HDCC 预测模型（5）：混频数据模型

随着信息技术的快速发展，各种类型数据的发布、存储与获取越来越便捷。数据维度和数据频度的丰富对统计建模提出了新的挑战和发展方向。其中，基于包含了丰富的地理位置信息的数据衍生出了空间统计模型；而基于不同频度的数据，衍生出了混频数据模型。

将高频数据与低频数据结合起来预测低频数据指标的方法，从 Ghysels 等(2004)提出以来，得到了快速发展。目前来看，国外对混频数据进行建模有四种方法：一是混频数据取样方法（简称 MIDAS 法）（这类的研究主要有 Ghysels 等，2004；等等）。二是混频——向量自回归模型（简称为 MF-VAR 模型）（这类的研究有 Mariano 和 Murasawa，2010；等等）。三是混频因子模型，该类模型又分为小规模混频因子模型（Mariano 和 Murasawa，2003；等）和大规模混频因子模型（Giannone 等，2008；等）。四是因子——MIDAS 模型（这类研究主要有 Marcellino 和 Schumacher，2010；等）。

上述四类模型中，相对来说 MIDAS 模型由于其估计的相对简单性和对混频数据处理的合理性，使得其应用最为广泛，相关研究文献也最多。因此，我们这里将主要介绍 MIDAS 模型。

1.MIDAS 模型简介

首先给出分布滞后模型，其表示方法：

$$y_t = \beta_0 + B(L)x_\tau + \varepsilon_t$$

其中，$B(L)$是由有限或者无限滞后多项式算子，模型假设数据具有相同的频率。MIDAS 模型不是严格意义上的分布滞后模型，其最显著的特点是能够处理混频数据，并能获得优于分布滞后模型的参数估计结果。

假定在 MIDAS 回归模型中，等式左边的低频数据为$y_t(t=1,\cdots,T)$，等式右边的高频数据为$x_\tau(\tau=1,\cdots,mT)$，令$x_t^{(m)}=x_\tau$，其中 m 表示混频数据的频率倍差。则 MIDAS 回归模型即可表示为如下的形式：

$$y_t=\beta_0+\beta_1 B\left(L^{1/m};\theta\right)x_t^{(m)}+\varepsilon_t^{(m)}$$

其中，滞后算子多项式$B\left(L^{1/m};\theta\right)=\sum_{k=0}^{K}B(k;\theta)L^{k/m}$是参数向量 θ 的一个函数，$L^{1/m}$是高频数据的滞后算子，如$L^{1/m}x_t^{(m)}=x_{t-1/m}^{(m)}$，$K$ 是高频数据的滞后阶数。

假设宏观经济中，y_t是季度数据序列，$x_t^{(m)}$是与y_t在同一样本期间内抽样 m 次的高频数据，若 m=3，则$x_t^{(m)}$相对来说就是一个月度数据序列，若 y_t为 2010 年第一季度的数据，则$x_t^{(3)}$表示 2010 年 3 月的数据，$x_{t-1/3}^{(3)}$表示 2010 年 2 月的数据，$x_{t-1}^{(3)}$表示 2009 年 12 月的数据。一个滞后阶数 K=12 的 MIDAS 宏观经济模型为：

$$\begin{aligned}y_t&=\beta_0+\beta_1 B\left(L^{1/3};\theta\right)x_t^{(3)}+\varepsilon_t^{(3)}\\&=\beta_0+\beta_1\left[B(0;\theta)x_t^{(3)}+B(1;\theta)x_{t-1/3}^{(3)}+\cdots+B(12;\theta)x_{t-12/3}^{(3)}\right]+\varepsilon_t^{(3)}\end{aligned}$$

当涉及解释变量不止一个时，单变量 MIDAS 并不能满足研究要求，因此需要将其扩展至多变量 MIDAS 模型，具体表示如下：

$$Y_t = \beta_0 + \sum_{i=1}^{n} \beta_i B_i\left(L^{1/m};\theta_i\right) X_{i,t}^{(m)} + \varepsilon_t$$

这里 n 表示多变量的个数。

2. 模型的权重函数

MIDAS 模型估计中关键问题是权重函数 $B(k;\theta)$ 中的参数向量 θ 和滞后阶数 K 的选取，这涉及权重函数的选择。一般来说，权重函数分为有限制的权重函数形式和无限制的权重形式。

有限制的权重函数形式通常分为三种，即 Almon 多项式函数、指数 Almon 多项式函数、β 多项式函数。

第一种 Almon 多项式函数，其基本形式为：

$$B(k;\theta) = \frac{\theta_0 + \theta_1 k + \theta_2 k^2 \cdots + \theta_p k^p}{\sum_{k=1}^{K}\left(\theta_0 + \theta_1 k + \theta_2 k^2 \cdots + \theta_p k^p\right)}$$

第二种指数 Almon 多项式函数，可以理解为 Almon 多项式函数的变形。指数 Almon 多项式函数是目前使用最多的一种多项式函数形式，它不仅能构造出各种不同的权重函数，而且能够保证权重数为正数，同时能使方程获得零逼近误差的良好性质（Ghysels 和 Valkanov，2006）。其具体形式为：

$$B(k;\theta) = \frac{\exp\left(\theta_0 + \theta_1 k + \theta_2 k^2 \cdots + \theta_p k^p\right)}{\sum_{k=1}^{K}\exp\left(\theta_0 + \theta_1 k + \theta_2 k^2 \cdots + \theta_p k^p\right)}$$

第三种为β多项式函数，该多项式函数是仅带有两个参数的β多项式函数，它同样也能构造多种形态的权重函数，函数的具体形式可以表示为：

$$B(k;\theta_1,\theta_2)=\frac{f(k/K,\theta_1;\theta_2)}{\sum_{k=1}^{K}f(k/K,\theta_1;\theta_2)}$$

其中，

$$f(x,a,b)=\frac{x^{a-1}(1-x)^{b-1}\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

$$\Gamma(a)=\int_0^{\infty}e^{-x}x^{a-1}dx$$

MIDAS 模型中，上述三种权重函数均能保证高频滞后阶数的权重函数为正，且上述权重的函数定义中暗含了权重之和为 1 的假设。第一个多项式函数在金融市场波动的预测和分析中使用较多，第二和第三个多项式函数在宏观经济的分析与预测中应用比较多。

前面三种权重函数形式均受到不同形式的函数关系约束，除上述三种有限制的权重函数之外，还有非限制的权重形式（U-MIDAS），非限制的权重形式则是完全无约束的，此时的 MIDAS 其具体的函数形式如下：

$$y_t=\beta'+\sum_{k=0}^{K}\beta_k x_{t-k/m}^{(m)}+\varepsilon_t^{(m)}$$

在非限制的权重形式下，MIDAS 模型对高频解释变量的权重无限制性的约束，采用普通线性回归的方式对模型参数进行估计即可，故在后续混频模型的应用中，我们采取了非限制的权重函数。

3. 模型高频数据变量滞后阶的确定

MIDAS 模型设定的另一个关键问题是高频数据变量滞后阶的选择问题，该问题直接关系到在模型的估计和预测中，使用多长跨度的高频数据来预测低频数据。在高频变量滞后项权重函数形式确定的情况下， 高频变量滞后阶数 K 的改变不会影响模型估计的参数的个数。Ghysels 等(2004)认为通过非线性估计方法优化 MIDAS 回归方程中权重函数中的参数向量 θ，得出的参数所绘出的权重函数图形中可以获得滞后阶数的最优长度，这样确定的滞后阶数完全是数据驱动的，所以是最优的。

然而在宏观经济的实际运用中，滞后阶数的选择具有任意性和经验性的特点，也有部分的学者使用 AIC 和 BIC 准则作为滞后阶数的选择标准（刘金全等，2010）。因此，在确定高频解释变量滞后期的过程中，可以通过不断测试不同的 K 值来比较模型的拟合程度和预测效果，进而来确定最优的滞后期。

4. *h* 步向前预测的 MIDAS（m,K,h）模型

在基本 MIDAS (m, K) 模型下，预测第 t 期的低频被解释变量使用的数据信息一般为 t 期前的数据信息。例如，在用月度数据对日度数据进行混频建模时，如果想预测 1 月份的因变量的数据值，则需要用 2 月 1 日前的所有高频数据信息，这样只能等到 2 月 1 日当自变量所有的 1 月份的日度数据都出来之后才能进行预测。但在实际应用中，通常都需要提前进行预测，比如希望在月中当自变量 1 月 15 日的日度数据出来之后，就利用前 15 天的数据，通过混频数据模型，来对 1 月份因变量的值进行预测，这时 h 步向前预测的 MIDAS 模型设定为 MIDAS(m,K,h)，具体模型表达式如下：

$$y_t = \beta_0 + \beta_1 B\left(L^{1/m};\theta\right)x_t^{(m)} + \varepsilon_t^{(m)}$$
$$\Leftrightarrow y_t = \beta_0 + \beta_1 \sum_{k=h}^{K} B(k;\theta) L^{k/m} x_t^{(m)} + \varepsilon_t^{(m)}$$
$$\Leftrightarrow y_t = \beta_0 + \beta_1 \sum_{k=h}^{K} B(k;\theta) x_{t-k/m}^{(m)} + \varepsilon_t^{(m)}$$

5. 自回归混频模型 ADL-MIDAS

在前述的 MIDAS 模型中，自变量通常不包含因变量的滞后项，但在进行宏观经济变量预测时，应该充分考虑到因变量的惯性特征。一般来说，宏观经济变量的这种惯性特征体现为低频解释变量前后期之间存在自相关性。为了反映这种关系,可以在标准的 MIDAS 模型中加入低频被解释变量的自回归项。 具体模型形式如下：

$$y_t = \beta_0 + \delta_1 y_{t-1} + \delta_1 y_{t-2} + \cdots + \delta_p y_{t-p} + \beta_1 \sum_{k=h}^{K} B(k;\theta) x_{t-k/m}^{(m)} + \varepsilon_t^{(m)}$$

该类模型记为 ADL-MIDAS(p,h)，其中 p，h 是滞后阶数。

同样，这里留个问题给读者思考。第一，对 MIDAS 模型进行估计时，一般需要提前确定哪些因素？第二，在利用 MIDAS 模型做预测分析时，如果 Y 和 X 的数据频度分别是月度数据和日度数据，请问当手里 Y 的观测数据量不多时，该怎么进行 X 的滞后阶数的选择？

主要参考文献：

石峻驿：《基于混频数据模型的服务消费市场规模预测》，研究报告，2020.

欢迎扫码关注我们！