Model Selection

TOC

基础知识

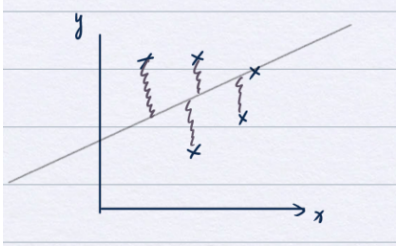
解释变量个数 k 对模型的影响（基于回归模型 $Y = X\beta + \varepsilon$）

- 回归系数估计：$b = (X^T X)^{-1} X^T Y$
- 估计量的协方差迹：$\mathrm{tr}(\mathrm{Cov}(b)) = \mathrm{tr}\left(\sigma^2 (X^T X)^{-1}\right)$
  当解释变量个数 $k \uparrow$ 时，$\mathrm{tr}(\mathrm{Cov}(b)) \uparrow$（即回归系数估计的整体方差增大）。
- 残差均方：$s^2 = \frac{\mathrm{SSE}}{n-k-1}$，当 $k \uparrow$ 时，用于估计 σ^2 的自由度 $n-k-1 \downarrow$，导致 s^2 对 σ^2 的估计精度下降。

两类拟合错误

欠拟合（under fitting）

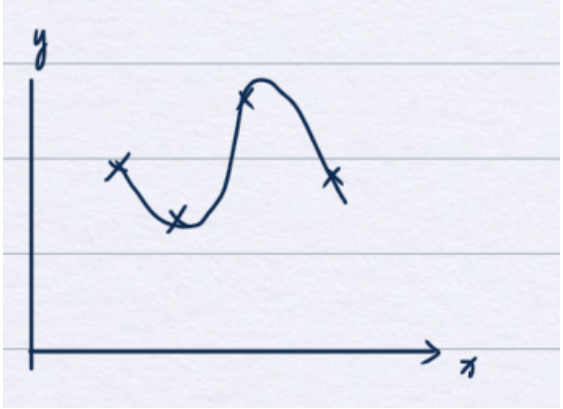
- 表现：模型过于简单，无法捕捉数据中的复杂模式。



- 图形特征（右侧示意图）：数据点围绕拟合线的波动大，拟合线无法贴合数据趋势。
- 偏差-方差特征：
  - 偏差（bias）大：模型对真实关系的近似误差大。
  - 方差（variance）小：更换数据点后，拟合线的变化幅度小（模型稳定性高，但拟合能力不足）。
- 本质原因：在更复杂的真实环境下，使用了过于简单的回归模型。

过拟合（over fitting）

- 表现：模型过于复杂，过度拟合了训练数据中的噪声，泛化能力差。



- 图形特征（右侧示意图）：拟合曲线极度贴合训练数据点，呈现不规则的波动。
- 偏差-方差特征：
  - 偏差（bias）小：模型对训练数据的拟合误差小。
  - 方差（variance）大：更换数据点后，拟合曲线的变化幅度大（模型稳定性差，泛化能力弱）。

Model Selection Principle（模型选择准则与步骤）

我们将从四个层面叙述如何选择模型

1. **判定系数**（R^2）
2. **调整后判定系数**（R_a^2）
3. C_p **准则**
4. **AIC（Akaike Information Criterion，赤池信息准则）**

R^2 选择

选择标准

R^2 衡量模型对因变量波动的解释程度，定义为

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

- R^2 **越大越好**：表示模型解释力越强。

  当 $R^2 = 1$ 时模型完全拟合；当 $R^2 = 0$ 时模型无解释力。图像表现为 R^2 随解释变量数量增加而递增，趋近于 1。

- **但不能无限制追求更大的** R^2：因为 R^2 随解释变量数量的增加永不下降，容易导致选择过拟合的模型。

- 因此 R^2 **适合用于单步比较（同一维度模型的比较）**，但不适合用于不同复杂度模型的最终决策。

选择方式

考虑三个解释变量 x_1, x_2, x_3，模型形式为

$$y = \beta_0 + \beta_1 x_i + \varepsilon, \quad i = 1, 2, 3,$$

设计矩阵为

$$X = (\mathbf{1}_n \mid x_i).$$

步骤回归的核心逻辑：
从包含最少变量的子集（如单变量）开始，逐步添加或删除变量，根据不同准则（如 R^2、C_p、AIC 等）在“拟合度”和“复杂度”之间实现平衡。

当 $R^2 = 1$ 时模型完全拟合；当 $R^2 = 0$ 时模型无解释力。图像表现为 R^2 随解释变量数量增加而递增，趋近于 1。

回归模型形式：

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, 3,$$

设计矩阵：

$$X = (\mathbf{1}_n, x_i).$$

回归平方和公式：

$$SSR = Y^T \left(H - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) Y.$$

其中 $H = X(X^T X)^{-1} X^T$。

单变量 x 得到 SSR：

- $x_1 = 0.3$
- $x_2 = 0.5$
- $x_3 = 0.6$

组合对应拟合优度：

- (x_1, x_3)：$R^2 = 0.9$
- (x_2, x_3)：$R^2 = 0.6$
- (x_1, x_2, x_3)：$R^2 = 0.95$

变量引入顺序：
x_3（$R^2 = 0.6$）$\rightarrow (x_1, x_3)$（$R^2 = 0.9$）$\rightarrow (x_1, x_2, x_3)$（$R^2 = 0.95$）

步骤回归（向前逐步选择）：从最小变量集开始，每次引入能最大提升拟合优度的变量。

R_a^2 选择

公式与关系

- 调整判定系数公式：

$$R_a^2 = 1 - \frac{MSE}{MST}$$

  其中 MSE 是均方误差，MST 是总均方。

- 与普通判定系数 R^2 的关系：

$$R_a^2 \leq R^2$$

R_a^2 对 R^2 进行了复杂度惩罚，解决了 R^2 **随解释变量数量增加而单调上升** 的问题，使其能够更客观地评价模型质量。

变化趋势

调整判定系数 R_a^2 随解释变量个数 k 的变化通常表现为：

- 当加入的新变量能显著提高拟合效果时，R_a^2 **上升**；
- 当加入的新变量并不能有效改善拟合（或只是噪声变量）时，R_a^2 **下降。**

因此 R_a^2 的趋势通常为：
先上升，达到峰值，再下降。

这体现出一个核心思想：

> **解释变量并非越多越好，存在使 R_a^2 最大的最优变量子集。**

应用

在变量子集选择（如前向逐步回归、后向逐步回归、最佳子集回归）中，常使用：

> **选择使 R_a^2 达到最大值的变量组合**

以实现以下平衡：

- **拟合效果**（模型的解释力）
- **模型复杂度**（变量数量、过拟合风险）

最终得到既简洁又有效的最优回归模型。

C_p 选择

C_p 公式及相关定义

- C_p 公式：

$$C_p \triangleq \frac{SSE(p)}{s^2} - [\,n - 2(p+1)\,]$$

- 变量子集：$X_p \triangleq \{x_1, x_2, \ldots, x_p\}$
- s^2：全模型下的均方误差（MSE）

- 回归模型：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$，对应残差平方和 $SSE(p)$
- 设计矩阵与帽子矩阵：
  $X_p = (\mathbf{1}_n, x_1, \ldots, x_p)$，

$$H_p = X_p (X_p^T X_p)^{-1} X_p^T$$

因此

$$SSE(p) = Y^T (I - H_p) Y$$

some derivation

若模型正确，$SSE(p) \sim \sigma^2 \chi^2(n-p-1)$，此时

$$C_p \sim \chi^2(n-p-1) - n + 2(p+1)$$

其 **均值为** $p+1$。

模型拟合良好时，$C_p \to p+1$；预测值均方误差 $MSE(\hat{Y}) \to \hat{\sigma}^2 \to 0$。

均方误差分解：

$$E(\theta - \hat{\theta})^2 = \text{bias}^2 + \text{Var}(\hat{\theta})$$

$$\begin{aligned} C_p &= \frac{SSE(p)}{s^2} - [\, n - 2(p+1) \,] \\ &\underline{E} \ \frac{E[SSE(p)]}{\sigma^2} - [\, n - 2(p+1) \,] \\ &= \frac{MSE(\hat{Y}_p)}{\sigma^2} \end{aligned}$$

即 C_p 本质为 **预测均方误差与噪声方差的比值。**

$E[SSE(p)]$ 与 $MSE(\hat{Y}_p)$ 的联系推导

1. 计算 $E[SSE(p)]$

$$
\begin{aligned}
E[SSE(p)] &= E[Y^T(I - H_p)Y] \\
&= \mathrm{tr}((I - H_p)E[YY^T])
\end{aligned}
$$

2. 分解 $E[YY^T]$

设 $Y = \mu + \varepsilon$，$\varepsilon \sim N(0, \sigma^2 I_n)$

$$
E[YY^T] = \mu\mu^T + \sigma^2 I_n
$$

代入得到：

$$
\mu^T(I - H_p)\mu + \sigma^2(n - p - 1)
$$

3. 计算 $MSE(\hat{Y}_p)$

$$
MSE(\hat{Y}_p) = E[\|\hat{Y}_p - \mu\|_2^2]
$$

因为 $\hat{Y}_p = H_p Y$，$Y = X_p\beta_p + \varepsilon$：

$$
MSE(\hat{Y}_p) = \mu^T(I - H_p)\mu + \sigma^2(p + 1)
$$

$$
MSE(\hat{Y}_p) - \sigma^2(p + 1) + \sigma^2(n - p - 1) = E[SSE(p)]
$$

即：

$$
\frac{MSE(\hat{Y}_p)}{\sigma^2} + (n - 2p - 2) = \frac{E[SSE(p)]}{\sigma^2}
$$

因此验证：

$$
C_p = \frac{MSE(\hat{Y}_p)}{\sigma^2}
$$

$MSE(\hat{Y}_p)$ 的下界

偏差-方差分解：

$$
\begin{aligned}
MSE(\hat{Y}_p) &= E\|\hat{Y}_p - \mu\|^2 \\
&= E\|\hat{Y}_p - E\hat{Y}_p + E\hat{Y}_p - \mu\|^2 \quad (\mu \text{ 为真实均值，} \hat{Y}_p \text{ 为预测值}) \\
&= E\|\hat{Y}_p - E\hat{Y}_p\|^2 + E\|E\hat{Y}_p - \mu\|^2 + 2E(\hat{Y}_p - E\hat{Y}_p)^T(E\hat{Y}_p - \mu) \\
&\quad \text{其中 } E\|E\hat{Y}_p - \mu\|^2 \text{ 为 bias}^2\text{（偏差平方和），最后一项交叉项期望为 0} \\
&= E\mathrm{tr}\left((\hat{Y}_p - E\hat{Y}_p)(\hat{Y}_p - E\hat{Y}_p)^T\right) + \mathrm{bias}^2 \\
&= \mathrm{tr}\left(\mathrm{Cov}(\hat{Y}_p)\right) + \mathrm{bias}^2 \\
&\quad \text{又 } \mathrm{Cov}(\hat{Y}_p) = H_p\sigma^2 I_n H_p^T = \sigma^2 H_p \text{ (H_p 为帽子矩阵)} \\
&= \sigma^2 \mathrm{tr}(H_p) + \mathrm{bias}^2 \\
&= (p+1)\sigma^2 + \mathrm{bias}^2 \quad (\text{因 } H_p \text{ 秩为 } p+1)
\end{aligned}
$$

因此：

$$
MSE(\hat{Y}_p) \geq (p+1)\sigma^2
$$

当且仅当 **偏差为 0（即模型正确）** 时，等号成立，此时 $MSE(\hat{Y}_p)$ 的最小值为 $(p+1)\sigma^2$。

$$
MSE(\hat{Y}_p) = \mathrm{tr}(\mathrm{Cov}(\hat{Y}_p)) + \|E[\hat{Y}_p] - \mu\|_2^2
$$

因为 $\mathrm{Cov}(\hat{Y}_p) = \sigma^2 H_p$，且 $\mathrm{tr}(H_p) = p+1$：

$$
\mathrm{tr}(\mathrm{Cov}(\hat{Y}_p)) = (p+1)\sigma^2
$$

因此：

$$
MSE(\hat{Y}_p) \geq (p+1)\sigma^2
$$

当且仅当模型无偏时等号成立。

综上，C_p 本质上用

$$
C_p \approx p+1
$$

来判断模型是否既不过度拟合也不过度简化。

信息准则AIC的推导与解释

AIC的定义与模型设定

- **AIC公式**：$\mathrm{AIC} \triangleq \max \mathrm{loglikelihood}(k) - k$（$k$ 为模型中待估参数的数量）
- **数据假设**：$\{X_T\}$ 为独立同分布（i.i.d）序列，模型待选变量数为 k
- **变量子集与回归模型**：
  变量子集 $\hat{X}_p = \{x_1, \ldots, x_p\}$，回归模型形式为 $Y = X_p\beta_p + \varepsilon$，其中设计矩阵 $X_p = [\mathbf{1}_n : \hat{X}_p]$（$\mathbf{1}_n$ 为全1列向量）
  假设 $Y \sim N(X_p\beta_p, \sigma^2 I_n)$

似然函数与对数似然

- **概率密度函数（pdf）**：

$$f(Y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}\|Y - X_p\beta_p\|_2^2 \right\}$$

- **对数似然函数**：

$$\mathrm{loglikelihood} = -\frac{1}{2\sigma^2}\|Y - X_p\beta_p\|_2^2 - \frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2$$

- 回归系数的MLE：$\hat{\beta}_p = (X_p^T X_p)^{-1} X_p^T Y$
- 方差的MLE：$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\|Y - X_p\hat{\beta}_p\|_2^2 = \frac{SSE(p)}{n}$

将MLE代入对数似然，整理得：

$$\max \mathrm{loglikelihood} = -\frac{n}{2} - \frac{n}{2}\log 2\pi - \frac{n}{2}\log SSE(p) + \frac{n}{2}\log n$$

根据AIC定义 $\mathrm{AIC} \triangleq \max \mathrm{loglikelihood} - k$（此处 $k = p + 1$，即 β_p 的维度 + 方差参数），代入并化简（消去与变量子集无关的常数项后）：

$$\mathrm{AIC} \triangleq -\frac{n}{2}\log SSE(p) - p$$

（核心是“对数残差平方和 + 参数数量的惩罚项”）

AIC的应用与局限

- **应用**：寻找使AIC最大的变量子集 X_p，以此平衡模型拟合度与复杂度。

- **局限**：若变量数 k 较大，遍历所有子集（共 2^k 种可能）的计算量大，实际中需结合逐步选择、启发式算法等减少计算负担。

Selection Theory

后退逐步回归（变量从多到少剔除）Backward

基本设定及实现步骤

- 初始变量集：$X = \{X_1, \ldots, X_k\}$（所有可能的随机变量）

步骤①：全变量模型检验

- 全模型（Full Model）：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$，残差平方和 $SSE_F = Y^T(I - H)Y$（H 为帽子矩阵，变量个数 k）
- 原假设：对每个 $i = 1 \to k$，$H_0 : \beta_i = 0$
- 简化模型（Reduced Model）：剔除第 i 个变量后的模型，变量个数 $k-1$，残差平方和记为 $SSE_R^{(i)}$
- F 检验公式：

$$F_i^{(1)} = \frac{SSE_R^{(i)} - SSE_F}{SSE_F/(n-k-1)} \sim F_{1,n-k-1}$$

- 筛选规则：计算 k 个 F 值，找到最小的 $F_t^{(1)}$。若 $F_t^{(1)} \leq F_{1,n-k-1}(\alpha)$，则剔除对应变量。

步骤②：剔除一个变量后的模型检验

- 全模型：前 $k-1$ 个变量回归，变量个数 $k-1$
- 原假设：对每个 $i = 1 \to k-1$，$H_0 : \beta_i = 0$
- 简化模型：剔除第 i 个变量后的模型，变量个数 $k-2$，残差平方和记为 $SSE_R^{(i)}$
- F 检验公式：

$$F_i^{(2)} = \frac{SSE_R^{(i)} - SSE_F}{SSE_F/(n-k-2)} \sim F_{1,n-k-2}$$

- 筛选规则：计算 $k-1$ 个 F 值，找到最小的 $F_t^{(2)}$。若 $F_t^{(2)} \leq F_{1,n-k-2}(\alpha)$，剔除对应变量。

步骤③：递归终止条件

重复步骤②，直至某一步计算的 $F_t > F_{1,n-t-1}(\alpha)$，停止剔除变量，得到最优回归模型。

向前逐步回归（变量从少到多引入）

基本设定及实现步骤

初始变量集：$X = \{X_1, \ldots, X_k\}$（所有可能的随机变量）

步骤①：单变量模型筛选

- 待选单变量模型：$y = \beta_0 + \beta_i x_i + \varepsilon$，$i = 1, 2, \ldots, k$
- 全模型（Full Model）：包含某变量 x_T 的模型，残差平方和 $SSE_F^{(1)}$（变量个数 = 1）
- 原假设与简化模型：$H_0: \beta_i = 0$，简化模型 $y = \beta_0 + \varepsilon$，残差平方和 SSE_R
- F 检验公式：

$$F_i^{(1)} = \frac{SSE_R - SSE_F^{(1)}}{SSE_F^{(1)}/(n-2)} \sim F_{1,n-2}$$

- 筛选规则：找到最大 F 值 $F_t^{(1)}$，若 $F_t^{(1)} > F_{1,n-2,\alpha}$，将对应变量引入模型。

步骤②：双变量模型筛选（在已引入变量基础上引入第二个变量）

- 待选双变量模型：$y = \beta_0 + \beta_1 x_1 + \beta_i x_i + \varepsilon$，$i = 2, \ldots, k$
- 全模型残差平方和 $SSE_F^{(2)}$（变量个数 = 2）
- 简化模型：$y = \beta_0 + \beta_1 x_1 + \varepsilon$，残差平方和 SSE_R
- F 检验公式：

$$F_i^{(2)} = \frac{SSE_R - SSE_F^{(2)}}{SSE_F^{(2)}/(n-3)} \sim F_{1,n-3}$$

- 筛选规则：找到最大 F 值，若大于临界值，将对应变量引入模型。

后续步骤

重复上述过程，每次在已选变量基础上引入使 F 值最大且显著的新变量，直至无变量能通过检验，得到最优变量子集。

逐步回归方法的缺陷与改进(stepwise)

方法	缺陷说明
后退法	计算量大，变量一旦被剔除就不再考虑重新引入，可能丢失潜在贡献变量
前进法	未考虑解释变量间相互作用，仅单独评估变量对模型的贡献，可能引入冗余或忽略变量间协同效应

- Stepwise 是对前进法的优化：在前进法引入变量后，通过类似后退法的逻辑检查已引入变量，若某变量不再显著则剔除。
- 平衡变量引入与剔除，考虑变量单独贡献及相互影响，提升模型变量选择合理性。