一类自适应权重下的混频逻辑回归模型及其应用研究

呼可可，刘梦丽
广州大学 经济与统计学院，广东 广州 510006
DOI:10.61369/ASDS.2026010011

摘　　要： 混频数据在经济预测中具有重要价值，但传统处理方法往往因同频化操作而损失高频信息。本文研究了混合频率数据下的逻辑回归建模问题，为了提升对高频数据的信息利用率，我们提出了一种迭代加权优化算法，模拟结果表明，优化权重可以显著提高回归系数的估计精度，降低估计系数的偏差和标准差。实证结果显示，优化权重对经济衰退风险的识别灵敏度更强。本文为混频数据下的二分类预测提供了新的方法参考。

关 键 词： 混频数据；逻辑回归模型；自适应权重；经济风险评估

A Type of Mixed-frequency Logistic Regression Model under Adaptive Weights And Its Application

Hu Keke, Liu Mengli
School of Economics and Statistics, Guangzhou University, Guangzhou, Guangdong 510006

Abstract: Mixed frequency data has important value in economic forecasting, but traditional processing methods often lose high-frequency information due to co-frequency operation. In order to improve the information utilization of high-frequency data, we propose an iterative weighted optimization algorithm, and the simulation results show that the optimization weight can significantly improve the estimation accuracy of the regression coefficient, and reduce the deviation and standard deviation of the estimation coefficient. The empirical results show that the optimization weight has a stronger sensitivity in identifying recession risks. This paper provides a new method reference for binary prediction under mixed frequency data.

Keywords: mixed-frequency data; logistic regression model; adaptive weight; economic risk assessment

引言

经济预测中常面临数据频率不一致的问题，如季度因变量与月度自变量，传统回归模型基于同频数据研究，对混频数据需要进行同频化处理，简单的等权聚合往往会导致大量高频信息损失，由此带来了关于混频数据的影响模式探讨和精准预测研究。混频数据抽样模型的提出，为直接使用原始混频变量进行建模分析提供了可能。Logistic 回归是一种成熟的、功能强大的二元结果分类方法，我们在逻辑回归中引入混频数据模型，提出了一种基于 Logistic 回归的新的混频数据驱动的加权方法。本文旨在构建一种自适应权重的 Logistic-MIDAS 模型[5]，通过数据驱动方式动态优化权重，提升模型在经济风险评估中的预测精度。

一、模型概述

（一）MIDAS 模型

Ghysels 等[1]首次提出能够直接对原始混频数据进行建模分析的 MIDAS 模型，该模型通过参数化的权重多项式，将高频解释变量直接应用到线性模型的构建、估计与预测。假设 y_t 表示第 t 期低频被解释变量，$x_t^{(m)}$ 表示高频解释变量，即在 t 至 $t+1$ 时间间隔内存在 m 个观测值，m 为高频变量与低频变量的频率倍差，如果 y_t 是年度数据，$x_t^{(m)}$ 是月度数据，则有 $m=12$，q 是高频变量的滞后阶数，则单变量 MIDAS 回归模型的基本形式可以表示为：

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

其中 $W\left(L^{1/m},\theta\right)=\sum_{k=0}^{K} w(k,\theta)L^{k/m}$，$w(k,\theta)$ 为多项式权重函数 $L^{1/m}$ 为高频滞后算子，$L^{1/m}x_t^{(m)} = x_{t-1/m}^{(m)}$，$\varepsilon_t$ 是模型误差项。

作者简介：
呼可可，广州大学经济与统计学院，硕士研究生；
刘梦丽，广州大学经济与统计学院，硕士研究生。

传统的 MIDAS 回归模型通常依赖于参数化的权重函数，这大大增加了模型的复杂性。为了解决这个问题，我们的研究建立在这些基础上，提出了一个简化的权重函数，直接结合了观测数据的特征。我们采用了一种非参数的、数据驱动的方法来选择权重，这使得混合频率数据的建模更加精确和自适应。因此，我们改进了上述公式中的模型，该模型具有无参数的权重函数，公式如下：

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

（二）Logistic 模型

逻辑回归模型为一种非线性回归分析模型，采用逻辑函数研究事件发生概率，判断因变量与自变量之间影响关系，在预测二分类问题方面发挥了重要作用 [2]。

模型预测器部分：

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n = \beta' x$$

$\beta = (\beta_0, \beta_1, \cdots, \beta_n)^T$ 为参数向量，$x = (1, x_1, x_2, \cdots, x_n)^T$ 为包含常数项的特征向量。

逻辑函数转换机制（sigmoid 函数）：

$$\pi x = \frac{1}{1 + exp(-\beta' x)}$$

通过 sigmoid 函数将线性预测器 η 映射到 (0,1) 区间，得到事件发生的概率。对于输入样本 (x, y)，利用 Logit 模型可以得到它属于某一类别的概率：

$$p(y|x) = \pi(x)^y \left(1 - \pi(x)\right)^{1-y}$$

这种概率计算方式通过逻辑函数将线性组合映射到区间，使得输出结果能被解释为事件发生或不发生的概率，从而为分类预测提供基础。

（三）多项式权重函数类型

MIDAS 模型的核心思想是通过引入多项式权重函数 $w(k, \theta)$ 减少待估参数的数量，而权重函数的设定形式不同会对 MIDAS 模型的结果产生不同的影响。以下介绍六种权重为混频建模中常用的传统权重，将作为后续模拟和实证研究中，与本文提出的自适应权重对比的基准模型 [3]。

1. 等权重

等权重为所有滞后项分配相同的权重，假设各滞后期对当前值的影响程度完全相同。无参数需要估计，权重函数为常数，这是最简单的权重分配方式。权重表达式为：

$$w_i = \frac{1}{m}, i = 1, 2, 3 \cdots m$$

2. 指数 Almon 权重

指数 Almon 滞后多项式是多项式分布滞后模型的现代扩展。该方法的核心思想是通过指数变换的多项式函数来构造权重，确保权重非负且灵活可变。指数变换具有双重作用：一是保证权重非负，二是增加函数非线性度以捕捉复杂模式。指数 Almon 权重广泛应用于混频数据采样模型中，用于处理高频解释变量对低频被解释变量的影响。具体形式可以为：

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

其中 p 为多项式阶数（通常 $p = 2$ 或 $p = 3$），θ_0，$\theta_1 \cdots \theta_p$ 是多项式的参数。

常用二阶形式（$p = 2$）：

$$w(\theta_1, \theta_2) = \frac{exp\left(\theta_0 + \theta_1 k + \theta_2 k^2\right)}{\sum_{k=1}^{K} exp\left(\theta_0 + \theta_1 k + \theta_2 k^2\right)}$$

3.Step 权重

Step 权重是一种非参数化或半参数化的权重分配方法，它将整个滞后区间划分为若干个区间，在每个区间内权重为常数。这种权重设定允许滞后效应在不同滞后阶段有不同但恒定的影响，适用于滞后效应呈现阶段性的情况，例如政策变化前后、经济周期不同阶段等。在计量经济学中，Step 权重可以看作分段常数函数的离散近似。具体形式可以表示为：

$$w(i, \theta_1, \theta_2, \theta_3) = \begin{cases} \theta_1, 0 \le i < k \\ \theta_2, k_1 \le i < k_2 \\ \theta_3, k_2 \le i < i^{max} \end{cases}$$

其中 $\theta_1 > \theta_2 > \theta_3 > 0, \sum_{j=1}^{3} \theta_j = 1$

4.Beta 权重

Beta 权重方法 [4] 是基于标准化 Beta 概率密度函数构造的权重，利用 Beta 分布丰富的形状灵活性来刻画各种滞后模式。Beta 分布定义在 [0,1] 区间上，恰好对应归一化的滞后时间指标。一般在金融市场波动的预测和分析中使用较多，Beta 多项式的具体形式可以表示为：

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

其中 $f(x, a, b) = \frac{x^{a-1}(1-x)^{b-1}\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$，$\Gamma(a) = \int_0^{\infty} e^{-x} x^{a-1} dx$

5. 非零 Beta 权重

非零 Beta 权重是在标准 Beta 权重的基础上进行修改，以确保所有权重都严格大于零，避免某些滞后项被赋予零权重。这在某些模型中可以提高数值稳定性或满足理论要求。其具体形式与 Beta 相似。

6. Legendre 权重

Legendre 权重基于 Legendre 正交多项式来构造权重函数。Legendre 多项式是在区间 [−1,1] 上定义的正交多项式，可以用于逼近任意光滑函数。通过将滞后指标映射到 [−1,1] 区间，并使用 Legendre 多项式的线性组合作为权重函数 [4]，我们可以获得高度灵活的权重形式。Legendre 权重函数的形式为：

$$w(k; \theta) = \sum_{i=0}^{p} \theta_i p_i(k)$$

其中$w(k;\theta)$是第k期权重，k是滞后时间期数，$p_i(k)$是第i阶 Legendre 多项式，$\theta_0,\theta_1\cdots,\theta_p$是多项式的参数，$p$是多项式阶数，Legendre 多项式的递推公式为：

$$p_0(k)=1, p_1(k)=k, p_{i+1}(k)=\frac{(2i+1)kp_i(k)-ip_{i-1}(k)}{i+1}$$

二、基于自适应权重的 Logistic-MIDAS 模型

基于以上简单逻辑回归模型和 MIDAS 模型的介绍，我们给出 Logistic-MIDAS 模型[5]以及自适应权重迭代优化算法。设响应变量y_t为二值变量，观测频率较低，协变量$X_t^{(m)}$观测频率较高，其中m表示高频期数。

Logistic-MIDAS 基本形式：

$$p(y_t=1|x)=\frac{1}{1+exp\left(-\left(\beta_0+\beta_1W\left(L^{1/m}\right)x_t^{(m)}\right)\right)}$$

线性预测项：$\eta_t=\beta_0+\beta_1W\left(L^{1/m}\right)x_t^{(m)}$，其中$\beta_0$为截距项，$\beta_1$为高频协变量的系数，$W\left(L^{1/m}\right)$为权重多项式，$L^{1/m}$为滞后算子。

逻辑变换后的概率预测：

$\pi_t=\frac{exp(\eta_t)}{1+exp(\eta_t)}=\sigma(\eta_t)$，其中$\sigma(z)=\frac{1}{1+e^{-z}}$是逻辑函数。

$$\pi x=\frac{1}{1+exp(-\beta' \mathrm{X})},\quad p(y|x)=\pi(x)^y\left(1-\pi(x)\right)^{1-y}$$

对于样本（x_i, y_i），其似然函数可以表示为：

$$L(\beta)=\prod_{i=1}^{m}\pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}$$

对数似然函数：

$$l(\beta)=logL(\beta)=\sum_{i=1}^{m}y_i ln\pi(x_i)+(1-y_i)\ln\left(1-\pi(x_i)\right)$$

给出设计矩阵

$$X_t^*=\begin{pmatrix}1 & 1 & \cdots & 1\\ x_{1,t-\frac{1}{m}} & x_{1,t-\frac{2}{m}} & \cdots & x_{1,t-\frac{q}{m}}\\ \vdots & \vdots & \ddots & \vdots\\ x_{p,t-\frac{1}{m}} & x_{p,t-\frac{2}{m}} & \cdots & x_{p,t-\frac{q}{m}}\end{pmatrix},\quad \mathrm{w}=\begin{pmatrix}w_1\\ w_2\\ \vdots\\ w_q\end{pmatrix}, \mathrm{X_t}=\mathrm{X_t^*w}=\begin{pmatrix}1\\ x_{1,t}\\ \vdots\\ x_{p,t}\end{pmatrix}$$

最小化目标函数：

$$-l(\beta)=-\sum_{t=1}^{m}y_t\ln\pi(\mathrm{X_t})+(1-y_t)\ln\left(1-\pi(\mathrm{X_t})\right)$$

权重需要满足约束条件$\sum_{i=1}^{k}w_i=1$和$w_i\geq 0$

该模型的参数估计为带约束的优化问题，权重需要满足约束条件$\sum_{i=1}^{k}w_i=1$和$w_i\geq 0$，通过交替迭代最小化负对数似然函数求解，具体流程如下：

i：初始化：给定一个初始权重w^t，满足$\sum_{i=1}^{k}w_i=1$和$w_i\geq 0$和初始系数β^t；

ii：交替迭代：

步骤1：固定当前系数β^t，通过梯度下降算法最小化负对数似然函数，得到优化权重w^{t+1}；

步骤2：固定新权重w^{t+1}，最小化逻辑回归负对数似然函数，得到更新值β^{t+1}

iii：收敛判断：若 $||\ \beta^{t+1}-\beta^t\ ||<\varepsilon$ 且 $||\ w^{t+1}-w^t\ ||<\varepsilon$（$\varepsilon$为某个预设阈值），迭代终止，输出最优参数$\beta^{new}$和权重$w^{new}$。

三、模型评估指标

（一）偏差和标准差

偏差衡量模型估计量的系统性偏离程度，它可以帮助我们判断估计量是否存在系统性偏差；标准差衡量预测概率的离散程度或波动性，体现了估计量的变异性，较小的样本标准差意味着更稳定的估计性能。

（二）均方误差

均方误差是衡量预测概率与真实结果之间差异的综合指标，同时考虑了偏差和方差的影响，在比较不同模型时，较低的 MSE 通常意味着更好的整体预测性能。

（三）AUC(Area Under the ROC Curve)

混淆矩阵

预测值 \ 真实值	$y=0$	$y=1$
$\hat{y}=0$	TN	FP
$\hat{y}=1$	FN	TP

$$TPR=\frac{TP}{TP+FN}, FPR=\frac{FP}{FP+TN}$$

真阳性率（True Positive Rate, TPR）与假阳性率（False Positive Rate, FPR）

ROC 曲线是通过变化分类阈值以 FPR 为横坐标、TPR 为纵坐标绘制出的曲线 AUC 则是 ROC 曲线下的面积。

（四）风险组（Risk Group, RG）

风险组是指在预测二分类问题时，按照预测概率从高到低排序后，取前α比例的观测样本构成的群体。RG[5]是一种针对性评估指标，特别适用于类别不平衡的场景。在预测“失业率对 GDP 的负面冲击”的研究中，风险组是指根据预测模型给出的概率值，从所有经济观测样本中筛选出预测概率最高的前α比例的观测单元。这些单元被认为最可能在未来特定时期内出现“高失业率伴随 GDP 显著下降”的复合负面经济事件。RG_α越高，表明模型识别极端风险事件的灵敏度越强。若$RG_{10\%}=0.8$则意味着模型仅通过关注其判定的风险最高的前10% 的时期，就能捕捉到历史上80% 的经济衰退。$RG_\alpha=\frac{\text{风险组中实际发生负面事件的观测数字}}{\text{总负面事件观测书数}}$。

（五）风险组下面积（Area Under Risk Group, AURG）

是评估模型在风险组内部对负面经济事件排序质量的指标。它通过绘制“风险组规模扩大过程中捕捉到的负面事件比例”曲线，并计算该曲线下面积占最大可能面积的比例，来反映模型在高风险群体中的分辨能力。计算方法如下：

从最小的风险组规模开始，逐步扩大至目标 α 比例。在每个规模点，计算当前风险组中实际发生负面事件的样本数占总负面事件样本数的比例（即该规模下的 RG 值）。以“风险组规模（横轴，从0到α%）”和“捕捉到的负面事件比例（纵轴，从0到

100%）”绘制曲线。计算该曲线下的面积 S_{shaded}，将 S_{shaded} 除以整个矩形面积（即 $\alpha\times100\%$），得到 $AURG_\alpha$。$AURG_\alpha=\frac{S_{shaded}}{\alpha\times100\%}$。

四、模拟

为了阐述自适应权重相较于其他权重在提取有效信息方面的优越性，本文构建了 Logistic-MIDAS 模型，并运用不同权重函数开展对比分析。我们采用第2章介绍的等权重、指数 Almon 权重、Beta 权重、Step 权重、非零 Beta 权重、Legendre 权重作为权重滞后多项式，将这六种权重以及我们构建的自适应权重函数代入到 Logistic-MIDAS 模型中去进行比较分析。蒙特卡罗模拟 [6] 设计基于 Logistic-MIDAS 回归模型的数据生成过程。

我们假设高频单变量遵循以下 ARCH(1,1) 过程：

$$x_{t/m}^{(m)}=\sqrt{\left(\sigma_{t/m}^{(m)}\right)^2}e_{t/m}^{(m)},\left(\sigma_{t/m}^{(m)}\right)^2=\alpha_0+\alpha_1\left(x_{t-1/m}^{(m)}\right)^2$$

$$\alpha_0=0.25,\alpha_1=0.85,e_{t/m}\sim N.i.i.d.(0,1)$$

在模型中，我们设定 $\beta_0=2,\beta_1=4,m=5$，考虑样本量大小为400，800，1000的情形，给出六种不同权重和自适应权重的模拟结果。$w_{equ},w_{exp},w_{beta},w_{step},w_{non-zero},w_{legendre}$，分别代表以上六种权重，$w_{opt}$ 代表我们提出的自适应优化权重。相应的系数估计分别表示为 $\hat{\beta}_{equ},\beta_{exp},\beta_{beta},\beta_{step},\beta_{non-zero},\hat{\beta}_{legendre},\beta_{opt}$，计算出系数估计值的偏差和标准差来评估他们的有限样本性能，同时为了评估相对效率 [7]，我们给出六种不同权重基于自适应权重的均方误差比，结果见表1。

表1：偏差、标准差、均方误差比

n	BS1	BS2	BS3	BS4	BS5	BS6	BS7
400	−0.5835	−0.3713	−0.5835	−0.5594	−0.5654	−0.3738	0.1316
800	−0.6305	−0.4367	−0.6305	−0.6271	−0.6205	−0.4515	0.0343
1200	−0.6400	−0.4518	−0.6400	−0.6364	−0.6236	−0.4686	0.0176
n	SD1	SD2	SD3	SD4	SD5	SD6	SD7
400	0.4652	0.5118	0.4652	0.5981	0.5989	0.4965	0.5544
800	0.3328	0.3771	0.3328	0.4889	0.4834	0.3703	0.3933
1200	0.2623	0.2999	0.2623	0.4128	0.3981	0.2946	0.3215
n	MSEequ / opt	MSEexp / opt	MSEbeta / opt	MSEstep / opt	MSEnon-zero / opt		MSElegendre / opt
400	1.7163	1.2316	1.7162	2.0668	2.0904		1.1901
800	3.2640	2.1376	3.2640	4.0594	3.9720		2.1891
1200	4.6190	2.8380	4.6190	5.5535	5.2831		2.9570

从以上数据可以看出，随着 n 值增加，偏差标准差在逐渐减小；自适应加权方法的偏差和标准差始终小于等权重（w_{equ}）、指数权重（w_{exp}）等六种不同权重方法，且均方误差比值都显著大于1，证明了该方法在精度和偏差方面的优越性。

此外为进一步证明该权重在分类性能上的优化效果，我们给出更多评估指标，以下使用 AUC，RG 作为评估指标并给出模拟结果，如表2所示：

表2：AUC、RG 评估

n	AUC1	AUC2	AUC3	AUC4	AUC5	AUC6	AUC7
400	0.8286	0.8347	0.8347	0.8301	0.8288	0.8346	0.8635
800	0.8284	0.8329	0.8329	0.8280	0.8275	0.8330	0.8612
1200	0.8285	0.8326	0.8326	0.8276	0.8274	0.8327	0.8609
n	RG1	RG2	RG3	RG4	RG5	RG6	RG7
400	0.0632	0.0633	0.0632	0.0633	0.0633	0.0633	0.0635
800	0.0632	0.0633	0.0632	0.0633	0.0633	0.0632	0.0635
1200	0.0631	0.0632	0.0631	0.0632	0.0632	0.0632	0.0635

五、实证研究

本研究使用了两种类型的宏观经济数据 [8]：被解释变量为二值变量 [9]，即低频季度 GDP 衰退指标，核心解释变量为高频月度失业率。选择季度 GDP 衰退指标作为低频被解释变量（二值变量：1= 衰退，0= 非衰退），符合 Logistic 模型的二分类预测场景；选择月度失业率作为高频解释变量，因为失业率与经济增长呈显著负相关，即失业率上升往往预示经济下行，且高频数据能及时捕捉经济波动信号，为 GDP 衰退预测提供时效性支撑。数据从1970年1月1日到2020年12月1日，涵盖了美国多个经济周期，数据来源于美联储经济在线数据（FRED）。由于高频数据和低频数据的频率不匹配，传统的回归方法无法直接应用 [10]。为充分利用高频失业率数据中的信息，以及对二值变量进行处理，本研究采用 Logistic-MIDAS 方法构建预测因子，使用六种不同权重与自适应权重进行对比，结果如表3：

表3：实证结果

Weight	*Metrics*	*Value*
w_{equ}	$RG_{0.1}$	0.1000
	AUC	0.6943

w_{exp}	$RG_{0.1}$	0.2000
	AUC	0.7299
w_{beta}	$RG_{0.1}$	0.3000
	AUC	0.7528
w_{step}	$RG_{0.1}$	0.3000
	AUC	0.7361
$w_{non-zero}$	$RG_{0.1}$	0.3000
	AUC	0.7356
$w_{legendre}$	$RG_{0.1}$	0.3000
	AUC	0.6722
w_{opt}	$RG_{0.1}$	0.4000
	AUC	0.7809

六、结论

针对传统混频数据建模中高频信息损失、参数化权重复杂的问题，本文将逻辑回归模型与混频数据相结合，提出了带自适应权重的 Logistic-MIDAS 模型。通过蒙特卡罗模拟对比6种传统权重与自适应权重的性能，发现自适应权重能显著降低估计偏差和标准差，且 MSE 比值均大于1，验证了其精度优势；基于美国1970-2020年宏观经济数据的实证研究表明，该模型在 AUC(0.7809) 和 $RG_{0.1}$(0.4000) 指标上优于传统权重模型，能更精准识别经济衰退风险。本文提出的模型为经济风险评估提供了新的有效工具，可为政策制定者提前部署干预措施提供参考。

参考文献

[1]Andreou, E., Ghysels, E. Regression models with mixed sampling frequencies[J]. Journal of Econometrics, 2010.

[2] 江丽 . 基于日内高频数据的 Logit 回归模型及其应用研究——以螺纹钢期货为例 [D]. 广州大学 , 2025.

[3] 刘营 . 中国经济增长的高维混频短期预测与精度提升机理研究 [D]. 吉林大学 , 2023.

[4] 许敏 . 基于自适应权重函数的 MIDAS-GARCH 模型及其应用研究 [D]. 广州大学 , 2025.

[5]Audrino, F., Kostrov, A., and Kostrov, J. Predicting U.S. Bank Failures with MIDAS Logit Models[J]. Journal of Financial and Quantitative Analysis, 2019.

[6] 祝子逸 , 朱敏 , 杨爱军等 . 基于稀疏组惩罚混频数据抽样模型的中国宏观经济预测 [J]. 数理统计与管理 , 2024.

[7] 卓杏轩 . 高维非线性混频数据模型及应用研究 [D]. 合肥工业大学 , 2019.

[8] 于洋 . 混频数据回归模型的建模理论、分析技术研究 [D]. 东北财经大学 , 2016.

[9] 刘汉 . 中国宏观经济混频数据模型的研究与应用 [D]. 吉林大学 , 2013.

[10]Ghysels, E. and Qian, H. Estimating MIDAS regressions via OLS with polynomial parameter profiling[J]. Econometrics and Statistics, 2019.

