2.2 用线性回归预测房价

线性模型 & 损失函数 & 目标函数

主讲人：李辉楚吴

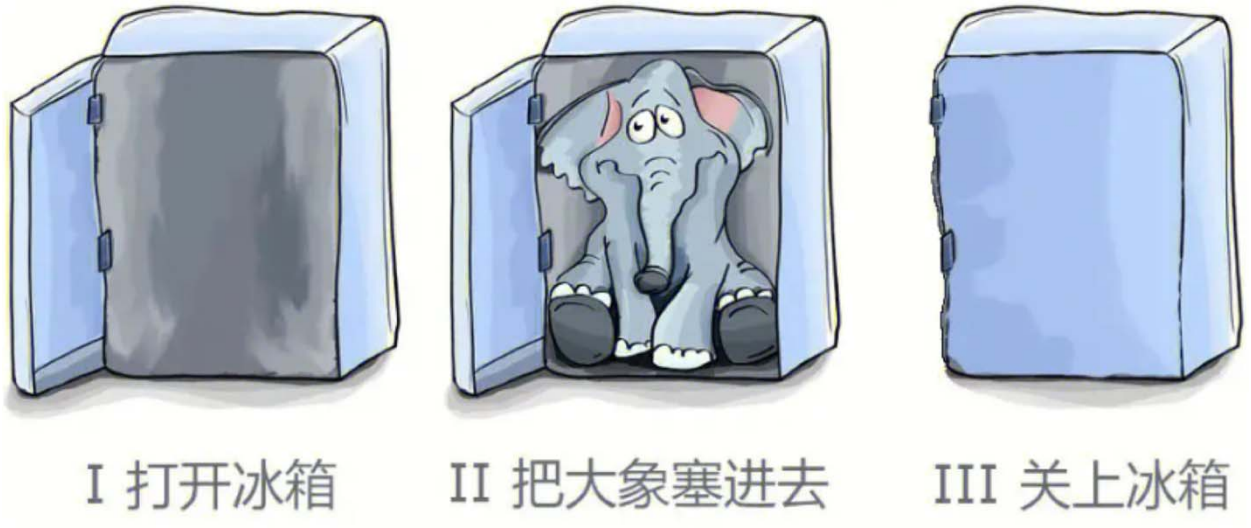
前情提要——构建预测模型的三个步骤

步骤1：一大堆数据

步骤3：选择最优模型

步骤2：一堆模型 $f_1, f_2, \cdots, f_n$

《钟点工》2000年春节联欢晚会



思考：仅考虑单个特征，应如何设计预测模型？如何选择最优模型？

如何构建最优的预测模型

步骤1

一大堆数据

继续使用波士顿的房价数据

1978年波士顿区域房屋价格表

x_1：犯罪率（%）	0.00632	0.02731	0.02729	0.03237	0.06905
x_2：大住宅用地占比（%）	18.00	0.00	0.00	0.00	0.00
x_3：非零售商业用地占比（%）	2.31	7.07	7.07	2.18	2.18
x_4：景观房（0/1）	0	0	0	0	0
x_5：氮氧化物浓度（ppm）	0.538	0.469	0.469	0.458	0.458
x_6：平均房间数（个）	6.575	6.421	7.185	6.998	7.147
x_7：老旧房屋占比（%）	65.2	78.9	61.1	45.8	54.2
x_8：离就业中心的加权距离	4.09	4.9671	4.9671	6.0622	6.0622
……	……	……	……	……	……
y：房价中位数（千美元）	**24.00**	**21.6**	**34.7**	**33.4**	**36.2**

步骤1

一大堆数据

继续使用波士顿的房价数据

1978年波士顿区域房屋价格表

x_1：犯罪率（%）					0.06905
x_2：大住宅用					0.00
x_3：非零售商					2.18
x_4：景观房					0
x_5：氮氧化物浓度（ppm）	0.538	0.469	0.469	0.458	0.458
x_6：平均房间数（个）	6.575	6.421	7.185	6.998	7.147
x_7：老旧房屋占比（%）	65.2	78.9	61.1	45.8	54.2
x_8：离就业中心的加权距离	4.09	4.9671	4.9671	6.0622	6.0622
……	……	……	……	……	……
y：房价中位数（千美元）	**24.00**	**21.6**	**34.7**	**33.4**	**36.2**

便于理解：仅考虑单个特征

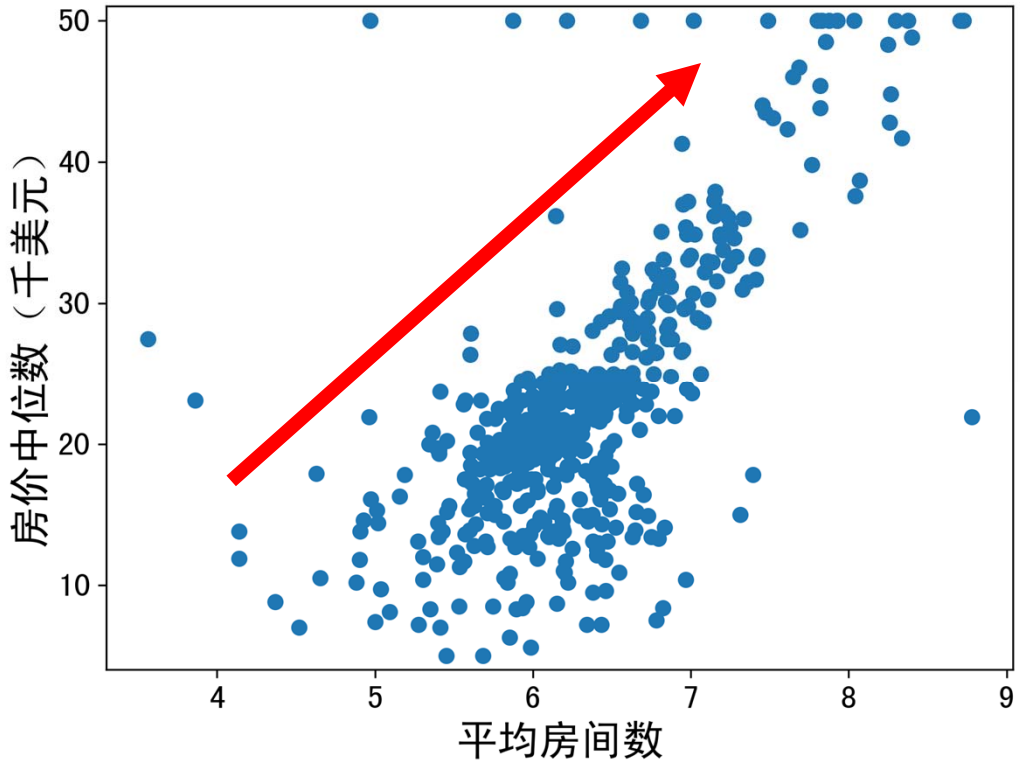
步骤1

一大堆数据

输入数据 x（特征，Feature）	输出结果 y（目标，Target | 标签，Label）
6.575	24.0
6.421	21.6
7.185	34.7
6.998	33.4
7.147	36.2
……	……

更直观的表示

房价与平均房间数正相关

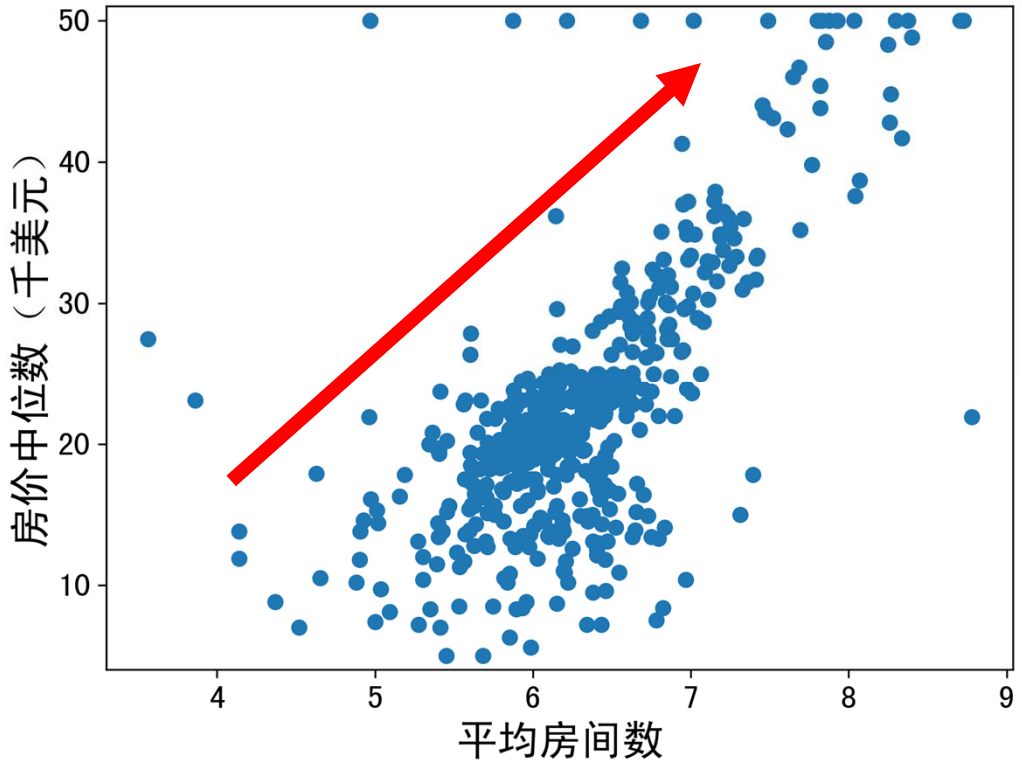


步骤2

一堆模型

$f_1, f_2, \cdots, f_n$

随意地构建正相关函数$y = f(x)$



步骤2

一堆模型
$f_1, f_2, \cdots, f_n$

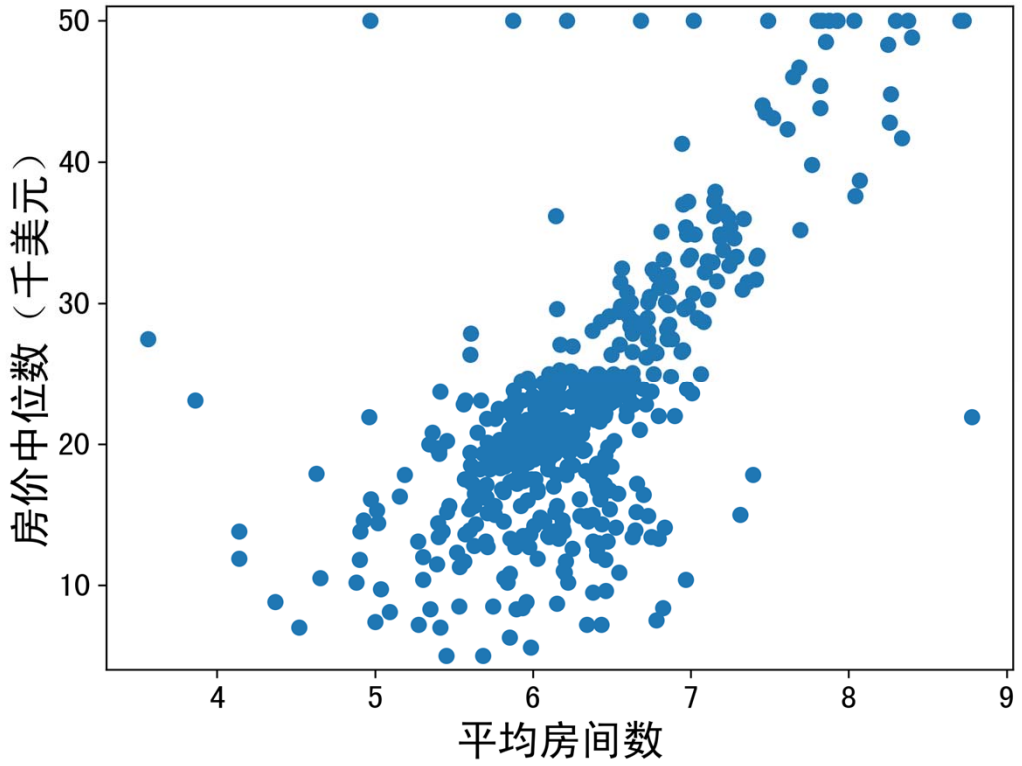
随意地构建函数$y = f(x)$

可以选择**线性函数** $y = wx + b$

w（weight）：权重 | 参数 | 系数

b（bias）：偏见 | 偏好

```
def pred_linear(x, w, b):
    return w * x + b
```



步骤2

一堆模型
$f_1, f_2, \cdots, f_n$

随意地构建函数$y = f(x)$

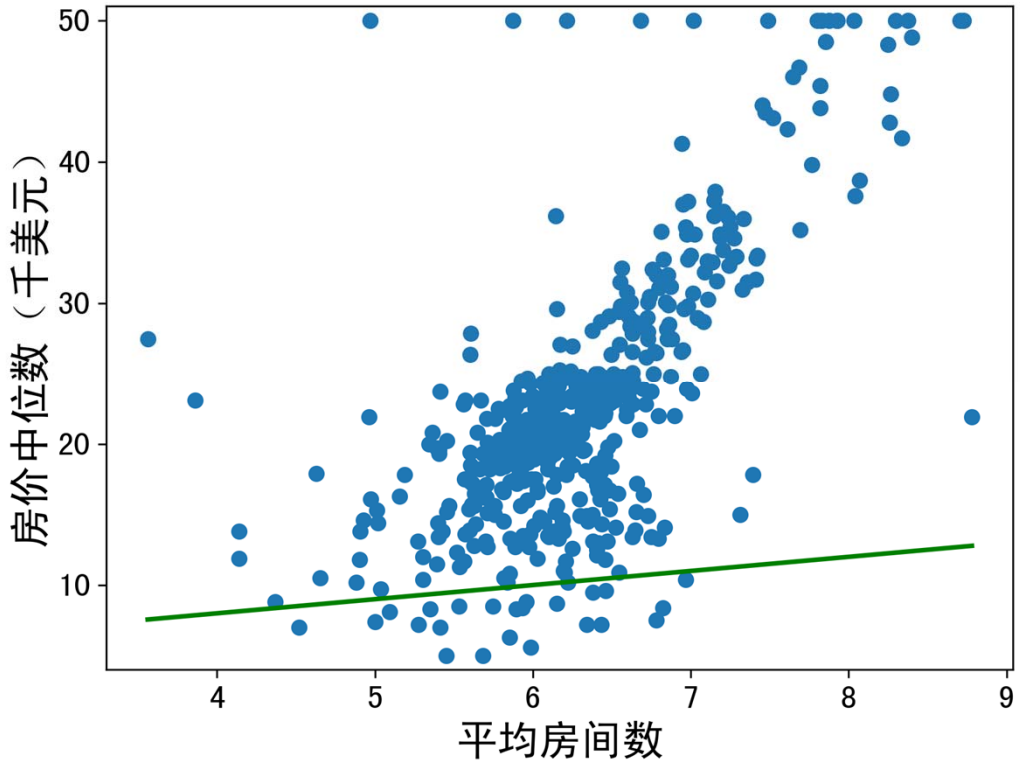
可以选择**线性函数** $y = wx + b$

w（weight）：权重 | 参数 | 系数

b（bias）：偏见 | 偏好

```
def pred_linear(x, w, b):
    return w * x + b
```

当 $(w, b) \leftarrow (1,4)$ 时……



步骤2

一堆模型
$f_1, f_2, \cdots, f_n$

随意地构建函数$y = f(x)$

可以选择**线性函数** $y = wx + b$

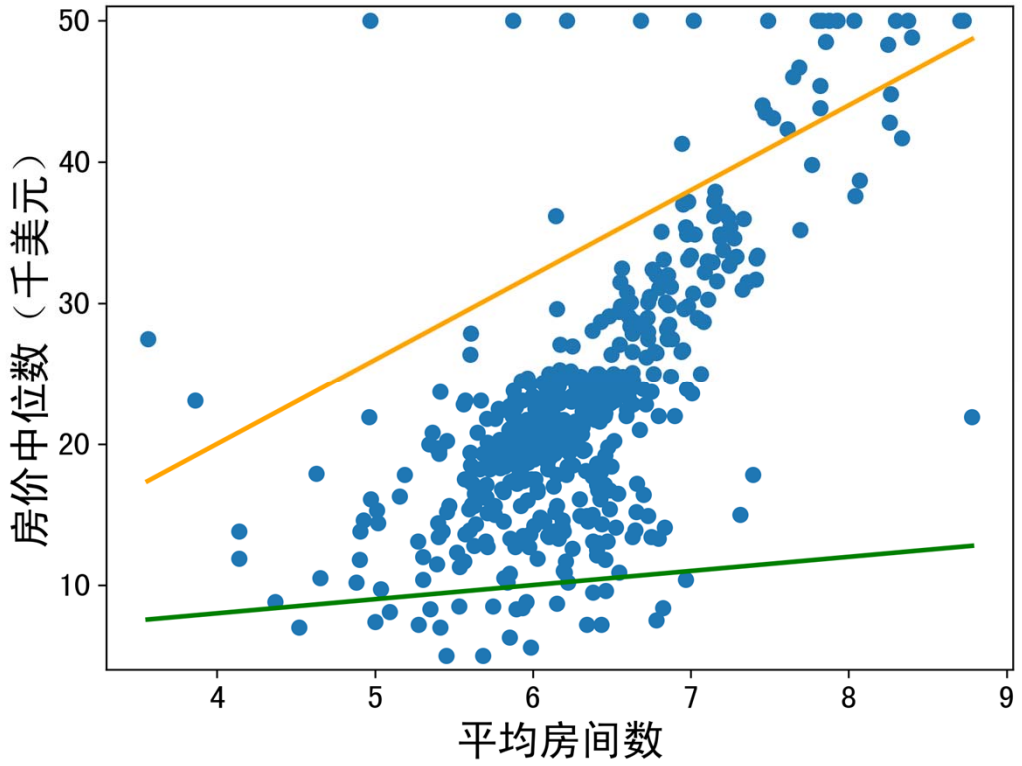
w（weight）：权重 | 参数 | 系数

b（bias）：偏见 | 偏好

```
def pred_linear(x, w, b):
    return w * x + b
```

当 $(w, b) \leftarrow (1,4)$ 时……

当 $(w, b) \leftarrow (6, -4)$ 时……



步骤2

一堆模型
$f_1, f_2, \cdots, f_n$

随意地构建函数$y = f(x)$

可以选择**线性函数** $y = wx + b$

w（weight）：权重 | 参数 | 系数

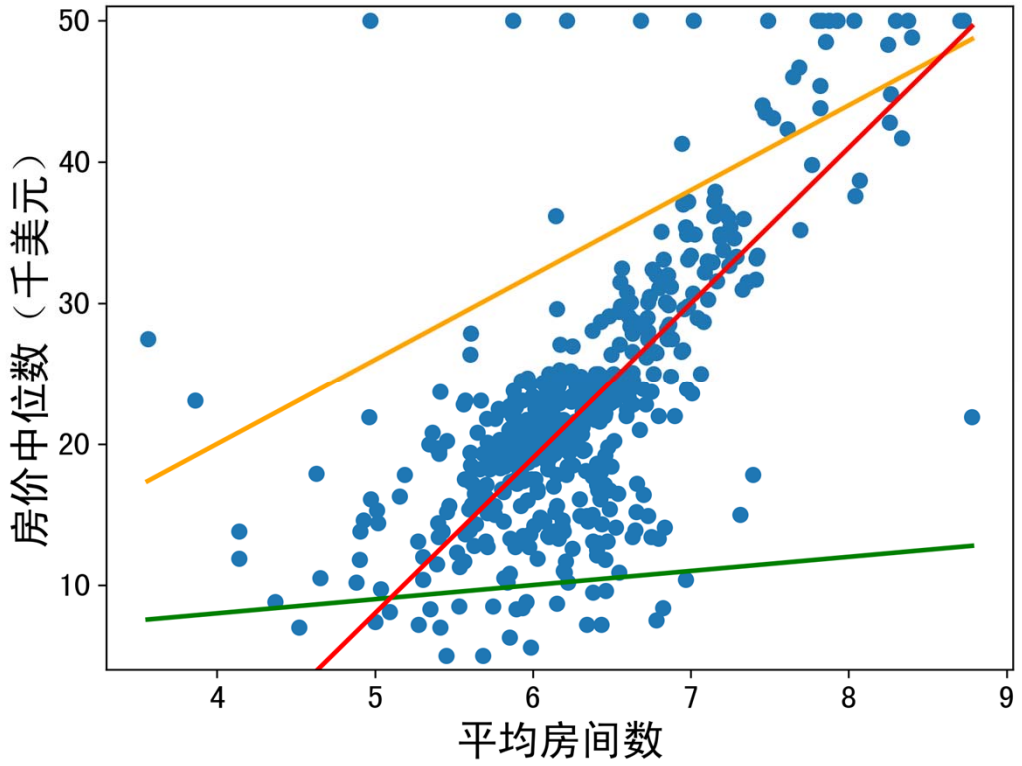
b（bias）：偏见 | 偏好

```
def pred_linear(x, w, b):
    return w * x + b
```

当 $(w, b) \leftarrow (1,4)$ 时……

当 $(w, b) \leftarrow (6, -4)$ 时……

当 $(w, b) \leftarrow (11, -47)$ 时……



步骤2

一堆模型
$f_1, f_2, \cdots, f_n$

随意地构建函数$y = f(x)$

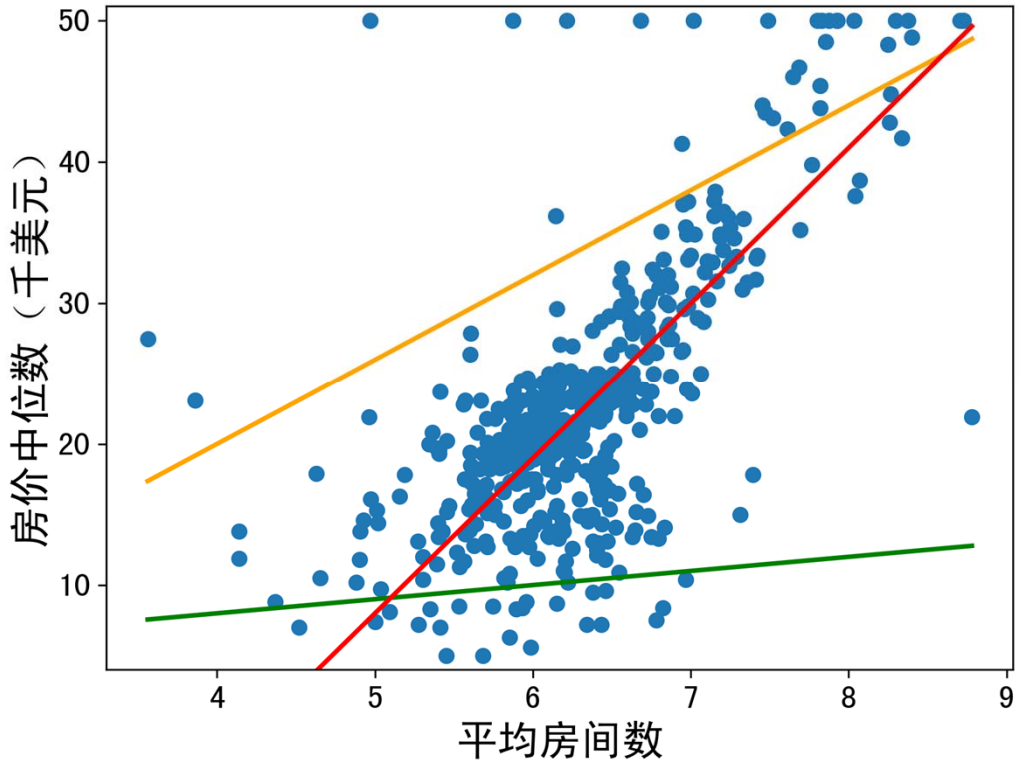
可以选择**线性函数** $y = wx + b$

练习：也可以尝试**非线性函数**

$$y = w\frac{1}{x} + b$$

$$y = we^x + b$$

$$y = wx^n + b$$



已经设计了一大堆不知道有没有用的函数

怎样选出最优的结果呢?

步骤3

选择最优模型

思考一下：如何判断预测模型的好坏？

和真实值越接近越准

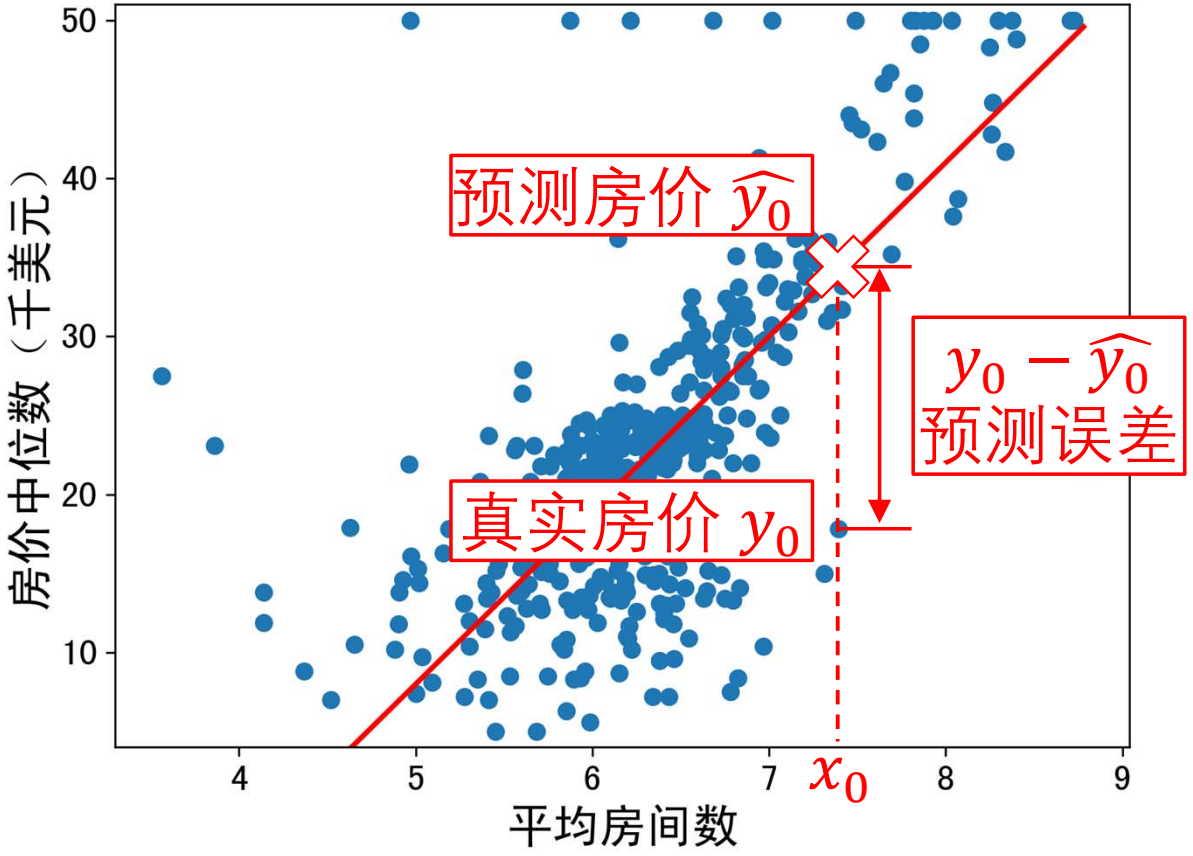
设 y 为真实值，$\hat{y}$ 为预测值

$x = x_0$时的预测误差可记为：

单个结果缺乏说服力

$$|y_0 - \widehat{y_0}|^2$$

“不管预测值高了或者低了都算作误差，所以应该消除正负号”



步骤3

选择最优模型

继续思考：如何对整体进行评价？

- **计算所有结果的和**

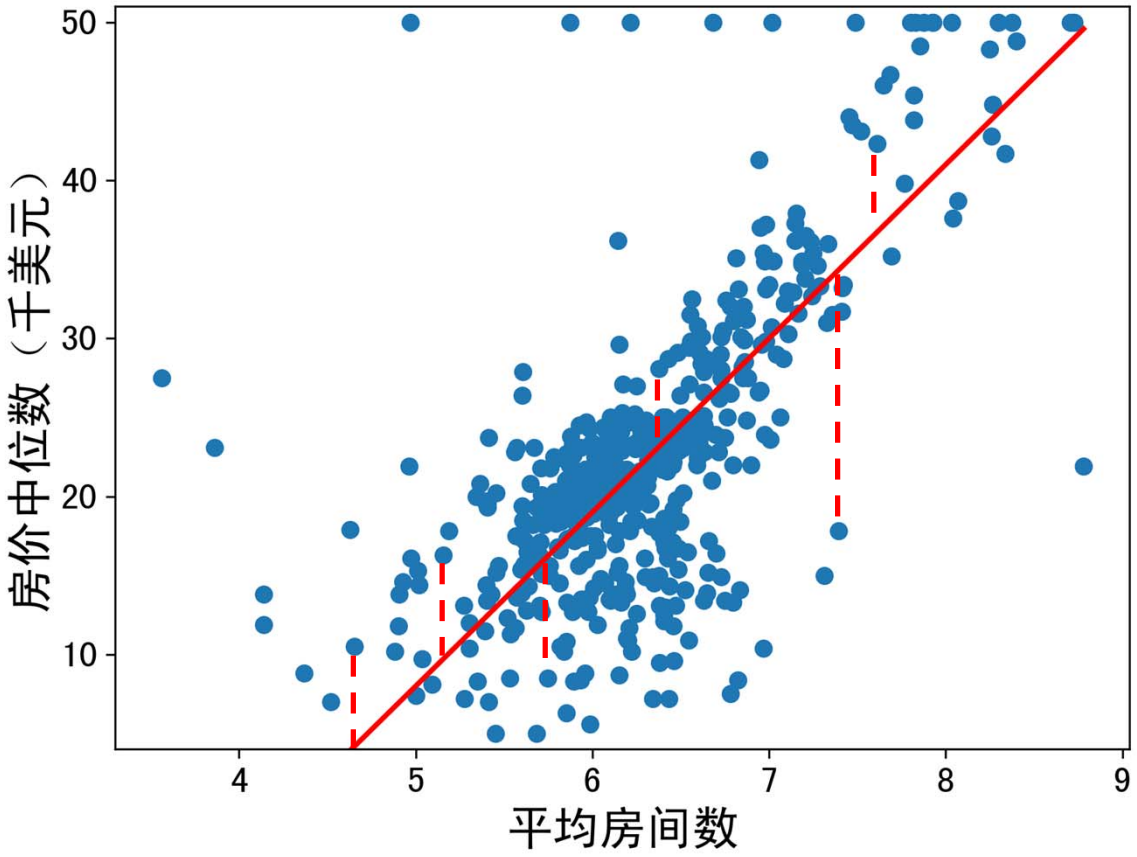
  误差会随样本数量增加而变大

- **取平均值**

  保证误差不随样本数量变化

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

$N \rightarrow$ 样本数量



步骤3

选择最优模型

继续思考：如何对整体进行评价？

- **计算所有结果的和**

  误差会随样本数量增加而变大

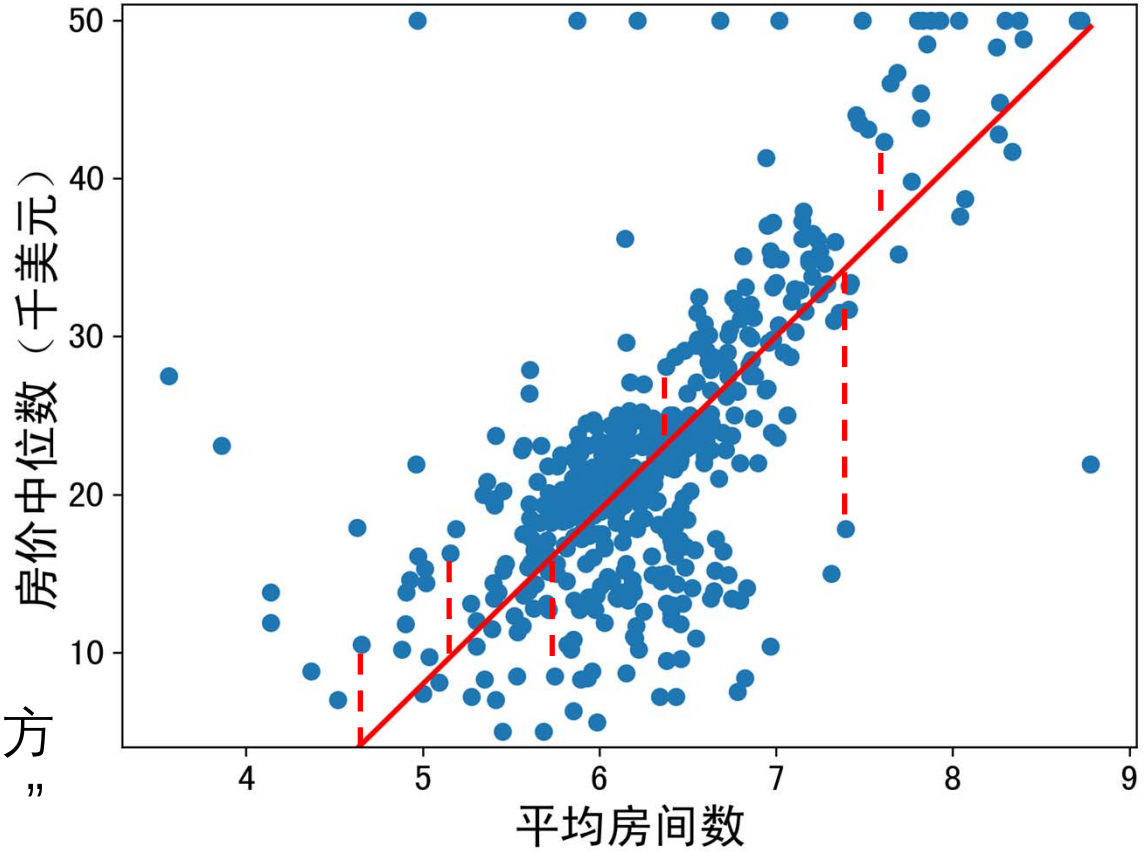
- **取平均值**

  保证误差不随样本数量变化

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

$N \rightarrow$ 样本数量

“通常我们会写成这样，因为方便计算。可以思考一下原因。”



步骤3

选择最优模型

继续思考：如何对整体进行评价？

- **计算所有结果的和**

  误差会随样本数量增加而变大

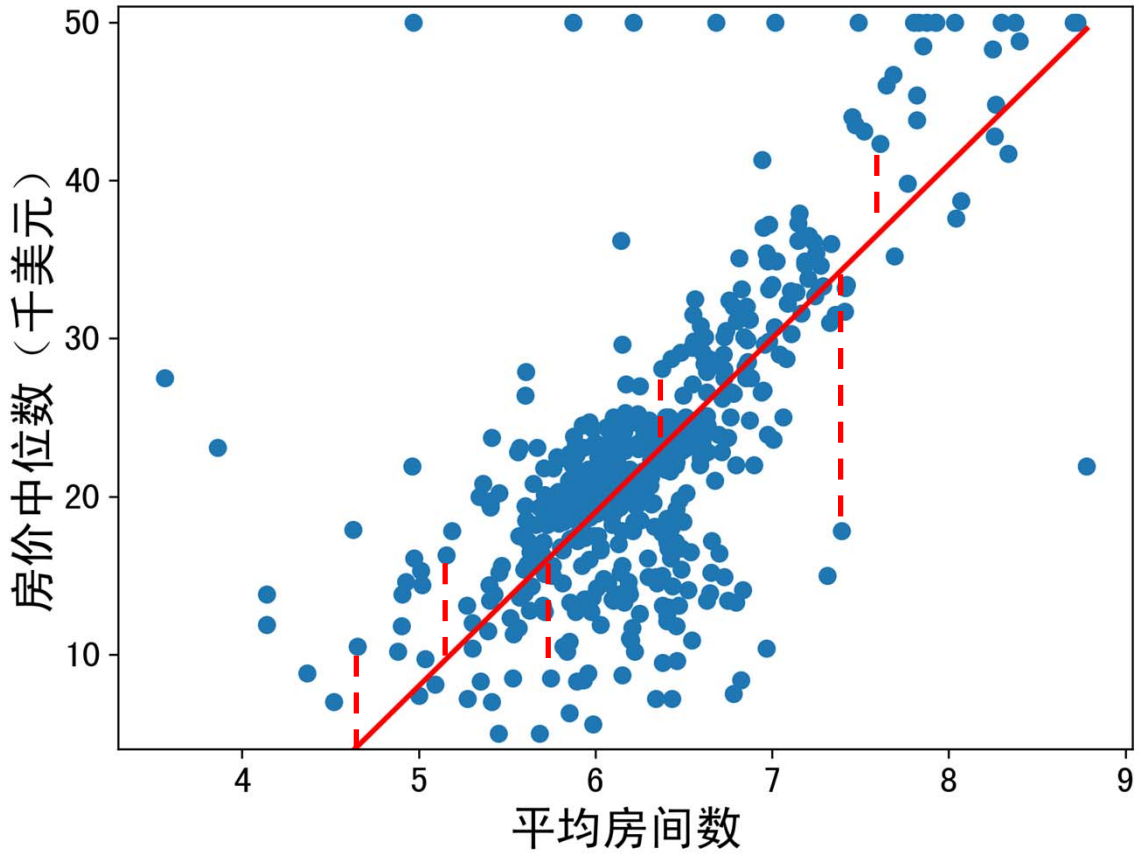
- **取平均值**

  保证误差不随样本数量变化

$$Loss = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

损失函数

衡量模型好坏



步骤3

选择最优模型

损失函数是描述预测函数好坏的函数

变换一下形式： $Loss = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ ← 代入 $\hat{y}_i$ $\hat{y}_i = wx_i + b$

$$Loss(w, b) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - (wx_i + b))^2$$

$Loss$的结果仅与预测函数的参数 w 和 b 相关

步骤3

选择最优模型

怎样找出最好的预测模型呢?

屠呦呦，中国首位诺贝尔医学奖获得者

- 查阅大量中医古籍找出2000+中草药制剂
- 筛选出640可能的治疟药方
- 从200种草药种得到380种提取物
- 经历190失败
- 对虐原虫的抑制率达到100%

向着确定的目标不断努力，
是成功的必要条件

人民网, 屠呦呦自述:190次失败之后的成功. 2015. http://politics.people.com.cn/n/2015/1005/c70731-27664603.html

步骤3

选择最优模型

怎样找出最好的预测模型呢？

- **给出w和b“所有”的值**

  $w \in [w_{min}, w_{max}]$

  $b \in [b_{min}, b_{max}]$

- **尝试所有的情况**

```
for w in w_list:
    for b in b_list:
        loss(x, y, w, b)
```

- **选择最优的结果**

$$w^*, b^* = arg \min_{w,b} L(w, b)$$

- 查阅大量中医古籍找出2000+中草药制剂
- 筛选出640可能的治疟药方
- 从200种草药种得到380种提取物
- 经历190失败
- 对虐原虫的抑制率达到100%

目标函数（Objective Function）

指导算法选出最优的预测模型。

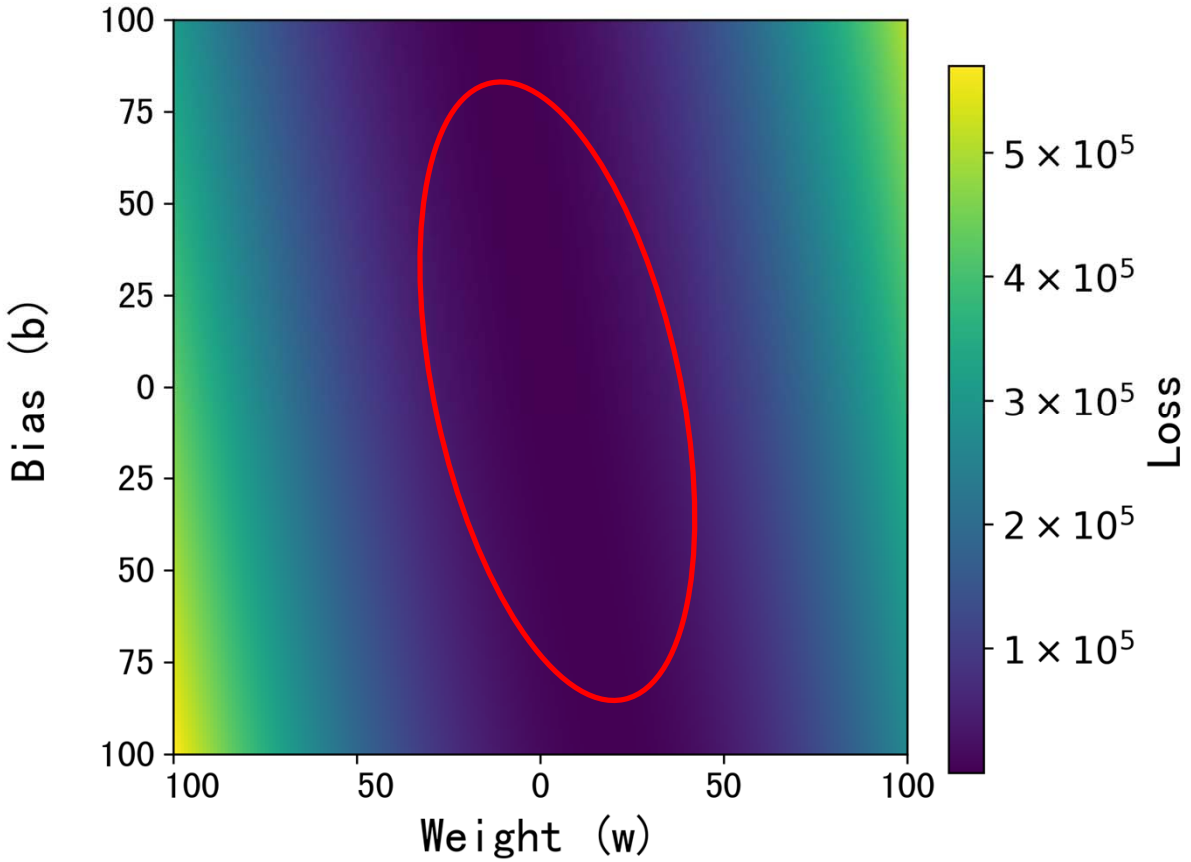
步骤3

选择最优模型

- 给出w和b“所有”的值

  $w \in [-100,100]$

  $b \in [-100,100]$



颜色越深，损失越小，模型预测得越准确

步骤3

选择最优模型

- 给出w和b“所有”的值

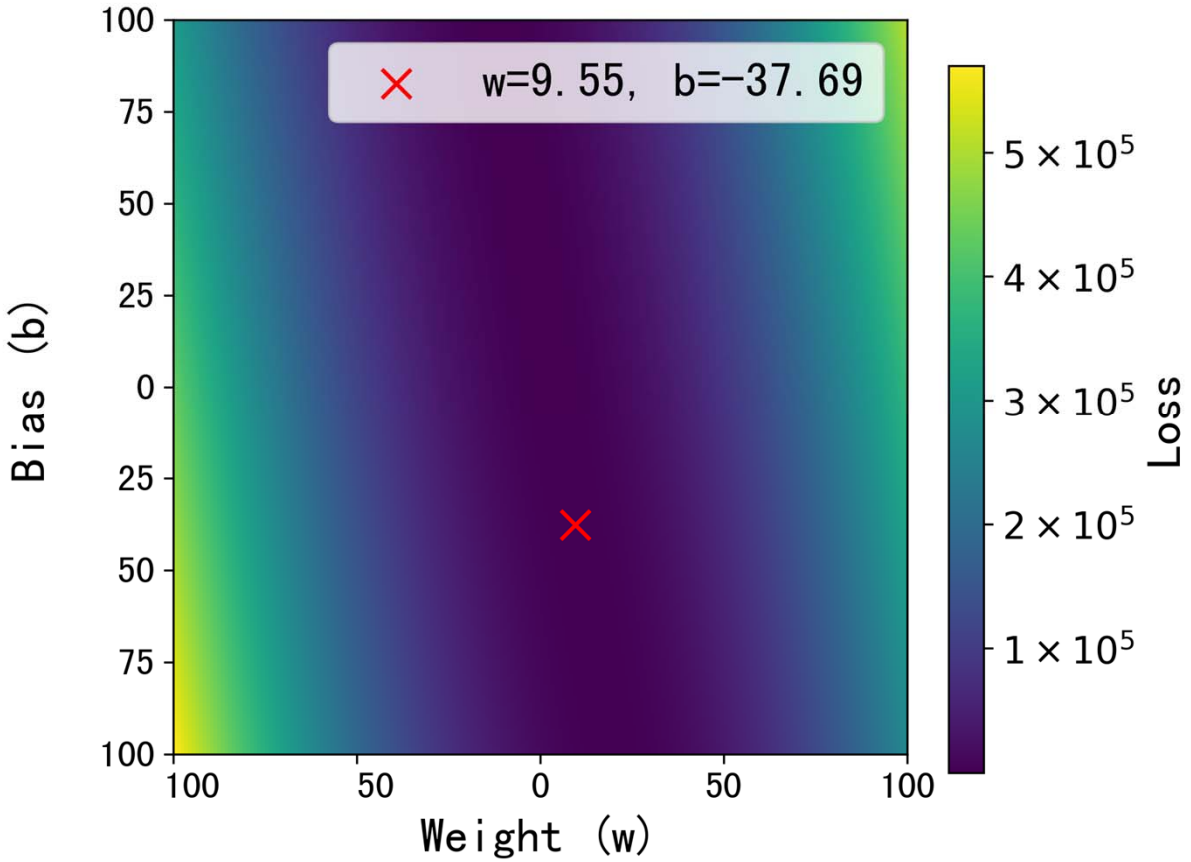
  $w \in [-100,100]$

  $b \in [-100,100]$

- 根据目标函数得到最优解

  $w = 9.55$

  $b = -37.69$



步骤3

选择最优模型

- 给出w和b“所有”的值

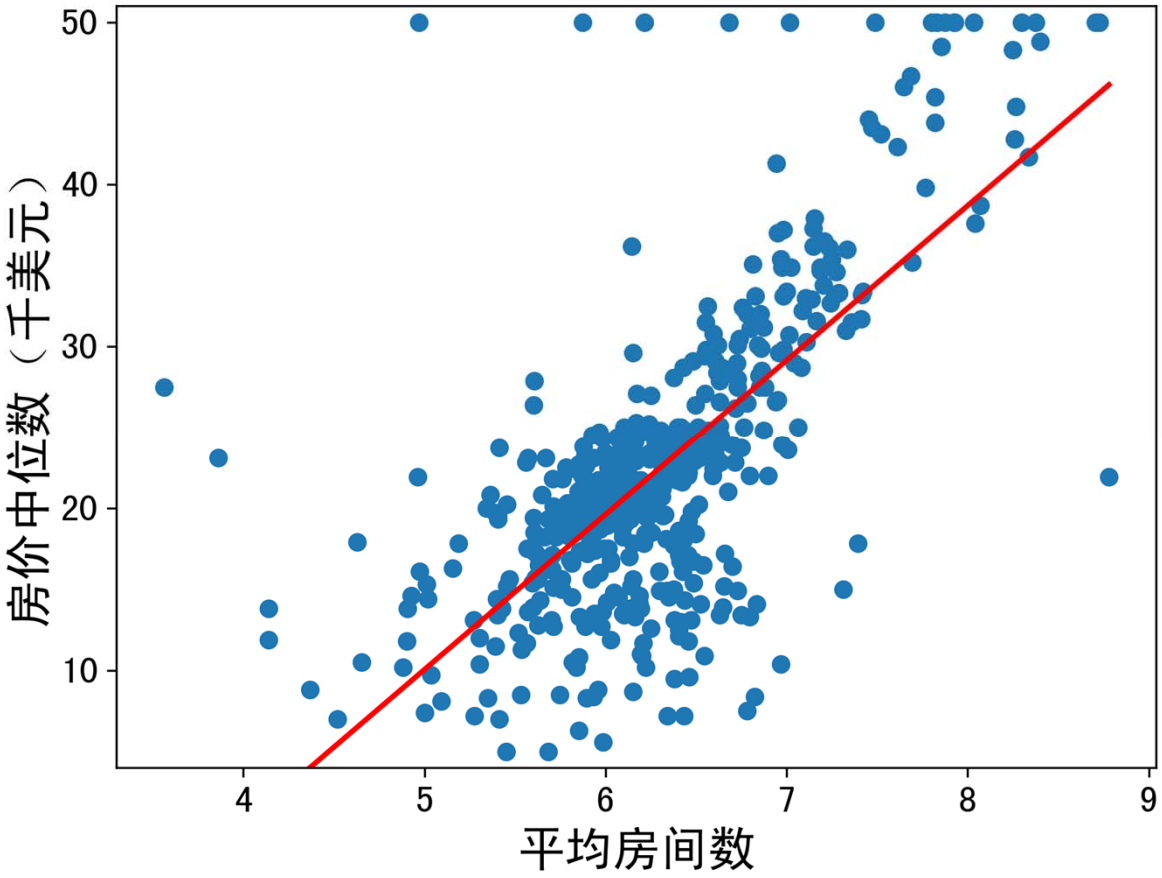
  $w \in [-100,100]$

  $b \in [-100,100]$

- 根据目标函数得到最优解

  $w = 9.55$

  $b = -37.69$



仅参考“平均房间数”时的最优预测函数

更加复杂的情况下是否还能如此呢?

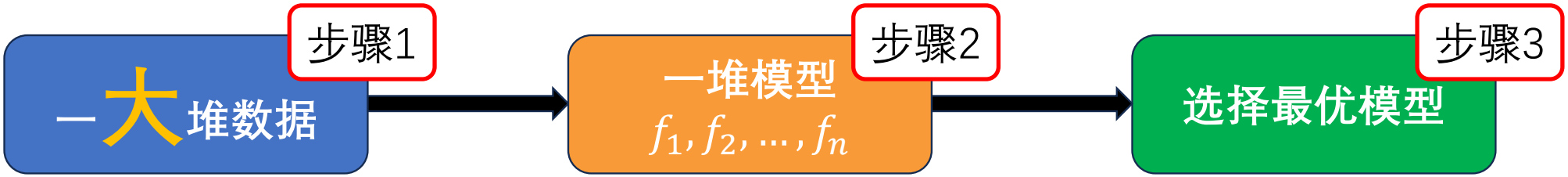
- 当房价相关特征的数量为10时
- 线性模型至少由10个参数决定
- 损失函数的计算次数N_{Loss}:

$$N_{Loss} = 100^{10} = 10^{20}\ (次)$$

- 计算一次Loss平均耗时 0.382（毫秒）
- 计算10^{20}次Loss耗时约 12亿年

“资源有限的情况下，还是不要蛮干。”

“T博士建议考虑一下Gradient Descend，也就是梯度下降。”



最简单的情况　　线性回归模型　　损失函数

目标函数

蛮力法面对复杂问题时的不足

下节内容

用**梯度下降**的方式找到最优解