

# 线性模型



# 目录

CONTENT

01

线性回归

02

预测能力

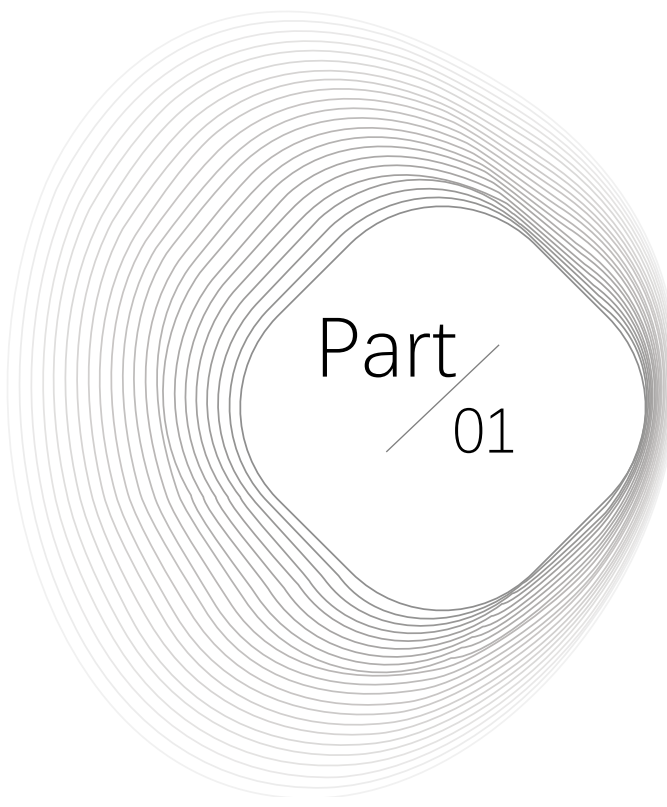
03

模型选择

04

逻辑回归





Part  
01

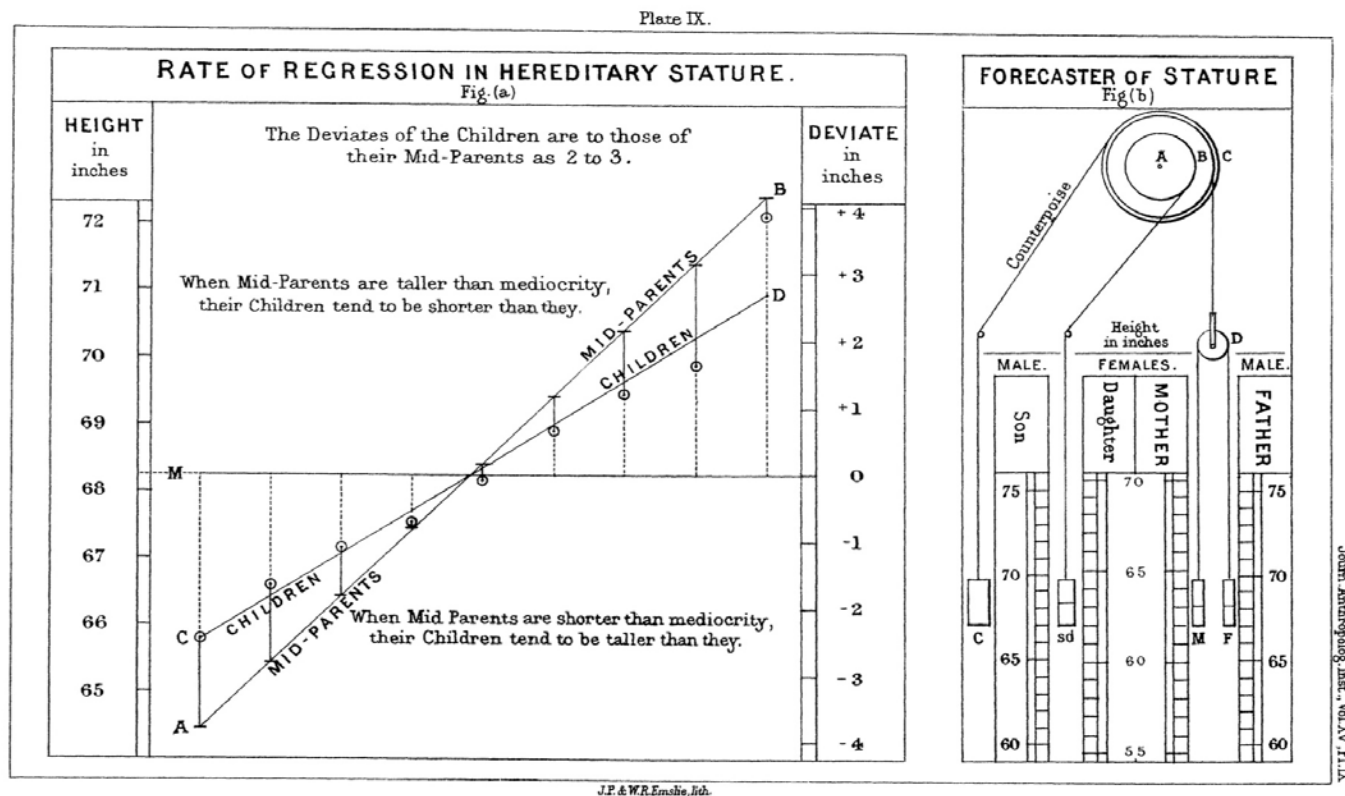
# 线性回归

- 为什么叫“回归”
- 从“数据生成角度来看”
- 如何回归
- 更高维度



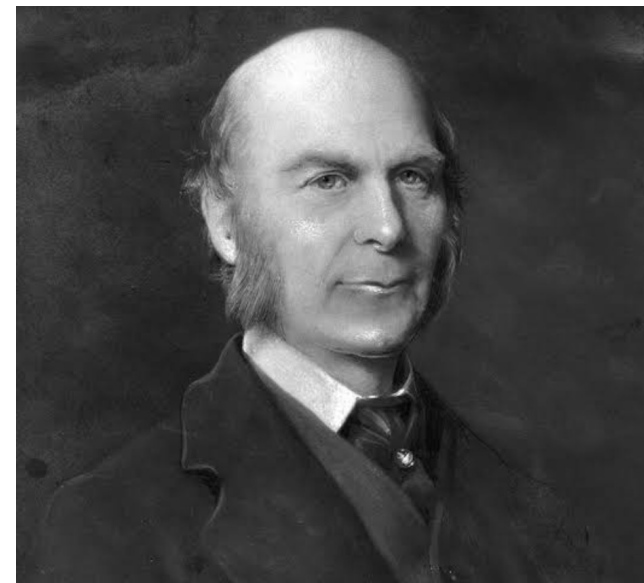
# 为什么叫“回归”

- Regression: a return to a former or less developed state.



Francis Galton 1886

$$y = 0.8567 + 0.516 * X, \quad X = \frac{1}{2}(\text{height}_{\text{father}} + 1.08 * \text{height}_{\text{mother}})$$



Sir Francis Galton  
(1822-1911)

高尔顿比10个生物学家中的9个更懂数学和物理  
比20个数学家中的19个更懂生物  
而比50个生物学家中的49个更懂疾病和畸形儿的知识。

—— Pearson

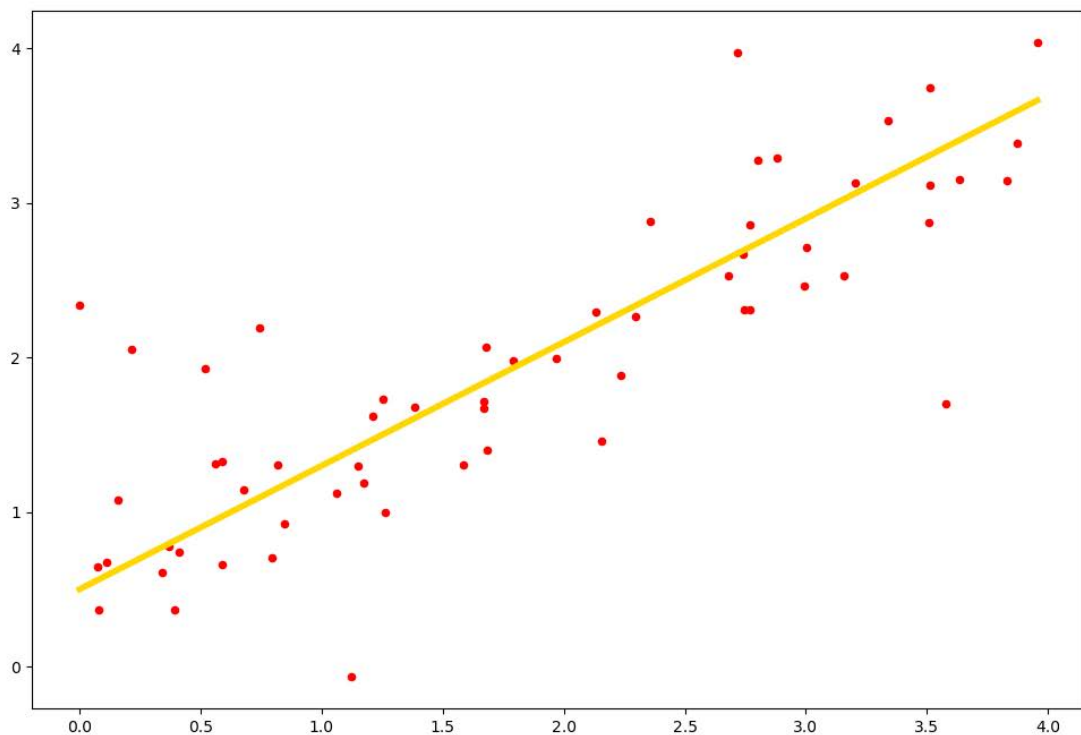


## 数据生成过程

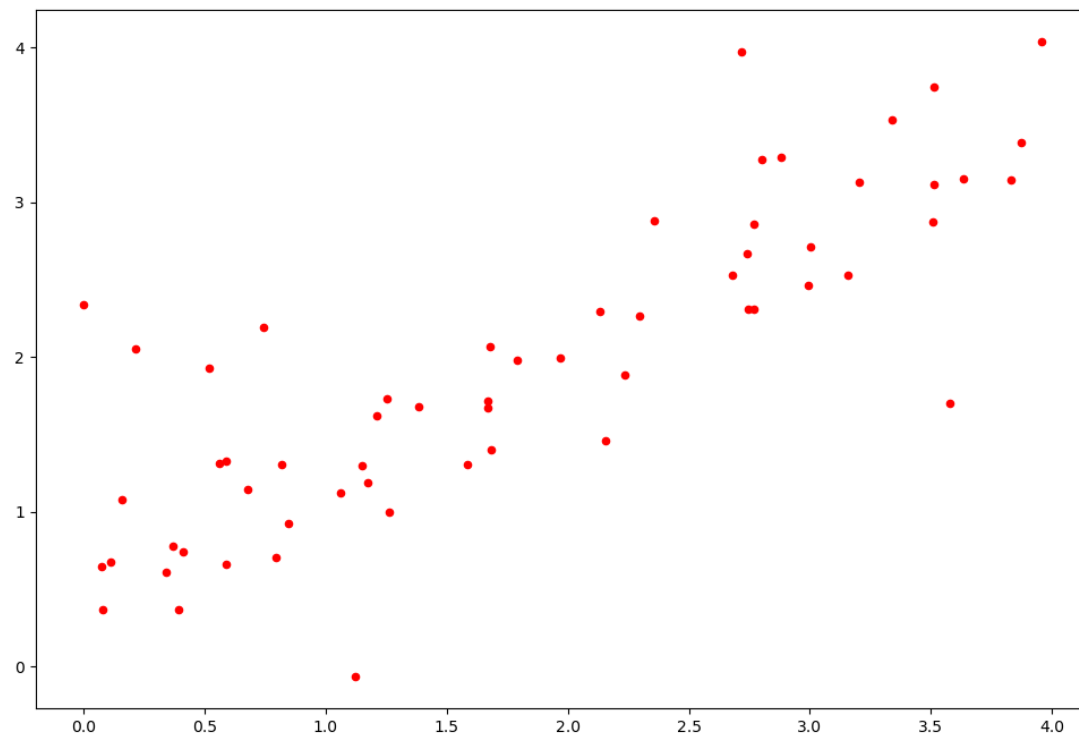
- 孩子的身高被父母的平均身高所决定
- 那么其他数据是不是可能也保持这种关系？

未知目标函数  
 $f: \mathcal{X} \rightarrow \mathcal{Y}$

部分数据=训练集  
 $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$

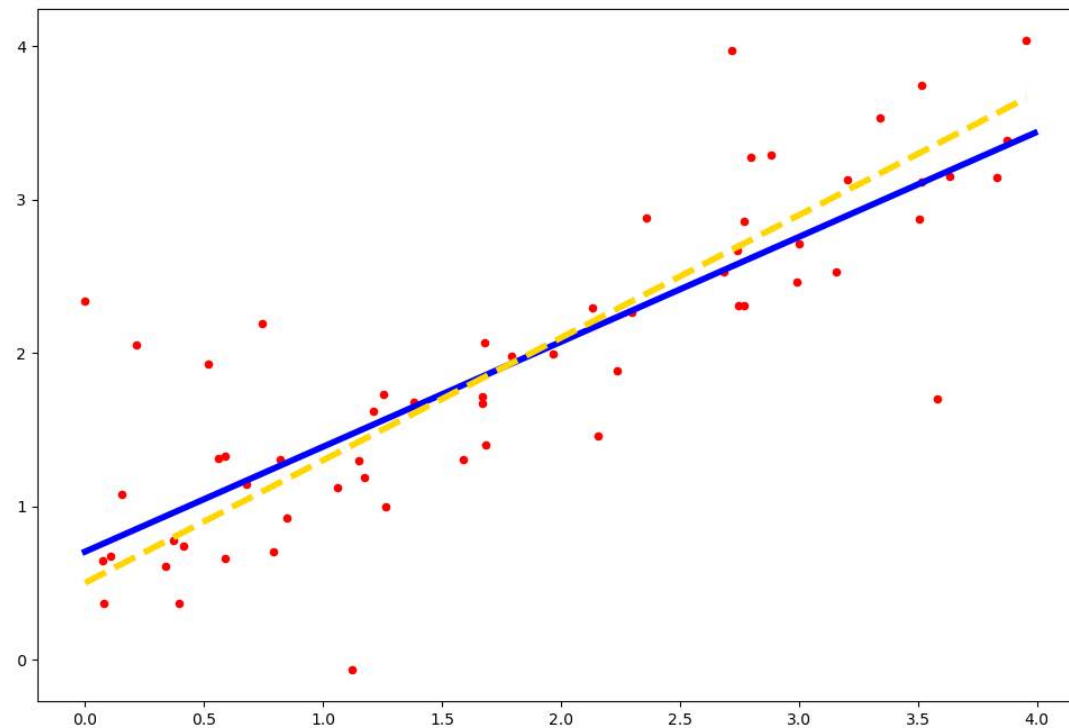
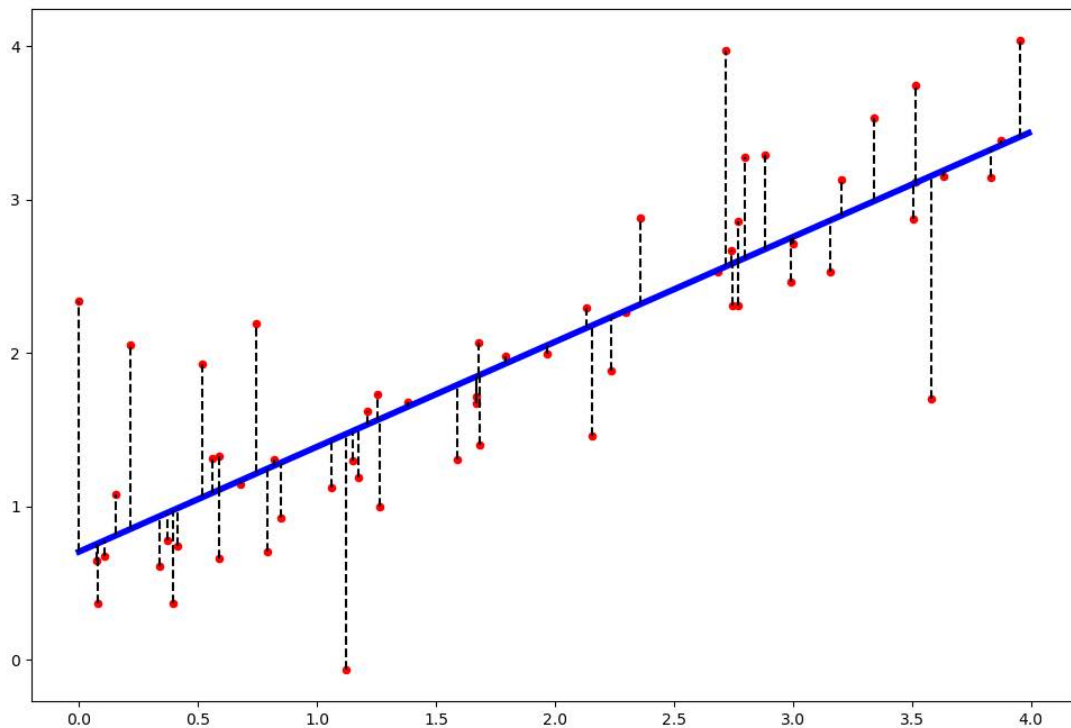


$$y_i = 0.8 * x_i + 0.5 + \varepsilon_i$$



## 1.3

## 如何回归：最小化错误

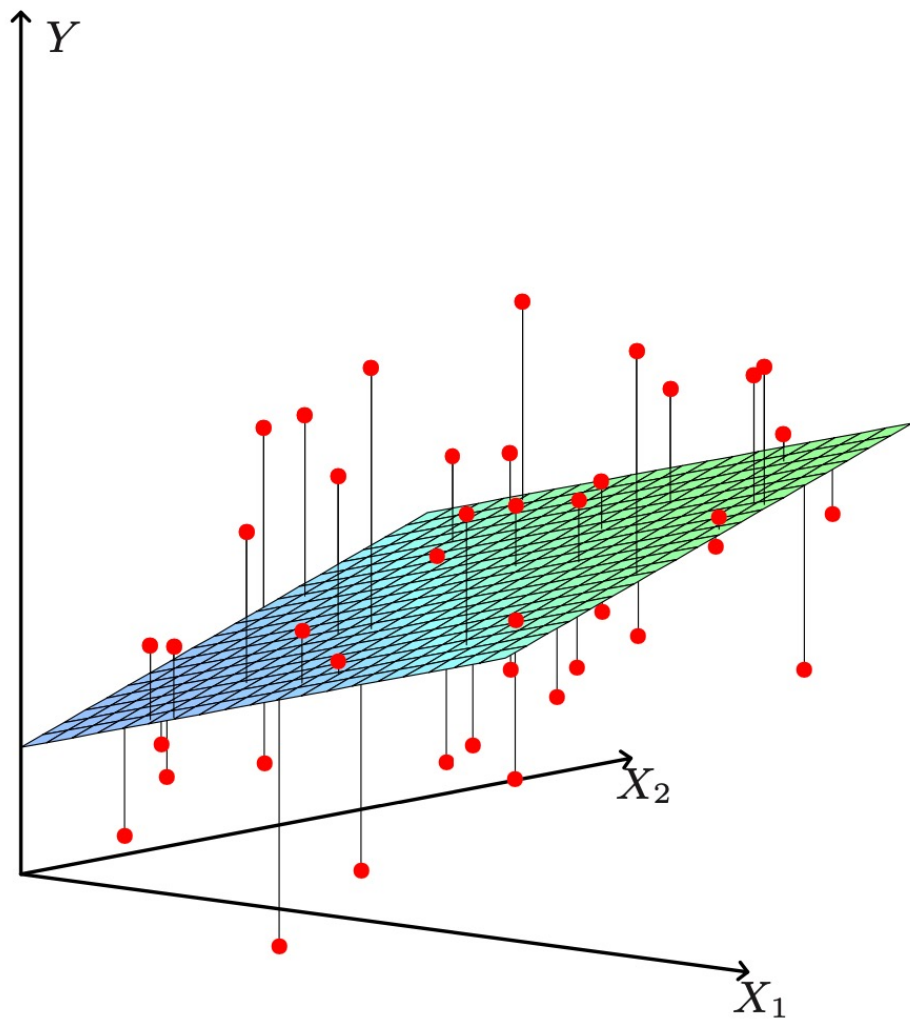


$y_i = \beta_1 * x_i + \beta_0 + \varepsilon_i$ ,  $\beta_1$ 与 $\beta_0$ 共同决定一条线，线可以很多，那么选哪条呢？

$\min(\sum_{i=1}^N (y_i - (\widehat{\beta}_1 * x_i + \widehat{\beta}_0))^2$  找到一条线，犯最小的错误，即可以最好的拟合数据生成



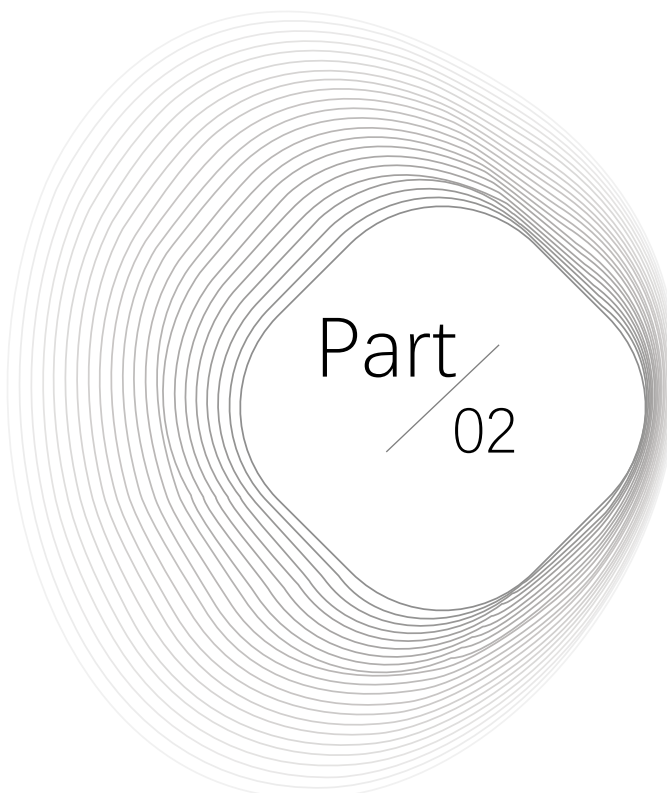
# 线性回归：更高维度



- 如何度量模型的准确性？

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- 残差减小了多少
- 为什么这是一个好指标？



Part  
02

# 预测能力

- 当我们追求预测能力
- 我们为什么追求预测能力
- 那么如何变得更好？





## 线性模型做预测

- 如果我们想尽可能地提升预测性？

[返回](#)

微博正文



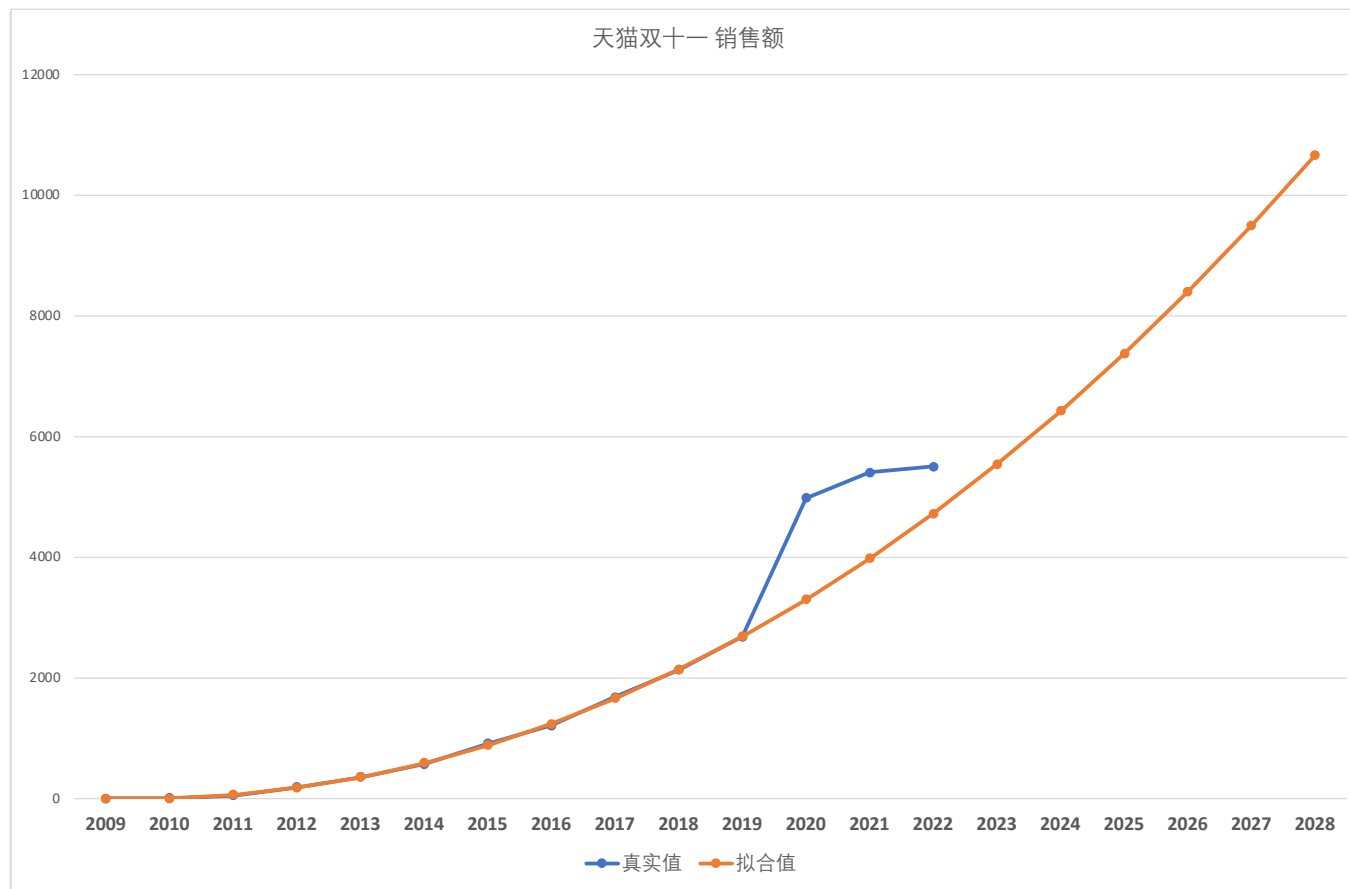
尹.

4-24 22:43 来自微博 weibo.com

+关注

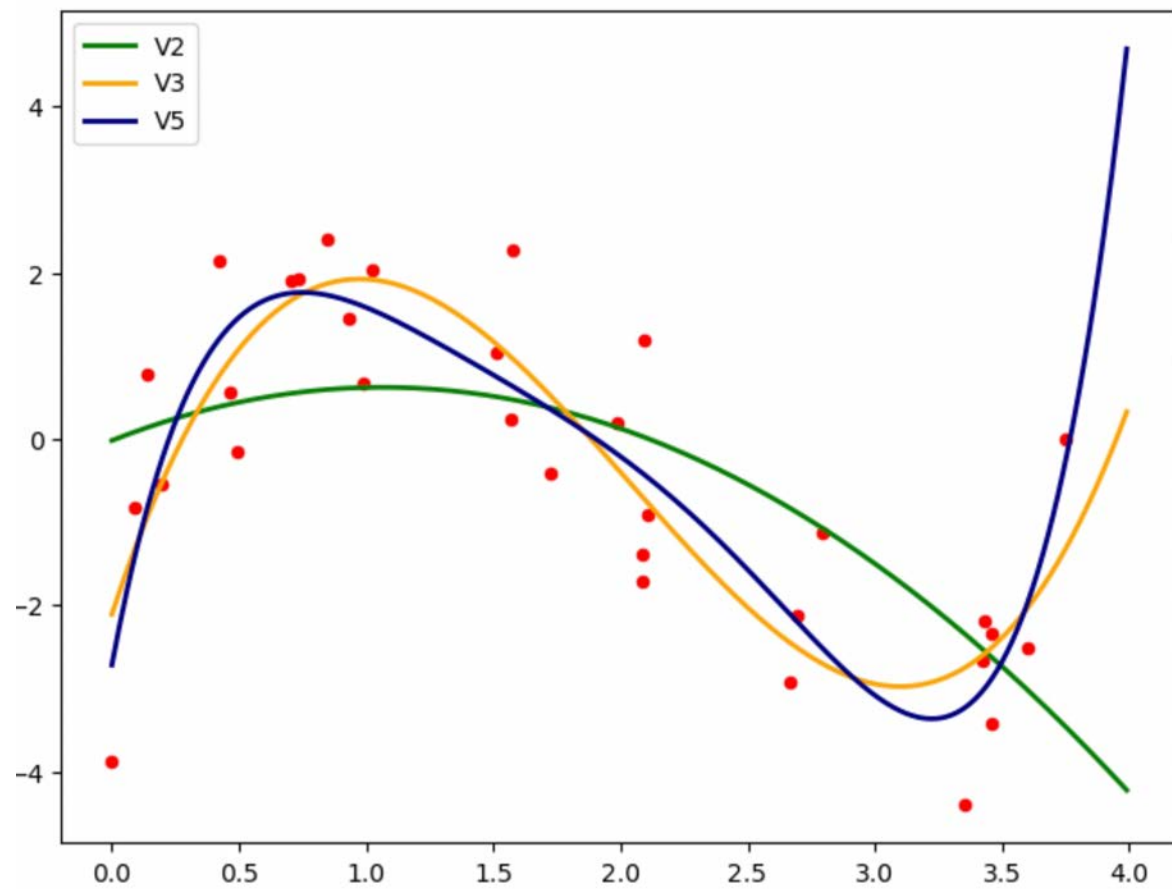
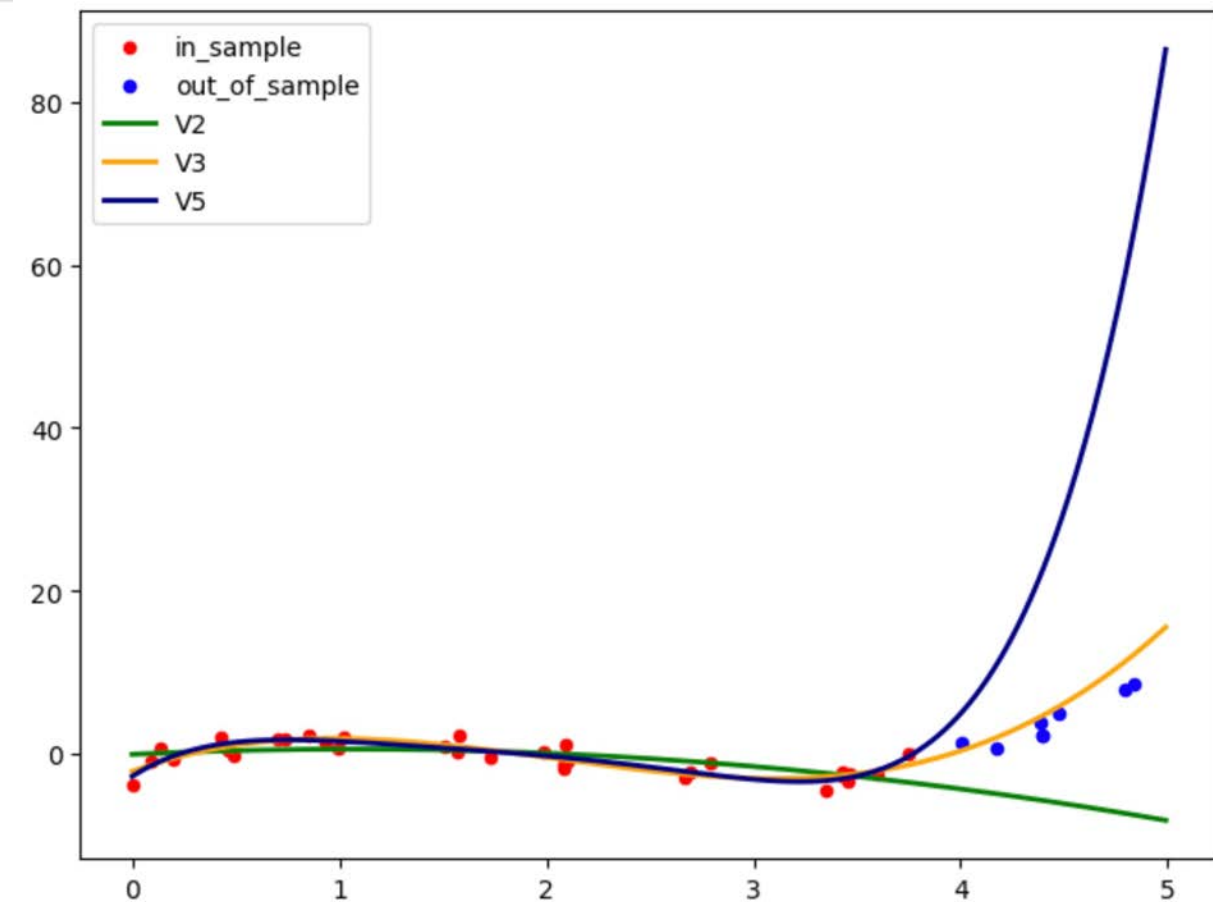
#淘宝双11骗局# 从天猫双十一的全天销售额来看，实际生产数据几乎完美地分布在三次回归曲线上，拟合度均超过99.94%，几乎为1，而且生产数据有10年之久，每一年的数据都这么高度拟合，数据过于完美，销售额与年份的增长趋势仿佛按预期设定的线性公式发展，属于小概率事件，在实际生活中几乎是不可能发生的事。因此可以断定，阿里为了吸引双十一的购物热度，对销售额数据进行了人工修饰，存在造假事实。可断定淘宝历年双11全天销售额数据存在假造，并且从一开始就在造假。马云真的是个大骗子，骗了全世界人民，并且骗了十年。如果继续如此造假，可预测2019年淘宝双11当天销售额为2675.37亿或者2689.00亿。

淘宝 2009 至 2018 年历年双 11 销售额数据造假

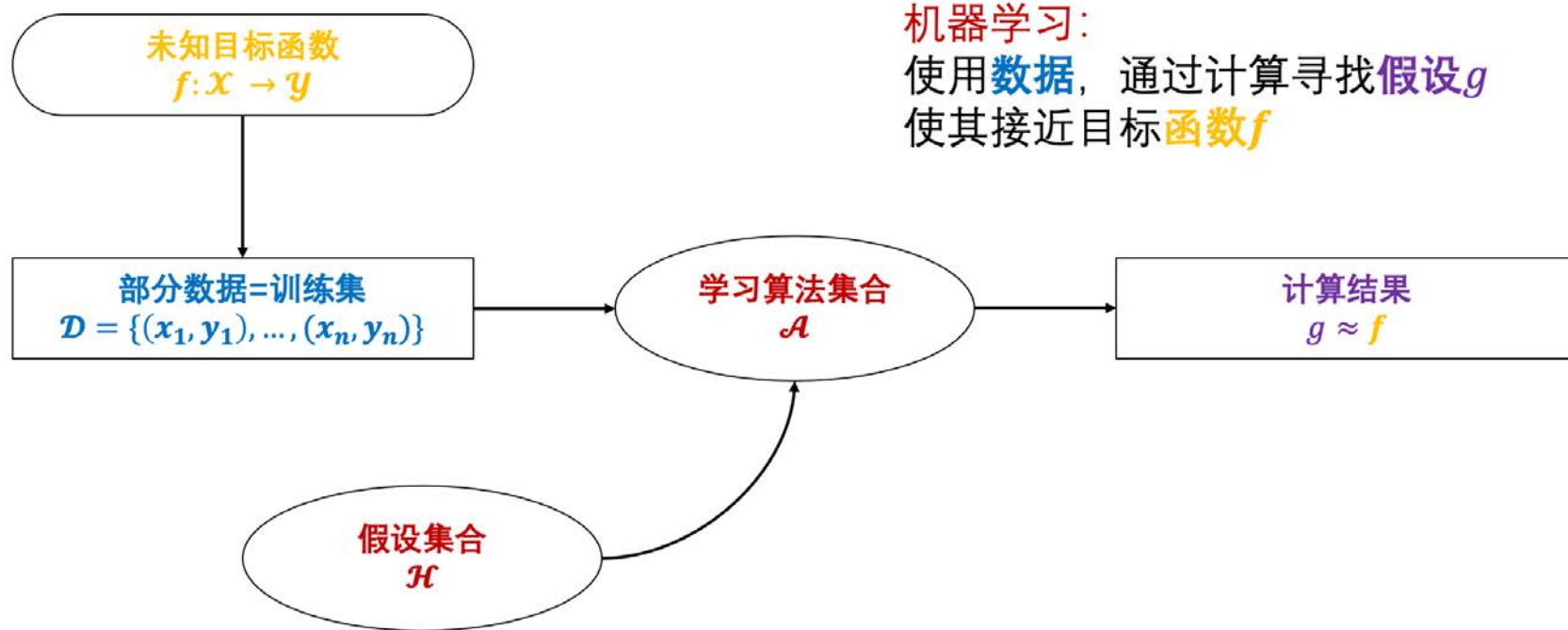


## 2.1

## 一个更可控的例子



## 我们为什么追求预测？



几个核心矛盾:

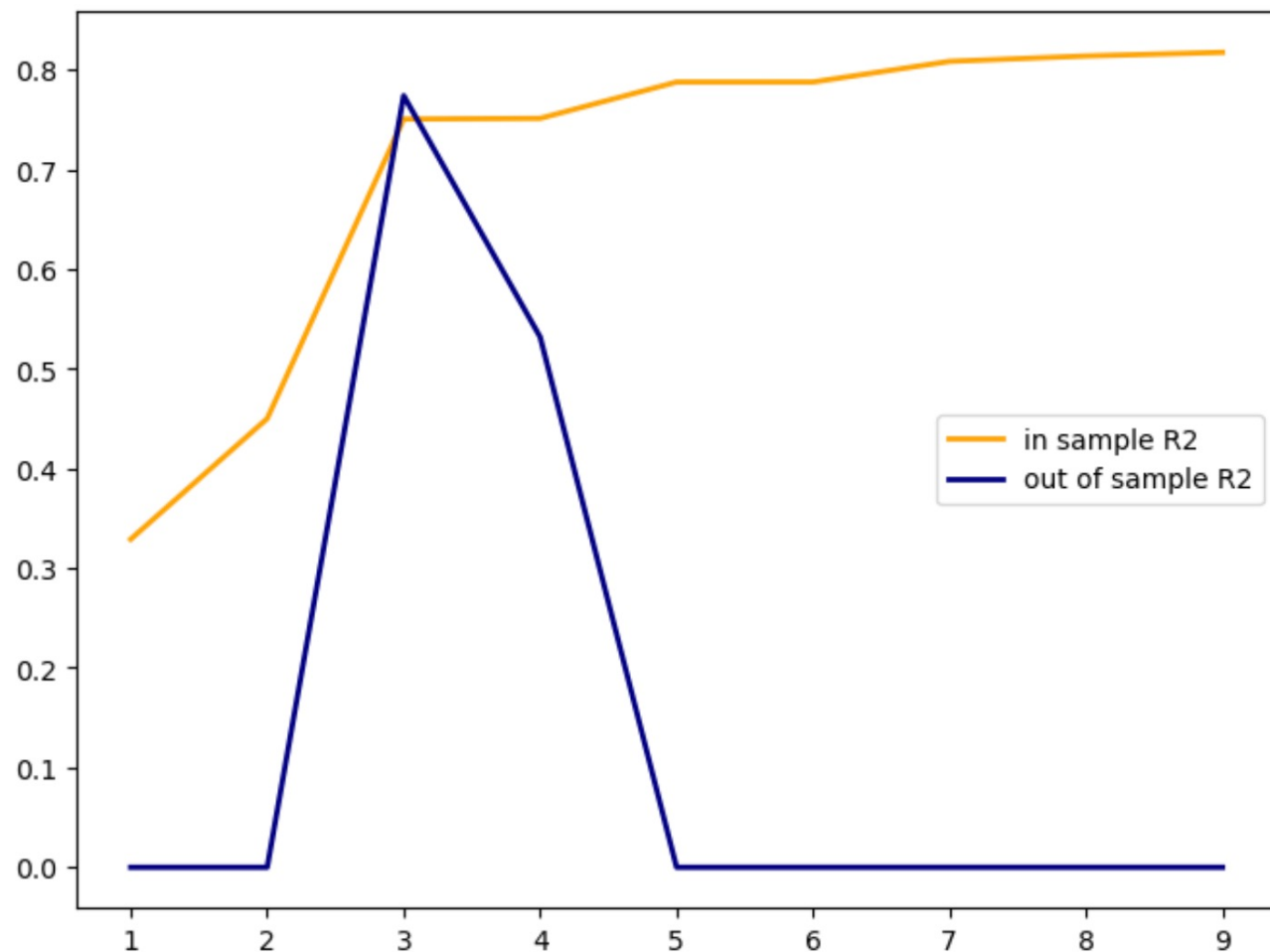
数据残缺  
且不知道缺多少

算法很多  
算法形态很多

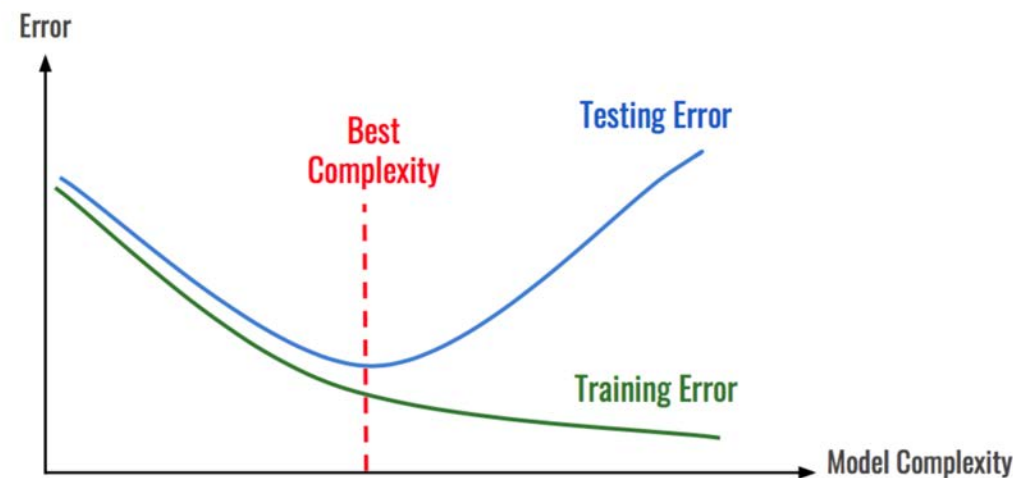
我们想要逼近  $f$   
但只有  $g$

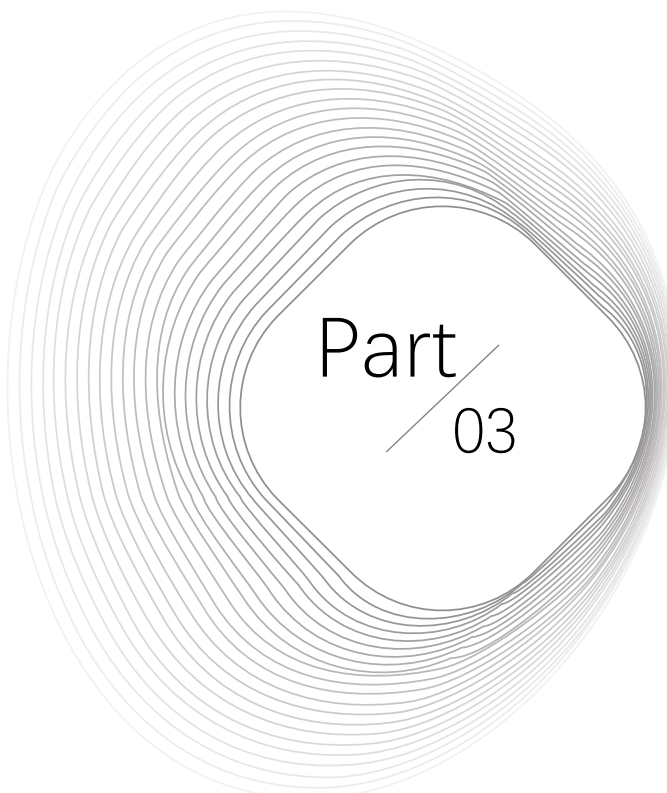
## 2.2

## 准确度的含义



- 样本内与样本外
  - 样本内：算法见过；样本外：算法没见过
  - 样本外：特意留的 而非真的
- 我们该关注哪个？
- 为什么之前只关注样本内？





Part  
03

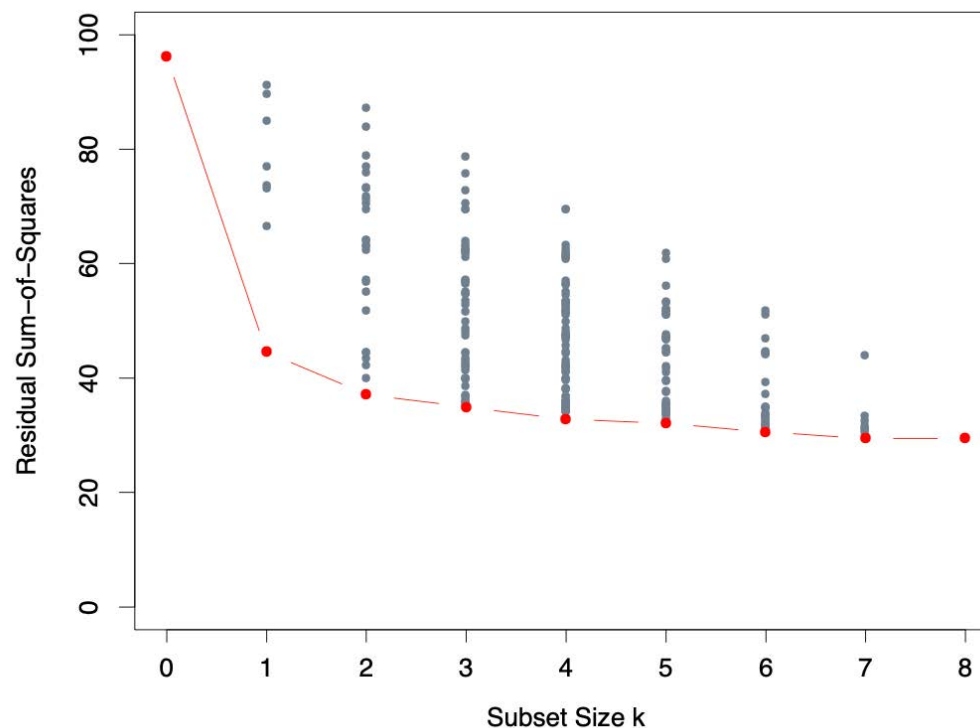
# 模型选择

- 特征数量
- 理论与现实
- 正则化



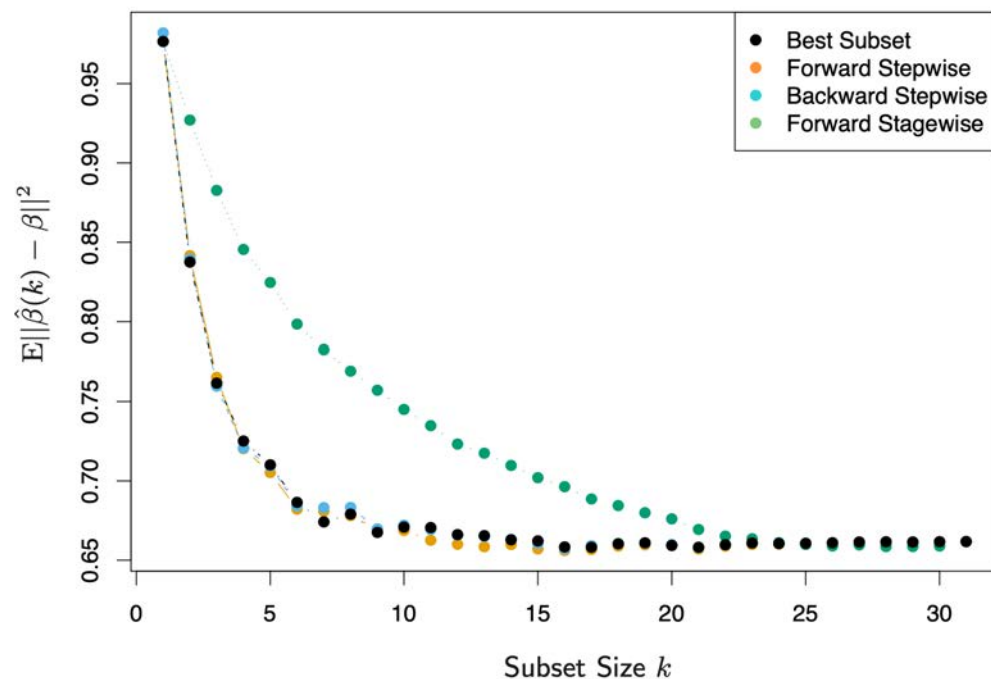
## 如何限制模型复杂度：限制变量个数

- 如何决定是否包含某个变量
- 数据里的 $b$ 个变量就是回归里的 $b$ 个变量？
- 如果候选变量个数 $b$ 大于样本量 $N$ ？
- 最佳子集：
  - 每多一个变量，预测能力都会提升
  - 确定size下最佳的subset
  - 那最佳的size呢？



## 可行方案：前向与后向逐步回归

- 前向逐步回归 (forward stepwise regression)
- 执行方式
  - 从截距项开始，逐步加入解释变量
  - 在每一步，加入一个变量，使得新模型较老模型有最大的准确度提升
- 特点
  - 计算可行性，甚至 $b > N$ ；可能陷入局部最优
- 后向逐步回归 (backward stepwise regression)
- 执行方式
  - 从完整模型开始，每次删去一个影响最小的解释变量
- 特点
  - 需要 $N > b$ ，预先设定完整模型，局部最优问题稍轻





# 如何限制模型复杂度：限制变量系数

OLS

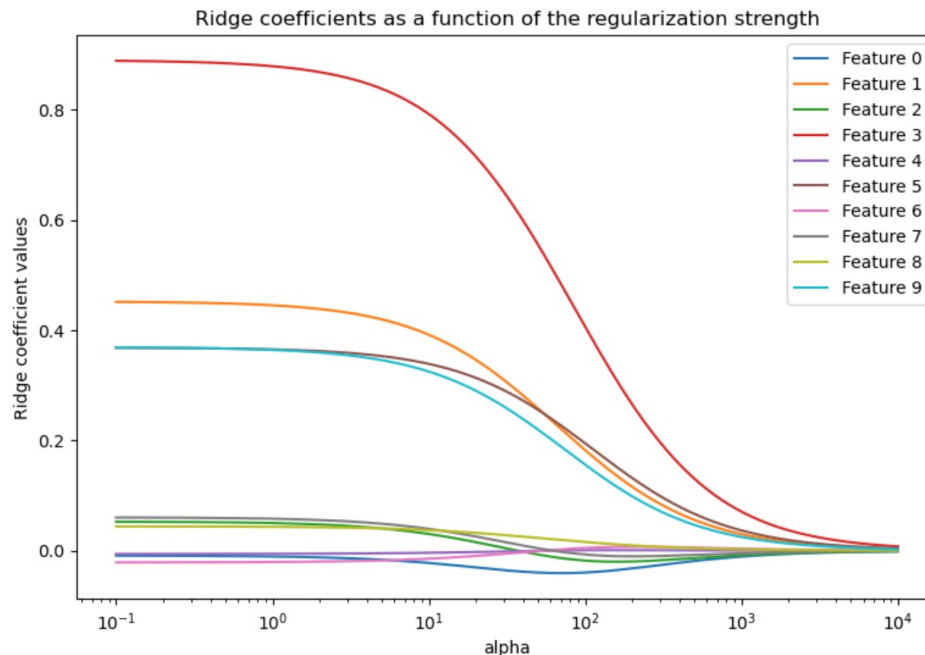
$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j.$$

$$\hat{\beta} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \right\}.$$

Ridge Regression

$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j.$$

$$\hat{\beta}^{\text{ridge}} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}.$$



- $\lambda$ 的含义是什么？
- 一种权重、一种价格
- 系数帮你降低残差
- 但是你要向系数支付
- 预算约束下的抉择
- **超参数**
- 要人来设定



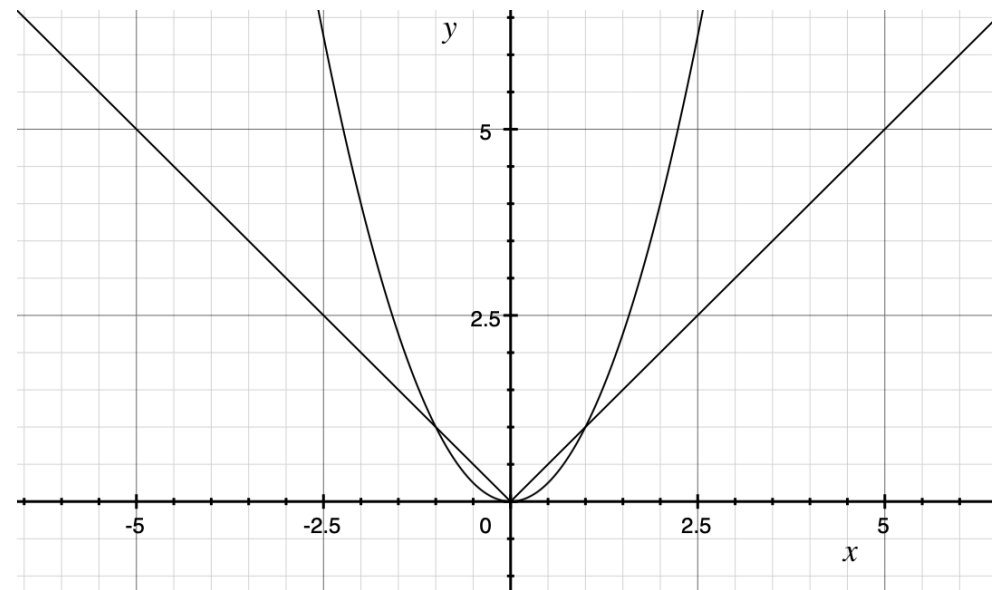


## Ridge的局限

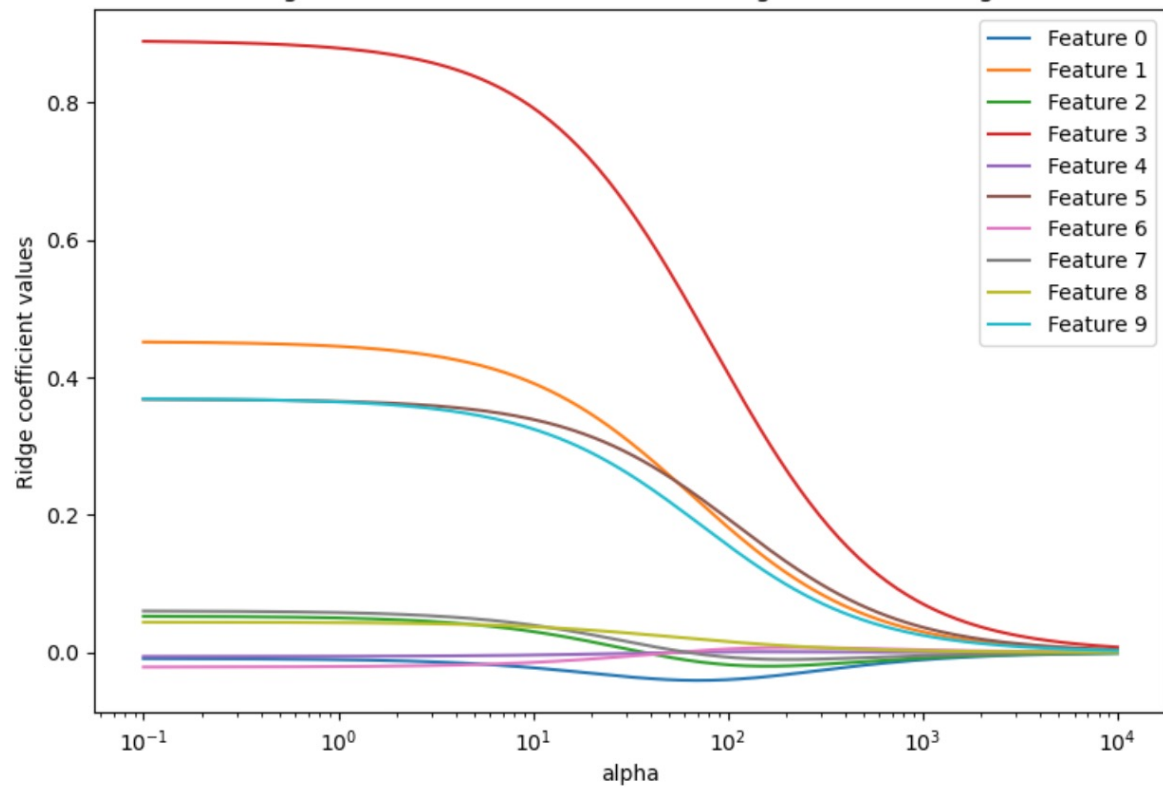
- 正则项的次数是2次，则在趋近于0时，惩戒很轻
- Beta可能会停留在一个很小但是非0的值

$$\hat{\beta}^{ridge} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

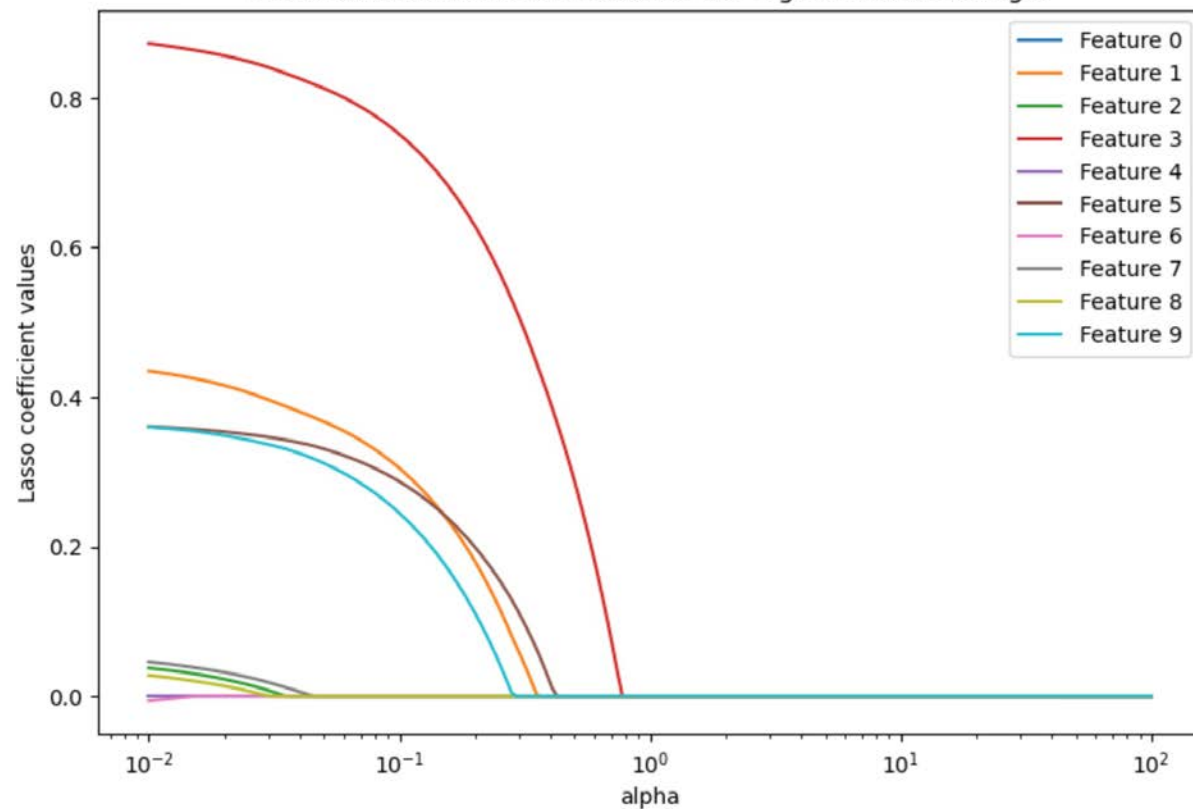
$$\hat{\beta}^{Lasso} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$



Ridge coefficients as a function of the regularization strength

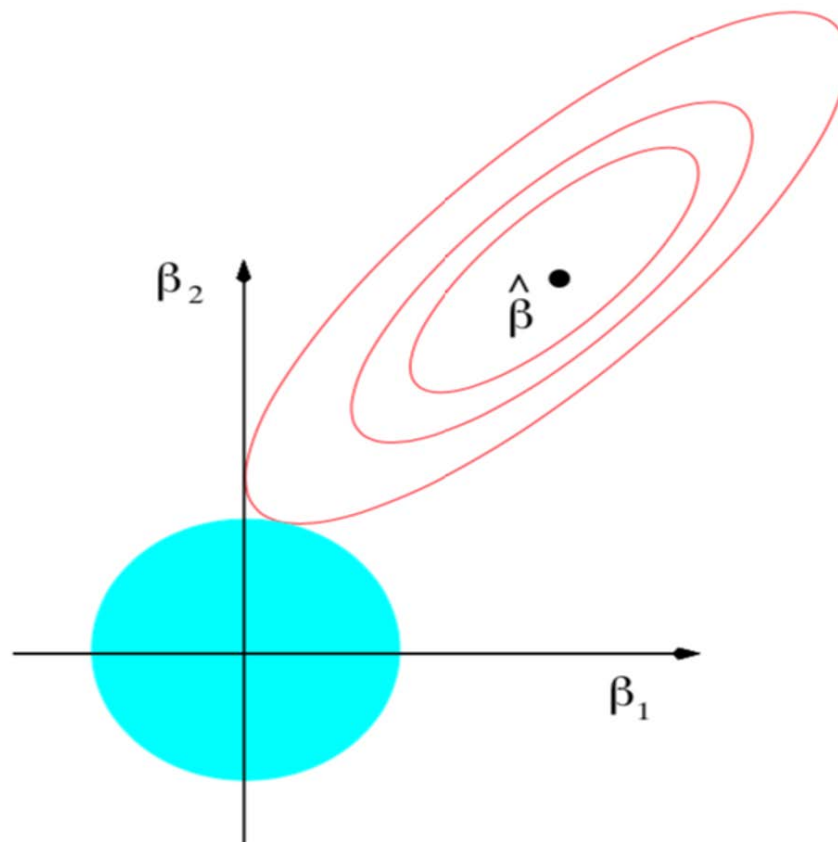
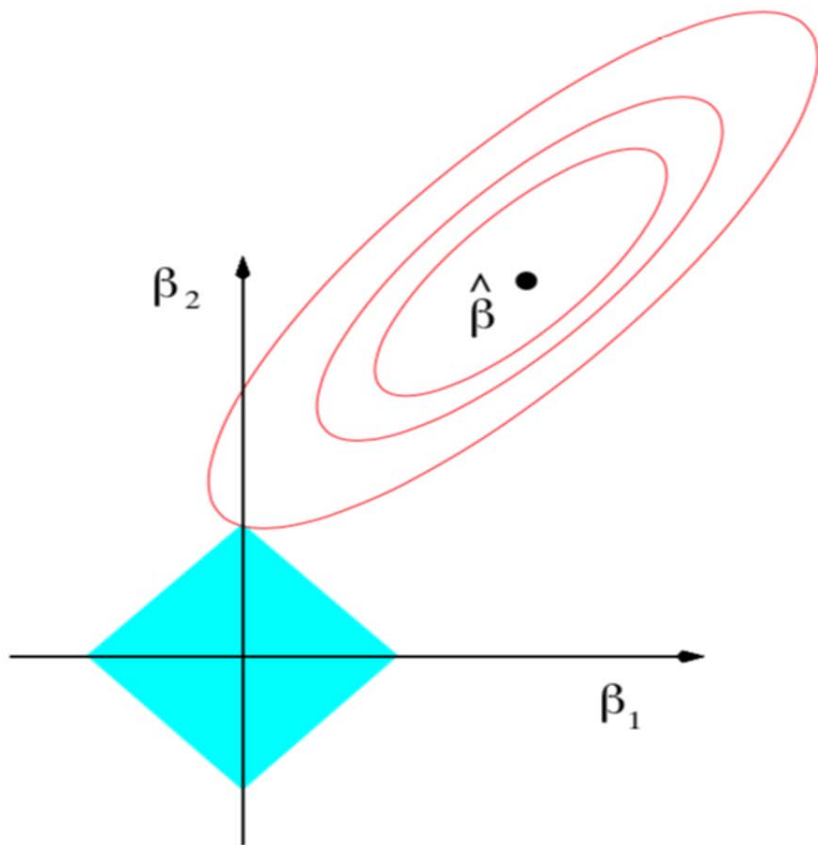


Lasso coefficients as a function of the regularization strength



3.5

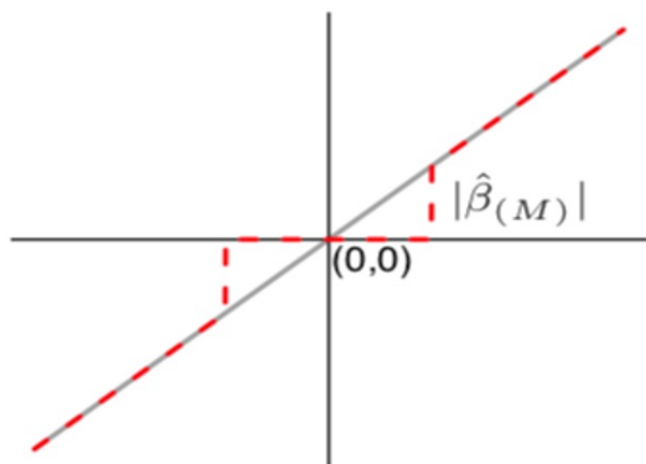
## Ridge 对比 Lasso



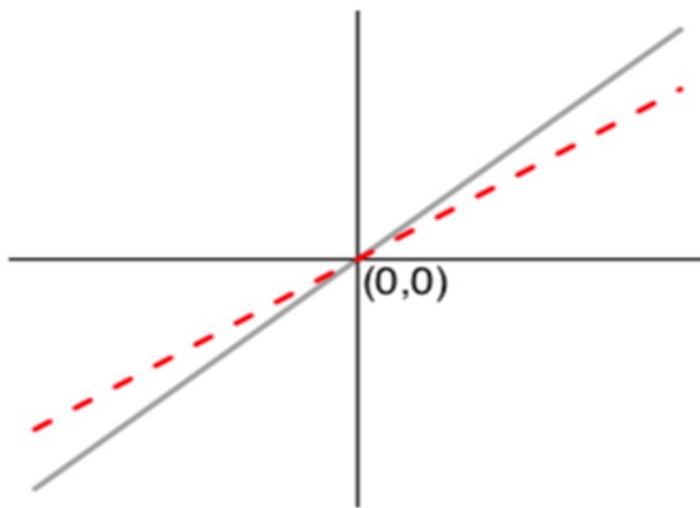
3.6

## 讨论：上述三种方法

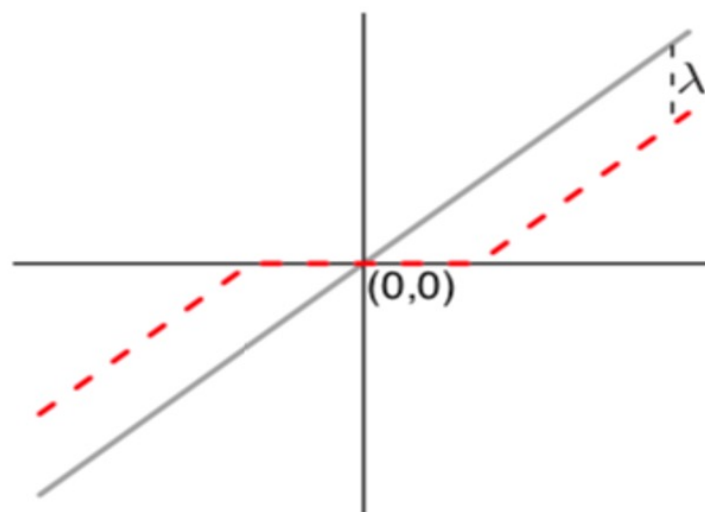
Best Subset



Ridge



Lasso



## 应用：Ridge与Lasso在经济金融中的应用

**Ridge：**当要求预测的有效性（方差小）时

**Lasso：**可以作为指标筛选的方法、 $b > N$ 时的方法

**超参数：**惩罚项的系数，越大则越追求有效性（小方差）

为什么计量经济学仍使用OLS？

# 我们为什么使用线性回归

计量经济学

机器学习

目的

研究某个变量影响

预测

变量重要性

理论、偏好  
约定俗成

无偏好

要求的性质

易于解释、理解

最小化 样本外误差

1.一致性

2.无偏性

3.有效性

表示力

预测区间

解释性



## L1正则 与 L2正则

- 更一般地，我们定义L1-正则（L1-Regulation）和L2-正则
- 是指针对系数 $w$ 进行“收费”，是一种**通用的正则化手段**
- L1根据 $|w|$ 的大小进行收费，L2根据 $w^2$ 进行收费
- 费率（**超参数**）还是人为设定的
- L2 高效、便宜、泛用性好
- L1 可以使得模型保持**稀疏性**
- 那么有没有可能综合二者同时使用？

## Elastic-Net 弹性网络

- 我全都要

$$\min_w \frac{1}{2n_{\text{samples}}} \|Xw - y\|_2^2 + \alpha\rho \|w\|_1 + \frac{\alpha(1-\rho)}{2} \|w\|_2^2$$

$\alpha$  衡量正则力度

$\rho$  衡量L1和L2权重

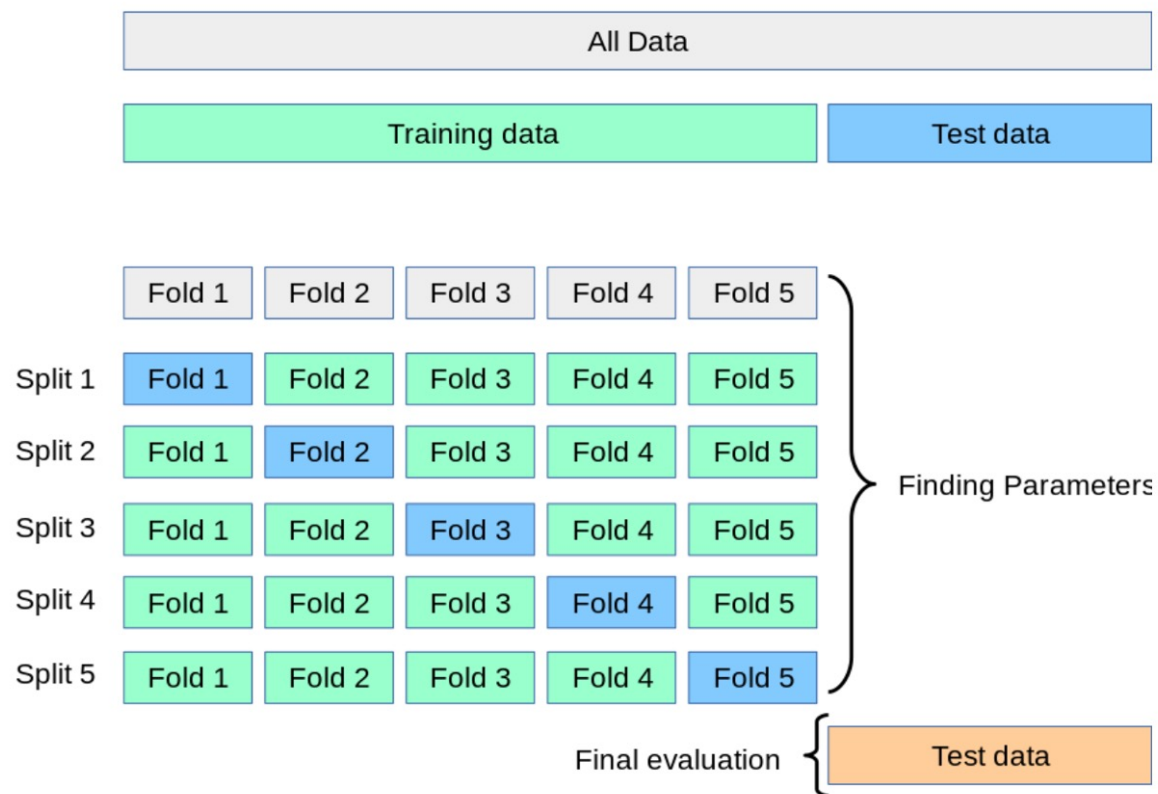
为什么这么设计？

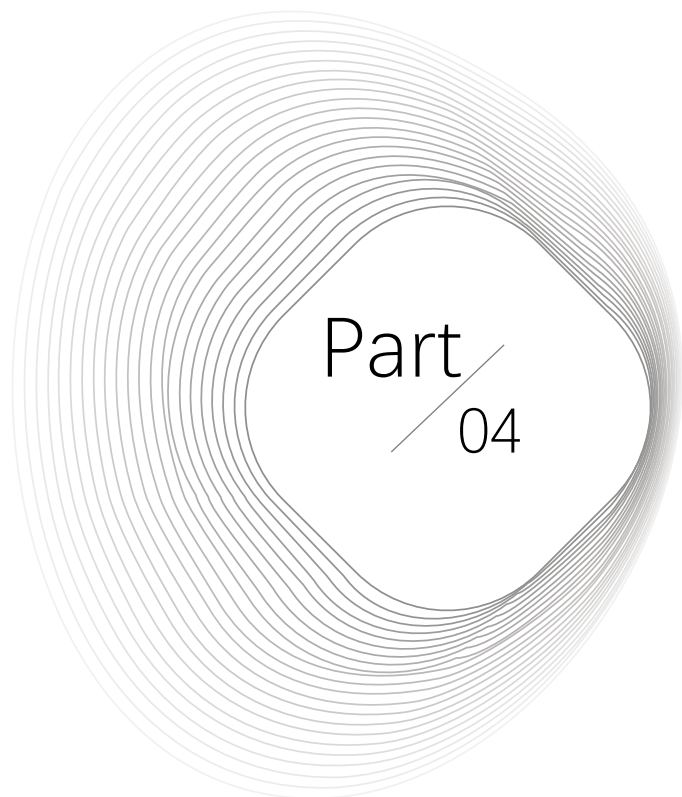




## 如何选择超参数?

- 参数是模型自动训练生成的, 但是超参数是人挑的
- 如何挑一个好的超参数
- 实践是检验真理的唯一标准, 多试几次
- 我们该如何高效的使用数据?
  - 留出多少数据做测试?
  - 能不能多用几次?
  - 交叉验证: k-fold cross validation
- 如何告诉别人你的结果有多好?
- 如何挑一个好的超参数?
  - 不要忽视你的大脑





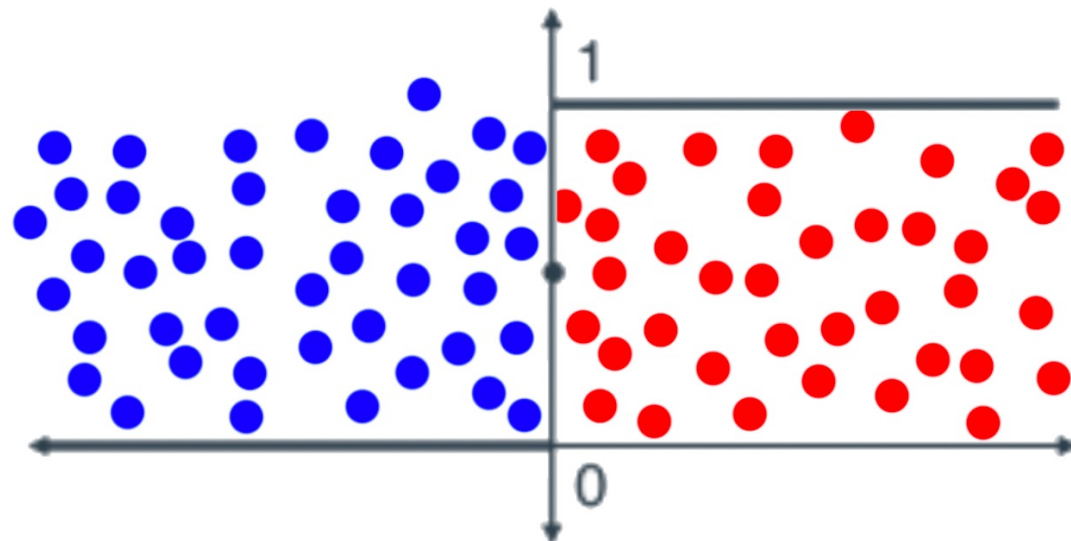
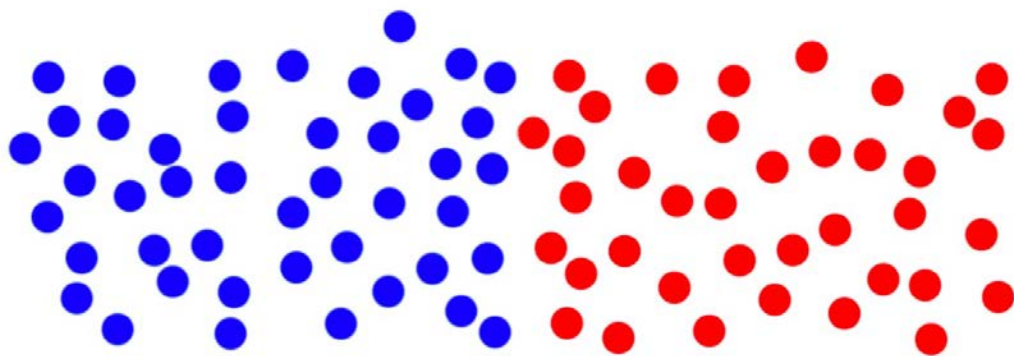
Part  
/ 04

# 逻辑回归

- 披着羊皮的狼
- 如何正则化



## 伪装成回归的分类问题



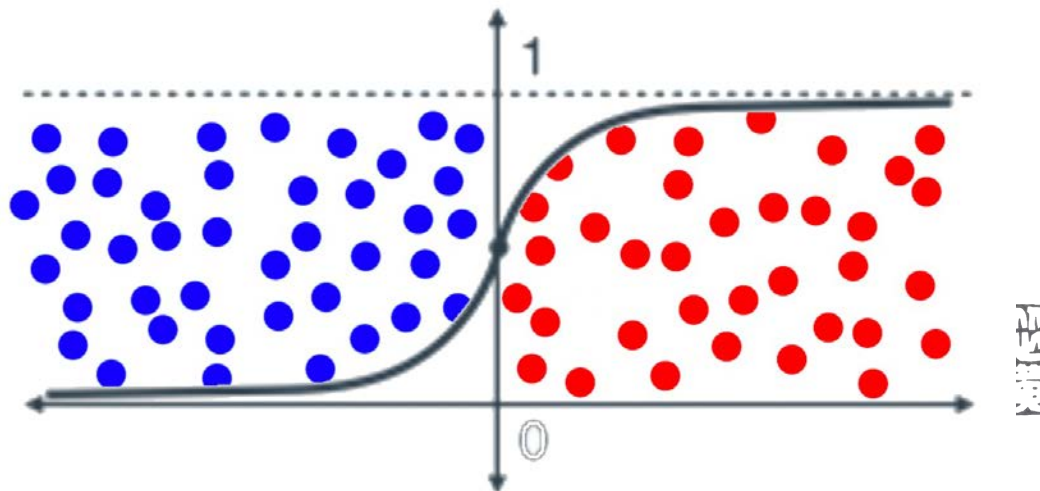
- 这种分割没法求导，能不能模拟一下？

$$\begin{cases} y \equiv 1 & \text{if } x \geq 0 \\ y \equiv 0 & \text{if } x \leq 0 \end{cases} \quad z = wx + b$$

$$\frac{\Pr(y=1)}{1-\Pr(y=1)} = e^{wx+b}$$

- 增长的要快，取值范围  $(0, +\infty)$

- $\Pr(y = 1) \equiv y = \frac{1}{1+e^{-(wx+b)}}$  逻辑回归



- $\Pr(y = 1) \equiv y = \frac{1}{1+e^{-(wx+b)}}$  逻辑回归
- 如何度量这种损失?
- $Loss = \sum_{i=1}^N (-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i))$
- **交叉熵** Cross-Entropy
- 如何正则呢
- $Loss = \textcolor{red}{C} \sum_{i=1}^N (-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i)) + \textcolor{blue}{r(w)}$
- L1 L2 正则化的普遍性
- 有几个“超参数”呢?

penalty	$r(w)$
None	0
$\ell_1$	$\ w\ _1$
$\ell_2$	$\frac{1}{2} \ w\ _2^2 = \frac{1}{2} w^T w$
ElasticNet	$\frac{1-\rho}{2} w^T w + \rho \ w\ _1$





2<sup>0</sup><sub>2</sub>3

THANKS

