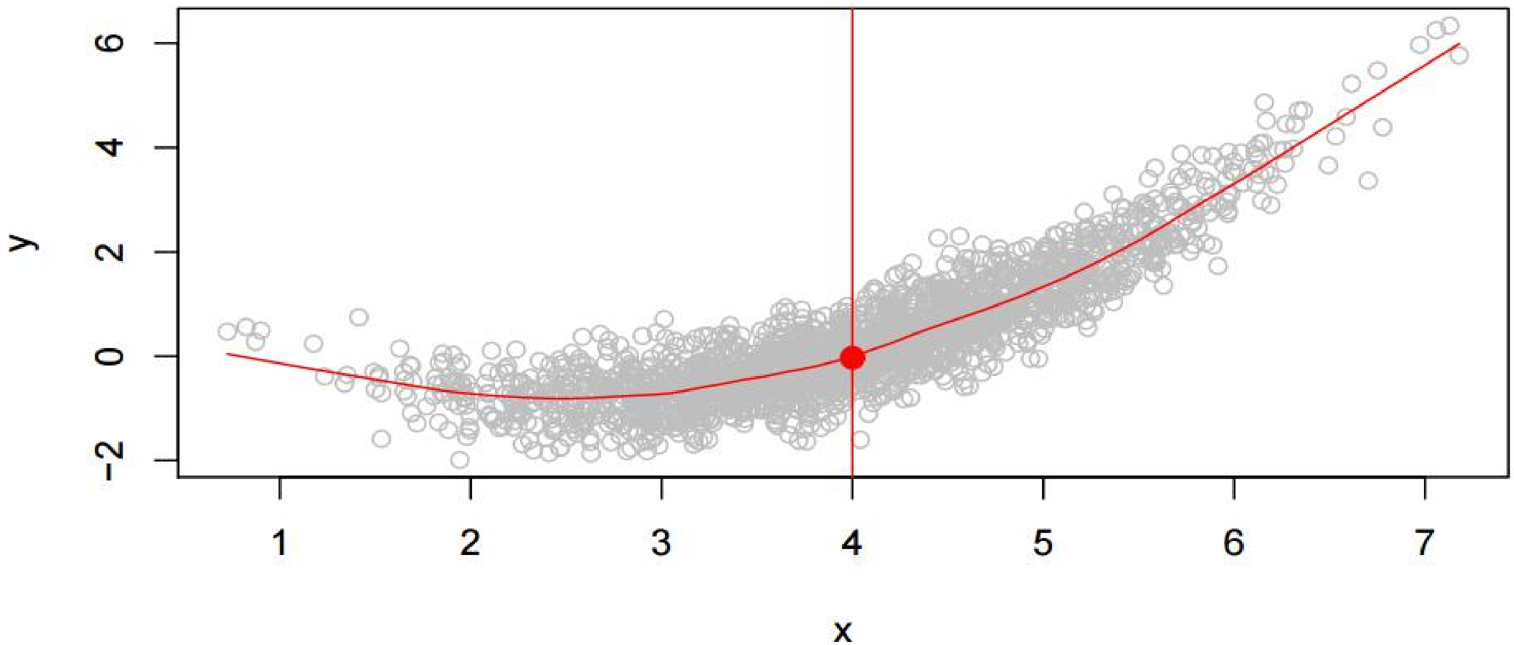
Regressão Simples

Existe um f(X) ideal?



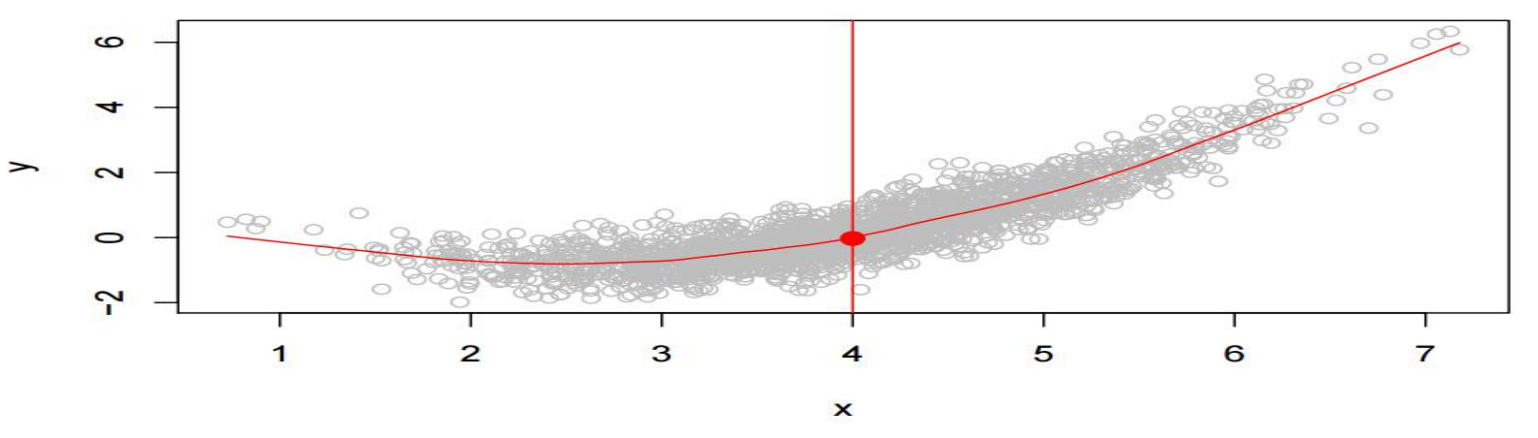
Regressão Simples

O que seria uma bom valor de Y quando X = 4.



OBS: há vários Y's para X = 4 !!!

Regressão Simples



Uma possibilidade é:

$$f(4) = E(Y|X = 4)$$

Aqui E(Y|X=4) é o valor esperado de Y quando

X = 4.

Função de Regressão

A função de regressão

$$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

é um preditor **ideal** de Y se eu minimizar a função

$$E[(Y - g(X))^2|X = x]$$

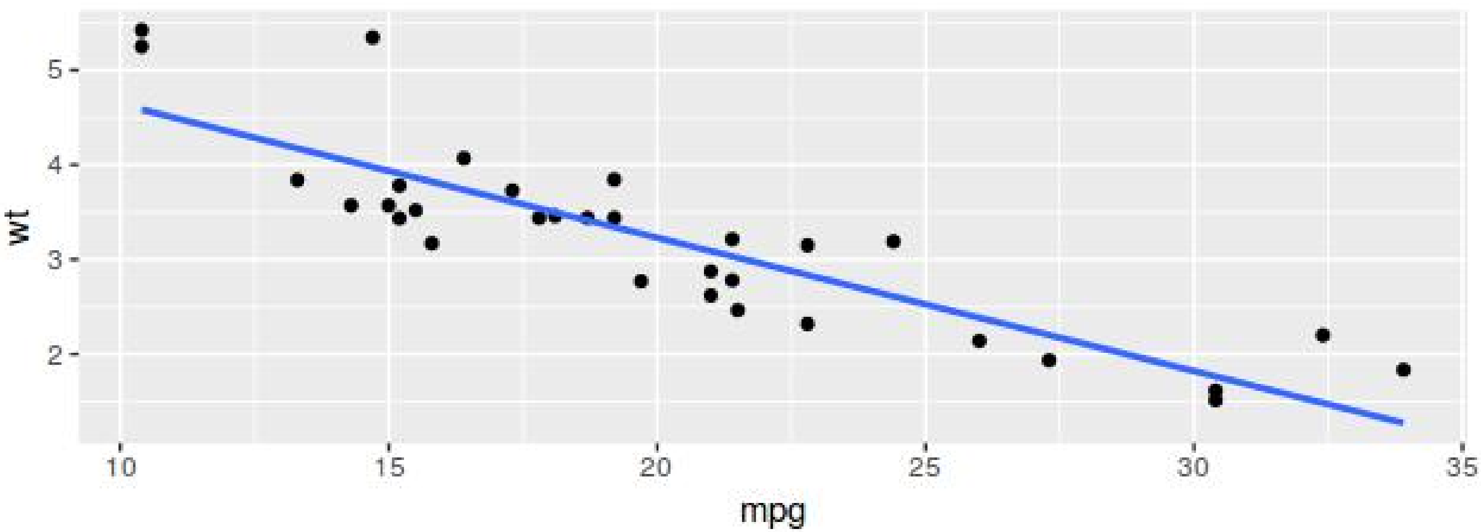
com relação a todos os possíveis g(X) para todos os pontos X = x.

Função Custo

Assumimos um modelo

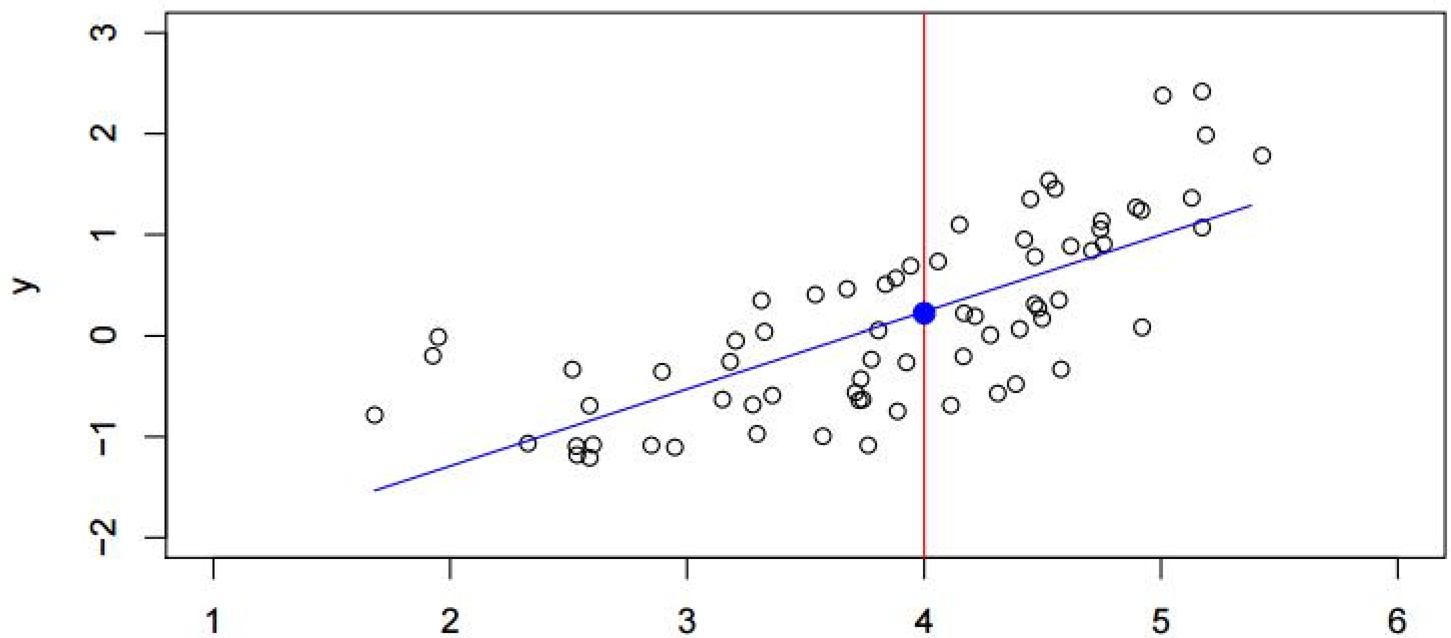
$$Y = \beta_0 + \beta_1 X + \epsilon$$

tal que queremos escolher β_0 e β_1 tal que a reta fica o mais próximo possível dos pontos.



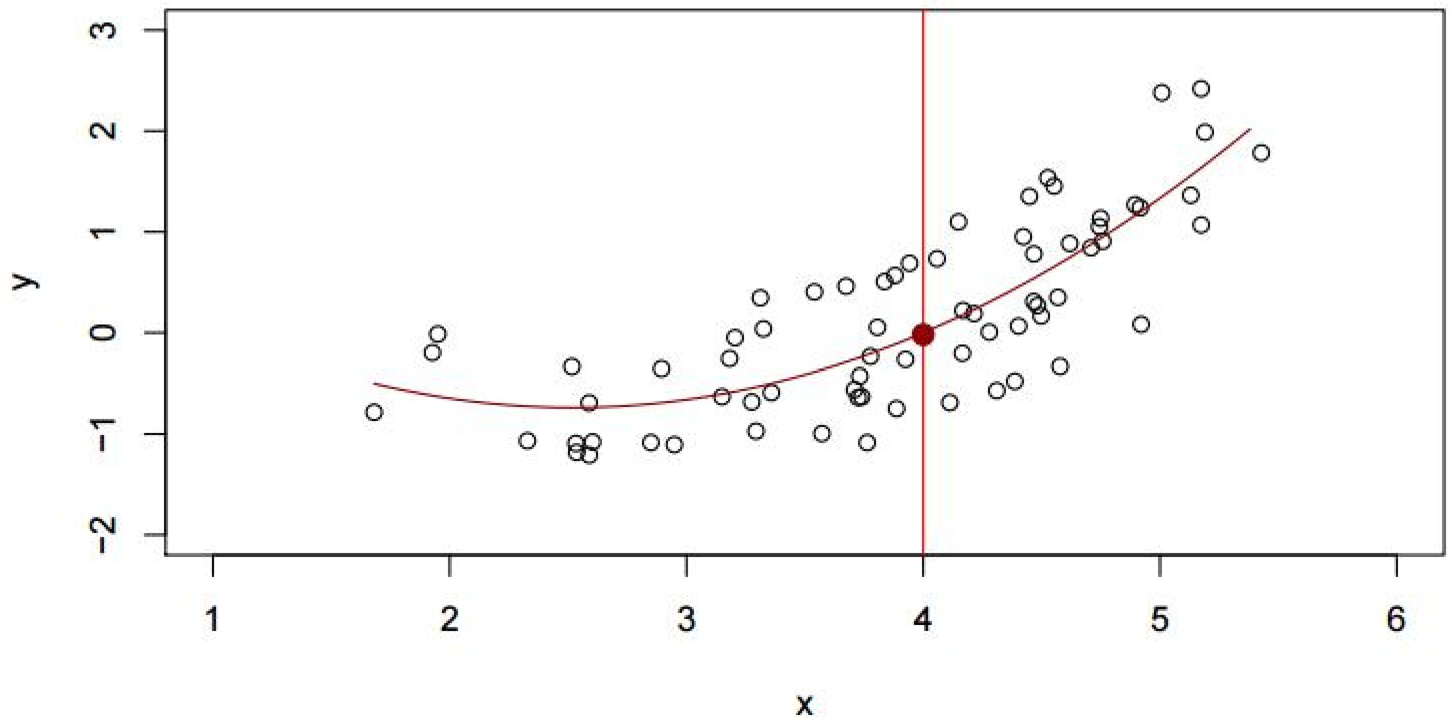
Função Custo

$$\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$



Função Custo

$$\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$$



Função Custo

$$min_{\beta_0,\beta_1}(f(X)-Y)^2$$

ou pensando que $f(x^{(i)})$ é função para a observação i e y^i é a i-ésima observação de Y, então:

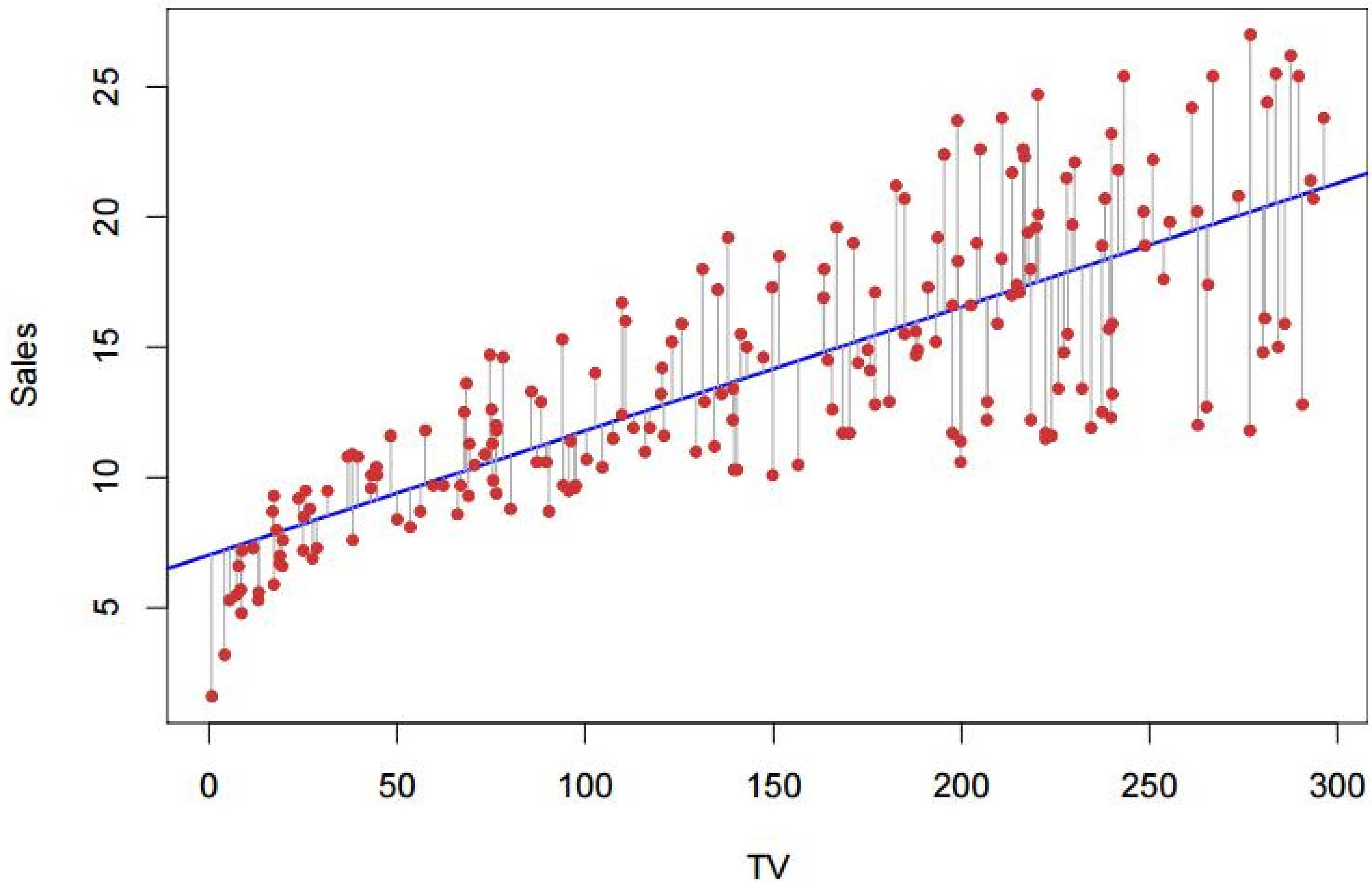
$$J(\beta_0,\beta_1)=\frac{1}{n}\sum_{i=1}^{n}(f(x^{(i)})-y^{(i)})^2$$

$J(\beta_{0,}\ \beta_1)$ é a função custo que deve ser minimizada.

$$min_{\beta_0,\beta_1}J(\beta_0,\beta_1)$$

Função Custo

$$min_{\beta_0,\beta_1} J(\beta_0,\beta_1)$$



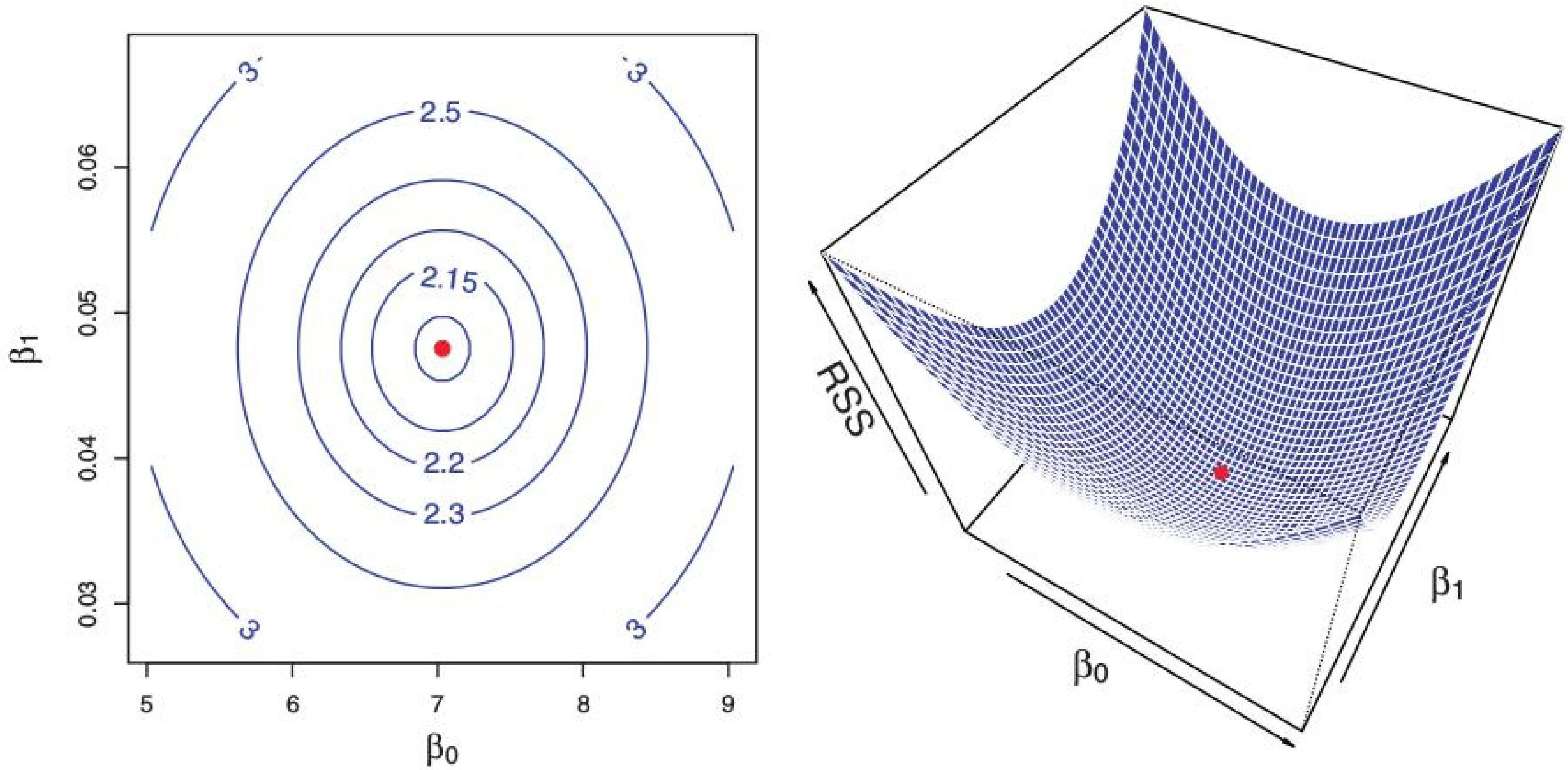
Minimização

$$J(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x^{(i)} - y^{(i)})^2$$

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Minimização



Minimização

Quero $min_{\beta} \, J(\vec{\beta})$

Repita {

$$\beta_j = \beta_j \quad \alpha \frac{\quad}{\beta_j} J(\vec{\beta})$$

simultaneamente para todos os β's.

}

Estimativa de σ

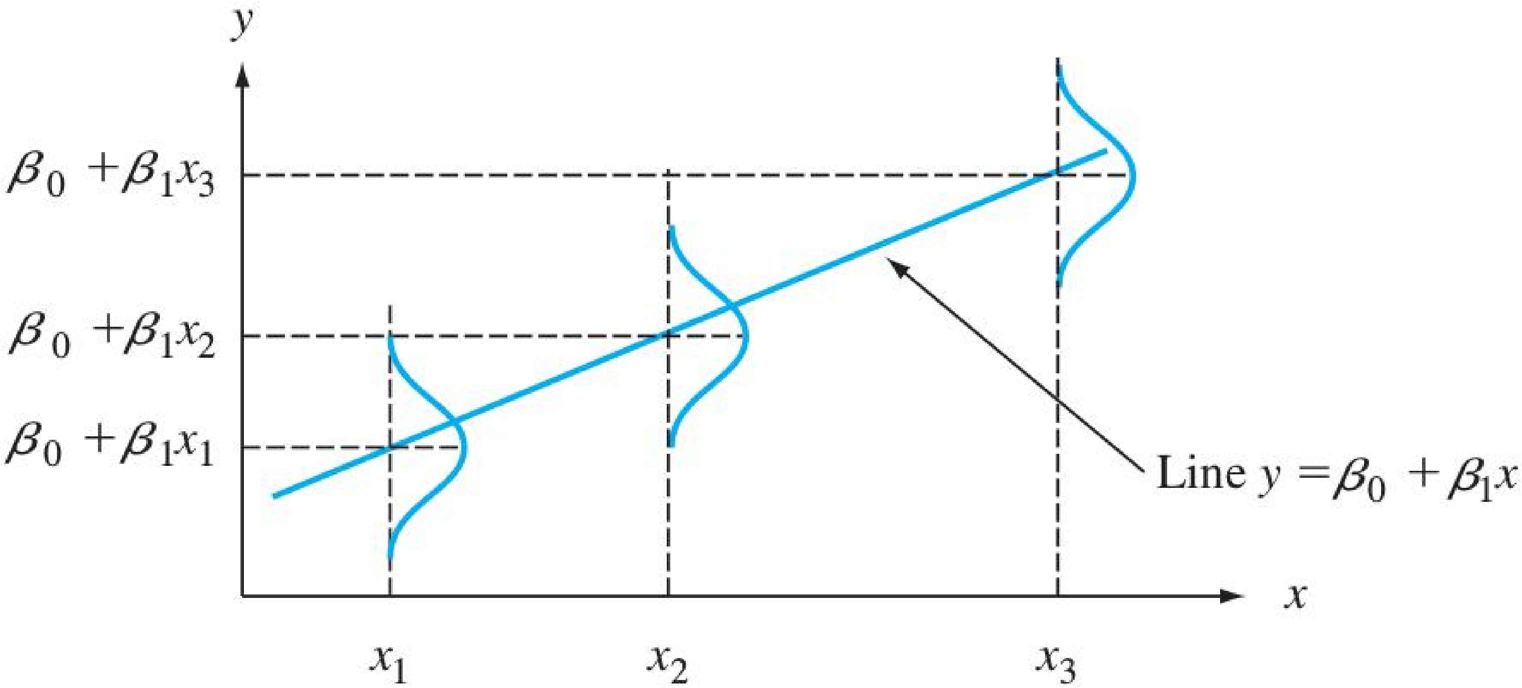
É possível obter σ a partir de SSE ou SE que é a soma dos erros do modelo:

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

tal que podemos estimar σ^2 com

$$\sigma^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

Regressão Simples (hipóteses)



Acurácia das estimativas do βs

Temos que

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

assim posso fazer um IC 95% por meio de

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

Acurácia das estimativas do βs

Isto é, existe aproximadamente 95% de chance que o intervalo

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

vai conter o verdadeiro valor de β_1.

Teste de Hipóteses

Com as mesmas informações que fizemos o IC podemos testas as hipósteses:

H_0: Não existe relação entre X e Y

H_0: Existe relação entre X e Y

O que corresponde matemáticamente a

H_0: $\beta_1 = 0$

H_0: $\beta 1 \neq 0$

Teste de Hipóteses

Para testar a hipótese computamos a estatística

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

aque tem distribuição t de student com n-2 graus de liberdade, sob a hipótese de que $\beta_1 = 0$.

Utilizando qualquer software estatístico é fácil obter a probabilidade de observar qualquer valor igual ou maior a |t|, o que chamamos de p-valor.

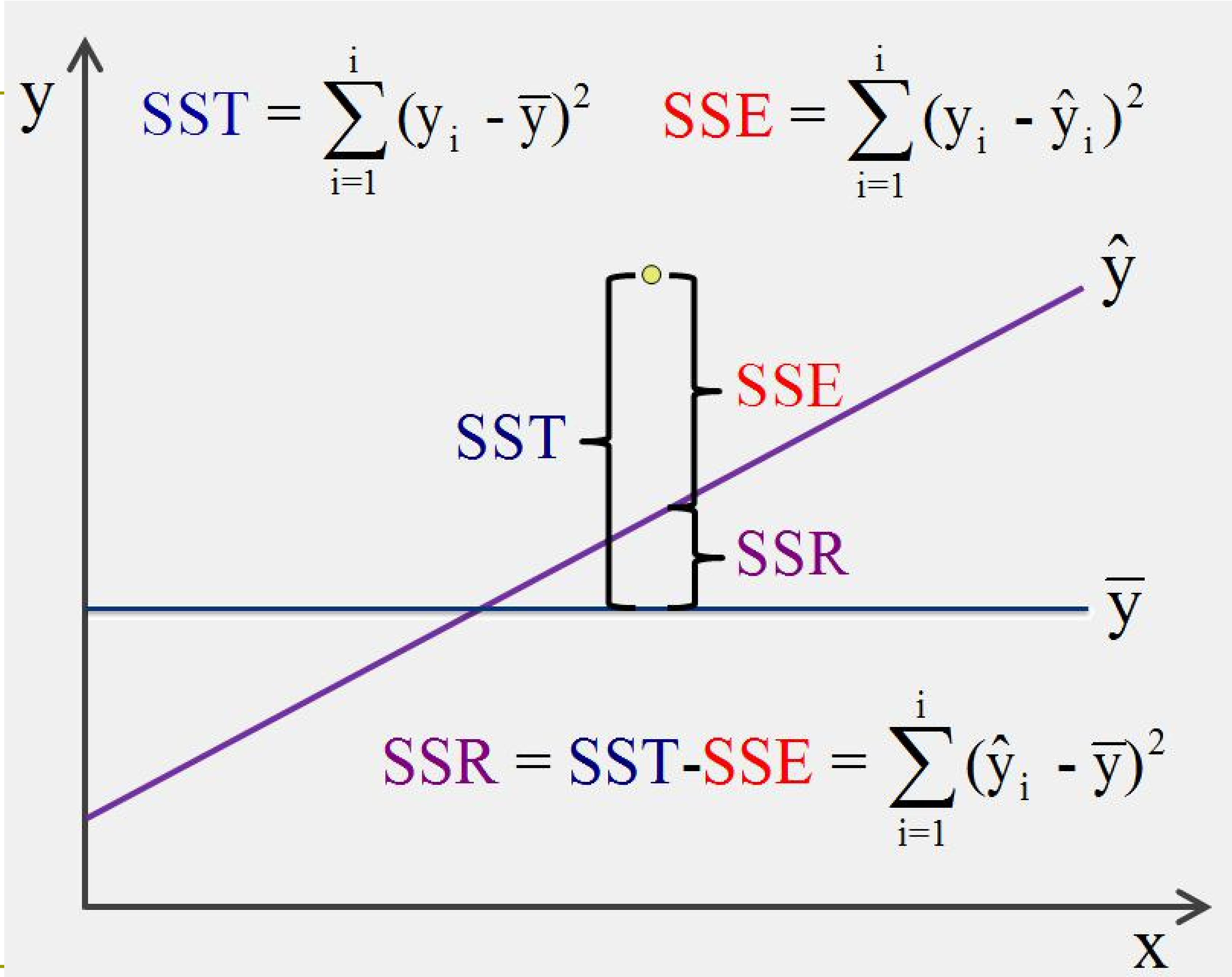
Coeficiente de Determinação

Soma dos quadrados totais

$$SST = S_{yy} = \sum (y_i - \hat{y}_i)^2$$

podemos utilizar

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$



$SST = \sum_{i=1}^{i} (y_i - \bar{y})^2$
$SSE = \sum_{i=1}^{i} (y_i - \hat{y}_i)^2$
$SSR = SST - SSE = \sum_{i=1}^{i} (\hat{y}_i - \bar{y})^2$
y
x
$\hat{y}$
$\bar{y}$
SST
SSE
SSR

Saída do Minitab

```
The regression equation is
cet num = 75.2 − 0.209 iod val
```

$\hat{\beta}_0$ $\hat{\beta}_1$

Predictor	Coef	SE Coef	T	P
Constant	75.212	2.984	25.21	0.000
iod val	−0.20939	0.03109	−6.73	0.000

$100r^2$

```
s = 2.56450   R-sq = 79.1%   R-sq(adj) = 77.3%
```

Analysis of Variance

SSE

SOURCE	DF	SS	MS	F	P
Regression	1	298.25	298.25	45.35	0.000
Error	12	78.92	6.58		
Total	13	377.17			

SST

Regressão Múltipla

No caso de mais do que uma variável explanatória, nosso modelo será

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

onde intrepetamos β_j como o **efeito médio** em Y da variação de uma unidade de X_j, **mantendo todas as outras variáveis fixas**.

O **cenário ideal** é quando os preditores **não são correlacionados**.

Regressão Múltipla

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Novamente para encontrar os βs minimizamos

$$min_{\vec{\beta}}(f(X)-Y)^2$$

tal que

$$J(\vec{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x^{(i)} + \ldots + \beta_p x^{(i)} - y^{(i)})^2$$

Minimização

Quero $min_{\beta} J(\vec{\beta})$

Repita {

$$\beta_j = \beta_j \quad \alpha \frac{}{\beta_j} J(\vec{\beta})$$

simultaneamente para todos os β's.

}

Minimização

Se o modelo for $Y = \beta_0 + \beta_1 X + \epsilon$

então

$$J(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x^{(i)} - y^{(i)}\right)^2$$

e

$$\frac{\partial}{\partial \beta_0} J(\beta_0, \beta_1) = \frac{2}{n} \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x^{(i)} - y^{(i)}\right)$$

$$\frac{\partial}{\partial \beta_1} J(\beta_0, \beta_1) = \frac{2}{n} \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x^{(i)} - y^{(i)}\right) x^{(i)}$$

Minimização

Assim o algoritmo fica

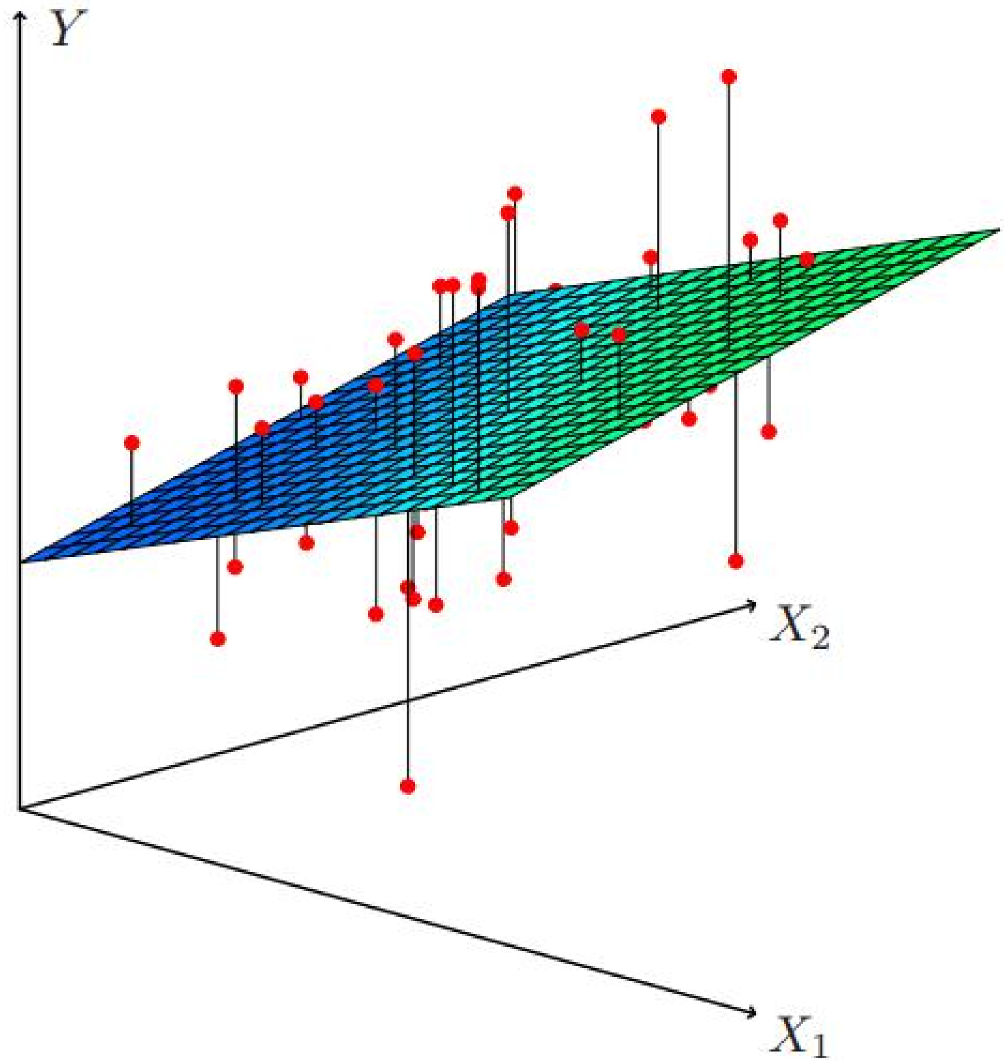
repita até a convergência {

$$\beta_0 \leftarrow \beta_0 - \alpha \frac{2}{n} \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x^{(i)} - y^{(i)} \right)$$

$$\beta_1 \leftarrow \beta_1 - \alpha \frac{2}{n} \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x^{(i)} - y^{(i)} \right) x^{(i)}$$

}

Minimização



Perguntas

1) Pelo menos algum dos preditores é útil para prever a variável resposta?

2) Todas as variáveis preditoras são necessárias ou somente um subconjunto delas?

3) Quão bem o modelo se ajusta aos dados?

4) Dado um conjunto de valores das variáveis preditoras, qual é o valor que vamos prever e quão precisa é esta predição?

Perguntas

1) Pelo menos algum dos preditores é útil para prever a variável resposta?

A estatística de teste será

$$F = \frac{(SST - SSE)/p}{SSE/n-p-1} \sim F_{p,n-p-1}$$

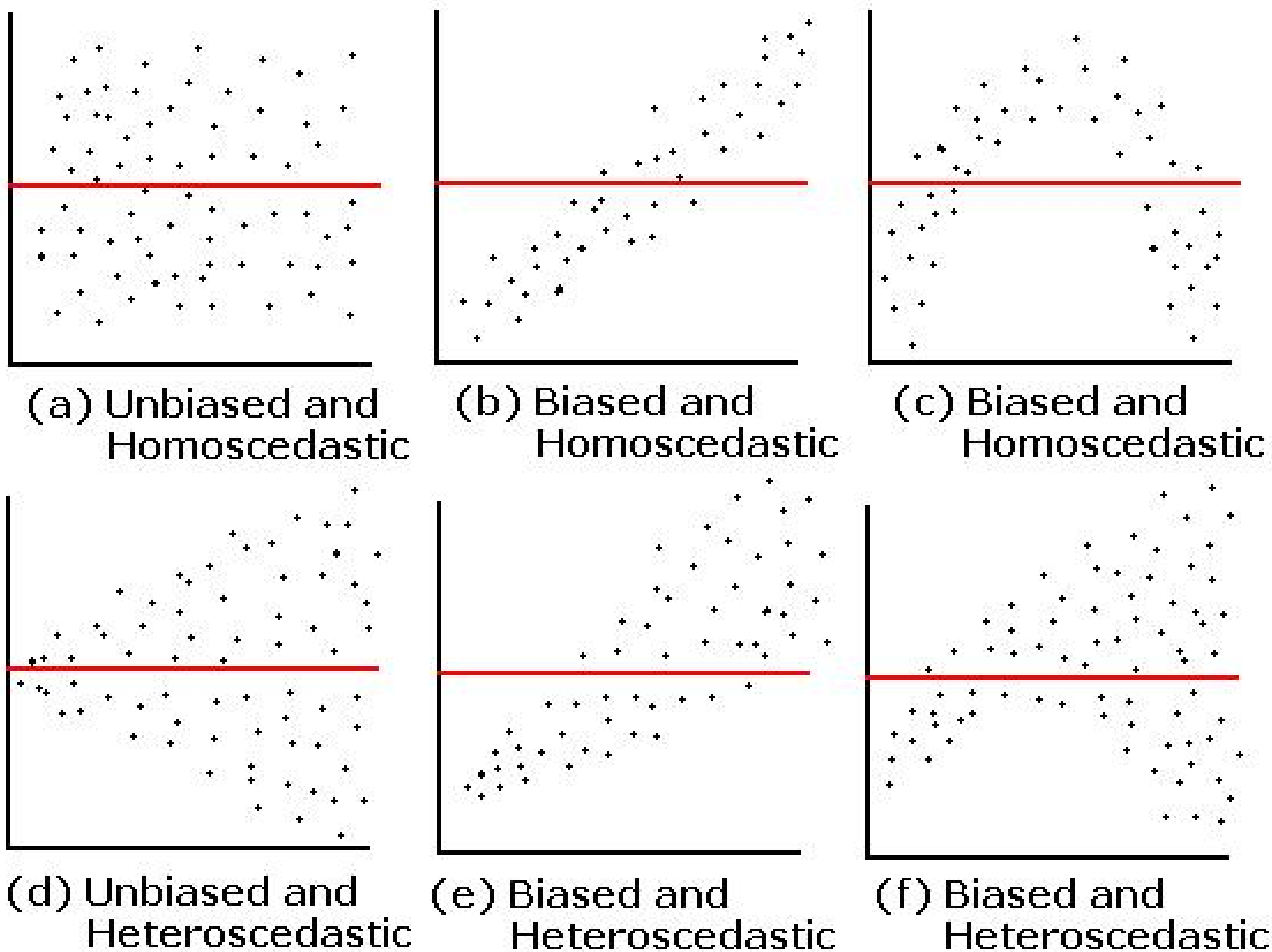
Perguntas

2) Todas as variáveis preditoras são necessárias ou somente um subconjunto delas?

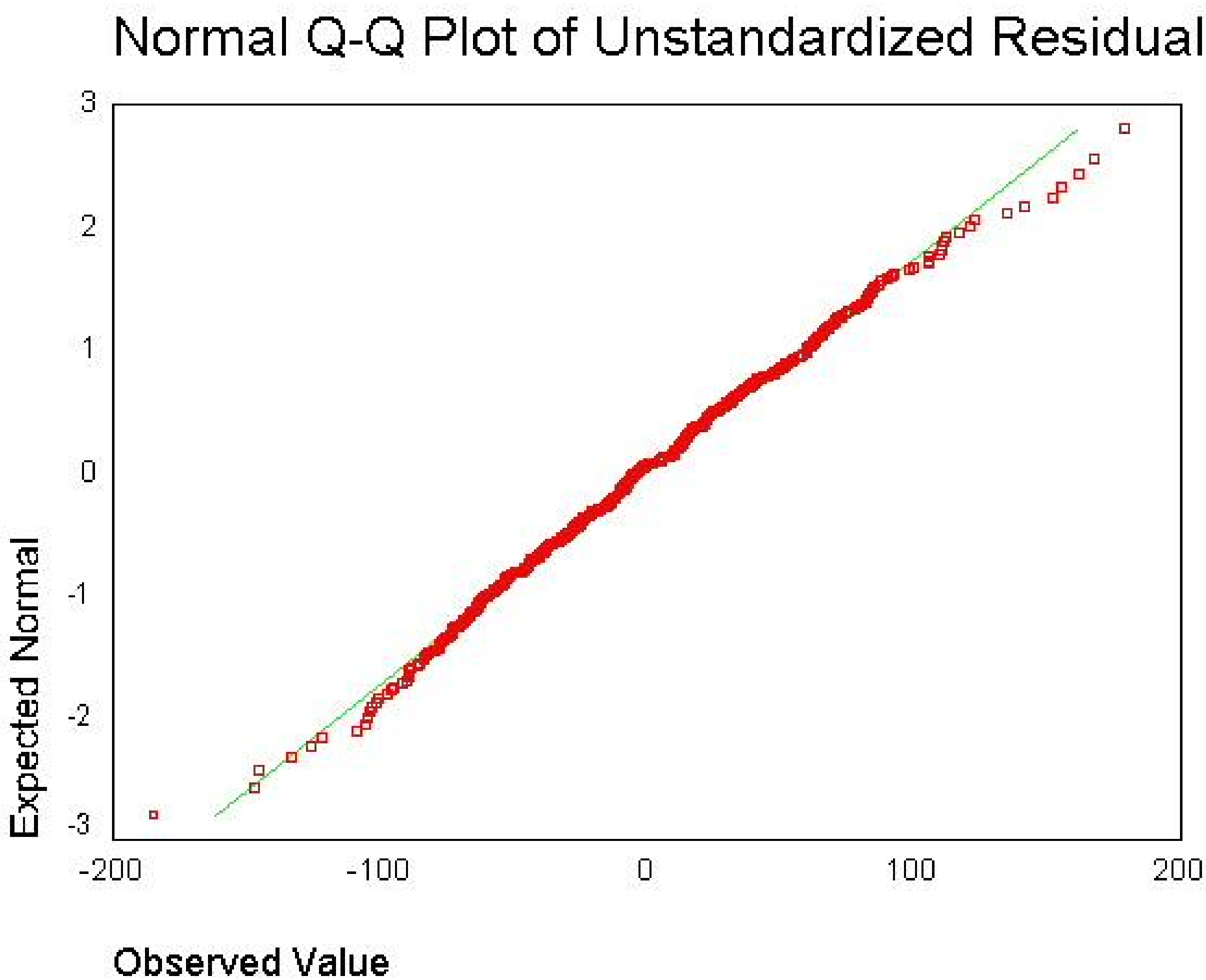
O mais direto é testar todos os subconjuntos de especificações e compara-los por meio de algum critério (C_p de Mallows, BIC, AIC e R^2 ajustado.

Usar stepwise.

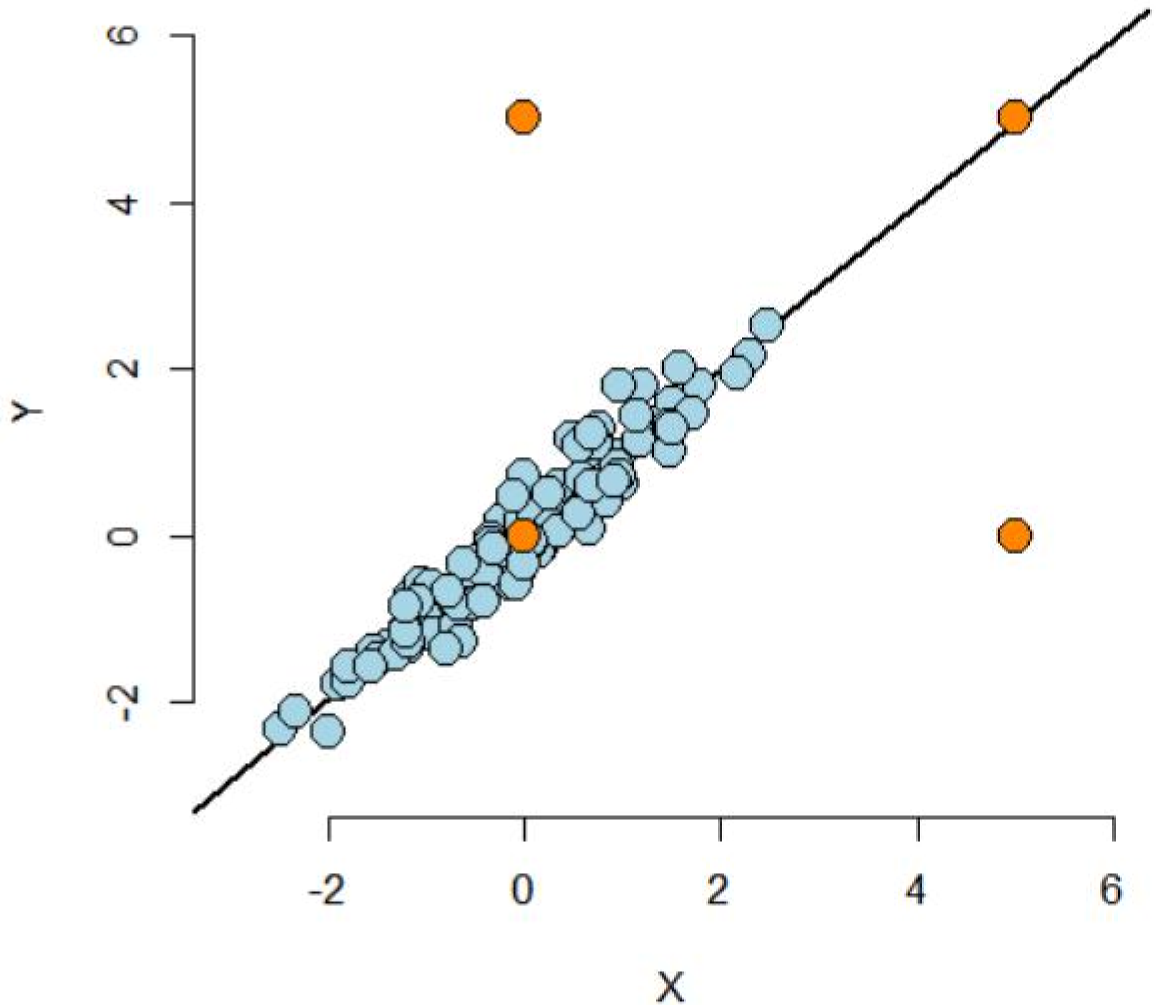
Análise de Resíduos



Normalidade dos resíduos



Leverage e influência



Diagnóstico no R

Use o comando ?influence.measures

- Resíduo padronizado

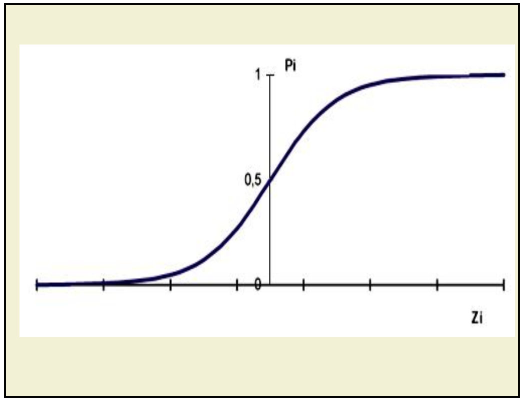
rstandar e rstudent

- Leverage

dffit, dfbeta, cook.distance

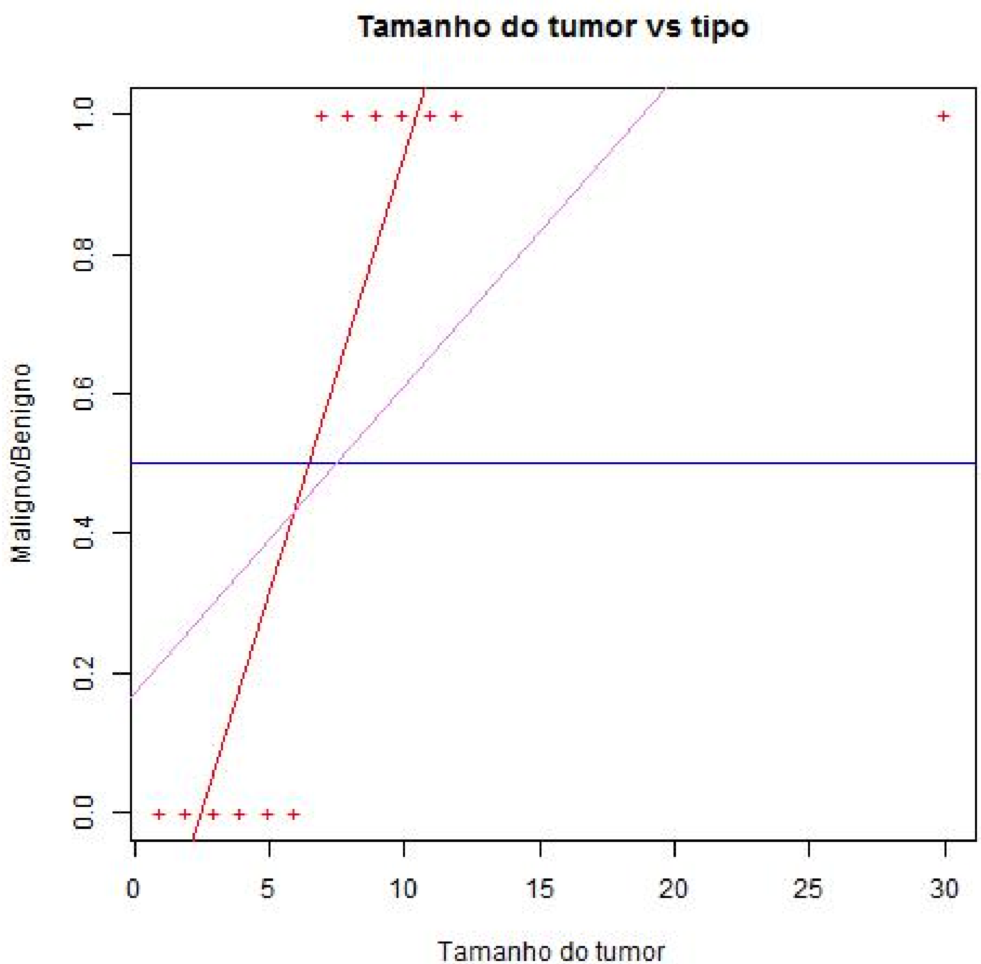
- PRESS = resid(fit)/ (1 - hatvalues(fit))

Regressão Logística



$$P(\theta) = \frac{1}{1 + e^{-\theta}}$$

Regressão Linear



Se p(x) ≥ 0,5 então y = 1

Se p(x) < 0,5 então y = 0

Grandes tumores mudam o corte.

P(X) pode ser > 1 e < 0

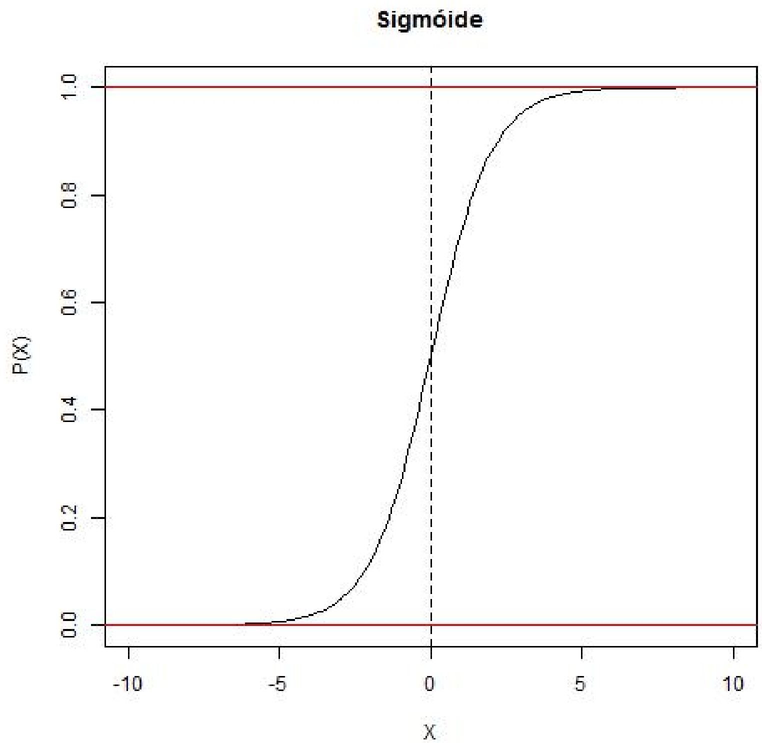
Regressão Logística:

0 ≤ P(X) ≤ 1

Regressão Logística

Queremos: 0 ≤ P(X) ≤ 1

Utilizaremos a função sigmoide:



$$P(\theta) = \frac{1}{1 + e^{-\theta}}$$

$$\theta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Interpretação de P(θ)

P(θ) é a **probabilidade estimada** que Y = 1 quando θ é a **entrada**.

P(Y=1| θ) = 1 – P(Y=0| θ).

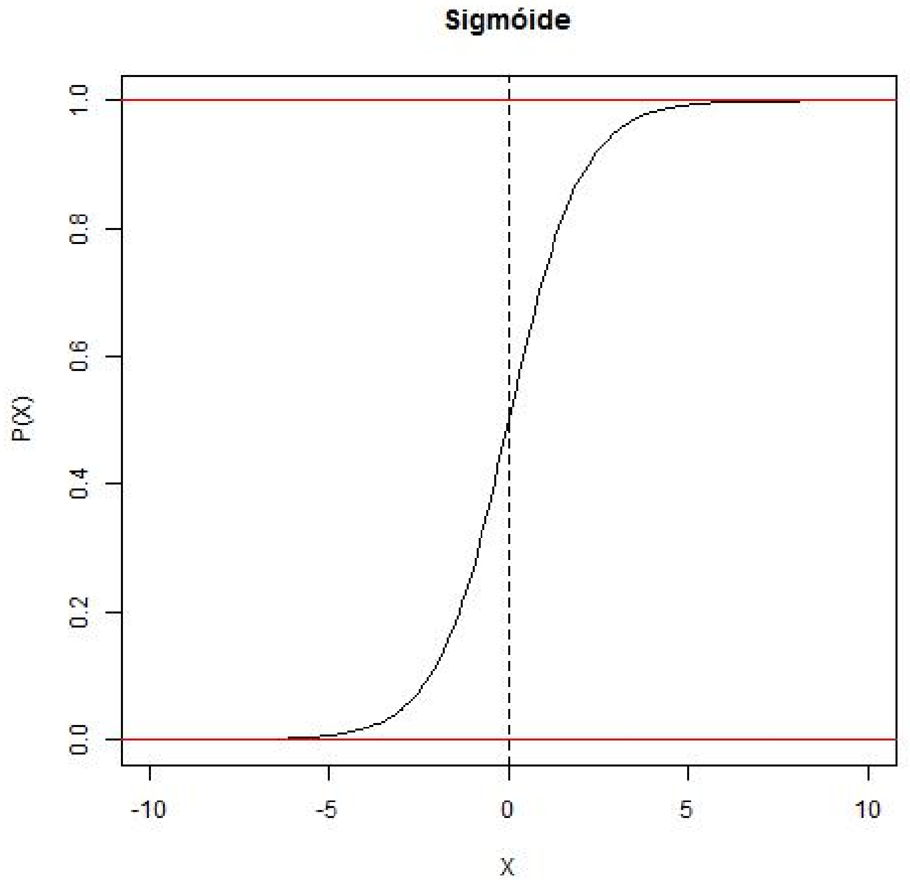
Probabilidade de Y = 1 dado θ.

Para classificação, define-se um corte, por exemplo 0,5.

P(θ) ≥ 0,5 => Y=1 OU **P(θ) < 0,5 => Y=0**.

Fronteira de Decisão

Seja $\theta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ então P(θ) ≥ 0,5 quando θ > 0.

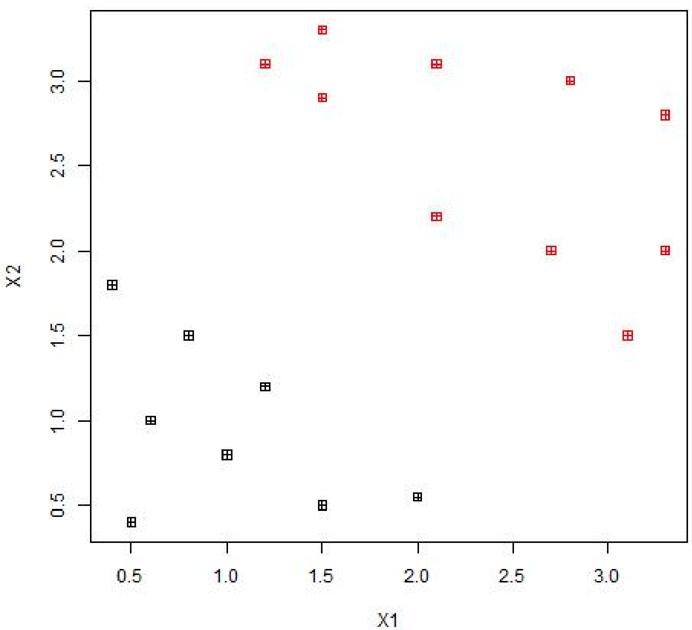


Fronteira de Decisão ...

Seja $\theta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ então P(θ) ≥ 0,5 quando θ > 0

$$\vec{\beta} = \begin{matrix} 3 \\ 1 \\ 1 \end{matrix}$$

então



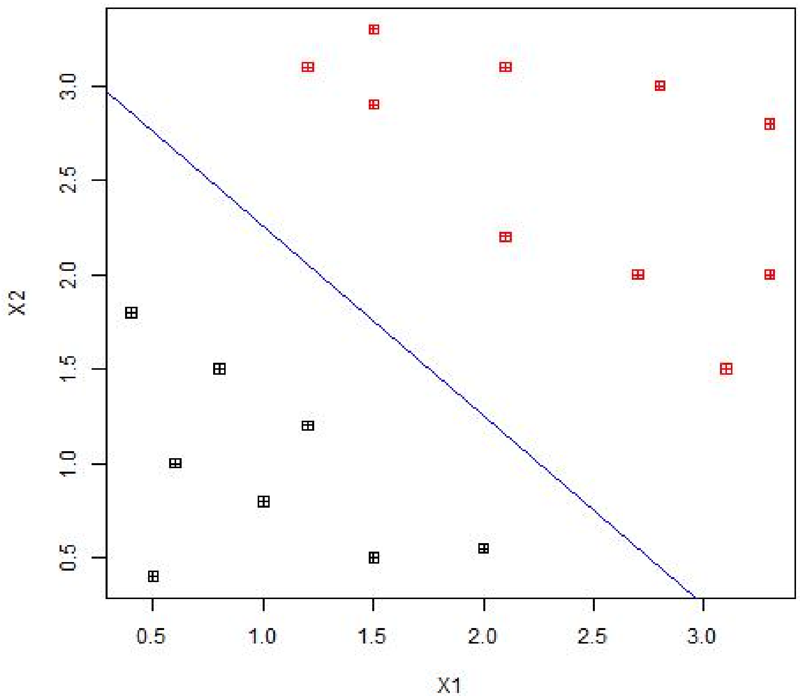
$$P(\theta) = g(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$$

$$P(\theta) = g(-3 + X_1 + X_2)$$

Prediz Y = 1 quando: $3 + X_1 + X_2 > 0$

Fronteira de Decisão ...

Seja $\theta = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ então P(θ) ≥ 0,5 quando θ > 0



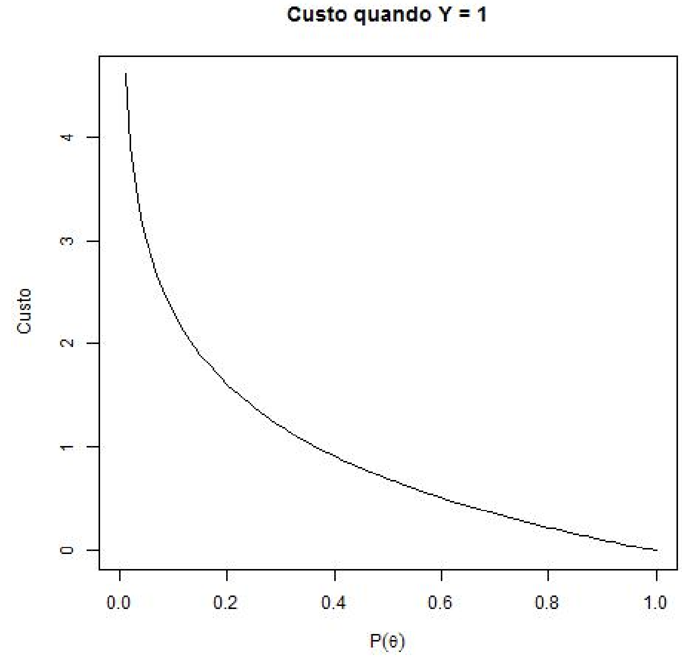
$$3 + X_1 + X_2 \quad 0$$

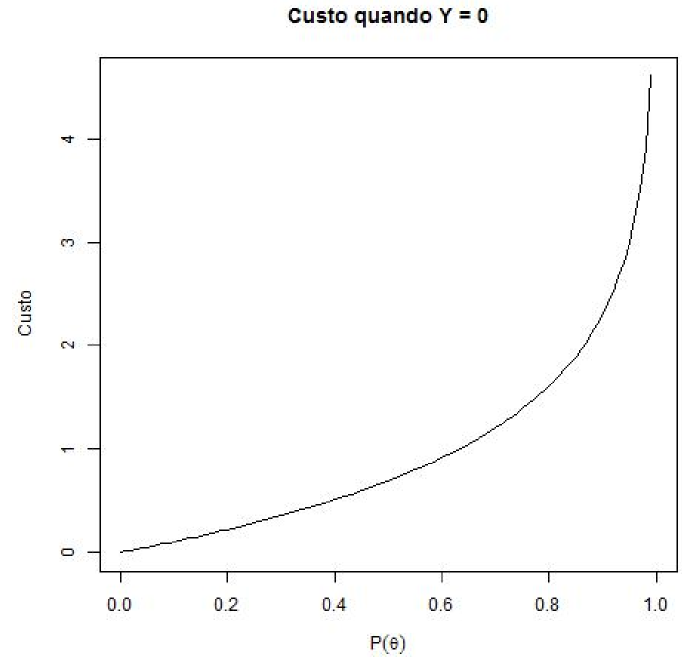
$$X_1 + X_2 \quad 3$$

Mas como escolher os β's ?

Função Custo

$$Custo\ (P(\theta), y) = \begin{cases} -\log(P(\theta)), & se\ y = 1 \\ -\log(1 - P(\theta)), & se\ y = 0 \end{cases}$$





Função Custo ...

$$Custo\ (P(\theta), y) = \begin{cases} -\log\big(P(\theta)\big), & se\ y = 1 \\ -\log\big(1\ -\ P(\theta)\big), & se\ y = 0 \end{cases}$$

$$J(\vec{\beta}) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log(P(\theta)) + \left(1 - y^{(i)}\right)\log(1\ - P(\theta))\right]$$

onde ***m*** é o número de instâncias no dataset.

Minimização

Quero $min_\beta \, J(\vec{\beta})$

Repita {

$$\beta_j = \beta_j \quad \alpha \frac{\quad}{\beta_j} J(\vec{\beta})$$

simultaneamente para todos os β's.

}

Sumário

Dado $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$:

Minimizar a função custo;

Obter os β's ;

Por fim, dado um ponto X = 1 por exemplo, calcular P(θ|X=1):

Se β = [1 2] então:

$$P(\beta_0 + \beta_1 X_1) = \frac{1}{1 + e^{\beta_0 + \beta_1 X_1}} = \frac{1}{1 + e^{\beta_0 + \beta_1}}$$

$$P(\beta_0 + \beta_1) = \frac{1}{1 + e^{1+2}} \quad 0{,}05$$

Exemplo de Regressão Logística

Classes:
compra_computador = 'sim'
compra_computador = 'nao'

Amostra:
X = (Idade <= 30,
Renda = media,
Aluno = 'sim'
Credito = 'normal') ?

P(X) ~ 1

Idade	Renda	Aluno	Credito	Classe
<=30	alta	nao	normal	nao
<=30	alta	nao	excelente	nao
31…40	alta	nao	normal	sim
>40	media	nao	normal	sim
>40	baixa	sim	normal	sim
>40	baixa	sim	excelente	nao
31…40	baixa	sim	excelente	sim
<=30	media	nao	normal	nao
<=30	baixa	sim	normal	sim
>40	media	sim	normal	sim
<=30	media	sim	excelente	sim
31…40	media	nao	excelente	sim
31…40	alta	sim	normal	sim
>40	media	nao	excelente	nao

Regressão Logística

Sensível à **colinearidade**.

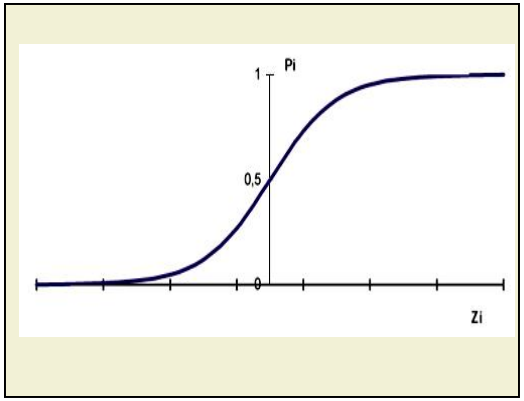
Fornece probabilidades como saída; pode-se variar o ***cut off***.

Não é robusto com relação aos **atributos irrelevantes**.

Permite testar estatisticamente a importância das variáveis.

Custo computacional **baixo**.

Regressão Penalizada



$$P(\theta)=\frac{1}{1+e^{-\theta}}$$

Função de minimização

Ridge

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

LASSO

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

Função de minimização

Ridge

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

LASSO

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

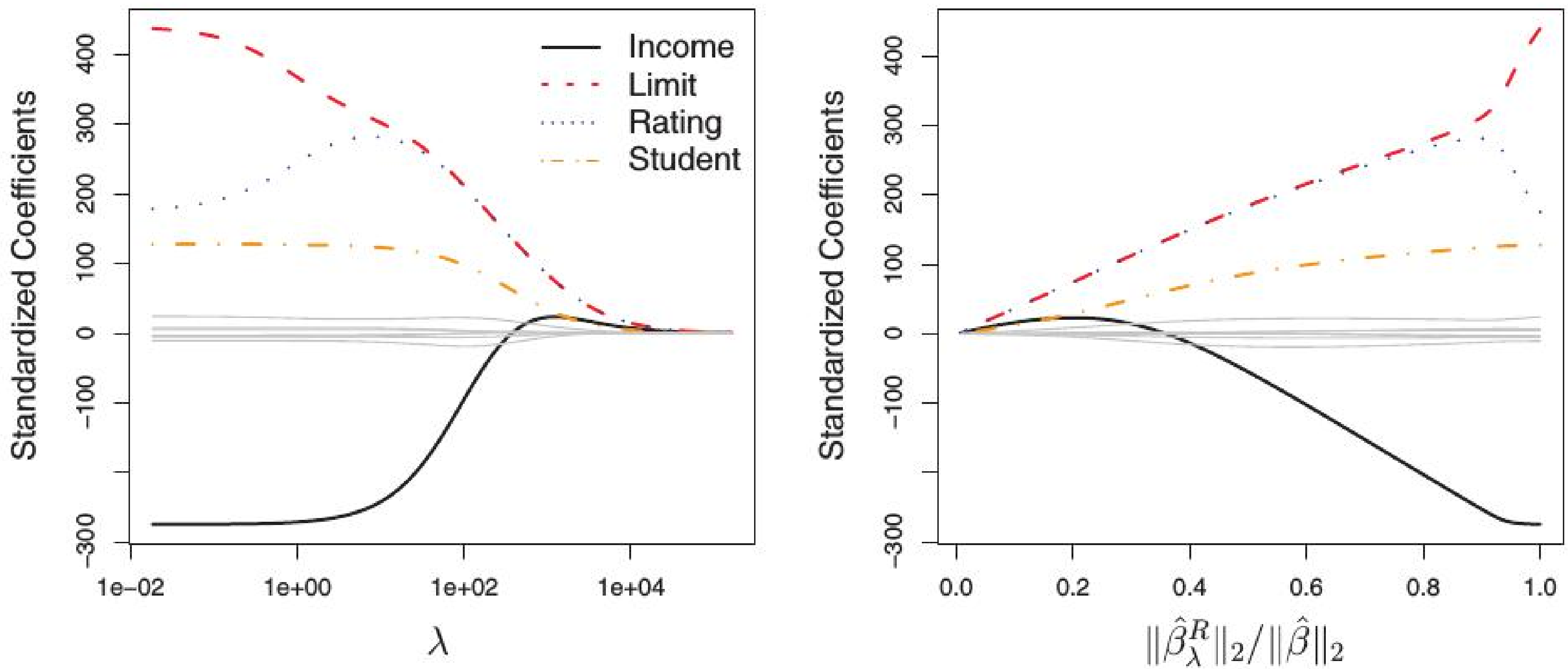
Função de minimização

O termo $\lambda \sum_{j=1}^{p} \beta_j^2$ tem a função de forçar os valores de β para zero na minimização. O λ é o fator de **penalização**.

Se λ = 0 é como uma regressão normal

Se λ = infinito os β tendem a zero.

Efeitos



Income
Limit
Rating
Student
Standardized Coefficients
−300
−100
0
100
200
300
400
1e−02
1e+00
1e+02
1e+04
λ
Standardized Coefficients
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$

Pontos importantes

- Os betas dependem de **lâmbida** mas também da **escala**.
- É importante normalizar os atributos!
- Mãos a obra!