Foundations of Data Science

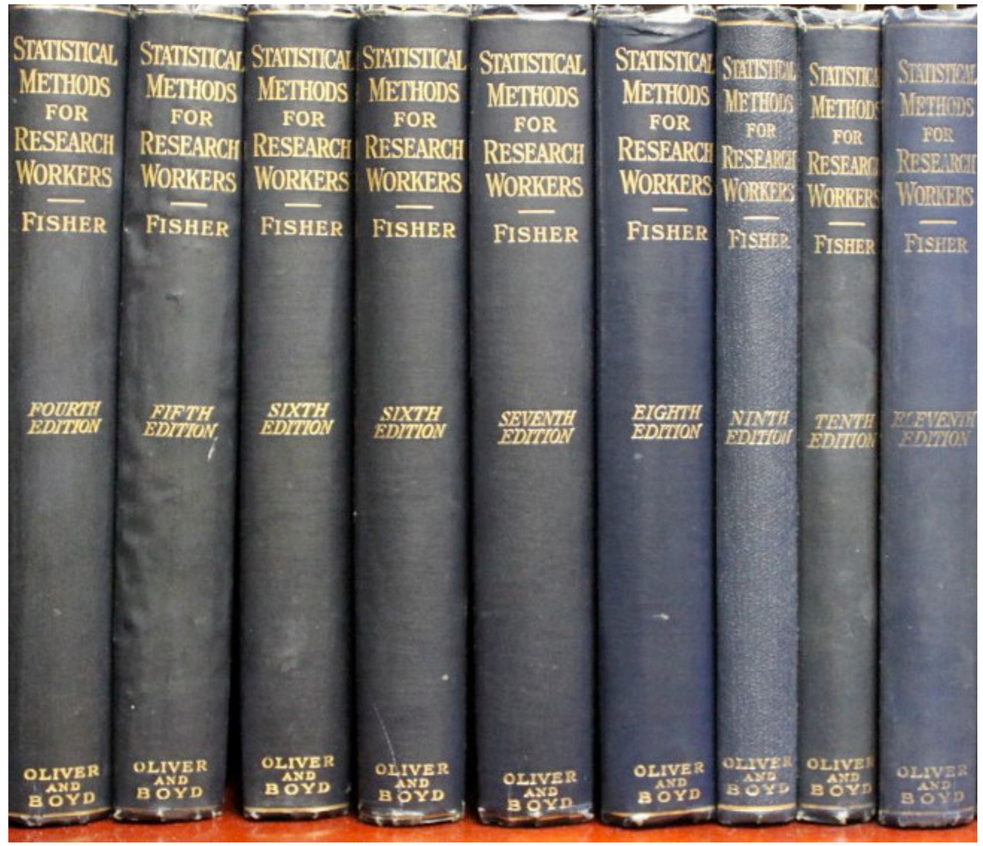
Statistical Analyses

尹一通、刘明谋 Nanjing University, 2024 Fall

方差分析

Analysis of Variance (ANOVA)

Ronald Fisher
(1890-1962)

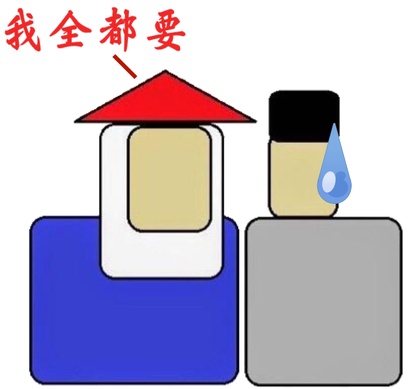


Statistical Methods for Research Workers (1925)

析因设计 / 析因分析

Factorial design / Factorial analysis

- 多个因素会影响实验结果吗？
  - 独立性检验（列联表、Pearson卡方检验）
- 不同因素对实验结果的影响如何？是否存在 “1+1>2” 的 “化学反应”？
  - 析因设计、析因分析
- 方差分析（ANOVA）、回归分析（regression）



单因素方差分析

one-way / single-factor analysis of variance

- 单个因素的各个水平对实验结果的影响（是否一样？最好的？最差的？）

- 模型假设（m 个水平，每个水平 n_i 次试验）
  - 正态性：水平 i 对结果的效应(effect)为 μ_i，每次试验有噪声 $\epsilon_{ij} \sim N(0,\sigma_i^2)$
  - 方差齐性：噪声同分布 $\sigma_1^2 = \ldots = \sigma_m^2 = \sigma^2$
  - 独立性：噪声之间相互独立

All models are wrong,
but some are useful.

- 现实：不同实验对象特质不同，明显影响试验结果
  - 更大的样本量 + 随机分组，每一组整体看起来都差不多

分析方差

- 总离差平方和（Sum of Squares）：$SS_T = \Sigma_{i,j}(X_{ij} - \bar{X}_{..})^2$
  - 组内离差平方和（误差平方和）：$SS_W = \Sigma_{i,j}(X_{ij} - \bar{X}_{i.})^2$
  - 组间离差平方和（效应平方和）：$SS_B = \Sigma_{i,j}(\bar{X}_{i.} - \bar{X}_{..})^2$

- 直觉：效应 $\mu_i \approx$ 组内均值 $\bar{X}_{i.}$，如果效应有差距，组间离差平方和会较大

  计算：$\mathbb{E}[SS_B] = \mathbb{E}[\Sigma_i n_i \bar{X}_{i.}^2 - n\bar{X}_{..}^2]$

$$
\begin{aligned}
&= \Sigma_i n_i \left[\mathrm{Var}(\bar{X}_{i.}) + \mathbb{E}[\bar{X}_{i.}]^2\right] - n\left[\mathrm{Var}(\bar{X}_{..}) + \mathbb{E}[\bar{X}_{..}]^2\right] \\
&= \Sigma_i n_i \left[\sigma^2/n_i + \mu_i^2\right] - n\left[\sigma^2/n + \bar{\mu}^2\right] \\
&= (m-1)\sigma^2 + \Sigma_i n_i(\mu_i - \bar{\mu})^2
\end{aligned}
$$

分析方差
组间离差平方和

- 组间离差平方和：$SS_B = \Sigma_{i,j}(\bar{X}_{i.} - \bar{X}_{..})^2$

  计算：$\mathbb{E}[SS_B] = \mathbb{E}[\Sigma_i n_i \bar{X}_{i.}^2 - n\bar{X}_{..}^2]$

  $$= \Sigma_i n_i \left[\mathrm{Var}(\bar{X}_{i.}) + \mathbb{E}[\bar{X}_{i.}]^2\right] - n\left[\mathrm{Var}(\bar{X}_{..}) + \mathbb{E}[\bar{X}_{..}]^2\right]$$

  $$= \Sigma_i n_i \left[\sigma^2/n_i + \mu_i\right] - n\left[\sigma^2/n + \bar{\mu}^2\right]$$

  $$= (m-1)\sigma^2 + \Sigma_i n_i(\mu_i - \bar{\mu})^2$$

- 如果效应相同 $\mu_i = \bar{\mu}$，组间离差平方和 $SS_B/(m-1)$ 是方差的无偏估计，

  $$\text{且 } SS_B/\sigma^2 \sim \chi^2(m-1)$$

正态总体的样本方差的分布

- 若 $X_1, \ldots, X_n \sim N(0,1)$, 则 $\sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2(n-1)$.
- **Proof**: 样本方差 $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$. 注意到 $\sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$

  又因为 $\bar{X} \sim N(0,1/n)$, 所以 $\sqrt{n}\bar{X} \sim N(0,1)$, 且 $n\bar{X}^2 \sim \chi^2(1)$

  改写 $\sum_{i=1}^{n} X_i^2 = (n-1)S^2 + n\bar{X}^2$, 记作 $\chi^2(n) = \underline{(n-1)S^2 + \chi^2(1)}$

  相互独立

矩生成函数 (MGF) : $M_{\chi_n^2}(t) = M_{(n-1)S^2}(t) \cdot M_{\chi_1^2}(t)$

卡方分布的矩生成函数是

$$M_{\chi_n^2}(t) = (1-2t)^{-n/2}$$

分析方差
组内离差平方和

- 组内离差平方和：$SS_W = \sum_{i,j} (X_{ij} - \bar{X}_{i.})^2$
  - $\sum_{j} (X_{ij} - \bar{X}_{i.})^2/\sigma^2 \sim \chi^2(n_i - 1)$
  - $SS_W/\sigma^2 \sim \chi^2(n - m)$
- 如果效应相同 $\mu_i = \bar{\mu}$，$SS_B/(m-1)$ 是方差的无偏估计，且

$$SS_B/\sigma^2 \sim \chi^2(m-1)$$

- F 检验?

正态总体的样本均值与样本方差

- 若 $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$，则 $\bar{X}$ 与 S^2 相互独立。
- **Proof**: 严格一点：$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$，$\bar{X} - X_j = \frac{1}{n}\sum_{i \neq j} X_i - \frac{n-1}{n} X_j$

$$\bar{X} - X_j \sim N\left(\frac{(n-1)\mu}{n} - \frac{(n-1)\mu}{n}, (n-1)\frac{\sigma^2}{n^2} + (n-1)^2\frac{\sigma^2}{n^2}\right) = N\left(0, \frac{n-1}{n}\sigma^2\right)$$

两两之间的协方差：$\mathrm{Cov}(\bar{X} - X_j, \bar{X}) = \mathrm{Cov}(X_j, \bar{X}) - \mathrm{Cov}(\bar{X}, \bar{X}) = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0$

$$\| $$

$$\mathrm{Cov}(X_j, X_j/n) = \sigma^2/n$$

两两独立不等于相互独立，但在联合正态分布中意味着相互独立

假设检验

- 总离差平方和（Sum of Squares）：$SS_T = \Sigma_{i,j}(X_{ij} - \bar{X}_{..})^2$
  - 组内离差平方和：$SS_W = \Sigma_{i,j}(X_{ij} - \bar{X}_{i.})^2$
  - 组间离差平方和：$SS_B = \Sigma_{i,j}(\bar{X}_{i.} - \bar{X}_{..})^2$

- 如果效应相同 $\mu_i = \bar{\mu}$，组间离差平方和 $SS_B/(m-1)$ 是方差的无偏估计，

$$且\ SS_B/\sigma^2 \sim \chi^2(m-1)$$

- 无关因素，组内离差平方和 $SS_W/\sigma^2 \sim \chi^2(n-m)$

- F 检验：如果效应相同 $\mu_i = \bar{\mu}$，$\dfrac{SS_B/(m-1)}{SS_W/(n-m)} \sim F(m-1, n-m)$

方差分析表

- 总离差平方和（Sum of Squares）：$SS_T = \Sigma_{i,j}(X_{ij} - \bar{X}_{..})^2$
  - 组内离差平方和：$SS_W = \Sigma_{i,j}(X_{ij} - \bar{X}_{i.})^2$
  - 组间离差平方和：$SS_B = \Sigma_{i,j}(\bar{X}_{i.} - \bar{X}_{..})^2$

- F 检验：如果效应相同 $\mu_i = \bar{\mu}$, $\dfrac{SS_B/(m-1)}{SS_W/(n-m)} \sim F(m-1, n-m)$

方差来源	平方和	自由度	均方(MS)	F比率
因素	SS (between)	m-1	SS(between) / (m-1)	$\dfrac{\text{MS(between)}}{\text{MS(within)}}$
误差	SS (within)	n-m	SS(within) / (n-m)	
总和	SS (total)	n-1		

快捷计算*

- 总离差平方和（Sum of Squares）：$SS_T = \Sigma_{i,j}(X_{ij} - \bar{X}_{..})^2$
  - 组内离差平方和：$SS_W = \Sigma_{i,j}(X_{ij} - \bar{X}_{i.})^2$
  - 组间离差平方和：$SS_B = \Sigma_{i,j}(\bar{X}_{i.} - \bar{X}_{..})^2$

- F 检验：如果效应相同 $\mu_i = \bar{\mu}$, $\dfrac{SS_B/(m-1)}{SS_W/(n-m)} \sim F(m-1, n-m)$

- 快捷计算：令 $T_i = \Sigma_j X_{ij}$, $T = \Sigma_i T_i$
  - $SS_T = \Sigma_{i,j} X_{ij}^2 - T^2/n$, $\quad SS_B = \Sigma_i T_i^2/n_i - T^2/n$, $\quad SS_W = SS_T - SS_B$

参数估计

- 如果效应不同，估计效应 μ_i：（最大似然估计）

$$L(\mu_1, \ldots, \mu_m, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum\sum_{i\le m\ j\le n_i} \frac{(X_{ij}-\mu_i)^2}{2\sigma^2}\right)$$

- 估计方差：无关因素，组内离差平方和 $SS_W/\sigma^2 \sim \chi^2(n-m)$

- 估计两个水平下的总体 $N(\mu_i, \sigma^2)$ 和 $N(\mu_j, \sigma^2)$ 的效应差 $\mu_i - \mu_j$

  - $\mathbb{E}[\bar{X}_{i.} - \bar{X}_{j.}] = \mu_i - \mu_j$，且 $\mathrm{Var}(\bar{X}_{i.} - \bar{X}_{j.}) = \sigma^2(1/n_i + 1/n_j)$

  - $$\frac{(\bar{X}_{i.} - \bar{X}_{j.}) - (\mu_i - \mu_j)}{\sigma\sqrt{1/n_i + 1/n_j}} \Big/ \sqrt{\frac{SS_W}{\sigma^2}/(n-m)} = \frac{(\bar{X}_{i.} - \bar{X}_{j.}) - (\mu_i - \mu_j)}{\sqrt{SS_W(1/n_i + 1/n_j)/(n-m)}} \sim t(n-m)$$

双因素方差分析

two-way analysis of variance

- 两个因素的各个水平对实验结果的影响（是否一样？最好的？最差的？）

- 模型假设（a 和 b 个水平，每个水平组合一次试验）
  - 线性：$X_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$
  - 正态性：每次试验噪声 $\epsilon_{ij} \sim N(0, \sigma_i^2)$
  - 方差齐性：噪声同分布 $\sigma_1^2 = \ldots = \sigma_m^2 = \sigma^2$
  - 独立性：噪声之间相互独立

All models are wrong,
but some are useful.

分析方差

- 总离差平方和：$SS_T = \Sigma_{i,j}(X_{ij} - \bar{X}_{..})^2$
  - 因素A的效应平方和：$SS_A = b \cdot \Sigma_{i \le a}(\bar{X}_{i.} - \bar{X}_{..})^2$
  - 因素B的效应平方和：$SS_B = a \cdot \Sigma_{j \le b}(\bar{X}_{.j} - \bar{X}_{..})^2$
  - 误差平方和：$SS_E = \Sigma_{i,j}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$
- 直觉：因素各水平的效应都相同的话 $\alpha_i = 0$ 或 $\beta_i = 0$，效应平方和不会大
  - 如果因素A效应相同 $\alpha_i = 0$，$SS_A/\sigma^2 \sim \chi(a-1)$，否则 $\mathbb{E}[SS_A/(a-1)] > \sigma^2$
  - 如果因素B效应相同 $\beta_i = 0$，$SS_B/\sigma^2 \sim \chi(b-1)$，否则 $\mathbb{E}[SS_B/(b-1)] > \sigma^2$
- 无关因素，误差平方和 $SS_E/\sigma^2 \sim \chi^2((a-1)(b-1))$

假设检验

- 直觉：因素各水平的效应都相同的话 $\alpha_i = 0$ 或 $\beta_i = 0$，效应平方和不会大
  - 如果因素A效应相同 $\alpha_i = 0$，$SS_A/\sigma^2 \sim \chi(a-1)$，否则 $\mathbb{E}[SS_A/(a-1)] > \sigma^2$
  - 如果因素B效应相同 $\beta_i = 0$，$SS_B/\sigma^2 \sim \chi(b-1)$，否则 $\mathbb{E}[SS_B/(b-1)] > \sigma^2$
- 无关因素，误差平方和 $SS_E/\sigma^2 \sim \chi^2((a-1)(b-1))$
- 检验因素A：$F_A = \dfrac{SS_A/(a-1)}{SS_E/[(a-1)(b-1)]}$
- 检验因素B：$F_B = \dfrac{SS_B/(b-1)}{SS_E/[(a-1)(b-1)]}$

双因素方差分析表

- 直觉：因素各水平的效应都相同的话 $\alpha_i = 0$ 或 $\beta_i = 0$，效应平方和不会大
  - 如果因素A效应相同 $\alpha_i = 0$，$SS_A/\sigma^2 \sim \chi(a-1)$，否则 $\mathbb{E}[SS_A/(a-1)] > \sigma^2$
  - 如果因素B效应相同 $\beta_i = 0$，$SS_B/\sigma^2 \sim \chi(b-1)$，否则 $\mathbb{E}[SS_B/(b-1)] > \sigma^2$
- 无关因素，误差平方和 $SS_E/\sigma^2 \sim \chi^2((a-1)(b-1))$

方差来源	平方和	自由度	均方(MS)	F比率
因素A	SS(A)	a-1	SS(A) / (a-1)	MS(A) / MS(E)
因素B	SS(B)	b-1	SS(B) / (b-1)	MS(A) / MS(E)
误差	SS(E)	(a-1)(b-1)	SS(E) / ((a-1)(b-1))	
总和	SS (total)	ab-1		

双因素方差分析和交互作用（interaction）

two-way analysis of variance

- 两个因素的各个水平对实验结果的影响（是否一样？最好的？最差的？）

- 模型假设（a 和 b 个水平，每个水平组合 c 次试验）
  - 线性：$X_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}$
  - 正态性：每次试验噪声 $\epsilon_{ij} \sim N(0, \sigma_i^2)$
  - 方差齐次性：噪声同分布 $\sigma_1^2 = \ldots = \sigma_m^2 = \sigma^2$
  - 独立性：噪声之间相互独立

All models are wrong,
but some are useful.

分析方差

- 总离差平方和：$SS_T = \Sigma_{i,j,k}(X_{ijk} - \bar{X}_{...})^2$
  - 因素A的效应平方和：$SS_A = bc \cdot \Sigma_{i \le a}(\bar{X}_{i..} - \bar{X}_{...})^2$
  - 因素B的效应平方和：$SS_B = ac \cdot \Sigma_{j \le b}(\bar{X}_{.j.} - \bar{X}_{...})^2$
  - 交互作用的效应平方和：$SS_{AB} = c \cdot \Sigma_{i,j}(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2$
  - 误差平方和：$SS_E = \Sigma_{i,j,k}(X_{ijk} - \bar{X}_{ij.})^2$
- 直觉：如果交互作用不存在，$\bar{X}_{ij.} \approx \bar{X}_{i..} + \bar{X}_{.j.} - \bar{X}_{...}$ 且 $\gamma_{ij} \approx 0$，效应平方和不会大
  - 如果没有交互作用 $\gamma_{ij} = 0$，$SS_{AB}/\sigma^2 \sim \chi^2((a-1)(b-1))$，否则 $\mathbb{E}[SS_{AB}/(a-1)(b-1)] > \sigma^2$
- 无关因素，误差平方和 $SS_E/\sigma^2 \sim \chi^2(ab(c-1))$

假设检验

- 直觉：如果交互作用不存在 $\gamma_{ij} = 0$，效应平方和不会大
  - 如果没有交互作用 $\gamma_{ij} = 0$，$SS_{AB}/\sigma^2 \sim \chi^2((a-1)(b-1))$，否则 $\mathbb{E}[SS_{AB}/(a-1)(b-1)] > \sigma^2$
- 无关因素，误差平方和 $SS_E/\sigma^2 \sim \chi^2(ab(c-1))$

- 检验交互作用：

$$F_{AB} = \frac{SS_{AB}/[(a-1)(b-1)]}{SS_E/[ab(c-1)]} \sim F((a-1)(b-1), ab(c-1))$$

双因素方差分析表

- 直觉：如果交互作用不存在 $\gamma_{ij} = 0$，效应平方和不会大
  - 如果没有交互作用 $\gamma_{ij} = 0$，$SS_{AB}/\sigma^2 \sim \chi^2((a-1)(b-1))$，否则 $\mathbb{E}[SS_{AB}/(a-1)(b-1)] > \sigma^2$
- 无关因素，误差平方和 $SS_E/\sigma^2 \sim \chi^2(ab(c-1))$

方差来源	平方和	自由度	均方(MS)	F比率
因素A	SS(A)	a-1	SS(A) / (a-1)	MS(A) / MS(E)
因素B	SS(B)	b-1	SS(B) / (b-1)	MS(A) / MS(E)
交互作用	SS(AB)	(a-1)(b-1)	SS(AB) / ((a-1)(b-1))	MS(AB) / MS(E)
误差	SS(E)	ab(c-1)	SS(E) / (ab(c-1))	
总和	SS (total)	abc-1		

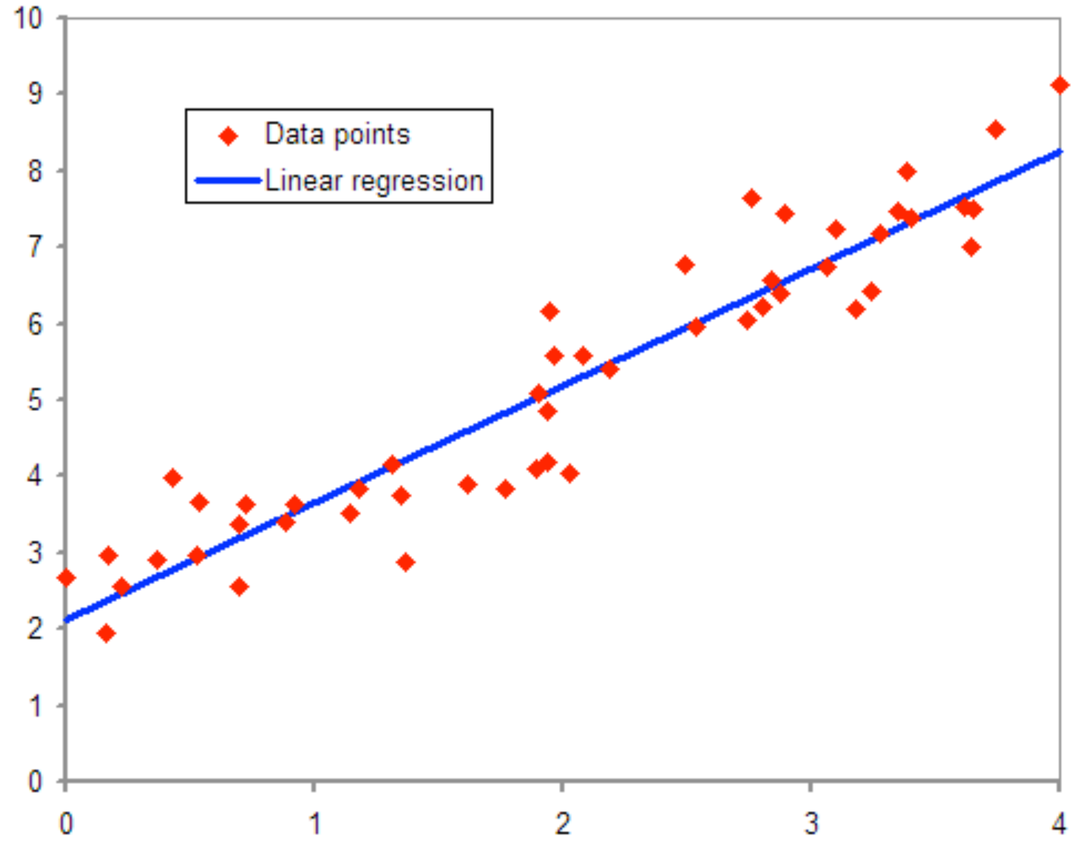
快捷计算*

- 总离差平方和：$SS_T = \Sigma_{i,j,k}(X_{ijk} - \bar{X}_{...})^2$
  - 因素A的效应平方和：$SS_A = bc \cdot \Sigma_{i \le a}(\bar{X}_{i..} - \bar{X}_{...})^2$
  - 因素B的效应平方和：$SS_B = ac \cdot \Sigma_{j \le b}(\bar{X}_{.j.} - \bar{X}_{...})^2$
  - 交互作用的效应平方和：$SS_{AB} = c \cdot \Sigma_{i,j}(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2$
  - 误差平方和：$SS_E = \Sigma_{i,j,k}(X_{ijk} - \bar{X}_{ij.})^2$

- 快捷计算：令 $T_{ij} = \Sigma_k X_{ijk}$, $T_{i.} = \Sigma_j T_{ij}$, $T_{.j} = \Sigma_i T_{ij}$, $T = \Sigma_{i,j} T_{ij}$
  - $SS_T = \Sigma_{i,j,k} X_{ijk}^2 - T^2/abc, \quad SS_E = SS_T - SS_A - SS_B - SS_{AB}$
  - $SS_A = \Sigma_{i,k} T_{i.}^2/ac - T^2/abc, \quad SS_B = \Sigma_{j,k} T_{.j}^2/bc - T^2/abc$
  - $SS_{AB} = \Sigma_{i,j} T_{ij}^2/c - T^2/abc - S_A - S_B$

回归

Regression



推断、预测

- 变量之间是有关联的
  - 父母的身高和子女的身高（Galton 1870s-1890s)
  - 人的身高与体重
  - 人的血压与年龄
  - 房价与房子位置、房屋面积

- 推断趋势：一些变量会随着另一些变量的变化如何变化？
- 预测：给定一个或者多个变量，另一个变量可能是?
  - 条件期望、条件概率
  - 是否存在简单的规律？
    - 假设、建模、拟合、检验

回归（regression）

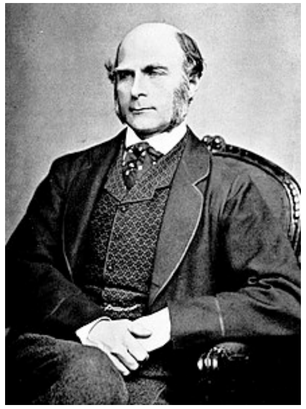
“回归到平均值”（regression toward the mean）

- 父母的身高和子女的身高（Galton 1870s-1890s)
  - 父母高，子女也会高
    - 但是不会太高
  - 父母矮，子女也会矮
    - 但是不会太矮

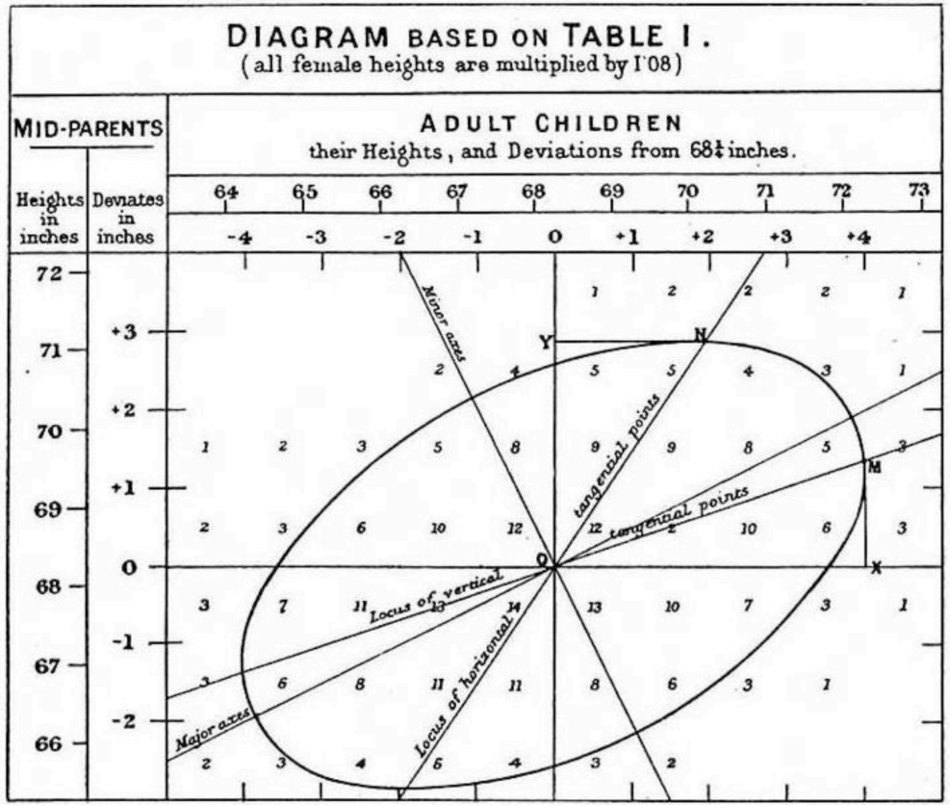
- 规律 + 误差

All models are wrong, but some are useful.

  - $Y = f(X) + \epsilon$
  - $\epsilon \sim N(0, \sigma^2)$



Francis Galton
（1822-1911)

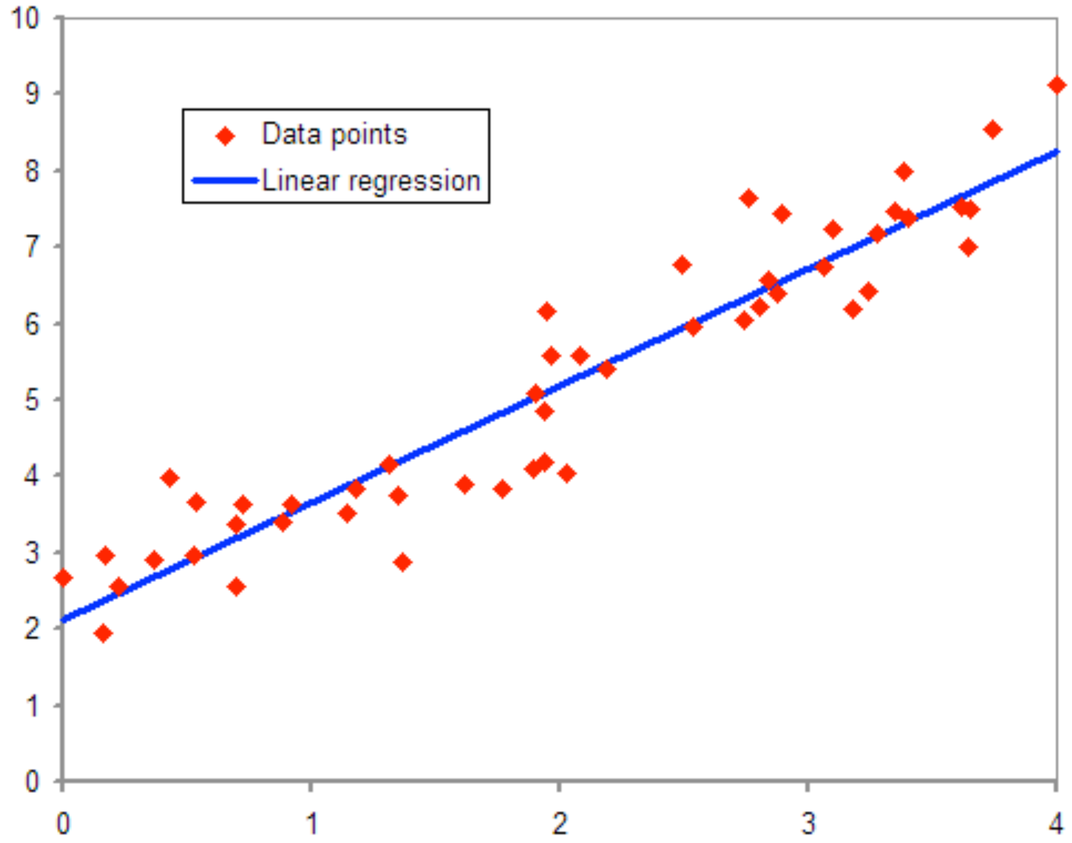


线性回归（linear regression）

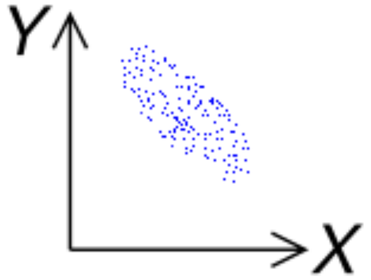
- 最简单最直接的规律
- 模型假设
  - 规律符合线性

All models are wrong,
but some are useful.

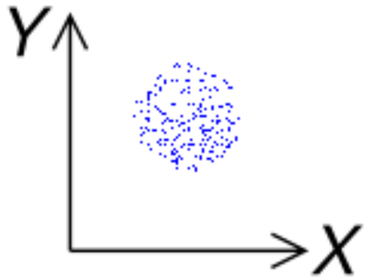
    - $Y_i = a + b \cdot X_i + \epsilon_i$
    - $Y_i = a + b \cdot X_i + c \cdot X_i^2 + \epsilon_i$
    - $Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + \epsilon$
  - 噪声正态性、齐性、独立性
    - $\epsilon_i \sim N(0,\sigma^2)$



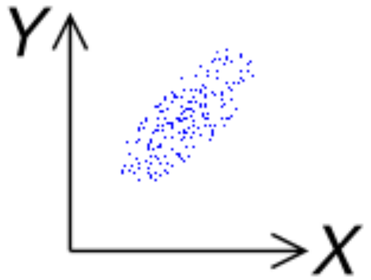
参数估计



- 简单线性回归：$Y_i = a + b \cdot X_i + \epsilon_i$。模型有多好？如何选择 a, b？
  - 最大似然估计（LME）：
    - $L(a, b, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_i (y_i - a - bx_i)^2 \right)$



    - 最小化 $\sum_i (y_i - a - bx_i)^2$：最小二乘(least squares)

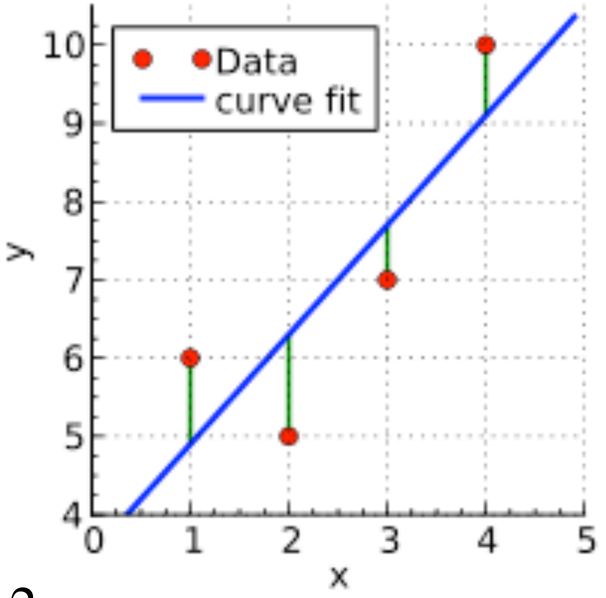


参数估计

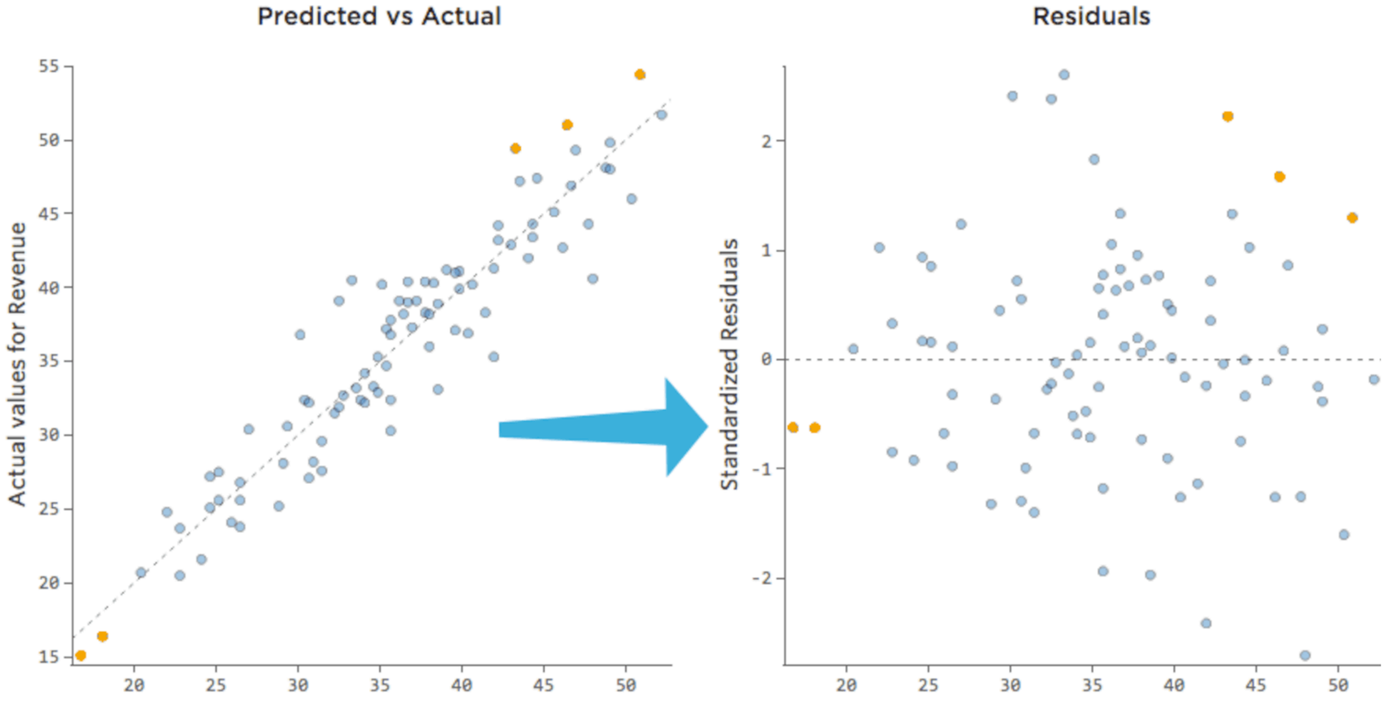
- 最小化 $\sum_i (y_i - a - bx_i)^2$：最小二乘(least squares)
- 计算:
  - 求驻点 $\begin{cases} 2\sum_i (y_i - a - bx_i) = 0 \\ 2\sum_i (y_i - a - bx_i)x_i = 0 \end{cases} \Longrightarrow \begin{cases} \sum_i y_i = na + b\sum_i x_i \\ \sum_i x_i y_i = a\sum_i x_i + b\sum_i x_i^2 \end{cases}$
  - 解方程组

参数估计

估计方差 σ^2

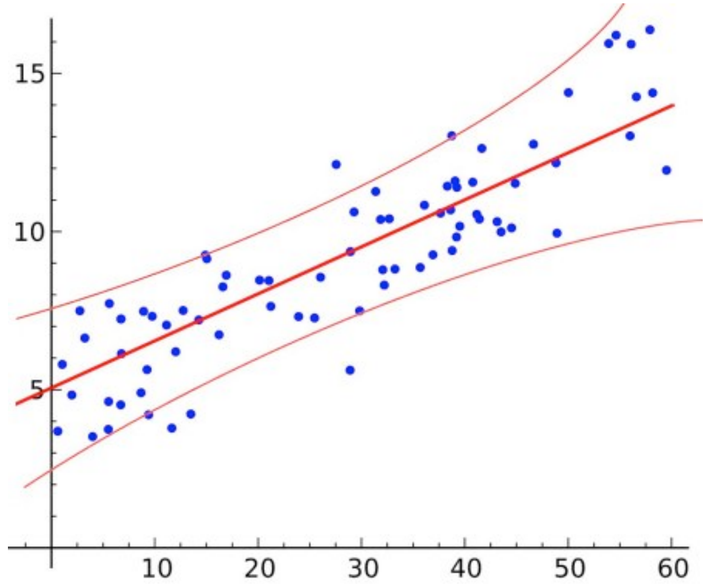


- 估计 $\hat{y}_i = \hat{a} + \hat{b}x_i$ 会有偏差
  - 残差（residual）：$\hat{y}_i - y_i$；残差平方和：$Q = \Sigma_i(\hat{y}_i - y_i)^2$
  - 残差图：$\hat{y}_i - y_i$ v.s. x_i



- $\frac{Q}{\sigma^2} \sim \chi^2(n-2)$
  - $\mathbb{E}\left[\frac{Q}{n-2}\right] = \sigma^2$

点预测和预测区间



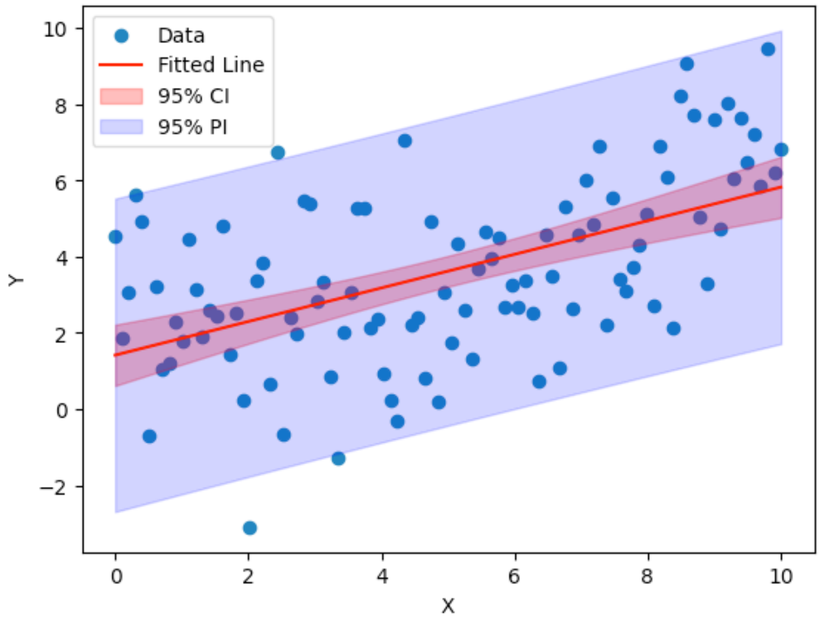
- 简单线性回归：$Y = a + b \cdot X + \epsilon$，其中 $\epsilon \sim N(0, \sigma^2)$
- 给定 X，预测 Y？
- 误差永远存在，精确估计不现实
  - 预测关于 $X_1, \ldots, X_n$ 的 $Y_1, \ldots, Y_n$ 的平均值，n 越大越精确（大数定理）
  - 预测区间：给定 X，预测 Y 可能的范围
- 控制问题：要求 Y 以 $1 - \alpha$ 概率处于范围 $[y_L, y_U]$，X 应该是多少?

其他话题

- 其实就不是线性关系？
  - 残差图： $\hat{y}_i - y_i$ v.s. x_i
  - 参数假设检验： $b = 0$？
  - 拟合优度：判定系数（coefficient of determination） $R \in [0,1]$
    - $SS_T = \Sigma_i(y_i - \bar{y})^2 = \Sigma_i(\hat{y}_i - \bar{y})^2 + \Sigma_i(y_i - \hat{y}_i)^2 = SS_R + SS_E$
    - $R^2 = SS_R/SS_T$
  - 线性回归的方差分析

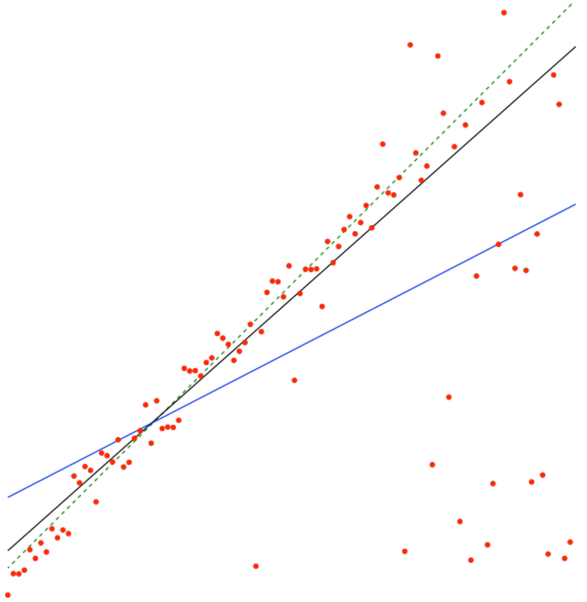
方差来源	平方和	自由度	均方(MS)	F比率
回归	SS (regression)	1	SS (regression)	$\frac{\text{MS(regression)}}{\text{MS(error)}}$
残差	SS (error)	n-2	SS(error) / (n-2)	
总和	SS (total)	n-1		

其他话题

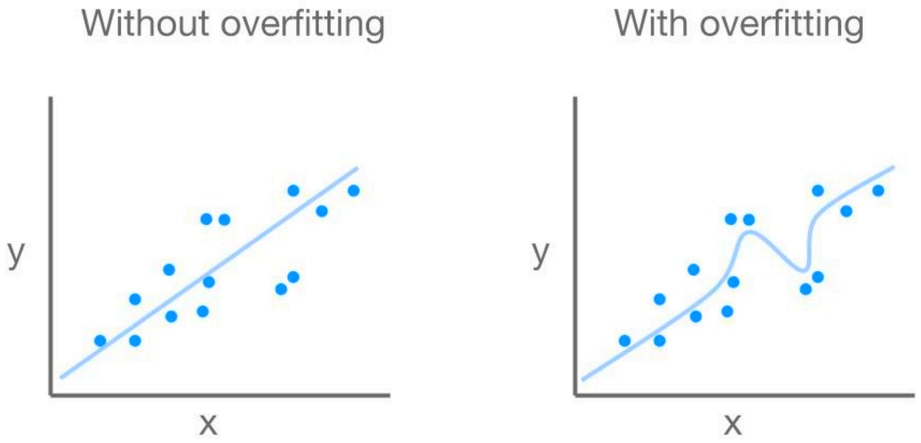


- 区间估计
- 某些模型可转化为线性回归
  - $Y = \alpha \cdot \exp(\beta x) \cdot \epsilon$
    - $\ln Y = \ln \alpha + \beta x + \ln \epsilon$
  - $Y = \alpha \cdot x^{\beta} \cdot \epsilon$
    - $\ln Y = \ln \alpha + \beta \ln x + \ln \epsilon$
  - $Y = \alpha + \beta \cdot h(x) + \epsilon$
    - $x \mapsto h(x)$
- 非线性回归
  - 对数线性回归、Logistic回归、支持向量回归、岭回归， 等等

其他话题



- 异常值（离群值，outlier）
  - 最小化 $\sum_i (y_i - a - bx_i)^2$
  - 少数离群值对均方误差贡献极大，普通最小二乘对此非常敏感
- 过拟合（overfitting）
  - 足够多的参数可以拟合任何数据，包括噪点
  - $y = a + bx + \epsilon$ v.s. $y = a + bx + cx^2 + dx^3 + \epsilon$



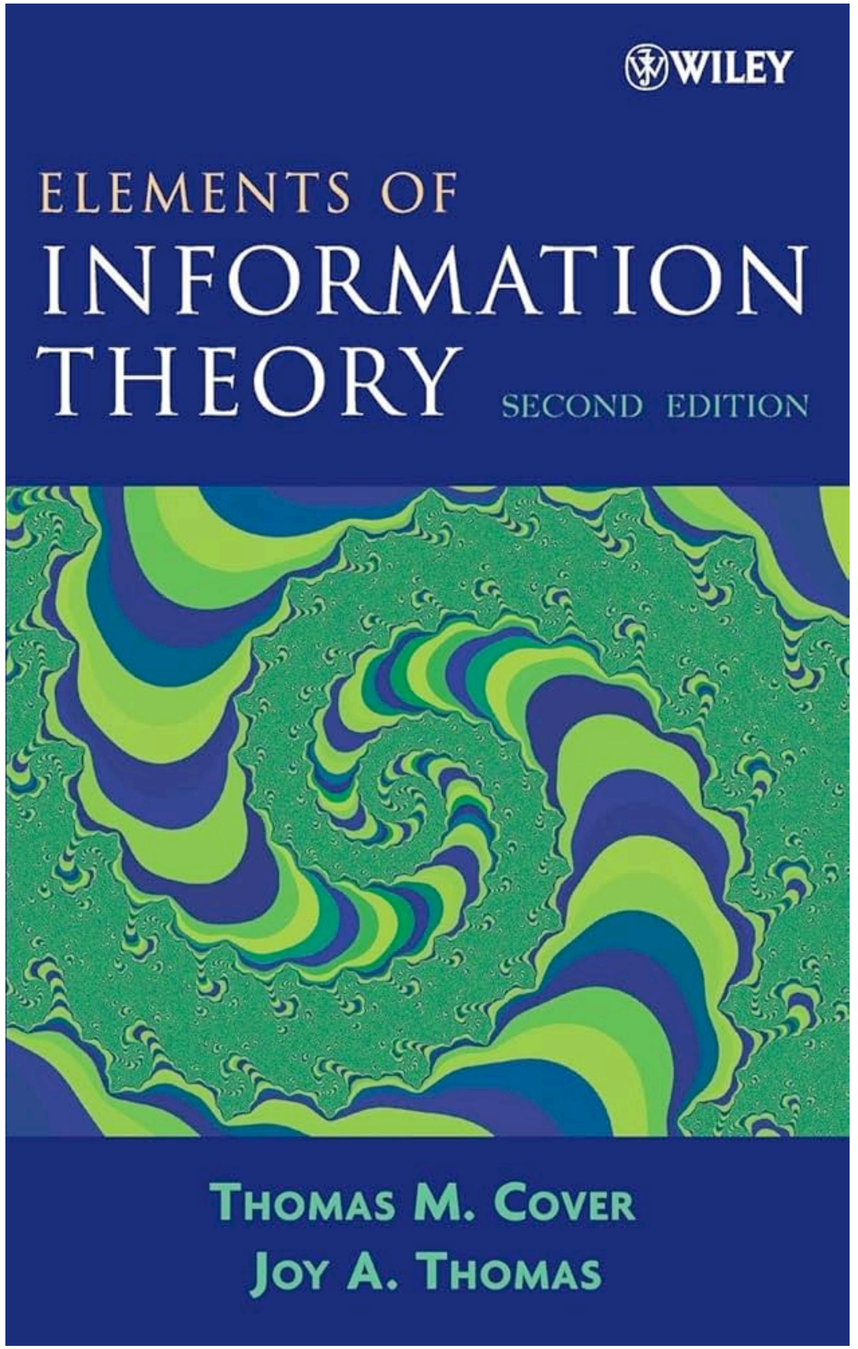
其他话题

多元线性回归

- $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$
- $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$
- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_\beta \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 \approx (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
  - 高维空间最优化问题，较为复杂 （**广告**：详情参见《高级算法》）

信息论初步

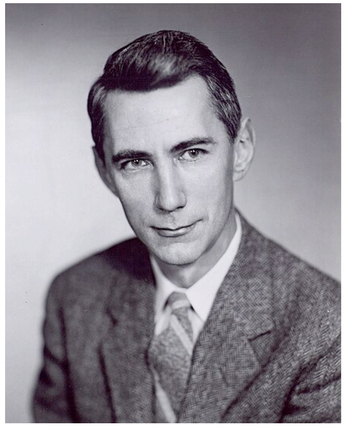
Information Theory



信息熵（Entropy）

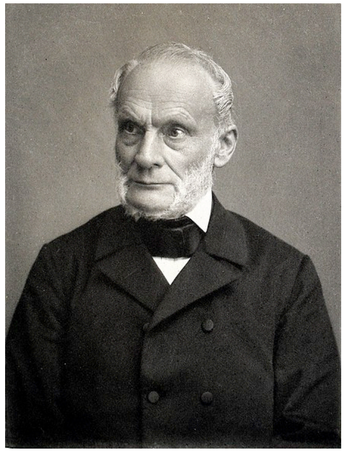
- 刻画随机变量“有多随机”

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$



Claude Shannon
(1916-2001)

- 度量单位 / 参照物：底数 2 。可以替换成任意参照物（e , 10 等等）



Rudolf Clausius
(1822-1888)

- 为什么这么定义？公理化*
  - 对称性：$H(p_1, p_2, \ldots) = H(p_2, p_1, \ldots)$
  - 最大熵：$H(p_1, p_2, \ldots, p_n) \le H(1/n, 1/n, \ldots, 1/n) = \log n$
  - 连续性
  - 递归性：$H(p_1, p_2, \ldots, p_n) = H(p_1 + p_2, \ldots, p_n) + (p_1 + p_2) H(p_1/(p_1 + p_2), p_2/(p_1 + p_2))$

信息熵（Entropy）

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

简单的性质

- 非负：对于任意随机变量 $H(X) \geq 0$
  - 证明：$\forall X, x, 0 \leq \Pr[X = x] \leq 1 \Rightarrow -\log(\Pr[X = x]) \geq 0$
- 可加：如果 X, Y 相互独立，那么联合分布 $H(X, Y) = H(X) + H(Y)$

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y) \log_2 p(x,y) = -\sum_{x,y} p(x)p(y)\left(\log_2 p(x) + \log_2 p(y)\right) \\ &= -\left(\sum_y p(y)\right) \sum_x p(x) \log_2 p(x) - \left(\sum_x p(x)\right) \sum_y p(y) \log_2 p(y) \\ &= -\sum_x p(x) \log_2 p(x) - \sum_y p(y) \log_2 p(y) = H(X) + H(Y) \end{aligned}$$

二项式系数

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

- 假设 $nq \in [0,n]$ 是整数，那么

$$\frac{2^{nH(q)}}{n+1} \leq \binom{n}{nq} \leq 2^{nH(q)}$$

- （上界）二项式定理：

$$1 = (q + (1-q))^n = \sum_{k=0}^{n} \binom{n}{k} q^k (1-q)^{n-k} \geq \binom{n}{nq} q^{qn} (1-q)^{(1-q)n}$$

$$\binom{n}{nq} \leq q^{-qn} (1-q)^{-(1-q)n} = 2^{-qn \log_2 q} 2^{-(1-q)n \log_2 (1-q)} = 2^{nH(q)}$$

二项式系数

- 假设 $nq \in [0,n]$ 是整数，那么

$$\frac{2^{nH(q)}}{n+1} \leq \binom{n}{nq} \leq 2^{nH(q)}$$

- （下界）：已知 $nq = \operatorname{argmin}_k \binom{n}{k} q^k (1-q)^{n-k}$

二项式定理 $\sum_{k=0}^{n} \binom{n}{k} q^k (1-q)^{n-k} = (q + (1-q))^n = 1$，最多 $n+1$ 项:

$$(n+1) \cdot \binom{n}{nq} q^{qn} (1-q)^{(1-q)n} \geq 1 \iff \binom{n}{nq} \geq \frac{q^{qn}(1-q)^{(1-q)n}}{n+1} \geq \frac{2^{nH(q)}}{n+1}$$

二项式系数

- 假设 $nq \in [0,n]$ 是整数，那么

$$\frac{2^{nH(q)}}{n+1} \leq \binom{n}{nq} \leq 2^{nH(q)}$$

- （下界）：目标 $q = \operatorname{argmin}_k \binom{n}{k} q^k (1-q)^{n-k}$

$$\binom{n}{k} q^k (1-q)^{n-k} - \binom{n}{k+1} q^{k+1} (1-q)^{n-k-1} = \binom{n}{k} q^k (1-q)^{n-k} \left(1 - \frac{n-k}{k+1} \frac{q}{1-q}\right)$$

非负，当 $1 - \frac{n-k}{k+1} \frac{q}{1-q} \geq 0 \iff k \geq qn - 1 + q$

压缩

- 压缩方案: $f:\Omega^n \to \{0,1\}^*$
- 一枚硬币正面概率 $p > 1/2$。对任意小的常数 $\delta > 0$, 当 n 足够大：
  - 存在一种压缩使用期望至多 $(1+\delta)nH(p)$ 个比特压缩长度 n 的抛硬币序列
  - 任意压缩长度 n 的抛硬币序列的压缩方案其期望至少是 $(1-\delta)nH(p)$个比特

压缩

- 压缩方案: $f: \Omega^n \to \{0,1\}^*$
- 一枚硬币正面概率 $p > 1/2$。对任意小的常数 $\delta > 0$, 当 n 足够大：
  - 存在一种压缩使用期望至多 $(1+\delta)nH(p)$ 个比特压缩长度 n 的抛硬币序列
- 令 $\epsilon > 0$ 是足够小的常数满足 $p - \epsilon > 1/2$。如果序列中有 $\leq n(p-\epsilon)$ 正面, 编码第一位为1、使用 $n+1$ 比特编码；否则第一位为0、对每一种情况使用一种不同的编码（如, m 种序列 $\lceil \log_2 m \rceil$ 比特）

压缩

- 一枚硬币正面概率 $p > 1/2$。对任意小的常数 $\delta > 0$, 当 n 足够大：
  - 存在一种压缩使用期望至多 $(1+\delta)nH(p)$ 个比特压缩长度 n 的抛硬币序列
- 令 $\epsilon > 0$ 是足够小的常数满足 $p - \epsilon > 1/2$。如果序列中有 $\leq n(p-\epsilon)$ 正面, 编码第一位为1、使用 $n+1$ 比特编码
  - (Chernoff) 第一种情况的概率 $\leq \exp(-n\epsilon^2/2p)$, 全期望 $(n+1)e^{-n\epsilon^2/2p}$

压缩

$$\binom{n}{nq} \leq 2^{nH(q)}$$

- 一枚硬币正面概率 $p > 1/2$。对任意小的常数 $\delta > 0$, 当 n 足够大：
  - 存在一种压缩使用期望至多 $(1+\delta)nH(p)$ 个比特压缩长度 n 的抛硬币序列

- 令 $\epsilon > 0$ 是足够小的常数满足 $p - \epsilon > 1/2$。如果序列中有 $> n(p-\epsilon)$ 正面, 第一位为0、对每一种情况使用一种不同的编码
  - 第二种情况数目 $\sum_{i=n(p-\epsilon)}^{n} \binom{n}{i} \leq \sum_{i=n(p-\epsilon)}^{n} \binom{n}{n(p-\epsilon)} \leq \frac{n}{2} 2^{nH(p-\epsilon)},$

    使用比特 $1 + \lceil \log_2 \frac{n}{2} 2^{nH(p-\epsilon)} \rceil \leq nH(p-\epsilon) + \log_2 n + 2$

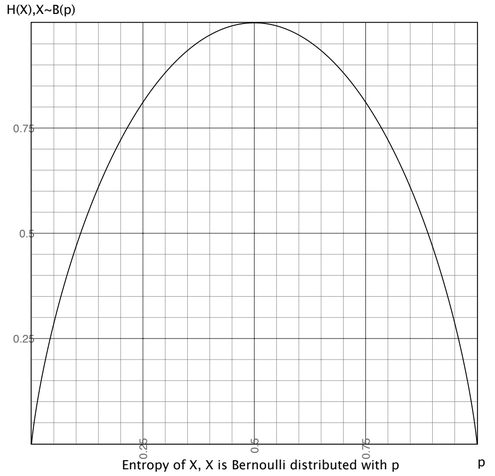
压缩

- 一枚硬币正面概率 $p > 1/2$。对任意小的常数 $\delta > 0$, 当 n 足够大：
  - 存在一种压缩使用期望至多 $(1+\delta)nH(p)$ 个比特压缩长度 n 的抛硬币序列

- 令 $\epsilon > 0$ 是足够小的常数满足 $p - \epsilon > 1/2$。如果序列中有 $\leq n(p-\epsilon)$ 正面，编码第一位为1、使用 $n+1$ 比特编码；否则第一位为0、对每一种情况使用一种不同的编码

$$e^{-n\epsilon^2/2p}(n+1) + (1 - e^{-n\epsilon^2/2p})(nH(p-\epsilon) + \log_2 n + 2)$$

$o(1)$ $\quad$ $1 - o(1)$ $\quad$ $o(n)$



压缩

- 一枚硬币正面概率 $p > 1/2$。对任意小的常数 $\delta > 0$，当 n 足够大：
  - 任意压缩长度 n 的抛硬币序列的压缩方案其期望至少是 $(1-\delta)nH(p)$个比特
- 若序列 S_1 比序列 S_2 出现概率更高，那么最优压缩方案满足 $|f(S_1)| \leq |f(S_2)|$
  - 包含 $< n(p+\epsilon)$个正面的序列至少使用跟 $n(p+\epsilon)$个正面的序列同样多的比特
- 对于任意大小为 s 的序列的集合，一定有一个序列 S 满足 $|f(S)| \geq \log_2 s - 1$
  - 包含 $n(p+\epsilon)$个正面的序列需要$\log_2 \binom{n}{n(p+\epsilon)} - 1 \geq \log_2 \frac{2^{nH((p+\epsilon))}}{n+1} - 1$
- $(1-\exp) \cdot (nH(p+\epsilon) - \log_2(n+1) - 1) \geq (1-\delta)nH(p)$

压缩

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

香浓信源编码定理*（source coding theorem）

- 压缩方案：前缀码 $f : \Omega^n \to \{0,1\}^*$
- 对于任意在 Ω 上的概率分布 D，对于足够大的 n：
  - 任意压缩方案都不可能使用期望低于 $nH(D)$ 个比特压缩 $(X_1, \ldots, X_n) \sim D$
  - 存在一种压缩方案使用期望 $< nH(D) + 1$ 个比特压缩 $(X_1, \ldots, X_n) \sim D$
- 对 $x \in \Omega$，使用长度为 $\lceil -\log_2 D(x) \rceil \leq -\log_2 D(x) + 1$ 个比特的编码
- 期望长度 $\leq \sum_x D(x)(-\log_2 D(x) + 1) = H(X) + 1$