计量经济分析

第四章　多元线性回归

徐秋华，xuqh@swufe.edu.cn

西南财经大学，金融学院

大纲

1. 古典线性回归模型
2. 普通最小二乘法
3. 拟合优度
4. OLS 估计量的小样本性质
5. 古典模型的假设检验
6. 多重共线性
7. 离散解释变量

古典线性回归模型

为什么使用多元回归？

在对数工资对受教育年限的一元回归中加入衡量个人能力的 IQ 测试成绩作为额外的解释变量就得到如下二元回归模型：

$$\log wage_i = \beta_0 + \beta_1 \times edu_i + \beta_2 \times IQ_i + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

考虑对数工资的变化 $\Delta \log wage$：

$$\Delta \log wage = \beta_1 \Delta edu + \beta_2 \Delta IQ + \Delta\varepsilon.$$

若 $\Delta\varepsilon = 0$ 且 $\Delta IQ = 0$，则有：

$$\frac{\Delta \log wage}{\Delta edu} = \beta_1.$$

因此，β_1 衡量了控制其他因素不变，尤其是控制了 IQ 测试成绩不变的情况下，受教育年限增加一年对对数工资的影响。

偏效应

即使在样本中没有任何两个个体的 IQ 测试成绩是相同的，我们依然可以利用式 (1) 来估计控制了 IQ 测试成绩不变的情况下受教育年限对对数工资的影响。这是因为回归 (1) 的样本回归函数可以表示为：

$$\widehat{\log wage} = \hat{\beta}_0 + \hat{\beta}_1 \times edu_i + \hat{\beta}_2 \times IQ_i.$$

易知 $\hat{\beta}_1$ 是 $\widehat{\log wage}$ 对 edu 的偏导数，即使样本中的所有个体都具有不同的 IQ 测试成绩，偏导数的定义决定了 $\hat{\beta}_1$ 可解释为控制了 IQ 测试成绩不变的情况下预计的受教育年限对对数工资的影响。类似地，对于一般的多元回归，回归系数也具有偏效应（partial effect）的解释。

非线性关系

多元回归模型还可以用于刻画变量间的非线性关系。例如，年龄对工资的影响可能存在边际效用递减的情况，为了在回归模型中体现这种可能性，可以在工资对年龄的一元回归模型中加入年龄的平方项：

$$wage_i = \beta_0 + \beta_1 \times age_i + \beta_2 \times age_i^2 + \varepsilon_i.$$

假设 $\Delta\varepsilon = 0$，我们有：

$$\frac{\partial wage}{\partial age} = \beta_1 + 2\beta_2 \times age.$$

因此，年龄对工资的偏效应是随着年龄的变化而变化的，若 $\beta_2 < 0$，则年龄的边际效用是递减的。

符号约定

一般地，具有 K 个解释变量的多元线性回归模型具有如下形式：

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + \varepsilon_i, \tag{2}$$

其中，对于大部分回归方程来说，$X_{i1} = 1$，即常数项。若定义向量：

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}, \quad \boldsymbol{X}_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{iK} \end{pmatrix},$$

根据向量内积的定义，式 (2) 可以进一步表示为：

$$Y_i = \boldsymbol{X}_i' \boldsymbol{\beta} + \varepsilon_i.$$

若再定义：

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1K} \\ X_{21} & \cdots & X_{2K} \\ \vdots & \cdots & \vdots \\ X_{n1} & \cdots & X_{nK} \end{bmatrix},$$

则有：

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

古典线性回归模型

古典线性回归模型满足如下假设：

- 假设 4.1（线性假设）：

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_K X_{iK} + \varepsilon_i = \boldsymbol{X}_i' \boldsymbol{\beta} + \varepsilon_i,$$

  其中，β 是待估计的未知参数；ε_i 是不可观测的误差项。
- 假设 4.2（严格外生性）：$\mathbb{E}(\varepsilon_i \mid \mathbf{X}) = 0, \quad (i = 1, 2, \ldots, n)$。
- 假设 4.3（无完全共线性）：矩阵 $\mathbf{X}$ 列满秩，即 $\operatorname{rank}(\mathbf{X}) = K$。
- 假设 4.4（球形扰动）：

$$\mathbb{E}(\varepsilon_i^2 \mid \mathbf{X}) = \sigma^2 > 0, \quad (i = 1, 2, \ldots, n) \tag{3}$$

$$\mathbb{E}(\varepsilon_i \varepsilon_j \mid \mathbf{X}) = 0, \quad (i, j = 1, 2, \ldots, n;\ i \neq j). \tag{4}$$

关于假设的说明 1

- 假设 4.1 要求回归方程关于参数 β 是线性的。因为解释变量 X_{ik} 可以是可观测变量的非线性变换，所以线性回归模型也可以刻画变量之间的非线性关系。

- 假设 4.2 可以表示为：

$$\mathbb{E}(\varepsilon_i \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) = 0, \quad (i = 1, 2, \ldots, n),$$

  即以所有的观测变量 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ 为条件，ε_i 的条件期望为 0。

- 在假设 4.2 成立的情况下，容易验证：$\mathbb{E}(Y_i \mid \boldsymbol{X}_i) = \boldsymbol{X}_i'\beta$，CEF 具有线性形式，$\beta$ 为 Y_i 对 $\boldsymbol{X}_i$ 做投影的线性投影系数。

关于假设的说明 2

严格外生性假设的推论

若假设 4.2 成立，则有：

1. 误差项的无条件期望是 0：$\mathbb{E}(\varepsilon_i) = 0$，$i = 1, 2, \ldots, n$；
2. 解释变量与误差项正交：$\mathbb{E}(X_{jk}\varepsilon_i) = 0$，
   $i, j = 1, 2, \ldots, n$；$k = 1, \ldots, K$；
3. 解释变量与误差项不相关：$\mathrm{Cov}(\varepsilon_i, X_{jk}) = 0$，
   $i, j = 1, 2, \ldots, n$；$k = 1, \ldots, K$。

- 对于时间序列模型来说，假设 4.2 通常是无法满足的。
- 投影误差正交于 $\boldsymbol{X}_i$，而 $\boldsymbol{X}_i$ 与回归模型的误差项 ε_i 正交是由假设 4.2 推出的，严格外生性比正交条件的约束更强。

关于假设的说明 3

- 假设 4.3 要求 $\mathbf{X}$ 列满秩，即 $\mathbf{X}$ 的任何一列都不能表示成其他列的线性组合。
- 因为 $\mathbf{X}$ 是 $n \times K$ 维矩阵，假设 4.3 成立需要 $n \geqslant K$（$n < K$ 怎么办？）。
- 若此假设不满足，我们称解释变量出现完全共线性问题，此时最小二乘估计值无法计算。
- 对于一元线性回归，假设 4.3 等价于要求解释变量的数据不恒为常数。
- 常见违反假设 4.3 的原因：虚拟变量陷阱、选择变量时的疏忽。
- 解释变量高度线性相关影响回归系数估计准确性的问题称为多重共线性。

关于假设的说明 4

由严格外生性假设，我们有：

$$\mathrm{Var}(\varepsilon_i \mid \mathbf{X}) = \mathbb{E}(\varepsilon_i^2 \mid \mathbf{X}) - \mathbb{E}(\varepsilon_i \mid \mathbf{X})^2 = \mathbb{E}(\varepsilon_i^2 \mid \mathbf{X}).$$

因此，式 (3) 说明误差项满足条件同方差。同理，

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j \mid \mathbf{X}) = \mathbb{E}(\varepsilon_i \varepsilon_j \mid \mathbf{X}) = 0, \quad (i, j = 1, 2, \ldots, n;\ i \neq j).$$

因此，式 (4) 说明不同个体误差项之间满足条件不相关。在时间序列模型中，上式表示不同时刻的误差项之间不存在条件相关性，这种情况被称为误差项无自相关（serial correlation）。上述两个条件可以统一记为：

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}) = \sigma^2 \boldsymbol{I}_n, \tag{5}$$

其中，$\boldsymbol{I}_n$ 为 $n \times n$ 的单位阵。式 (5) 表明 $\boldsymbol{\varepsilon}$ 的条件方差和单位阵 $\boldsymbol{I}_n$ 成正比，这是假设 4.4 被称为球形扰动假设的原因。

普通最小二乘法

估计的基本思想

估计 β，实际上是对线性投影系数进行估计。线性投影系数最小化如下均方预测误差：

$$S(\boldsymbol{b}) = \mathbb{E}\left((Y - \boldsymbol{X}'\boldsymbol{b})^2\right).$$

令 $\{(\boldsymbol{X}_i, Y_i) : i = 1, 2, \ldots, n\}$ 是样本量为 n 的一个样本。为了估计回归系数，可将上式中的期望换成样本均值，即最小化如下残差平方的样本均值：

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i'\boldsymbol{b}\right)^2,$$

或，等价地，最小化如下残差平方和：

$$RSS(\boldsymbol{b}) = \sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i'\boldsymbol{b}\right)^2.$$

普通最小二乘法

普通最小二乘法

给定样本 $(Y_i, \boldsymbol{X}_i)_{i=1}^n$，最小二乘法通过最小化如下残差平方和估计多元线性回归模型的系数 β：

$$RSS(\boldsymbol{b}) \equiv \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i' \boldsymbol{b})^2.$$

记 β 的最小二乘估计量为 $\hat{\beta}_{ols}$，我们有：

$$\hat{\beta}_{ols} = \underset{\boldsymbol{b} \in \mathbb{R}^K}{\arg\min}\ RSS(\boldsymbol{b}),$$

其中，arg 符号表示取其后最优化问题的最优解。

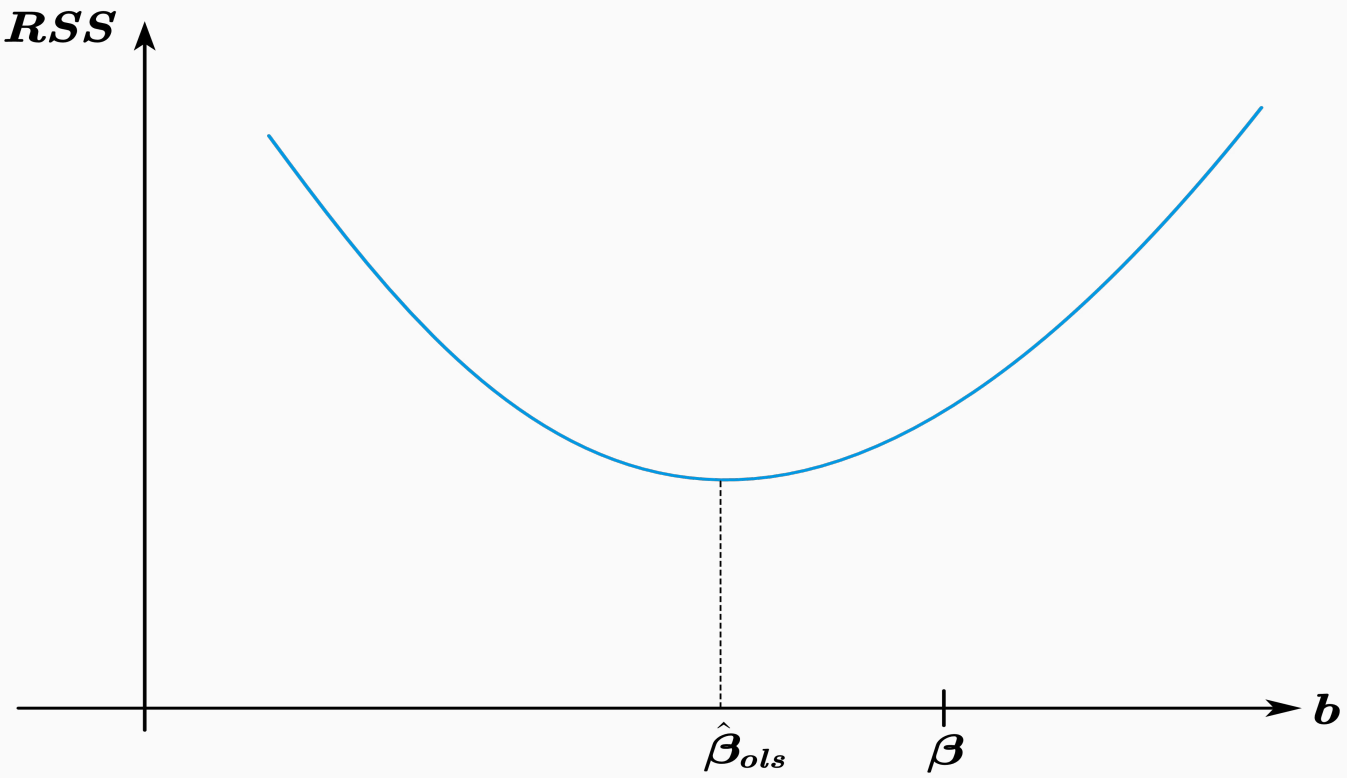

图 1: 最小二乘估计

OLS 估计量的推导 1

令 $\boldsymbol{e}$ 表示所有个体的残差构成的向量，即

$$\boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

样本中的正交条件可以表示为：

$$\sum_{i=1}^{n} \boldsymbol{X}_i e_i = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{bmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{bmatrix}' \boldsymbol{e} = \mathbf{X}'\boldsymbol{e} = \mathbf{0}. \tag{6}$$

由式 (6) 可以推出 OLS 估计量的表达式。

OLS 估计量的推导 2

将 $\boldsymbol{e}$ 表示成被解释变量的真实值构成的向量 $\boldsymbol{Y}$ 与拟合值向量 $\hat{\boldsymbol{Y}}$ 之差：

$$\begin{aligned}\boldsymbol{e} &= \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} - \begin{bmatrix} \boldsymbol{X}_1'\hat{\boldsymbol{\beta}}_{ols} \\ \boldsymbol{X}_2'\hat{\boldsymbol{\beta}}_{ols} \\ \vdots \\ \boldsymbol{X}_n'\hat{\boldsymbol{\beta}}_{ols} \end{bmatrix} = \boldsymbol{Y} - \begin{bmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{bmatrix} \hat{\boldsymbol{\beta}}_{ols} \\ &= \boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ols}.\end{aligned}$$

将此式代入到式 (6) 可得：

$$\mathbf{X}'\boldsymbol{e} = \mathbf{X}'(\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ols}) = \mathbf{0},$$

即

$$\mathbf{X}'\boldsymbol{Y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{ols}.$$

上式为多元情况下的正则方程。可以证明：在假设 4.3 成立的情况下，矩阵 $\mathbf{X}'\mathbf{X}$ 可逆，则有：

$$\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y}. \tag{7}$$

样本回归函数

我们称

$$\hat{Y}_i = \boldsymbol{X}_i'\hat{\boldsymbol{\beta}}_{ols} = X_{i1}\hat{\beta}_{1,ols} + \cdots + X_{iK}\hat{\beta}_{K,ols}$$

为样本回归函数（sample regression function, SRF），其中，$\hat{\beta}_{j,ols}$ 表示 OLS 估计量 $\hat{\boldsymbol{\beta}}_{ols}$ 的第 j 个元素。$\hat{Y}_i$ 称为个体 i 的拟合值。前面提到的所有个体拟合值构成的向量 $\hat{\boldsymbol{Y}}$ 可进一步表示为：

$$\hat{\boldsymbol{Y}} \equiv \mathbf{X}\hat{\boldsymbol{\beta}}_{ols} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{Y}.$$

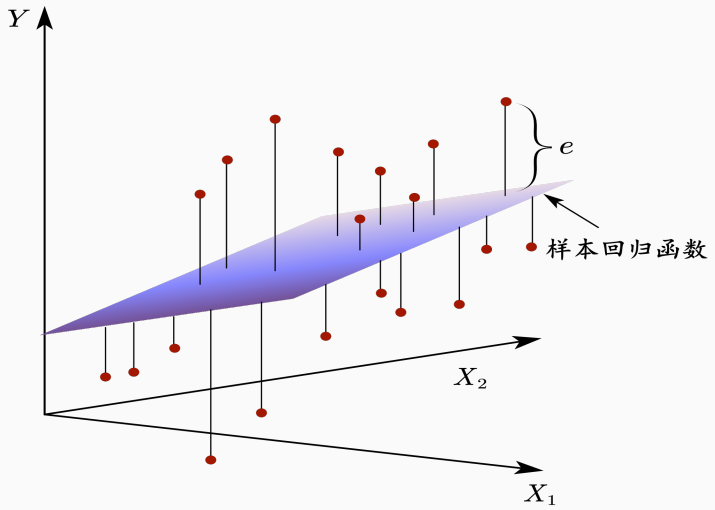

图 2: 样本回归函数

最小二乘法的几何意义 1

最小二乘法的几何意义是对 $\mathbf{X}$ 的列空间做投影。在图 3 中，拟合值（即投影）$\hat{\boldsymbol{Y}}$ 和 $\boldsymbol{Y}$ 满足：

$$\hat{\boldsymbol{Y}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{Y}.$$

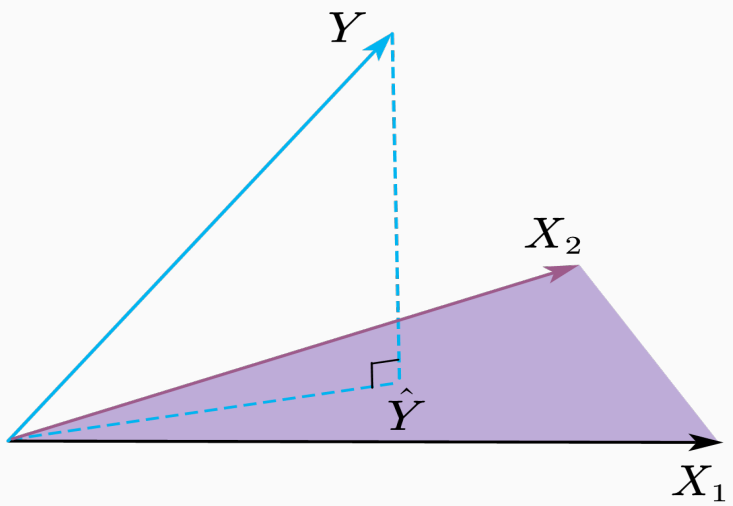

图 3: 最小二乘法的几何意义

所以，左乘矩阵 $\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ 起到对 $\mathbf{X}$ 的平面做投影的作用，我们将其记为 $\mathbf{P_X}$ 并称为投影矩阵。

最小二乘法的几何意义 2

$\boldsymbol{Y}$ 到平面的垂线即为残差 $\boldsymbol{e}$。对于残差向量 $\boldsymbol{e}$，我们有：

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \mathbf{P}_{\mathbf{X}} \cdot \boldsymbol{Y} = \left(\boldsymbol{I}_n - \mathbf{P}_{\mathbf{X}}\right) \boldsymbol{Y}.$$

所以，左乘矩阵 $\boldsymbol{I}_n - \mathbf{P}_{\mathbf{X}}$ 的作用是得到对 $\mathbf{X}$ 的平面做投影的残差，我们将其记为 $\mathbf{M}_{\mathbf{X}}$ 并称为消灭矩阵（annihilator）。容易验证 $\mathbf{P}_{\mathbf{X}}$ 和 $\mathbf{M}_{\mathbf{X}}$ 都是对称幂等阵，并且根据两个矩阵的几何意义容易知道：

$$\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}, \quad \mathbf{M}_{\mathbf{X}}\mathbf{X} = \mathbf{0}.$$

Frisch-Waugh-Lovell 定理

FWL 定理

考虑如下回归形式：

$$\boldsymbol{Y} = \begin{bmatrix} \mathbf{X}^{(-k)} & \boldsymbol{X}^{(k)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{-k} \\ \beta_k \end{bmatrix} + \varepsilon,$$

其中，数据矩阵 $\mathbf{X}$ 被分解成 $\begin{bmatrix} \mathbf{X}^{(-k)} & \boldsymbol{X}^{(k)} \end{bmatrix}$，$\boldsymbol{X}^{(k)}$ 是 $\mathbf{X}$ 的第 k 列，$\mathbf{X}^{(-k)}$ 是除 $\mathbf{X}$ 的第 k 列之外的所有列构成的矩阵。回归系数 β 也做了相应的分解，β_{-k} 是除 β_k 之外的所有回归系数构成的向量。为了得到 β_k 的估计量，可以按照以下步骤操作：

1. 将 $\boldsymbol{Y}$ 对 $\mathbf{X}^{(-k)}$ 回归，得到残差向量 $\boldsymbol{e}_{\boldsymbol{Y}}^{*}$；
2. 将 $\boldsymbol{X}^{(k)}$ 对 $\mathbf{X}^{(-k)}$ 进行回归，得到残差向量 $\boldsymbol{e}_{k}^{*}$；
3. $\boldsymbol{e}_{\boldsymbol{Y}}^{*}$ 对 $\boldsymbol{e}_{k}^{*}$ 回归，得到 β_k 的 OLS 估计量 $\hat{\beta}_{k,ols}$（此回归不用放常数项）。

拟合优度

非中心化 R^2

对于多元线性回归模型，如果回归方程中含有常数项，则平方和分解公式依然成立。因此，（中心化）R^2 仍可定义为：

$$R^2 \equiv 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}. \tag{8}$$

在线性回归模型中不包含常数项的情况下，Stata 汇报的是如下定义的非中心化 R^2：

非中心化 R^2

非中心化 R^2 的定义为：

$$R_{uc}^2 \equiv 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{Y}'\boldsymbol{Y}}.$$

它衡量的是被解释变量的变化由解释变量解释的比例。

调整的 R^2

R^2 的一个重要性质是如果回归中增加额外的解释变量，即使增加的解释变量与被解释变量无关，R^2 也只增不减。

调整的 R^2

调整的 R^2 定义如下：

$$\bar{R}^2 \equiv 1 - \frac{\sum_{i=1}^{n} e_i^2/(n-K)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)}. \tag{9}$$

调整的 R^2 通常也记为 $adj\text{-}R^2$。

OLS 估计量的小样本性质

OLS 估计量的小样本性质

(a) （无偏性）若假设 4.1-4.3 成立，则：$\mathbb{E}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) = \boldsymbol{\beta}$；

(b) 若假设 4.1-4.4 成立，则：$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) = \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$；

(c) （Gauss-Markov 定理）若假设 4.1-4.4 成立，则最小二乘估计量在所有线性无偏估计量中具有有效性，即对任意的关于 $\boldsymbol{Y}$ 线性并且无偏的估计量 $\hat{\boldsymbol{\beta}}$，$\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) \succeq \mathbf{0}$。[1]

[1] 因此，最小二乘估计量被称为最优线性无偏估计量（Best Linear Unbiased Estimator, BLUE）。

单个回归系数估计量的方差 1

我们考虑如下回归中系数 β_1 的最小二乘估计量的条件方差：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_K X_{iK} + \varepsilon_i.$$

事实上，由 FWL 定理，上式中 β_1 的最小二乘估计量表达式为：

$$\hat{\beta}_{1,ols} = \frac{\sum_{i=1}^{n} \hat{r}_{i1} Y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}, \tag{10}$$

其中，$\hat{r}_{i1}$ 是 X_{i1} 对 $X_{i2}, \ldots, X_{iK}$ 和常数项做回归得到的残差（注意 $\hat{r}_{i1}$ 的样本均值为 0）。

单个回归系数估计量的方差 2

基于 (10)，我们有如下定理：

单个系数估计量的方差

若假设 4.1-4.4 成立，我们有：

$$\mathrm{Var}(\hat{\beta}_{j,ols} \mid \mathbf{X}) = \frac{\sigma^2}{\mathrm{TSS}_j(1-R_j^2)}, \quad j = 1, \ldots, K, \tag{11}$$

其中，$\mathrm{TSS}_j = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$，$\bar{X}_j = \sum_{i=1}^{n} X_{ij}/n$，$R_j^2$ 是 X_{ij} 对所有其他解释变量和常数项做回归得到的（中心化）R^2。

σ^2 的估计

σ^2 的估计

σ^2 的估计量，记为 s^2，具有如下形式：

$$s^2 \equiv \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}.$$

只要 $n > K$，在假设 4.1-4.4 成立的情况下，我们有：$\mathbb{E}(s^2 \mid \mathbf{X}) = \sigma^2$，因此，$s^2$ 是 σ^2 的无偏估计。

古典模型的假设检验

CAPM

资本资产定价模型（Capital Asset Pricing Model, CAPM）是金融经济学中的重要模型，该模型指出：

$$\mathbb{E}\left[r_{jt} - r_f\right] = \beta_j \mathbb{E}\left[r_{mt} - r_f\right], \tag{12}$$

其中，r_{jt} 是 t 时刻风险资产 j 的收益率，r_{mt} 是 t 时刻市场组合的收益率，r_f 表示无风险资产的收益率。系统性风险（systematic risk）测度 β_j 满足：

$$\beta_j = \frac{\mathrm{Cov}(r_{jt}, r_{mt})}{\mathrm{Var}(r_{mt})}.$$

注意到 β_j 具有线性投影系数的形式，因此可以用最小二乘法进行估计。为此，定义

$$u_{jt} = r_{jt} - \mathbb{E}(r_{jt}), \quad u_{mt} = r_{mt} - \mathbb{E}(r_{mt}).$$

于是，式 (12) 可以表示为：

$$r_{jt} - r_f = \beta_j(r_{mt} - r_f) + \varepsilon_{jt}, \quad \varepsilon_{jt} = u_{jt} - \beta_j u_{mt}.$$

检验 CAPM 有效性

我们考虑如下回归：

$$r_{jt} - r_f = \alpha_j + \beta_j(r_{mt} - r_f) + \varepsilon_{jt}, \quad \varepsilon_{jt} = u_{jt} - \beta_j u_{mt}.$$

若 CAPM 对资产 j 有效，则有 $\alpha_j = 0$，此条件是否成立可以通过假设检验来分析。我们将想要检验的假设称为原假设（null hypothesis），通常记为 $\mathbb{H}_0$。在上述例子中，原假设为：

$$\mathbb{H}_0 : \ \alpha_j = 0.$$

正态假设

为了推导小样本下的假设检验理论，我们施加如下正态假设：

假设 4.5：（误差项正态假设）

$$\varepsilon \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n),$$

即以 $\mathbf{X}$ 为条件，ε 服从联合正态分布。

正态分布的性质

正态分布的密度函数只和分布的期望和方差有关，一般地，以 $\mathbf{X}$ 为条件的条件正态分布的期望和方差都是 $\mathbf{X}$ 的函数。但是，假设 4.5 要求 ε 的条件正态分布的期望和方差都与 $\mathbf{X}$ 无关，因此，ε 的无条件分布也是 $\mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$，并且 ε 与 $\mathbf{X}$ 相互独立。正态分布有一些良好的性质对于条件正态分布依然成立：

- 若两个随机变量以 $\mathbf{X}$ 为条件服从联合正态且不相关，则两个随机变量以 $\mathbf{X}$ 为条件相互独立；
- 若 ε 以 $\mathbf{X}$ 为条件服从正态分布，则 $\mathbf{A}\varepsilon$ 以 $\mathbf{X}$ 为条件也服从正态分布，其中 $\mathbf{A}$ 是 $\mathbf{X}$ 的函数。

OLS 估计量的分布

注意到

$$\hat{\beta}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon,$$

即

$$\hat{\beta}_{ols} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon = \mathbf{A}\varepsilon. \tag{13}$$

于是，在假设 4.1-4.5 成立的情况下，我们有：

$$(\hat{\beta}_{ols} - \beta) \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}). \tag{14}$$

又因为 $\boldsymbol{Y} = \mathbf{X}\beta + \varepsilon$，我们可以得到 $\boldsymbol{Y}$ 的条件分布为：

$$\boldsymbol{Y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \boldsymbol{I}_n).$$

由上式，在假设 4.5 成立的情况下，被解释变量服从条件正态分布。所以，若 Y_i 是离散变量，假设 4.5 不能成立。

单个回归系数的假设检验

我们关心第 k 个解释变量的系数 β_k 是否等于某一常数 β_k^0，则原假设为：$\mathbb{H}_0:\ \beta_k = \beta_k^0$，备择假设为 $\mathbb{H}_1:\ \beta_k \neq \beta_k^0$。由式 (14)，在假设 4.1-4.5 和原假设成立的情况下，

$$(\hat{\beta}_{k,ols} - \beta_k^0) \mid \mathbf{X} \sim \mathcal{N}\left(0, \sigma^2 \cdot \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}\right),$$

其中，$\left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}$ 是矩阵 $(\mathbf{X}'\mathbf{X})^{-1}$ 主对角线上的第 k 个元素。所以，若定义

$$z_k \equiv \frac{\hat{\beta}_{k,ols} - \beta_k^0}{\sqrt{\sigma^2 \cdot \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}},$$

则 z_k 服从标准正态分布。然而，虽然 z_k 的分布已知，但是，将 z_k 作为检验统计量依然是不可行的，这是因为 σ^2 是未知的。

t 统计量

我们可以将 z_k 表达式中的 σ^2 替换为 s^2，将替换后的检验统计量称为 t 统计量。

t 统计量的分布

考虑原假设 $\mathbb{H}_0: \beta_k = \beta_k^0$，此假设检验问题的 t 统计量被定义为：

$$t_k \equiv \frac{\hat{\beta}_{k,ols} - \beta_k^0}{\sqrt{s^2 \cdot \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}}, \tag{15}$$

其中，$\sqrt{s^2 \cdot \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}$ 称为 $\hat{\beta}_{k,ols}$ 的标准误（standard error），通常记为 $SE(\hat{\beta}_{k,ols})$。令假设 4.1-4.5 成立，在原假设 $\mathbb{H}_0$ 下，t 统计量服从自由度为 $n-K$ 的 t 分布，记为 t_{n-K}。

双边检验

基于 t 统计量的假设检验称为 t 检验，对于双边检验，其具体步骤如下：

1. 根据原假设 $\mathbb{H}_0$ 中参数的具体取值 β_k^0 计算由式 (15) 定义的 t 值；
2. 确定显著性水平 α，通常取为 5%，通过统计软件计算自由度为 $n-K$ 的 t 分布的临界值 $t_{\alpha/2}(n-K)$（也可以通过查表的方式找到临界值），如图 4 所示，$t_{\alpha/2}(n-K)$ 满足 t 分布在此临界值右侧的阴影面积为 $\alpha/2$。在 Stata 语言中，我们可以通过命令 invttail($n-K$, $\alpha/2$) 求得 $t_{\alpha/2}(n-K)$，例如，$\alpha=5\%$，$n-K=100$，我们可以用命令 dis invttail(100, 0.025) 显示临界值为 1.984；
3. 如果第一步计算的 t_k 满足 $|t_k| < t_{\alpha/2}(n-K)$，则不能拒绝 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

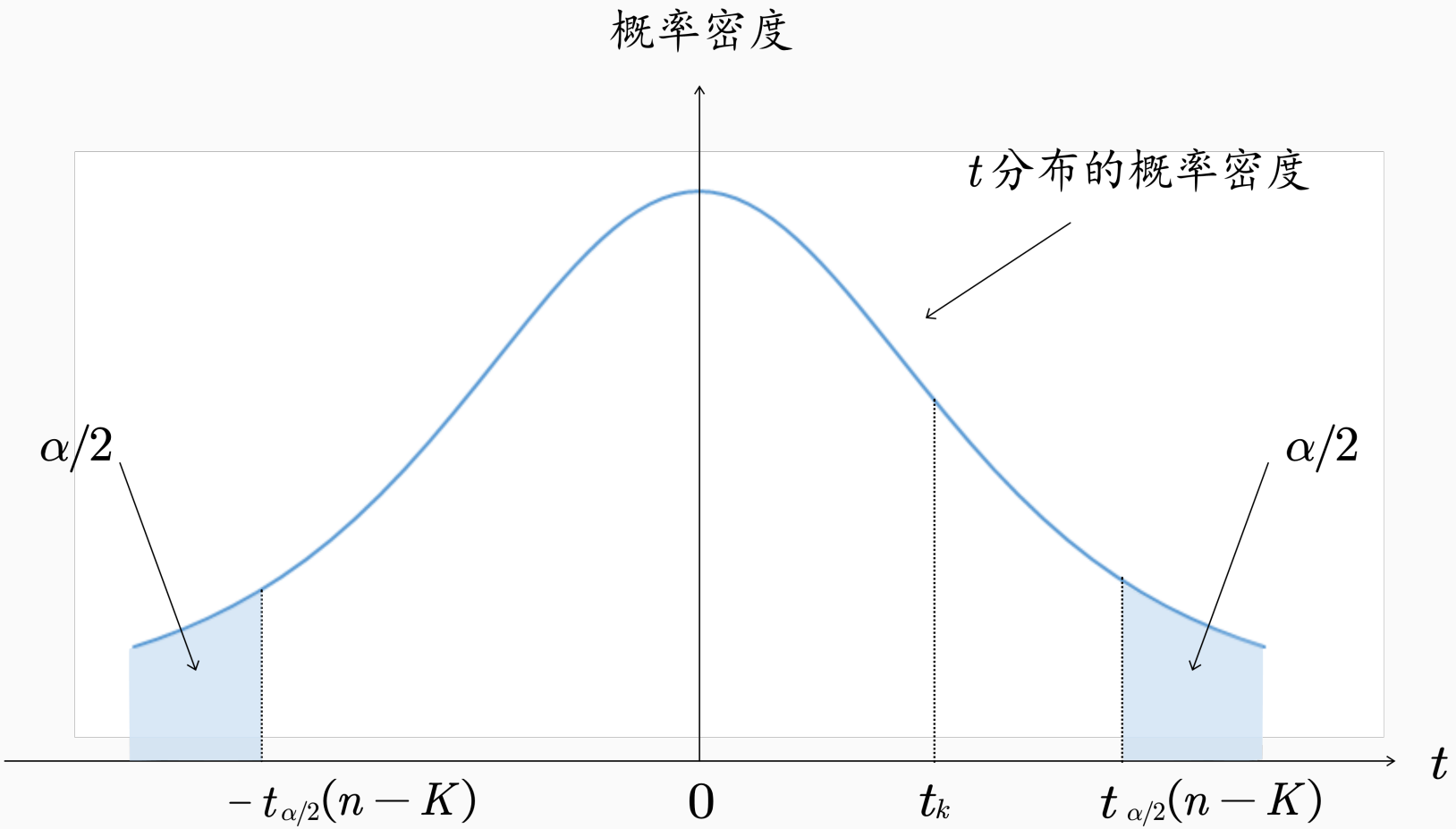

图 4: 基于 t 分布的双边假设检验

置信区间

我们还可以定义上述假设检验问题的置信水平为 $1-\alpha$ 的置信区间。t 检验的不拒绝区域为：

$$\mathcal{D}=\left\{(\mathbf{X},\boldsymbol{Y}) \ : \ \left|\frac{\hat{\beta}_{k,ols}-\beta_k^0}{SE(\hat{\beta}_{k,ols})}\right|<t_{\alpha/2}(n-K)\right\}.$$

通过求解上式中的不等式可知参数 β_k^0 的取值满足：

$$\hat{\beta}_{k,ols}-SE(\hat{\beta}_{k,ols})\cdot t_{\alpha/2}(n-K)<\beta_k^0<\hat{\beta}_{k,ols}+SE(\hat{\beta}_{k,ols})\cdot t_{\alpha/2}(n-K).$$

因此，我们不能拒绝 $\mathbb{H}_0$ 当且仅当原假设中的参数取值 β_k^0 落入区间：

$$[\hat{\beta}_{k,ols}-SE(\hat{\beta}_{k,ols})\cdot t_{\alpha/2}(n-K),\quad \hat{\beta}_{k,ols}+SE(\hat{\beta}_{k,ols})\cdot t_{\alpha/2}(n-K)],$$

这一区间被称为置信水平为 $1-\alpha$ 的置信区间，它覆盖真实参数的概率为 $1-\alpha$。

基于 p 值的检验

双边 t 检验也可以通过计算 p 值判断是否拒绝 $\mathbb{H}_0$，具体步骤如下：

1. 根据原假设 $\mathbb{H}_0$ 中参数的具体取值 β_k^0 计算由式 (15) 定义的 t 值 t_k；
2. 根据如下公式计算 p 值：

$$p = \mathbb{P}(t > |t_k|) \times 2, \tag{16}$$

   其中，$\mathbb{P}(t > |t_k|)$ 是服从 t_{n-K} 分布的随机变量 t 取值大于 $|t_k|$ 的概率；
3. 如果 $p > \alpha$，则不能拒绝 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

（见教材例子）

一般线性假设检验

考虑 Fama-French 三因子模型：

$$r_{p,t} = \alpha_p + \beta_{p_1} \times MKT_t + \beta_{p_2} \times SMB_t + \beta_{p_3} \times HML_t + \varepsilon_{p,t},$$

其中，$r_{p,t}$ 是资产组合 p 在 t 时段的超额收益率；MKT_t、SMB_t 和 HML_t 分别表示市场因子、规模因子和价值因子。三因子对资产组合 p 的超额收益率是否有解释能力可以通过如下假设检验进行分析：

$$\mathbb{H}_0 \ : \ \beta_{p_1} = \beta_{p_2} = \beta_{p_3} = 0.$$

备择假设为：

$$\mathbb{H}_1 \ : \ \beta_{p_1}，\ \beta_{p_2} \text{ 和 } \beta_{p_3} \text{ 至少有一个不为 } 0。$$

由于原假设是除常数项以外的所有解释变量的系数为 0，因此，此类假设检验问题被称为联合显著性检验。

我们可以把上述假设检验表示为一般形式，为此，定义如下矩阵：

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

以及向量 $\boldsymbol{r} = (0, 0, 0)'$，则有：

$$\mathbb{H}_0 \ : \ \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r},$$

其中，$\boldsymbol{\beta} = (\alpha_p,\ \beta_{p_1},\ \beta_{p_2},\ \beta_{p_3})'$。

事实上，所有关于多元线性回归系数的线性假设检验的原假设都可以表示为上述一般形式。例如，如果将 $\boldsymbol{R}$ 取为第 k 个位置上元素为 1 其余位置上元素为 0 的 K 维行向量以及取 $\boldsymbol{r} = \beta_k^0$，则单个回归系数的假设检验的原假设可以表示为上述一般形式。

F 统计量

线性假设检验的 F 统计量

考虑原假设 $\mathbb{H}_0: \boldsymbol{R}\beta = \boldsymbol{r}$，其中，$\boldsymbol{r}$ 是 $J \times 1$ 维列向量；$\boldsymbol{R}$ 是 $J \times K$ 维矩阵且满足 $\mathrm{rank}(\boldsymbol{R}) = J$，此假设检验问题的 F 统计量被定义为：

$$F \equiv \frac{(\boldsymbol{R}\hat{\beta}_{ols} - \boldsymbol{r})'[\boldsymbol{R}(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\beta}_{ols} - \boldsymbol{r})/J}{s^2} \tag{17}$$

令假设 4.1-4.5 成立，在原假设 $\mathbb{H}_0$ 下，F 统计量服从自由度为 J 和 $n-K$ 的 F 分布，记为 $F(J, n-K)$。

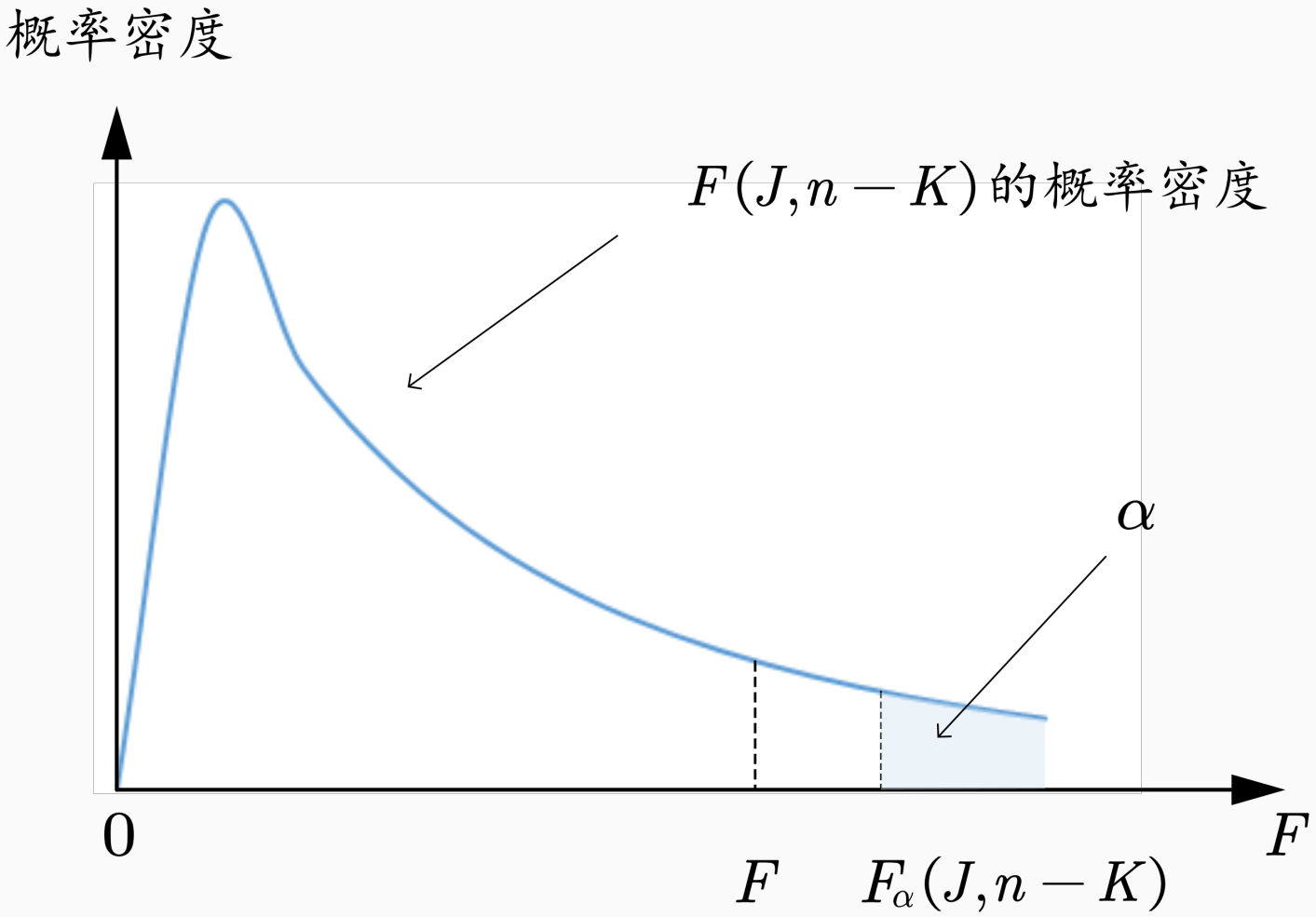

图 5: 基于 *F* 分布的假设检验

检验步骤

基于 F 统计量的假设检验称为 F 检验，其具体步骤如下：

1. 根据式 (17) 计算 F 统计量的值，简称 F 值；
2. 确定显著性水平 α，通常取为 5%，通过统计软件计算 $F(J, n-K)$ 的临界值 $F_\alpha(J, n-K)$（也可以通过查表的方式找到临界值），如图 5 所示，$F_\alpha(J, n-K)$ 满足 F 分布在此临界值右侧的阴影面积为 α。在 Stata 语言中，我们可以通过命令 invFtail(J, $n-K$, α) 求得 $F_\alpha(J, n-K)$，例如，$\alpha = 5\%$，$J = 3$，$n - K = 30$，我们可以用命令 dis invFtail(3, 30, 0.05) 显示临界值为 2.922；
3. 如果第一步计算的 F 值满足 $F < F_\alpha(J, n-K)$，则不能拒绝 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

基于 p 值的检验

我们也可以根据 p 值判断是否拒绝 $\mathbb{H}_0$，具体步骤如下：

1. 根据式 (17) 计算 F 值；
2. 计算分布 $F(J, n-K)$ 在 F 值右侧区域的面积 p；
3. 如果 $p > \alpha$，则不能拒绝 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

基于残差平方和的 F 统计量表达式

F 统计量的等价形式

F 统计量可以表示为：

$$F = \frac{(RSS_R - RSS_U)/J}{RSS_U/(n-K)}, \tag{18}$$

其中，RSS_U 是无约束回归的残差平方和，即 $\boldsymbol{e}'\boldsymbol{e}$；$RSS_R$ 是有约束回归的残差平方和，由求解如下最优化问题得到：

$$\min_{\boldsymbol{b}} RSS(\boldsymbol{b}) \quad \text{s.t.} \quad \boldsymbol{R}\boldsymbol{b} = \boldsymbol{r}.$$

（见教材例 4.3 和 4.4）

多重共线性

离散解释变量

Questions?