

第八章 虚拟变量回归学习辅导

一、本章的基本内容

(一) 基本内容

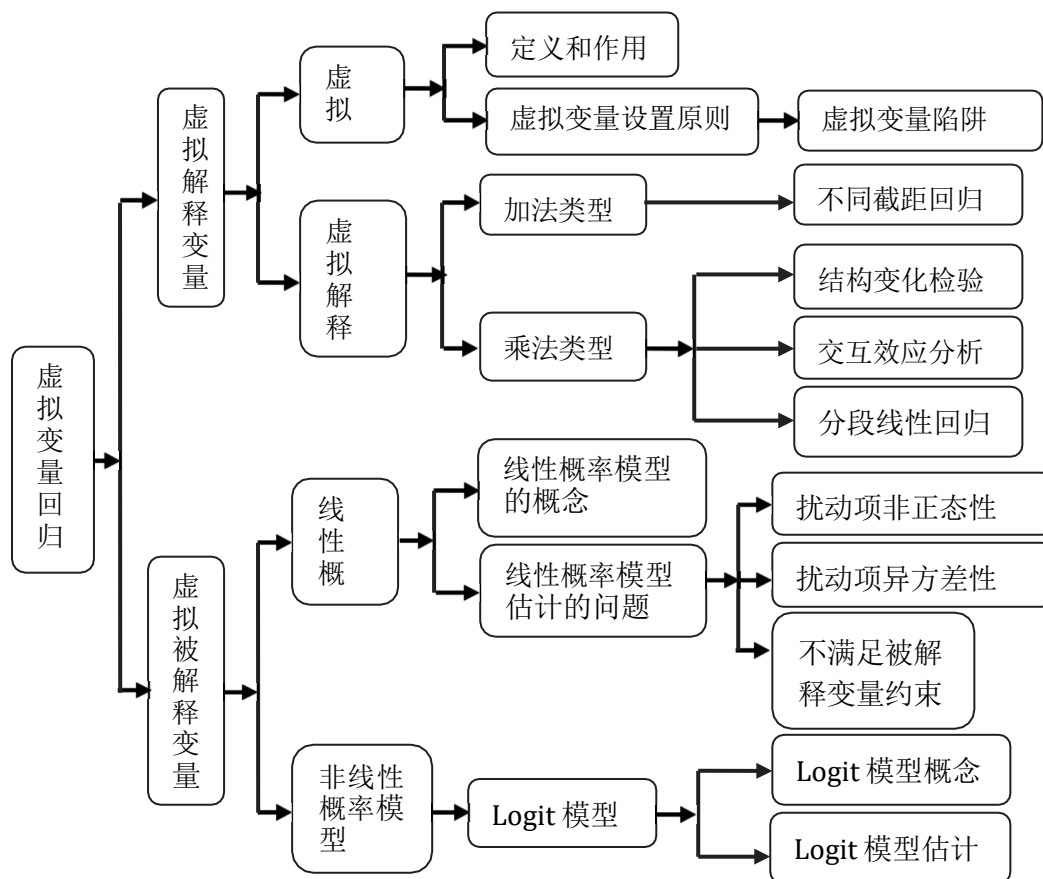


图 8.1 第八章基本内容

(二) 本章的教学目标

在现实经济生活中，影响被解释变量变动的因素，除了前几章讲的可以直接获得实际观测数据的定量变量外，还包括一些诸如性别、种族、职业、季节、受教育程度、政府政策的变动等定性因素（或称属性因素）的影响。在实际经济分析中，这些定性变量有时具有不可忽视的重要影响。为了在模型中反映定性因素的影响，可以将定性因素转化为虚拟变量去实现。简单而又常用的情况是，虚拟变量的取值为 1 和 0，代表“是”“否”或“有”“无”。这种做法实际上是一种变换或映射，将不能精确计量的定性因素的水平或状态变换为数字 0 和 1 来描述。

在计量经济学模型中，虚拟变量除了可以作为解释变量外，还可以作为被解释变量。当虚拟变量作为被解释变量时，其作用是对某一经济现象或活动进行“是”与“否”的判断或决策。例如，研究是否购买商品住房、是否参加人寿或财产保险、是否能按期偿还贷款、新产品在市场上是否畅销、对某一改革措施所持的态度等。这些问题的特征是被研究的对象（即被解释变量）在受到多种因素影响时，其取值只有两种状态：“是”与“否”。这在计量经济学中被称为“二元型响应”现象，这种现象常在市场研究或社会问题研究中遇到。如何处理二元型响应被解释变量模型的估计、推断问题，是本节要解决的问题。

本章的教学目标是：深刻理解定性因素在计量经济分析中的背景和含义。明确虚拟变量在建立和估计计量经济模型中的意义和作用。熟练掌握引入和应用虚拟变量的基本思想和方法。能够运用虚拟变量模型作相应的经济实证分析方面的应用。掌握 EViews 软件中相关内容的操作方法。

虚拟被解释变量的教学目标是了解什么是线性概率模型(LPM)；明确 OLS 法不适用与直接估计 LPM 的原因；了解非线性概率模型的基本思想以及 Logit 模型的基本概念和估计方法；能初步掌握运用 EViews 对 Logit 模型进行估计的方法。

二、重点与难点分析

1. 虚拟变量的设置原则

被解释变量的变动经常是定量因素和定性因素共同作用的结果。在计量经济模型中，应当同时包含定量和定性两种因素对被解释变量的影响作用。但在设置反映定性因素的虚拟变量时必须遵循一定的原则。即

若 1 个定性因素有 m 个相互排斥的类型（或属性、水平）时，按照模型设定中有无截距项，虚拟变量个数的设置规则分为两种情况：（1）在有截距项的模型中，只能引入 $m-1$ 个虚拟变量，否则会陷入所谓“虚拟变量陷阱”，产生完全的多重共线性；（2）在无截距项的模型中，可引入 m 个虚拟变量，不会导致完全的多重共线性。这种情形下虚拟变量参数的估计结果，实际上是 $D=1$ 时被解释变量的样本均值。

可以通过以下例子加以理解：

考虑季节因素（春夏秋冬，四个类型）对某种商品销量 Y 的影响，如果建立如下模型

$$Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{4i} + u_i$$

$$\text{其中: } D_{1i} = \begin{cases} 1 & \text{春季} \\ 0 & \text{其他} \end{cases}; \quad D_{2i} = \begin{cases} 1 & \text{夏季} \\ 0 & \text{其他} \end{cases}; \quad D_{3i} = \begin{cases} 1 & \text{秋季} \\ 0 & \text{其他} \end{cases}; \quad D_{4i} = \begin{cases} 1 & \text{冬季} \\ 0 & \text{其他} \end{cases}$$

即，在有截距项的模型中，引入 $m=4$ 个虚拟变量，则有

$$D_{1i} + D_{2i} + D_{3i} + D_{4i} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

或者 $-e + D_{1i} + D_{2i} + D_{3i} + D_{4i} = 0$ ，其中 $e = (1, 1, \dots, 1)'$ 。也就是说，构成 OLS 估计解释变量数据矩阵 X 的向量 $e, D_{1i}, D_{2i}, D_{3i}, D_{4i}$ 线性相关，即完全的共线性，OLS 估计无法获得。实际上此时只需设置 $m-1=3$ 个虚拟变量即可，比如

$$Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

因为 D_{1i}, D_{2i}, D_{3i} 已构成春夏秋冬四个类型的完备选择（当三者都为 0 时即代表冬季），而这样又能避免虚拟变量陷阱。

但是，在无截距项的模型中，引入 $m=4$ 个虚拟解释变量不会导致它们间的线性相关性（即不存在不全为零的数 $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ 使得 $\lambda_1 D_{1i} + \lambda_2 D_{2i} + \lambda_3 D_{3i} + \lambda_4 D_{4i} = 0$ ），因而不会引起多重共线性，不会落入虚拟变量设置陷阱。而稍后将证明，在这种情况下，虚拟变量参数的估计结果，实际上是 $D=1$ 时所对应的 Y 的均值。

2. 虚拟变量引入模型的具体方式及含义

在计量经济模型中，引入虚拟解释变量有两种基本方式：加法方式和乘法方式。不同的途径引入虚拟变量有不同的作用，加法方式引入虚拟变量考虑的是截距是否变化；乘法方式引入虚拟变量考虑的是斜率是否变化。当然，很多时候可以考虑加法方式和乘法方式同时引入虚拟变量，用以分析截距和斜率的变化，或者称为结构变化分析。

（1）加法方式

假设模型为
$$Y_i = \alpha_1 + \alpha_2 D_i + \beta X_i + \mu_i$$

条件期望
$$E(Y_i | X_i, D_i = 0) = \alpha_1 + \beta X_i$$

$$E(Y_i | X_i, D_i = 1) = (\alpha_1 + \alpha_2) + \beta X_i$$

表示的是虚拟变量的两种状态下，截距（平均水平）的差异。当然是否真的有差异，还要检验 α_2 是否为零。

（2）乘法方式

以乘法方式引入虚拟变量，是将虚拟解释变量与其他解释变量相乘（或反映不同定性因素的虚拟变量之间相乘）作为解释变量，以分析模型中斜率系数的变化。当然多数时候是乘法方式与加法方式合用，进行结构变化检验，交互效应分析及分段回归等。

$$\text{模型} \quad Y_i = \alpha_1 + \alpha_2 D_i + \beta_1 X_i + \beta_2 (D_i X_i) + u_i$$

可视为乘法方式引入虚拟变量的最基本最重要的形式，它用于结构变化检验。除了截距，可以判断斜率是否有变化，当然这依赖于对 β_2 是否为零的检验。

类似地，交互效应分析及分段回归的基本模型分别是

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 (D_{2i} D_{3i}) + \beta X_i + u_i$$

$$Y_i = \alpha_0 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i$$

同样需要检验 α_4 是否为零或 β_2 是否为零，以判断是否存在交互效应分析或分段回归。

*3. 线性概率模型与普通线性回归模型的联系¹

我们以前讨论的回归分析主要是研究条件均值 $E(Y_i | X_i)$ 轨迹的问题，而线性概率模型主要是研究条件概率 $P(Y_i | X_i)$ 轨迹的问题。两者看上去研究的对象不同，但是在二元响应被解释变量的问题中它们其实是一致的。需要注意的是，由于被解释变量的条件概率必须在 0~1 之间，所以当我们用条件均值 $E(Y_i | X_i)$ 来描述模型的时候，它的值也必须在 0~1 之间。

*4. OLS 不能直接适用于 LPM 的原因²

(1) 由于被解释变量 Y 只能取 0 和 1 两个值，所以扰动项 u_i 显然不服从正态分布，不满足线性回归模型的基本假设。因此，虽然 OLS 对参数估计的无偏性没有影响，但是在小样本下对其统计推断有所影响。这里我们需要注意的是，大样本条件下由中心极限定理，扰动项是趋近于正态的，所以如果样本足够大，扰动项的非正态性不会对参数估计产生太大的影响。

(2) 假设 $P(Y_i = 1 | X_i) = p_i$ ，我们可以容易的得到 $Var(u_i) = p_i(1 - p_i)$ 。因此扰动项存在异方差。这里我们可以利用加权最小二乘法进行修正。

(3) 由于线性概率模型中我们必须要求 $E(Y_i | X_i)$ 在 0~1 之间，但是实际分析中我们并不能保证这一点。所以这是用 LPM 的 OLS 法估计存在的实际问题。

¹ 这部分内容在本科教学中供选择使用。

² 这部分内容在本科教学中供选择使用。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/305224004101011301>