计量经济学STAT30021

第五讲：多元线性回归分析 (2)

肖志国

复旦大学管理学院

2025年10月

模型设定的选择

模型设定的选择，简称模型选择，指的是我们选择什么样的模型进行实证分析。

模型选择是实证研究中最重要的问题，也是最难的问题。

具体来说，模型选择主要涉及到以下三个部分的选择：

1. 第一，哪些变量作为解释变量
2. 第二，解释变量出现的函数形式
3. 第三，被解释变量出现的函数形式

这三个问题都无法（完美）解决，我们只能近似解决。

模型选择

对于第一个问题，这是模型选择的首要问题，也是实际情况中最难的问题。我们之前讨论过添加无关变量以及忽略有关变量的结果。因为这个问题和因果效应的估计密切相关，我们在后期课程中将会给出一些解决方案。

对于第二个问题，也就是说，当解释变量选定以后，它们在回归分析中出现的形式，也就是$E[Y|X_1, \ldots, X_k]$的函数形式，是接下来的根本问题。由于我们的实证研究主要关心的是特定解释变量的（因果）效应，我们首先选择用线性函数对其进行近似，然后通常只考虑那些系数解读具有明显直观含义的非线性函数形式作为备选。

对于第三个问题，我们一般只考虑一种非线性变换作为备选：对数变换。这主要也是因为对数变换的系数解读具有直观含义。

回归模型函数形式的设定

设Y为我们感兴趣的随机变量。给定任何一组变量$X_1, \ldots, X_m$:

- 我们都可以以Y为被解释变量，$X_1, \ldots, X_m$为解释变量做线性回归分析。
- 只要$E[Y|X_1, \ldots, X_m]$是$X_1, \ldots, X_m$的线性函数，那么我们的OLS估计量就是合理的。
- 但是现实问题中的关键问题在于：我们不知道$E[Y|X_1, \ldots, X_m]$的函数形式。
- 也就是说，我们不知道$E[Y|X_1, \ldots, X_m]$是不是$X_1, \ldots, X_m$的线性函数。

在实际问题中，我们需要找出一个用于数据分析的具体模型。

我们先看看通常的最小二乘估计到底估计的是什么。

OLS与条件期望函数的线性近似

记$\mathbf{X}=[1,X_1,X_2,\ldots,X_k]'$，$\hat{\beta}=[\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i']^{-1}\sum_{i=1}^n \mathbf{X}_i Y_i$的概率极限为

$$\beta^*=\left(E[\mathbf{XX}']\right)^{-1}E[\mathbf{X}Y]. \tag{1}$$

容易证明，β^* 满足：

$$\beta^*=\arg\min_b \; E\left[(E[Y|\mathbf{X}]-\mathbf{X}'b)^2\right]. \tag{2}$$

也就是说，作为一个函数，$\mathbf{X}'\beta^*$是所有$\mathbf{X}$的线性函数中最接近条件期望函数$E[Y|\mathbf{X}]$ 的那个函数。

由于$\mathbf{X}'\hat{\beta}$是我们可以通过数据分析得到的$\mathbf{X}'\beta^*$的近似，所以$\mathbf{X}'\hat{\beta}$是我们可以通过数据分析得到的条件期望函数$E[Y|\mathbf{X}]$的最佳线性近似。也就是说，最小二乘估计得到的是条件期望函数的最佳线性近似。

OLS与条件期望函数的非线性近似

条件期望函数$E[Y|\mathbf{X}]$很可能不是$\mathbf{X}$的线性函数。这时候OLS的结果就不具有理论合理性了。

但是非线性函数形式有不可数无穷多种，我们不可能使用穷举法进行搜索得到。

我们探索$E[Y|\mathbf{X}]$的非线性函数的形式，主要是对一些典型的情景进行尝试：

1. 对数形式
2. 交叉项形式（包括平方项形式）
3. 倒数形式
4. 线性组合的多层非线性变换（神经网络等机器学习算法）

神经网络模型

一个单层的经典神经网络模型形如：

$$
\begin{aligned}
E[Y|\mathbf{X}] &= f(X_1,\ldots,X_k,\theta) \\
&= \beta_0+\beta_1 Z_1+\cdots+\beta_M Z_M \\
&= \beta_0+\beta_1\sigma\left(\alpha_{01}+\alpha_{11}X_1+\cdots+\alpha_{k1}X_k\right) \\
&+\ldots \\
&+\beta_M\sigma\left(\alpha_{0M}+\alpha_{1M}X_1+\cdots+\alpha_{kM}X_k\right)
\end{aligned}
\tag{3}
$$

其中$\sigma()$ 成为激活函数，典型的选择有：

$$
\begin{aligned}
\sigma(v) &= \frac{1}{1+e^{-v}} && (sigmoid) \\
\sigma(v) &= \max(v,0) && (ReLU) \\
\sigma(v) &= \frac{e^{v}-e^{-v}}{e^{v}+e^{-v}} && (tanh)
\end{aligned}
\tag{4}
$$

神经网络模型

一个双层的神经网络模型形如：

$$
\begin{aligned}
E[Y|\mathbf{X}] &= f(X_1, \dots, X_k, \theta) \\
&= \beta_0^{(2)} \\
&+ \beta_1^{(2)} \sigma\Big(\beta_0^{(1)} + \beta_1^{(1)} \sigma(\alpha_{01} + \alpha_1' \mathbf{X}) + \cdots + \beta_M^{(1)} \sigma(\alpha_{0M} + \alpha_M' \mathbf{X})\Big) \\
&+ \dots \\
&+ \beta_L^{(2)} \sigma\Big(\beta_0^{(1)} + \beta_1^{(1)} \sigma(\alpha_{01} + \alpha_1' \mathbf{X}) + \cdots + \beta_M^{(1)} \sigma(\alpha_{0M} + \alpha_M' \mathbf{X})\Big)
\end{aligned} \tag{5}
$$

其中

$$
\alpha_j' \mathbf{X} = \alpha_{1j} X_1 + \cdots + \alpha_{kj} X_k, \quad j = 1, \dots, M. \tag{6}
$$

真实模型已知

当我们知道真实模型的时候。我们可以对一些实际选择的合理性进行判断。

假设我们利用如下的分布

$$\begin{bmatrix} Y \\ \log X_1 \\ X_2^{1/3} \\ X_3 \end{bmatrix} \sim N\left(\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 0.6 & 0.6 & 0 \\ 0.6 & 1 & 0.6 & 0 \\ 0.6 & 0.6 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right)$$

随机生成一组数据$(Y_i, X_{i1}, X_{i2}, X_{i3}), i = 1, \ldots, n$。然后我们利用这一组数据进行线性回归。

那么我们的线性回归分析有没有意义？

真实模型已知

基于真实的分布，我们可以推导出：

$$E[Y_i|X_{i1}, X_{i2}, X_{i3}] = -0.875 + 0.375 \log X_{i1} + 0.375 X_{i2}^{1/3}. \quad (7)$$

也就是说，我们如果以Y_i为被解释变量，$\log X_{i1}, X_{i2}^{1/3}$为解释变量进行线性回归分析的话，我们得到的截距项、$\log X_{i1}$的系数以及$X_{i2}^{1/3}$的系数的OLS估计量将会分别是-0.875，0.375和0.375的无偏估计量。

而我们直接进行$Y_i \sim X_{i1} + X_{i2} + X_{i3}$回归分析所得到的结果，是$E[Y_i|X_{i1}, X_{i2}, X_{i3}]$的最佳线性近似。其对应的系数估计，并不是以上系数的合理估计。

只有当这个线性近似属于比较好的近似时，也就是说，条件期望函数越近似是线性函数时，$Y_i \sim X_{i1} + X_{i2} + X_{i3}$的回归系数的直观解读才越具有合理性。

OLS估计量的方差

Theorem (OLS估计量的方差)

在假设*MLR.1-MLR.5*均满足的情况下，对于任意$j = 1, \dots, k$，有

$$Var(\hat{\beta}_j|\mathbb{X}) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \tag{8}$$

其中

- $\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$代表全部解释变量的数据
- $SST_j = \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2$为$X_j$的样本总波动
- R_j^2为X_j对所有其他解释变量X_m，$m \neq j$做回归的R^2

OLS估计量的方差的组成

$Var(\hat{\beta}_j|\mathbb{X})$由三部分组成：

- 回归误差项的方差 σ^2：回归方程的噪音越大，越难估计任何一个解释变量对Y的局部影响–此因素不随样本量变化而变化
- x_j的样本总波动 SST_j: x_j的总波动越大，$Var(\hat{\beta}_j|\mathbb{X})$越小–当样本量增加时，此因素会随之增加，从而其他因素不变时，$Var(\hat{\beta}_j|\mathbb{X})$会随样本量增加而减少
- 解释变量之间的线性关系 R_j^2: R_j^2越大，则$Var(\hat{\beta}_j|\mathbb{X})$越大

多重共线性

当至少存在一个R_j^2的值很接近于1时，我们说回归模型具有多重共线性。

- 多重共线性表示解释变量之间存在比较强的线性相关关系
- 严格来说，多重共线性并不是一个问题，因为它并不违反假设条件MLR.3
- 如何检查变量之间是否存在多重共线性：计算相关系数矩阵，当有些相关系数比较大时，说明存在多重共线性
- 多大叫"大"？–没有绝对标准(VIF 并不是一个好标准)
- 多重共线性的负面影响：涉及到的解释变量的系数估计很可能非常不准确，因为其方差很大
- 多重共线性的解决办法：去掉相关系数很大的变量组中一个或者多个变量–这时又可能会出现忽略相关变量的偏差问题
- 有些情况下多重共线性的存在不影响主要结果

错误设定模型中的估计量的方差

假设我们的真实模型是

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, \tag{9}$$

其中$E[u|X_1, X_2] = 0$。考虑β_1的两个估计量：

- $\hat{\beta}_1$是以Y为被解释变量，X_1, X_2为解释变量进行线性回归分析得到的X_1的系数的估计量
- $\tilde{\beta}_1$是以Y为被解释变量，X_1为解释变量进行线性回归分析得到的X_1的系数的估计量

当$\beta_2 \neq 0$时，$\tilde{\beta}_1$为β_1的忽略相关变量情形下的估计量，故而是有偏的，而此时$\hat{\beta}_1$是无偏的。

当$\beta_2 = 0$时，$\hat{\beta}_1$为β_1的添加相关变量情形下的估计量，故而是无偏的，而此时$\tilde{\beta}_1$是无偏的。

偏差 vs 方差

如果我们以无偏性作为唯一标准，则$\hat{\beta}_1$是优于$\tilde{\beta}_1$的。但是注意到：

$$Var(\hat{\beta}_1|X_1, X_2) = \frac{\sigma^2}{SST_1(1 - R_1^2)}, \tag{10}$$

而

$$Var(\tilde{\beta}_1|X_1, X_2) = \frac{\sigma^2}{SST_1}, \tag{11}$$

也就是说$\tilde{\beta}_1$的方差总是比$\hat{\beta}_1$的方差小的。

所以

- 当$\beta_2 = 0$时，$\tilde{\beta}_1$优于$\hat{\beta}_1$
- 当$\beta_2 \neq 0$时，当样本量比较大时，$\hat{\beta}_1$优于$\tilde{\beta}_1$

参数σ^2的估计

线性回归模型的另外一个重要参数是σ^2。令

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2, \tag{12}$$

其中$n-k-1$称作是自由度。

Theorem (定理3.3)

在*Gauss-Markov*假设*MLR.1-MLR.5*都满足的条件下，$\hat{\sigma}^2$是σ^2的无偏估计量：

$$E[\hat{\sigma}^2|\mathbb{X}] = \sigma^2.$$

标准差 vs 标准误

我们称σ为回归模型的标准差(standard deviation)，$\hat{\sigma}$为回归模型的标准误(standard error)。
$\hat{\beta}_j$的标准差为

$$sd(\hat{\beta}_j|\mathbb{X}) = \frac{\sigma}{\sqrt{SST_j(1-R_j^2)}},$$

$sd(\hat{\beta}_j|\mathbb{X})$的估计量，也就是$\hat{\beta}_j$的标准误为

$$se(\hat{\beta}_j|\mathbb{X}) = \frac{\hat{\sigma}}{\sqrt{SST_j(1-R_j^2)}}, \tag{13}$$

同方差性假设

- 多元线性回归模型中，$\hat{\beta}_j$,$j = 0, 1, \ldots, k$的无偏性不需要同方差性假设MLR.5
- 但是，$\hat{\sigma}^2$的无偏性依赖于同方差性假设MLR.5
- 当$u_1, \ldots, u_n$存在序列相关性时，或者$u_1, \ldots, u_n$互相独立但是方差不相同时：
  - $\hat{\sigma}^2$不再是σ^2的无偏估计量
  - $\hat{\beta}_j$的方差公式(8)不再成立
  - 相应地，$\hat{\beta}_j$的标准误公式(13)也不再成立
  - 注意：绝大多数统计软件在做回归分析的时候，汇报的标准误公式是同方差性假设MLR.5成立时候的公式
  - 如果我们怀疑假设MLR.5可能不成立，那么我们需要给出纠正后的标准误公式(第8章)

Gauss-Markov 定理

我们称β的估计量$\tilde{\beta}$为**一个线性估计量**，若存在一个只依赖于$\mathbb{X}$的$n \times (k+1)$矩阵A，使得$\tilde{\beta} = AY$，其中

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbb{X} = \begin{bmatrix} X_1' \\ \vdots \\ X_n' \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1k} \\ \vdots & \vdots & \vdots & \\ 1 & X_{n1} & \dots & X_{nk} \end{bmatrix}.$$

Theorem (Gauss-Markov 定理)

在*Gauss-Markov*假设*MLR.1-MLR.5*都满足的条件下，β的*OLS*估计量$\hat{\beta}$是β的所有线性无偏估计量中方差最小的一个。具体而言，对于β的任何一个线性无偏估计量$\tilde{\beta}$，$Var(\tilde{\beta}|\mathbb{X}) - Var(\hat{\beta}|\mathbb{X})$是一个非负定矩阵。

补充知识

- 二次型与正定，正半定，负定，负半定矩阵
- 幂等矩阵及其性质
- 多元正态分布的性质
- χ^2-分布，t-分布，以及F-分布及其性质
- 多元线性回归的向量矩阵表示
- 多元线性回归的相关结果的推导

D.4 QUADRATIC FORMS AND POSITIVE DEFINITE MATRICES

DEFINITION D.12 (Quadratic Form)

Let $\mathbf{A}$ be an $n \times n$ *symmetric* matrix. The **quadratic form** associated with the matrix $\mathbf{A}$ is the real-valued function defined for all $n \times 1$ vectors $\boldsymbol{x}$:

$$f(\boldsymbol{x}) = \boldsymbol{x}'\mathbf{A}\boldsymbol{x} = \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} a_{ij}x_ix_j.$$

DEFINITION D.13 (Positive Definite and Positive Semi-Definite)

(i) A symmetric matrix $\mathbf{A}$ is said to be **positive definite** (p.d.) if

$$\boldsymbol{x}'\mathbf{A}\boldsymbol{x} > 0 \text{ for all } n \times 1 \text{ vectors } \boldsymbol{x} \text{ except } \boldsymbol{x} = \mathbf{0}.$$

(ii) A symmetric matrix $\mathbf{A}$ is **positive semi-definite** (p.s.d.) if

$$\boldsymbol{x}'\mathbf{A}\boldsymbol{x} \geq 0 \text{ for all } n \times 1 \text{ vectors.}$$

If a matrix is positive definite or positive semi-definite, it is automatically assumed to be symmetric.

PROPERTIES OF POSITIVE DEFINITE AND POSITIVE SEMI-DEFINITE MATRICES:

(1) A positive definite matrix has diagonal elements that are strictly positive, while a p.s.d. matrix has nonnegative diagonal elements; (2) If $\mathbf{A}$ is p.d., then $\mathbf{A}^{-1}$ exists and is p.d.; (3) If $\mathbf{X}$ is $n \times k$, then $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ are p.s.d.; (4) If $\mathbf{X}$ is $n \times k$ and $\text{rank}(\mathbf{X}) = k$, then $\mathbf{X}'\mathbf{X}$ is p.d. (and therefore nonsingular).

D.5 IDEMPOTENT MATRICES

DEFINITION D.14 (Idempotent Matrix)

Let $\mathbf{A}$ be an $n \times n$ symmetric matrix. Then $\mathbf{A}$ is said to be an **idempotent matrix** if and only if $\mathbf{AA} = \mathbf{A}$.

PROPERTIES OF IDEMPOTENT MATRICES: Let $\mathbf{A}$ be an $n \times n$ idempotent matrix. (1) $\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A})$; (2) $\mathbf{A}$ is positive semi-definite.

We can construct idempotent matrices very generally. Let $\mathbf{X}$ be an $n \times k$ matrix with $\text{rank}(\mathbf{X}) = k$. Define

$$\mathrm{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$\mathrm{M} \equiv \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I}_n - \mathbf{P}.$$

Then $\mathbf{P}$ and $\mathbf{M}$ are symmetric, idempotent matrices with $\text{rank}(\mathbf{P}) = k$ and $\text{rank}(\mathbf{M}) = n - k$. The ranks are most easily obtained by using Property 1: $\text{tr}(\mathbf{P}) = \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}]$ (from Property 5 for trace) $= \text{tr}(\mathbf{I}_k) = k$ (by Property 1 for trace). It easily follows that $\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{P}) = n - k$.

The Chi-Square Distribution

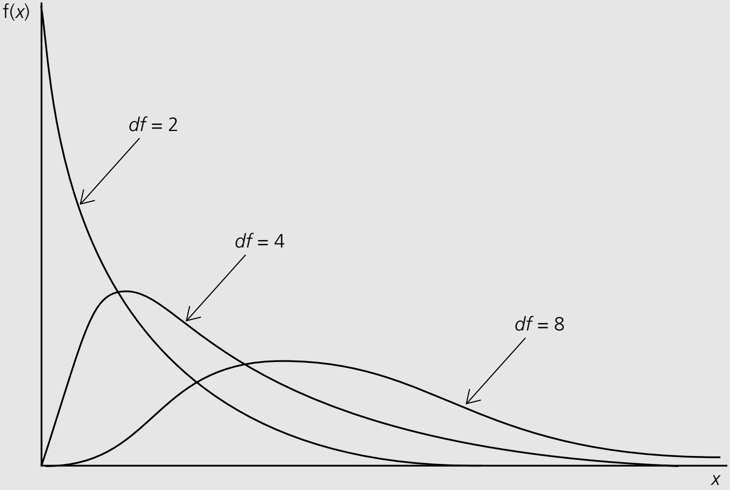
The chi-square distribution is obtained directly from independent, standard normal random variables. Let Z_i, $i = 1,2,\ldots,n$, be independent random variables, each distributed as standard normal. Define a new random variable as the sum of the squares of the Z_i:

$$X = \sum_{i=1}^{n} Z_i^2. \qquad \textbf{(B.41)}$$

Then, X has what is known as a **chi-square distribution** with n **degrees of freedom** (or *df* for short). We write this as $X \sim \chi_n^2$. The *df* in a chi-square distribution corresponds to the number of terms in the sum (B.41). The concept of degrees of freedom will play an important role in our statistical and econometric analyses.

Figure B.9

The chi-square distribution with various degrees of freedom.



The t Distribution

The t distribution is the workhorse in classical statistics and multiple regression analysis. We obtain a t distribution from a standard normal and a chi-square random variable.

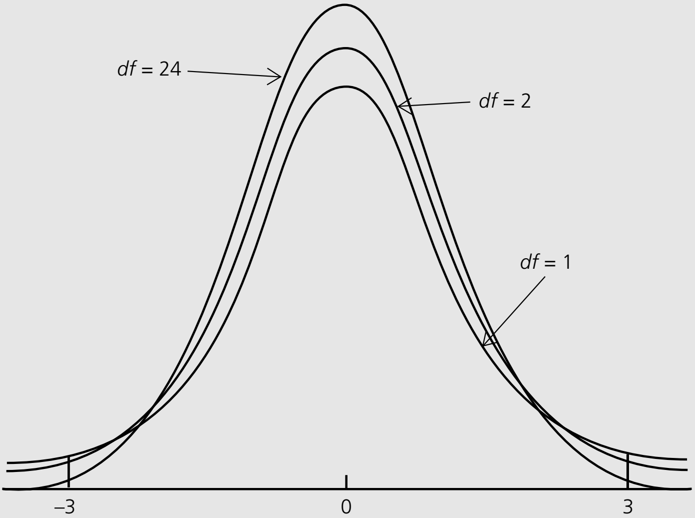
Let Z have a standard normal distribution and let X have a chi-square distribution with n degrees of freedom. Further, assume that Z and X are independent. Then, the random variable

$$T = \frac{Z}{\sqrt{X/n}} \tag{B.42}$$

has a **t distribution** with n degrees of freedom. We will denote this by $T \sim t_n$. The t distribution gets its degrees of freedom from the chi-square random variable in the denominator of (B.42).

Figure B.10

The *t* distribution with various degrees of freedom.



The F Distribution

Another important distribution for statistics and econometrics is the F distribution. In particular, the F distribution will be used for testing hypotheses in the context of multiple regression analysis.

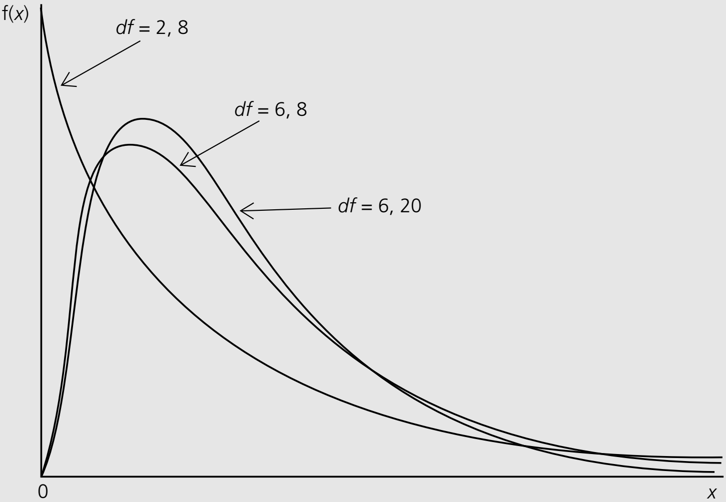
To define an F random variable, let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ and assume that X_1 and X_2 are independent. Then, the random variable

$$F = \frac{(X_1/k_1)}{(X_2/k_2)} \quad \textbf{(B.43)}$$

has an **F distribution** with (k_1,k_2) degrees of freedom. We denote this as $F \sim F_{k_1,k_2}$. The pdf of the F distribution with different degrees of freedom is given in Figure B.11.

Figure B.11

The F_{k_1,k_2} distribution for various degrees of freedom, k_1 and k_2.



PROPERTIES OF THE MULTIVARIATE NORMAL DISTRIBUTION: (1) If $\boldsymbol{y} \sim \text{Normal}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then each element of $\boldsymbol{y}$ is normally distributed; (2) If $\boldsymbol{y} \sim \text{Normal}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then y_i and y_j, any two elements of $\boldsymbol{y}$, are independent if and only if they are uncorrelated, that is, $\sigma_{ij} = 0$; (3) If $\boldsymbol{y} \sim \text{Normal}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then $\mathbf{A}\boldsymbol{y} + \boldsymbol{b} \sim \text{Normal}(\mathbf{A}\boldsymbol{\mu} + \boldsymbol{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$, where $\mathbf{A}$ and $\boldsymbol{b}$ are nonrandom; (4) If $\boldsymbol{y} \sim \text{Normal}(\mathbf{0},\boldsymbol{\Sigma})$, then, for nonrandom matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A}\boldsymbol{y}$ and $\mathbf{B}\boldsymbol{y}$ are independent if and only if $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' = \mathbf{0}$. In particular, if $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_n$, then $\mathbf{AB}' = \mathbf{0}$ is necessary and sufficient for independence of $\mathbf{A}\boldsymbol{y}$ and $\mathbf{B}\boldsymbol{y}$; (5) If $\boldsymbol{y} \sim \text{Normal}(\mathbf{0},\sigma^2\boldsymbol{I}_n)$, $\mathbf{A}$ is a $k \times n$ nonrandom matrix, and $\mathbf{B}$ is an $n \times n$ symmetric, idempotent matrix, then $\mathbf{A}\boldsymbol{y}$ and $\boldsymbol{y}'\mathbf{B}\boldsymbol{y}$ are independent if and only if $\mathbf{AB} = \mathbf{0}$; (6) If $\boldsymbol{y} \sim \text{Normal}(\mathbf{0},\sigma^2\boldsymbol{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are nonrandom symmetric, idempotent matrices, then $\boldsymbol{y}'\mathbf{A}\boldsymbol{y}$ and $\boldsymbol{y}'\mathbf{B}\boldsymbol{y}$ are independent if and only if $\mathbf{AB} = \mathbf{0}$.

Chi-Square Distribution

In Appendix B, we defined a **chi-square random variable** as the sum of *squared* independent standard normal random variables. In vector notation, if $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$, then $\boldsymbol{u}'\boldsymbol{u} \sim \chi^2_n$.

PROPERTIES OF THE CHI-SQUARE DISTRIBUTION: (1) If $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ is an $n \times n$ symmetric, idempotent matrix with $\text{rank}(\mathbf{A}) = q$, then $\boldsymbol{u}'\mathbf{A}\boldsymbol{u} \sim \chi^2_q$; (2) If $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ symmetric, idempotent matrices such that $\mathbf{AB} = \mathbf{0}$, then $\boldsymbol{u}'\mathbf{A}\boldsymbol{u}$ and $\boldsymbol{u}'\mathbf{B}\boldsymbol{u}$ are independent, chi-square random variables.

t Distribution

We also defined the ***t* distribution** in Appendix B. Now we add an important property.

PROPERTY OF THE *t* DISTRIBUTION: If $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$, $\boldsymbol{c}$ is an $n \times 1$ nonrandom vector, $\mathbf{A}$ is a nonrandom $n \times n$ symmetric, idempotent matrix with rank q, and $\mathbf{A}\boldsymbol{c} = \mathbf{0}$, then $\{\boldsymbol{c}'\boldsymbol{u}/(\boldsymbol{c}'\boldsymbol{c})^{1/2}\}/(\boldsymbol{u}'\mathbf{A}\boldsymbol{u})^{1/2} \sim t_q$.

F Distribution

Recall that an ***F* random variable** is obtained by taking two *independent* chi-square random variables and finding the ratio of each standardized by degrees of freedom.

PROPERTY OF THE *F* DISTRIBUTION: If $\boldsymbol{u} \sim \text{Normal}(\mathbf{0},\mathbf{I}_n)$ and $\mathbf{A}$ and $\mathbf{B}$ are $n \times n$ nonrandom symmetric, idempotent matrices with $\text{rank}(\mathbf{A}) = k_1$, $\text{rank}(\mathbf{B}) = k_2$, and $\mathbf{AB} = \mathbf{0}$, then $(\boldsymbol{u}'\mathbf{A}\boldsymbol{u}/k_1)/(\boldsymbol{u}'\mathbf{B}\boldsymbol{u}/k_2) \sim F_{k_1,k_2}$.

THEOREM E.4 (UNBIASEDNESS OF $\hat{\sigma}^2$)
Under Assumptions E.1 through E.4, $\hat{\sigma}^2$ is unbiased for σ^2: $\mathrm{E}(\hat{\sigma}^2|\mathbf{X}) = \sigma^2$ for all $\sigma^2 > 0$.

PROOF: Write $\hat{\boldsymbol{u}} = \boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = \mathbf{M}\boldsymbol{y} = \mathbf{M}\boldsymbol{u}$, where $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and the last equality follows because $\mathbf{MX} = \mathbf{0}$. Because $\mathbf{M}$ is symmetric and idempotent,

$$\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} = \boldsymbol{u}'\mathbf{M}'\mathbf{M}\boldsymbol{u} = \boldsymbol{u}'\mathbf{M}\boldsymbol{u}.$$

Because $\boldsymbol{u}'\mathbf{M}\boldsymbol{u}$ is a scalar, it equals its trace. Therefore,

$$\begin{aligned}\mathrm{E}(\boldsymbol{u}'\mathbf{M}\boldsymbol{u}|\mathbf{X}) &= \mathrm{E}[\mathrm{tr}(\boldsymbol{u}'\mathbf{M}\boldsymbol{u})|\mathbf{X}] = \mathrm{E}[\mathrm{tr}(\mathbf{M}\boldsymbol{u}\boldsymbol{u}')|\mathbf{X}] \\ &= \mathrm{tr}[\mathrm{E}(\mathbf{M}\boldsymbol{u}\boldsymbol{u}'|\mathbf{X})] = \mathrm{tr}[\mathbf{M}\mathrm{E}(\boldsymbol{u}\boldsymbol{u}'|\mathbf{X})] \\ &= \mathrm{tr}(\mathbf{M}\sigma^2\mathbf{I}_n) = \sigma^2\mathrm{tr}(\mathbf{M}) = \sigma^2(n-k).\end{aligned}$$

The last equality follows from $\mathrm{tr}(\mathbf{M}) = \mathrm{tr}(\mathbf{I}_n) - \mathrm{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = n - \mathrm{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = n - \mathrm{tr}(\mathbf{I}_k) = n - k$. Therefore,

$$\mathrm{E}(\hat{\sigma}^2|\mathbf{X}) = \mathrm{E}(\boldsymbol{u}'\mathbf{M}\boldsymbol{u}|\mathbf{X})/(n-k) = \sigma^2.$$