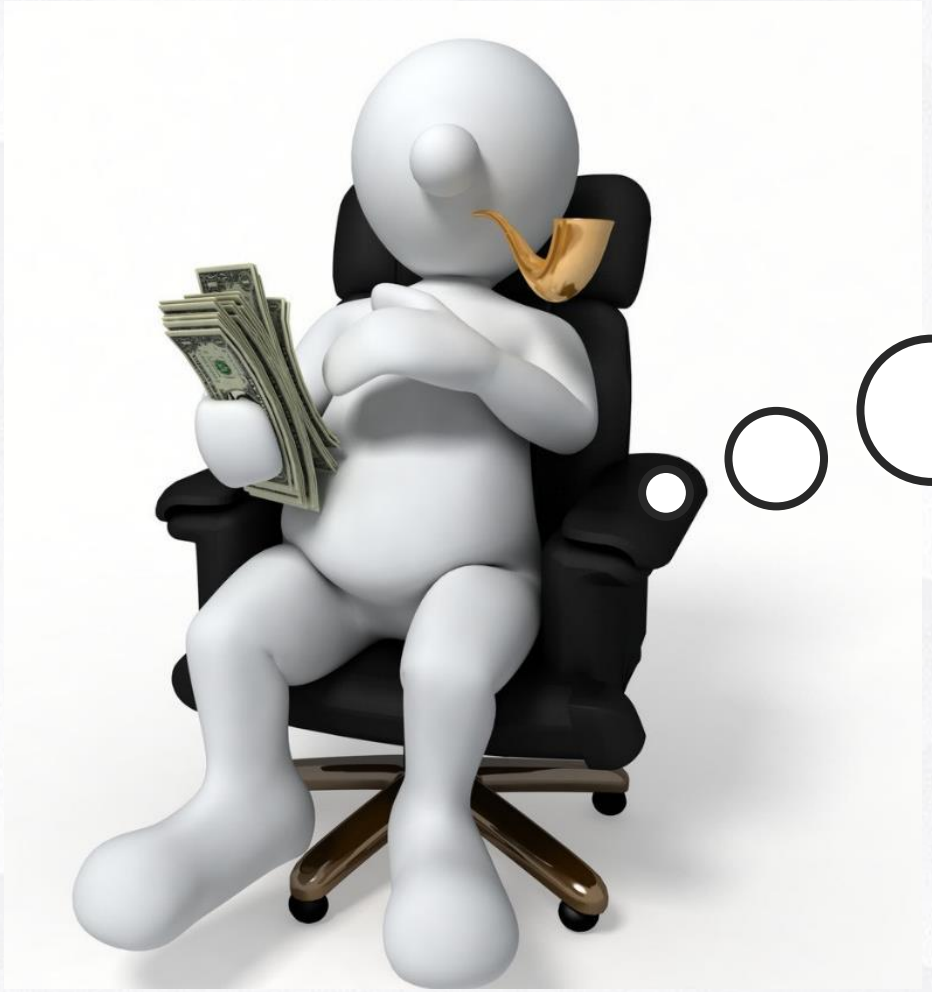
老王的淘宝店



隔壁老王开了一家淘宝店，他有1万元广告预算，他应该把钱投到哪呢?

他有三个选择

Baidu 百度

CCTV.
中国中央电视台
CHINA CENTRAL TELEVISION

And 一个问题

销售业绩

?

Baidu 百度

CCTV.
中国中央电视台
CHINA CENTRAL TELEVISION

And 一个问题

因变量：连续型

自变量

?

明确因变量

Y = UBI车险

Y = 消费价格

Y = 房价

目标（有监督学习）

- 因变量：Y
- 自变量：$X = (\mathrm{X}_1, \mathrm{X}_2, \cdots, X_p)'$

$$\mathrm{Y} = f(X) + \epsilon$$

误差项

为何要估计f？

- 目标1：**预测（Prediction）** $\hat{Y} = \hat{f}(X)$

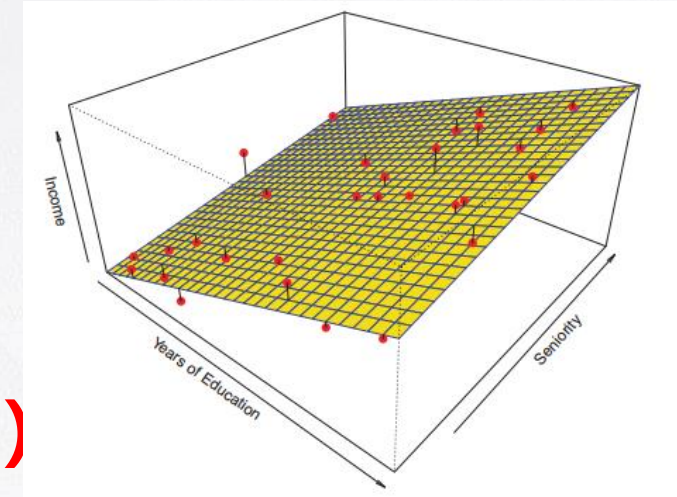
$$\mathrm{E}\left(Y - \hat{Y}\right)^2 = \mathrm{E}\left[f(X) - \hat{f}(X)\right]^2 + Var(\epsilon)$$

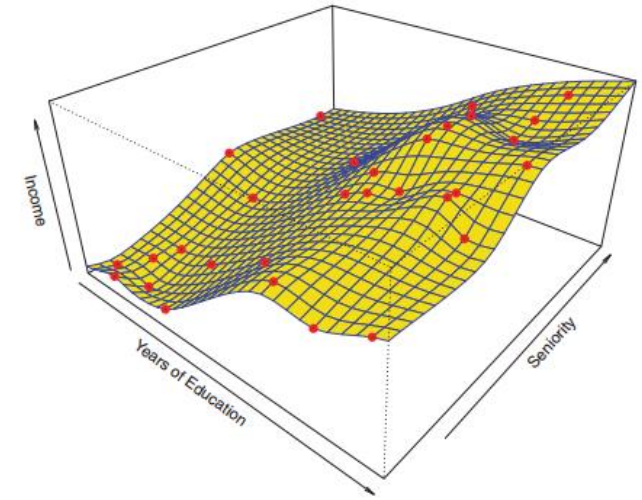
Reducible **Irreducible**

- 问题：当Y是离散型时，如何量化预测误差？
- 目标2：**推断（Inference）**
  - 哪些自变量与Y有关？
  - 具体是什么关系？能否线性表达？

如何估计f？

- 方法1：**参数方法（Parametric Methods）**
  - 例如：线性回归模型
- 方法2：**非参数方法（Non-parametric Methods）**
  - 不假设关于f的参数形式
  - 较好的拟合性和光滑性：Spline，Kernel Smoothing
  - 缺点：容易过拟合





模型评价

- 连续型因变量：MSE（Mean Squared Error）

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

- Training MSE：在训练集评估
- **Test MSE**: 在预测集评估

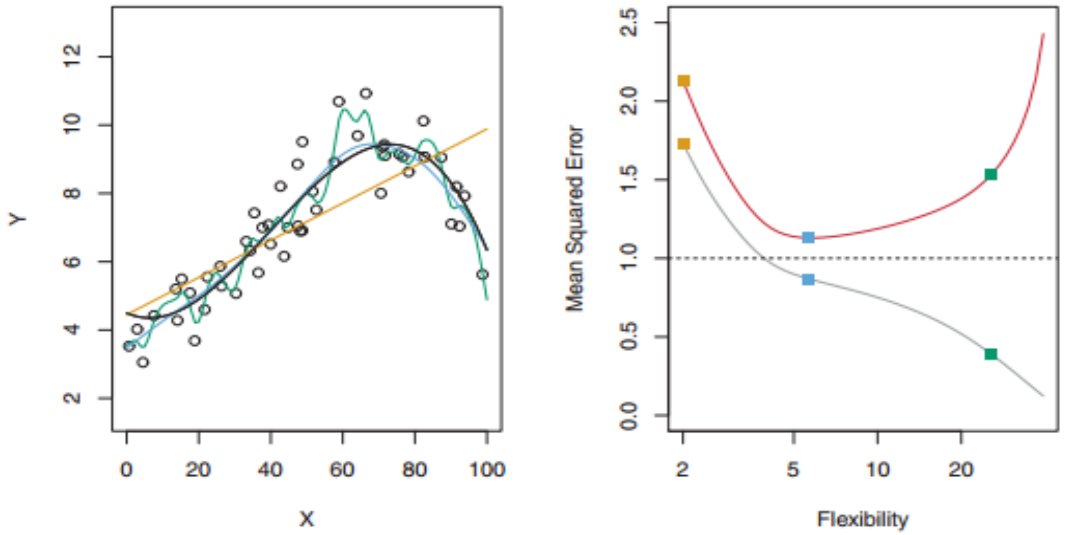


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

Bias-Variance Trade-off

- Expected test MSE:

$$\mathrm{E}\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\epsilon)$$

Y：回归问题

Y：个人收入

教育程度
性别
年龄
…

Y：登录时长

好友发帖数
粉丝数
转发数
…

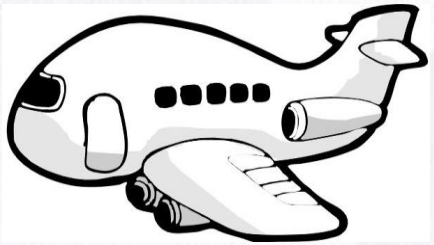
Y：汽车保养花费

汽车品牌
价格
车型
…

Y：分类问题

Y：是否流失

当月花费
好友个数
满意度
…



Y：是否延误

天气状况
机型
目的地
…

Y：是否被ST

资产规模
资产周转率
资产收益率
…


谢谢！