诊断检验（Diagnostic Checking）

TOC

- 自相关的修正方法（以AR(1)为例）
  - 情形a：ρ 已知——迭代法（Cochrane-Orcutt 方法）
- ρ 未知时的迭代估计方法（可行广义最小二乘思路）
  - Step 1：初始残差与 ρ 的估计
  - Step 2：加权最小二乘修正模型
  - Step 3：迭代优化
- 特殊情形：$\rho \approx 1$（差分法）
  - ρ未知时的自相关迭代估计（情形d）
  - 目标函数与参数关系
  - ρ的区间搜索与迭代

- 违反线性假设的修正：Box-Cox变换
  - Box-Cox变换的定义
  - 变换后的模型与似然函数
  - 确定最优λ

- Deny Assumption：解释变量矩阵列秩不足，线性相关（or多重共线性）
  - 多重共线性（multicollinearity）的病态性：基于瑞利商的严谨解释
    - 瑞利商与矩阵特征值的关系
    - 多重共线性下的病态性推导
  - 多重共线性的判断方法整理
    - 基于矩阵特征值与行列式的判断
  - VIF（方差膨胀因子）
    - VIF的计算逻辑
    - VIF的判断标准
    - VIF等价于方差比较：方差比较因子
    - 回归系数的方差与 VIF 的联系

前提模型假设

在线性回归模型 $Y = X\beta + \varepsilon$ 中，需要对以下假设进行诊断检验，以验证模型设定与估计结果的可靠性。

误差项的分布假设

假设：$\varepsilon \sim N(0_n, \sigma^2 I_n)$

该假设包含以下三个核心条件：

- **同方差性**：
  $\mathrm{var}(\varepsilon_i) = \sigma^2$
  即每个误差项的方差相等，残差的波动幅度应随样本点一致。
- **无自相关性**：
  $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0,\ \forall i \neq j$
  不同误差项之间应相互独立，无系统性相关。
- **正态性**：
  误差项服从正态分布，便于后续的显著性检验与置信区间推断。

线性假设

要求解释变量与被解释变量之间的关系可被线性表达：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

若真实关系为非线性而误设为线性，则会造成模型设定错误（Specification Error），导致估计偏差。

解释变量矩阵的假设

解释变量矩阵 X 为 $n \times (k+1)$ 维（其中 n 为样本量，k 为解释变量个数），需满足以下条件：

- **样本量充足**：
  $(k+1) \leq n$，否则无法进行参数估计。
- **矩阵可逆性与稳定性**：
  $(X^T X)^{-1}$ 必须存在，以确保回归系数估计量

  $$b = (X^T X)^{-1} X^T Y$$

  可计算且数值稳定。
- **无完全多重共线性**：
  各解释变量之间不能存在线性依赖，否则 $X^T X$ 不可逆，模型估计失效。

假设的验证推导

为验证线性回归模型假设是否成立，我们基于残差 e 进行推导：

残差的定义

残差 $e = Y - \hat{Y} = Y - HY = (I - H)Y$，其中 H 为帽子矩阵，且 $Y = X\beta + \varepsilon$。

1. 残差的分布

残差 e 服从正态分布：

$$e \sim N\left(0, (I-H)\sigma^2 I_n (I-H)\right) \implies e \sim N\left(0, (I-H)\sigma^2\right)$$

2. 残差与拟合值的协方差

计算残差 e 与拟合值 $\hat{Y}$（即 HY）的协方差：

$$\begin{aligned}
\operatorname{cov}(e, \hat{Y}) &= \operatorname{cov}((I-H)Y, HY) \\
&= (I-H)\operatorname{cov}(Y)H^T \\
&= \sigma^2 (I-H)H^T \\
&= \sigma^2 (H - H^2) \\
&= 0 \quad (\text{因 } H \text{ 是对称且幂等矩阵，} \ H^2 = H)
\end{aligned}$$

故残差与拟合值不相关。

3. 残差的和为 0

残差向量与全 1 向量 $\mathbf{1}_n$ 的内积为 0：

$$\mathbf{1}_n^T e = \mathbf{1}_n^T (I-H)Y = 0$$

这是因为 $(I-H)$ 与 X 正交，而 X 的第一列通常为全 1 向量 $\mathbf{1}_n$，即 $(I-H)\mathbf{1}_n = 0$，从而 $\mathbf{1}_n^T e = 0$。

2. 违反假设(1)的推导：异方差情形下的估计

$\{\sigma_i^2\}$ 已知

模型设定

考虑线性回归模型 $Y = X\beta + \varepsilon$，其中：

- $E(\varepsilon_i) = 0$；
- 异方差：$\mathrm{var}(\varepsilon_i) = \sigma_i^2$；
- 无自相关：$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0\ (i \neq j)$。

记 $W = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$，则 $\mathrm{cov}(\varepsilon) = W^2$。

此时最小二乘估计（LSE）

最小二乘估计为

$$\hat{\beta}_{\mathrm{LSE}} = (X^T X)^{-1} X^T Y.$$

其性质：

- **无偏性**：

$$E(\hat{\beta}_{\mathrm{LSE}}) = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T X\beta = \beta.$$

- **不再是 BLUE**：
  在存在异方差时，LSE 虽然仍然无偏，但不再是方差最小的线性无偏估计（即不再满足 BLUE）。
- **协方差矩阵**：

$$\mathrm{cov}(\hat{\beta}_{\mathrm{LSE}}) = (X^T X)^{-1} X^T \mathrm{cov}(Y) X (X^T X)^{-1} = (X^T X)^{-1} X^T W^2 X (X^T X)^{-1}.$$

此时加权最小二乘（WLS）与 BLUE

为获得方差最小的线性无偏估计（BLUE），对方程进行权变换：

定义加权矩阵与加权变量

$$\tilde{X} = W^{-1} X = \begin{bmatrix} X_1^T/\sigma_1 \\ X_2^T/\sigma_2 \\ \vdots \\ X_n^T/\sigma_n \end{bmatrix}, \qquad \tilde{Y} = W^{-1} Y = \begin{bmatrix} y_1/\sigma_1 \\ y_2/\sigma_2 \\ \vdots \\ y_n/\sigma_n \end{bmatrix}.$$

对加权模型做普通最小二乘，得到加权最小二乘估计

$$\hat{\beta}_{\text{BLUE}} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y} = (X^TW^{-2}X)^{-1}X^TW^{-2}Y.$$

该估计即为加权最小二乘估计（WLS），在误差方差已知或可估的情况下满足 BLUE 条件。

此时BLUE 的证明（要点）

设任意线性无偏估计写成

$$\tilde{\beta} = \left[(X^TW^{-2}X)^{-1}X^TW^{-2} + \Delta\right]Y.$$

- **无偏性** 要求

$$E(\tilde{\beta}) = \beta \quad \Longrightarrow \quad \Delta X = 0.$$

- **协方差比较**:

$$\begin{aligned}
\text{cov}(\tilde{\beta}) &= \left[(X^TW^{-2}X)^{-1}X^TW^{-2} + \Delta\right]\text{cov}(Y)\left[(X^TW^{-2}X)^{-1}X^TW^{-2} + \Delta\right]^T \\
&= \left[(X^TW^{-2}X)^{-1}X^TW^{-2} + \Delta\right]W^2\left[(X^TW^{-2}X)^{-1}X^TW^{-2} + \Delta\right]^T \\
&= (X^TW^{-2}X)^{-1} + \Delta W^2\Delta^T
\end{aligned}$$

由于 W^2 为非负定矩阵，得出

$$\text{cov}(\tilde{\beta}) \succeq (X^TW^{-2}X)^{-1},$$

且当且仅当 $\Delta = 0$ 时取得等号。因此 $\hat{\beta}_{\text{BLUE}}$ 在线性无偏估计中方差最小，即为 BLUE。

- 当存在异方差时，LSE 仍无偏但效率受损。若能估计或知道误差的方差结构（即 W），应采用 WLS 得到 BLUE:

$$\hat{\beta}_{\text{WLS}} = (X^TW^{-2}X)^{-1}X^TW^{-2}Y.$$

- 在实际中若 W 未知，通常用两步估计（估计残差方差结构，再作加权）或使用稳健标准误（如 White 异方差稳健标准误）来进行推断。

$\{\sigma_i^2\}$ 未知

模型设定

考虑线性回归模型：

$$\begin{cases} y_i = x_i^T \beta + \varepsilon_i \\ E(\varepsilon_i) = 0 \\ \mathrm{Var}(\varepsilon_i) = \sigma_i^2 \ (\text{unknown}) \\ \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0 \ (i \neq j) \end{cases}$$

最小二乘估计（LSE）形式

最小二乘估计为：

$$\hat{\beta}_{\text{LSE}} = (X^T X)^{-1} X^T Y$$

大样本下LSE的相合性证明（趋近于BLUE的性质）

我们需要证明：当样本量 $n \to \infty$ 时，$\hat{\beta}_{\text{LSE}}$ 具有**相合性**（即 $\hat{\beta}_{\text{LSE}} \xrightarrow{p} \beta$），且在大样本下近似具有BLUE的性质。

步骤1：分解LSE的方差

LSE第 i 个分量的方差为：

$$\begin{aligned} \mathrm{Var}\left(\hat{\beta}_{\text{LSE},i}\right) &= \left[(X^T X)^{-1} X^T W^2 X (X^T X)^{-1}\right]_{[i,i]} \\ &= \sum_j \left((X^T X)^{-1} X^T\right)_{[i,j]} \sigma_j^2 \left(X (X^T X)^{-1}\right)_{[j,i]} \\ &\leq \max_j \sigma_j^2 \sum_j \left((X^T X)^{-1} X^T\right)_{[i,j]} \left(X (X^T X)^{-1}\right)_{[j,i]} \\ &= \max_j \sigma_j^2 \left((X^T X)^{-1}\right)_{[i,i]} \end{aligned}$$

其中 $W = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$。

步骤2：假设 $X^T X$ 为对角矩阵的情形

若 $X^T X$ 是对角矩阵，则 $(X^T X)^{-1}$ 的第 i 个对角元为 $\frac{1}{(X^T X)_{[i,i]}}$。

而 $(X^TX)_{[i,i]} = \sum_{j=1}^n x_{ji}^T x_{ji} = \sum_{j=1}^n x_{ji}^2$，当 $n \to \infty$ 时，若解释变量满足一定的正则性条件（如平方和发散），则 $(X^TX)_{[i,i]} \to \infty$，进而：

$$((X^TX)^{-1})_{[i,i]} = \frac{1}{(X^TX)_{[i,i]}} \to 0$$

因此，$\text{Var}\left(\hat{\beta}_{\text{LSE},i}\right) \to 0$。

步骤3：相合性的判定

相合性需满足两个条件：

- 无偏性：$E(\hat{\beta}_{\text{LSE},i}) = \beta_i$
- 方差收敛到 0：$\text{Var}\left(\hat{\beta}_{\text{LSE},i}\right) \overset{n\to\infty}{\to} 0$

由上述推导，LSE满足这两个条件，故具有**相合性。**

故而当 $\{\sigma_i^2\}$ 未知时，尽管LSE不再是小样本下的BLUE，但在大样本下具有相合性，因此仍可使用LSE进行估计。

违反无自相关假设（$\text{cov}(\varepsilon_i, \varepsilon_j) \neq 0, i \neq j$）的分析

自相关的定义与简单情形

自相关即误差项之间的协方差不为0，定义为：

$$\text{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i - E\varepsilon_i)(\varepsilon_j - E\varepsilon_j)$$

我们讨论**相邻项自相关**的简单情形：$j = i - 1$，且 ε 满足**平稳性（stationary）**。此时，自相关系数为：

$$\text{Cor}(\varepsilon_i, \varepsilon_j) = \frac{E\varepsilon_i\varepsilon_j}{\sqrt{\text{var}\varepsilon_i \text{var}\varepsilon_j}} = \frac{E\varepsilon_i\varepsilon_{i-1}}{\sqrt{\text{var}(\varepsilon_i)\text{var}(\varepsilon_{i-1})}}$$

平稳性（Stationary）的假设

平稳序列满足：

$$
\begin{cases} E\varepsilon_i = a \text{ (常数)} \\ \operatorname{cov}(\varepsilon_i, \varepsilon_j) = R(|i-j|) \text{ (仅与间隔有关)} \\ \text{取 } i = j\text{时，} \quad \operatorname{var}(\varepsilon_i) = b \text{ (常数)} \end{cases}
$$

自相关的推断方法

一阶自协方差估计

用残差 e_i（$e_i \approx \varepsilon_i$）估计**一阶自协方差（auto-covariance lag-1）**：

$$
\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\varepsilon_{i-1} \approx \frac{1}{n}\sum_{i=1}^{n}e_ie_{i-1} \triangleq \hat{\rho}(1)
$$

D-W检验（Durbin-Watson Test）

D-W统计量用于检验一阶自相关，公式为：

$$
\text{D-W test} = \frac{\sum_{i=1}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n}e_i^2}
$$

对其变形推导：

$$
\begin{aligned}
\text{D-W test} &= \frac{2\sum_{i=1}^{n}e_i^2 - 2\sum_{i=1}^{n}e_ie_{i-1}}{\sum_{i=1}^{n}e_i^2} \\
&= 2\left(1 - \frac{\sum_{i=1}^{n}e_ie_{i-1}}{\sum_{i=1}^{n}e_i^2}\right) \\
&= 2(1 - \hat{\rho}(1))
\end{aligned}
$$

D-W统计量与自相关系数 $\rho(1)$ 的关系：

- 若 $\rho(1) = 0$（无自相关），则 $\text{D-W test} = 2$；
- 若 $\rho(1) = 1$（完全正自相关），则 $\text{D-W test} = 0$；
- 若 $\rho(1) = -1$（完全负自相关），则 $\text{D-W test} = 4$。

自相关的模型与估计（以AR(1)模型为例）

假设误差项服从**一阶自回归模型（AR(1)）**：

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t, \quad |\rho| < 1$$

其中 v_t 满足经典假设：$Ev_t = 0$，$\mathrm{var}(v_t) = \sigma_v^2$，$\mathrm{cov}(v_t, v_j) = 0 \ (t \neq j)$。

AR(1)模型的矩性质

- 期望：$E\varepsilon_t = \rho E\varepsilon_{t-1} + Ev_t = 0$（递归可得 $E\varepsilon_t = 0$）；
- 方差：

$$\begin{aligned} \mathrm{var}(\varepsilon_t) &= \rho^2 \mathrm{var}(\varepsilon_{t-1}) + \sigma_v^2 \\ &= \rho^2 \mathrm{var}(\varepsilon_t) + \sigma_v^2 \quad (\text{平稳性，} \mathrm{var}(\varepsilon_t) = \mathrm{var}(\varepsilon_{t-1})) \\ \Longrightarrow \quad \mathrm{var}(\varepsilon_t) &= \frac{\sigma_v^2}{1-\rho^2} \end{aligned}$$

- 协方差：$\mathrm{cov}(\varepsilon_t, \varepsilon_{t-1}) = \mathrm{cov}(\rho\varepsilon_{t-1} + v_t, \varepsilon_{t-1}) = \rho \mathrm{var}(\varepsilon_{t-1}) = \rho \cdot \frac{\sigma_v^2}{1-\rho^2}$

自相关的修正方法（以AR(1)为例）

将原模型 $y_t = x_t^T \beta + \varepsilon_t$ 与 AR(1) 模型结合，分两种情形讨论：

情形a：ρ 已知——迭代法（Cochrane-Orcutt 方法）

对原模型做迭代变换：

- 原模型：$y_t = x_t^T \beta + \varepsilon_t$
- 滞后一期：$y_{t-1} = x_{t-1}^T \beta + \varepsilon_{t-1}$，两边乘 ρ 得：$\rho y_{t-1} = \rho x_{t-1}^T \beta + \rho \varepsilon_{t-1}$

两式相减：

$$y_t - \rho y_{t-1} = (x_t^T - \rho x_{t-1}^T)\beta + (\varepsilon_t - \rho\varepsilon_{t-1})$$

令 $\tilde{y}_t = y_t - \rho y_{t-1}$，$\tilde{x}_t^T = x_t^T - \rho x_{t-1}^T$，$v_t = \varepsilon_t - \rho\varepsilon_{t-1}$，则模型变为：

$$\tilde{y}_t = \tilde{x}_t^T \beta + v_t$$

此时 v_t 满足经典回归假设（无自相关、同方差等），故可用普通最小二乘估计 β：

$$\hat{\beta} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y}$$

其中 $\tilde{X} = \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix}$，$\tilde{Y} = \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_n \end{bmatrix}$。

ρ 未知时的迭代估计方法（可行广义最小二乘思路）

当自相关系数 ρ 未知时，需通过**迭代法**估计 ρ 并修正模型，步骤如下：

Step 1：初始残差与 ρ 的估计

- 先通过普通最小二乘（OLS）估计原模型：

$$\hat{\beta}_{\text{LSE}} = (X^TX)^{-1}X^TY$$

- 计算残差 $e = Y - \hat{Y} = Y - X\hat{\beta}_{\text{LSE}}$，用残差代替误差项 ε，构造向量 $E_1 = \begin{bmatrix} e_1 \\ \vdots \\ e_{n-1} \end{bmatrix}$，$E_2 = \begin{bmatrix} e_2 \\ \vdots \\ e_n \end{bmatrix}$。
- 估计 ρ：

$$\hat{\rho} = (E_1^TE_1)^{-1}E_1^TE_2$$

Step 2：加权最小二乘修正模型

将 $\hat{\rho}$ 代入迭代变换（参考 ρ 已知时的方法），令：

$$\tilde{y}_i = y_i - \hat{\rho}y_{i-1}, \quad \tilde{x}_i^T = x_i^T - \hat{\rho}x_{i-1}^T$$

此时模型变为：

$$\tilde{y}_i = \tilde{x}_i^T\beta + v_i$$

对该模型用 OLS 估计 β：

$$\hat{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}$$

Step 3：迭代优化

将新估计的 $\hat{\beta}$ 代入 Step 1，重新计算残差并更新 $\hat{\rho}$，重复上述步骤直到 $\hat{\rho}$ 和 $\hat{\beta}$ 收敛。

特殊情形：$\rho \approx 1$（差分法）

当迭代中发现 $\rho \approx 1$ 时，原模型可近似为**差分模型**。推导如下：

原模型：

$$\begin{cases} y_i = x_i^T \beta + \varepsilon_i \\ y_{i-1} = x_{i-1}^T \beta + \varepsilon_{i-1} \end{cases}$$

两式相减（一阶差分）：

$$y_i - y_{i-1} = (x_i^T - x_{i-1}^T)\beta + (\varepsilon_i - \varepsilon_{i-1})$$

令 $\Delta y_i = y_i - y_{i-1}$，$\Delta x_i^T = x_i^T - x_{i-1}^T$，则模型变为：

$$\Delta y_i = \Delta x_i^T \beta + v_i \quad (\text{其中 } v_i = \varepsilon_i - \varepsilon_{i-1} \approx v_i \text{ 满足经典假设})$$

对差分模型用 OLS 估计 β：

$$\hat{\beta} = (\Delta X^T \Delta X)^{-1} \Delta X^T \Delta Y$$

这种方法通过差分消除了强自相关的影响，是 $\rho \approx 1$ 时的有效修正手段。

ρ未知时的自相关迭代估计（情形d）

当自相关系数ρ未知时，需通过**迭代最小化残差平方和（SSE）**来估计ρ和模型参数，步骤如下：

目标函数与参数关系

我们的目标是找到$\hat{\beta}$和$\hat{\rho}$，使得：

$$\hat{\beta}, \hat{\rho} = \arg\min_{\beta,\rho} \sum_{i=1}^{n} (\tilde{y}_i - \tilde{x}_i^T \beta)^2 = \arg\min_{\beta,\rho} \mathrm{SSE}(\beta, \rho)$$

其中$\tilde{y}_i, \tilde{x}_i$是经过ρ变换后的变量（如$\tilde{y}_i = y_i - \rho y_{i-1}$，$\tilde{x}_i^T = x_i^T - \rho x_{i-1}^T$），即$\tilde{y}_i, \tilde{x}_i$与$\rho$直接相关。

ρ的区间搜索与迭代

- 首先限定$\rho \in [0, 1]$，将区间离散化为若干点，例如：$\rho_0 = 0, \rho_1 = 0.1, \ldots, \rho_{10} = 1$。
- 对每个给定的ρ_k，确定对应的$\tilde{y}, \tilde{x}$，然后通过OLS估计β：

$$\hat{\beta}(\rho_k) = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}$$

- 计算该ρ_k下的$\mathrm{SSE}(\rho_k, \hat{\beta}(\rho_k))$，找到使SSE最小的$\rho_k$，记为$\hat{\rho}_1$。
- 基于$\hat{\rho}_1$进一步细化区间，重复上述步骤，直到$\hat{\rho}$足够精确（即迭代得到更accurate的ρ）。

违反线性假设的修正：Box-Cox变换

当模型的**线性假设不成立**时，可通过**Box-Cox变换**对被解释变量y进行变换，使其与解释变量满足线性关系。

Box-Cox变换的定义

对y的变换形式为：

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}$$

其中λ是变换参数，需通过极大似然估计（MLE）确定。

变换后的模型与似然函数

变换后模型为：

$$Y^{(\lambda)} = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

- 概率密度函数（pdf）：

$$\text{pdf} = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left\{ -\frac{1}{2\sigma^2} \|Y^{(\lambda)} - X\beta\|^2 \right\} J(\lambda)$$

其中雅克比项$J(\lambda) = \prod_{i=1}^n y_i^{\lambda-1}$（由变量变换的雅克比行列式推导而来）。

- 对β和σ^2进行OLS估计：

$$\hat{\beta} = (X^T X)^{-1} X^T Y^{(\lambda)}, \quad \hat{\sigma}^2 = \frac{1}{n} \|Y^{(\lambda)} - X\hat{\beta}\|^2$$

- 对数似然函数（Log-likelihood）：
将$\hat{\beta}$和$\hat{\sigma}^2$代入pdf并取对数，化简得：

$$\log L = -\frac{n}{2} \log \|Y^{(\lambda)} - X\hat{\beta}\|^2 + \log J(\lambda) + C$$

进一步整理为：

$$\log L = -\frac{n}{2} \log \frac{\|Y^{(\lambda)} - X\hat{\beta}\|^2}{J(\lambda)^{\frac{2}{n}}} + C = -\frac{n}{2} \log \text{SSE}^*(\lambda) + C$$

其中$\text{SSE}^*(\lambda)$是经过雅克比项调整后的残差平方和。

确定最优λ

通过搜索使对数似然函数最大（或等价地使$\text{SSE}^*(\lambda)$最小）的λ，即为最优变换参数。变换后的数据$(x_i, y_i^{(\lambda)})$将更接近线性关系，从而满足线性回归的假设。

Deny Assumption：解释变量矩阵列秩不足，线性相关（or 多重共线性）

$X \sim n \times (k+1)$ 列秩 < $k+1$

列线性相关 $\lambda_{\min}(X^T X) \approx 0$

- $X^T X \sim (\lambda, u)$（特征值与特征向量）
- $X^T X u = \lambda u \iff u^T X^T X u = \lambda u^T u = \lambda$（$u$ 为单位特征向量）

具体到含常数项的模型（$(\mathbf{1}_n, x_1, \ldots, x_k) = (x_0, x_1, \ldots, x_k)$）：

$$Xu = [\mathbf{1}_n, x_1, \dots, x_k] \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_k \end{bmatrix} = \sum_{i=0}^{k} u_i x_i^T = 0$$

多重共线性（multicollinearity）的病态性：基于瑞利商的严谨解释

多重共线性的“病态”本质可通过**瑞利商（Rayleigh Quotient）**对矩阵奇异性的刻画来严谨推导。

瑞利商与矩阵特征值的关系

对于实对称矩阵 A（如 X^TX），其瑞利商定义为：

$$R(A, u) = \frac{u^TAu}{u^Tu}$$

其中 u 是非零向量。瑞利商的取值范围满足：

$$\lambda_{\min}(A) \leq R(A, u) \leq \lambda_{\max}(A)$$

且当 u 是 A 对应于 $\lambda_{\min}(A)$ 的单位特征向量时，$R(A, u) = \lambda_{\min}(A)$。

多重共线性下的病态性推导

在多重共线性场景中，存在**单位向量** u 使得 $Xu \approx 0$（即解释变量存在近似线性组合）。此时，对矩阵 $A = X^TX$，计算其瑞利商：

$$R(X^TX, u) = \frac{u^T(X^TX)u}{u^Tu} = \frac{(Xu)^T(Xu)}{1} = \|Xu\|^2 \approx 0$$

结合瑞利商与最小特征值的关系 $\lambda_{\min}(X^TX) \leq R(X^TX, u)$，可得：

$$\lambda_{\min}(X^TX) \leq \|Xu\|^2 \approx 0$$

这表明 X^TX 的**最小特征值趋近于0**，即 X^TX 是**近似奇异矩阵**。而矩阵的逆 $(X^TX)^{-1}$ 的范数与最小特征值的倒数成正比（$\|(X^TX)^{-1}\| \propto \frac{1}{\lambda_{\min}(X^TX)}$），因此当 $\lambda_{\min}(X^TX) \approx 0$ 时，$(X^TX)^{-1}$ 的元素会急剧增大，导致回归系数估计量的方差剧烈膨胀——这就是多重共线性使模型“病态”的核心数学逻辑。

回归系数估计量的协方差矩阵：

$$\mathrm{cov}(b) = \sigma^2 (X^T X)^{-1}$$

若 $X^T X$ 奇异或近似奇异，则：

$$\mathrm{tr}(\mathrm{cov}(b)) = \sigma^2 \mathrm{tr}((X^T X)^{-1}) = \sum \mathrm{var}(b_i) \text{ 会特别大}$$

多重共线性的判断方法整理

基于矩阵特征值与行列式的判断

1. **最小特征值判断**：若 $\lambda_{\min}(X^T X) \approx 0$，说明存在多重共线性。
2. **特征值比值判断**：若 $\frac{\lambda_{\max}(X^T X)}{\lambda_{\min}(X^T X)} \approx \infty$，表明多重共线性严重。
3. **行列式判断**：若矩阵 $X^T X$ 的行列式

$$|X^T X| = \prod_{i=1}^{n} \lambda_i \approx 0$$

（λ_i 为 $X^T X$ 的特征值），则存在多重共线性。

4. VIF 判断：下会介绍

VIF（方差膨胀因子）

方差膨胀因子的定义为

$$VIF_j \triangleq \frac{1}{1 - R_j^2}$$

其中 $R_j^2 = \frac{SSR_j}{SST_j}$。

VIF的计算逻辑

- 设矩阵 $X \triangleq (1_n, x_1, ..., x_k)$，其中 x_j 是 X 的第 $j+1$ 列（$n \times 1$ 向量）。
- 构造矩阵 $X_{(j)} = (1_n, x_1, ..., x_{j-1}, x_{j+1}, ..., x_k)$（即去掉 x_j 后的矩阵）。
- 对模型 $x_j = X_{(j)}\beta + \varepsilon$ 进行回归，计算该回归的 **回归平方和**（SSR_j）与 **总平方和**（SST_j），进而得到

$$R_j^2 = \frac{SSR_j}{SST_j}$$

  最终计算

$$VIF_j = \frac{1}{1 - R_j^2}$$

VIF的判断标准

- 若 $R_j^2 \approx 1$，则 $VIF_j = \infty$，说明 X 的列（解释变量）线性相关，多重共线性严重。
- 若 $R_j^2 \approx 0$，则 $VIF_j = 1$，说明 X 的列线性无关，无多重共线性。
- 一般认为 $VIF_j > 10$ 时，存在较严重的多重共线性；有时也会用所有 VIF_j 的平均值

$$\frac{1}{k}\sum_{j=1}^{k} VIF_j$$

  辅助判断。

VIF等价于方差比较：方差比较因子

$$VIF_j = \frac{1}{1 - \frac{SSR_j}{SST_j}} = \frac{\frac{1}{n}SST_j}{\frac{1}{n}SSE_j}$$

其中：

- $\frac{1}{n}SST_j = \frac{1}{n}\sum(x_{ji} - \bar{x}_j)^2 = var(x_j)$（可理解为 x_j 的总方差）。
- $\frac{1}{n}SSE_j = \frac{1}{n}\sum(x_{ji} - \hat{x}_{ji})^2$（$\hat{x}_{ji}$ 是 $E(x_{ji}|X_{(j)})$ 的估计，即 x_j 被其他解释变量解释后的残差方差）。
- 分母实际反映了 x_j 对 $(1_n, x_1, ..., x_{j-1}, x_{j+1}, ..., x_k)$ 的条件方差 $var(x_j|(1_n, x_1, ..., x_{j-1}, x_{j+1}, ..., x_k))$。

所以 VIF 的等价定义为 **分子（无条件方差）/分母（有条件方差）**，即“方差膨胀因子”。
公式表达为：

$$VIF_j = \frac{SST_j}{SSE_j}$$

回归系数的方差与 VIF 的联系

对于回归系数 b_j，其方差为：

$$var(b_j) = \sigma^2 (X^T X)^{-1}[j+1, j+1]$$

矩阵 $X^T X$ 的构造为：

$$X^T X = \begin{bmatrix} 1_n^T \\ x_1^T \\ x_2^T \\ \vdots \\ x_k^T \end{bmatrix} [1_n, x_1, x_2, \ldots, x_k]$$

1. 矩阵元素对应关系：

$$(X^T X)[j+1, j+1] = x_j^T x_j$$

2. 定义向量：

$$\alpha_{j+1}^T = x_j^T X_{(j)}$$

其中 $X_{(j)}$ 表示去掉 x_j 后的矩阵。

$$
\begin{aligned}
var(b_j) &= \frac{\sigma^2}{x_j^T x_j - x_j^T X_{(j)}(X_{(j)}^T X_{(j)})^{-1} X_{(j)}^T x_j} \\
&= \frac{\sigma^2}{x_j^T \left(I - X_{(j)}(X_{(j)}^T X_{(j)})^{-1} X_{(j)}^T \right) x_j} \\
&= \frac{\sigma^2}{x_j^T (I-H) x_j} \qquad \text{（其中 } H = X_{(j)}(X_{(j)}^T X_{(j)})^{-1} X_{(j)}^T \text{ 为投影矩阵）} \\
&= \frac{\sigma^2}{SSE_j} \\
&= VIF_j \cdot \frac{\sigma^2}{SST_j}
\end{aligned}
$$

结合 $VIF_j = \frac{SST_j}{SSE_j}$，可知：

$$
var(b_j) = VIF_j \cdot \frac{\sigma^2}{SST_j}
$$

因此，**VIF 衡量了多重共线性对回归系数方差膨胀的程度。**当 VIF_j 较大时，说明变量 x_j 与其他解释变量高度相关，从而导致 b_j 的估计方差显著增大。