基于贝叶斯回归的浙江省各项贷款余额预测

方忠玲
（华中科技大学管理学院，湖北　武汉　430074）

贷款在经济社会发展中具有无可替代的作用，在促进经济增长，优化资源配置，支持中小企业的发展，提升居民生活质量，推动金融市场的进步等方面具有重要的意义。同时，信贷政策也政府对国民经济进行宏观调控的重要手段。贷款对促进经济的发展不言而喻。信款为企业的经济活动提供了必要的资金，保证企业在维持再生产、扩大再生产、产业升级、科技创新等方面得以进行，从而推动整体经济的增长。信贷还起到优化资源配置的作用。一方面，它能将闲置资金引导至有需求的领域，提高了资金的使用效率，另一方面，金融机构通过优化信贷结构，优先支持关键行业和新兴产业，促进了经济结构的调整和优化，推动了社会资源的优化配置。信贷对扩大内需、促进消费，让人民群众生活更加美好的作用显而易见。贷款能够使居民进行大额消费，如购房、购车等，不仅可以提升居民的个人生活水平，也刺激了市场需求，带动了相关产业的发展。贷款对金融机构自身的发展，以及金融系统稳定运行也至关重要。金融机构通过不断创新贷款产品和服务，不仅能使自己发展壮大，而且能使金融市场活力，促进金融市场的发展与进步，确保整个金融系统的健康运转。此外，政府可以运用贷款政策对经济进行宏观调控，如在经济低迷时，通过降低利率以刺激经济，在经济过热时，通过提高利率以抑制通胀。各项贷款余额是衡量金融机构信贷投放规模的核心指标，也是观察金融体系运行状况的晴雨表，更是推动经济社会发展的重要引擎。研究分析各项贷款余额的变动规律，对于完善宏观调控政策、优化信贷资源配置、防范化解金融风险、帮助相关部门制定合理的信贷措施、促进金融更好服务于经济高质量发展等方面具有重要的现实意义。贝叶斯回归从统计的角度出发，利用先验知识和观察数据确定参数的概率分布。即首先通过先验分布来表达对模型参数的初始信念，然后根据实际数据来更新这些信念，进而得到模型参数的后验分布，以量化参数的不确定性，且完成对预测值的估计。它不仅可以得到预测的分布，而且还可以确定其置信区间，与传统的回归相比，可获得更多的信息量。具有抗干扰能力强，灵活稳健，避免过拟合，适合小样本等优点，是一种对传统回归分析的有力补充[①②]。运用贝叶斯回归对浙江省各项贷款余额进行预测。

1　贝叶斯回归模型简介

贝叶斯回归是利用贝叶斯定理，通过结合先验分布和似然函数，推导出参数的后验分布，由后验分布推断回归系数的估计值。

若给定数据集 $D=\{(x_i,y_i)\}(i=1,2,\cdots,n),x_i\in R^d$，$y_i\in R$，则 y 对 X 的线性回归可表示为：

$$y=\beta\cdot X+\varepsilon \tag{1}$$

式中，β 为回归系数向量；ε 为回归误差项，通常假设 ε 服从正态分布：

$\varepsilon\sim N(0,\sigma^2 I)$。

传统的观点认为，参数 β 为未知固定值，一般可以通过最小二乘法求出。而贝叶斯方法则认为，

① 李咏影，李伦波.朴素贝叶斯与 Softmax 回归在文本分类上的对比研究[J].电脑知识与技术，2021，17(28)：131-133.
② 徐登可，田瑞琴.函数型空间自回归模型的贝叶斯估计[J].高校应用数学学报，2022，37(3)：323-336.

参数β并非未知固定值，而是服从某一概率分布的随机变量③④。

假设回归系数β的先验分布为多元正态分布：

$\beta \sim N(\mu_0, \Sigma_0)$

其中，μ_0为先验分布的均值向量，通常为0；Σ_0为先验分布协方差，通常为r^2I，r^2为先验分布的方差。

根据贝叶斯理论，回归系数β的后验分布为：

$$P(\beta|D)=\frac{P(D|\beta)\cdot P(\beta)}{P(D)} \tag{2}$$

式中，$P(\beta|D)$为给定输入和输出时模型参数的后验分布，表示在观察到D之后估计β的不确定性；$P(D|\beta)$为在给定回归系数β和自变量X时，因变量y的似然函数，它表达了在不同参数下观测数据出现的可能性的程度；$P(\beta)$为给定输入参数的先验分布；$P(D)$为给定数据集的边际似然或称证据，是一个与β无关的归一化常数⑤⑥。

通俗讲，先验分布$P(\beta)$是反映抽样前对β的认识，后验分布$P(\beta|D)$则是反映抽样后对β的重新认识，之间的差异是由于样本的出现后对β之前认识的一种调整。即后验分布$P(\beta|D)$可以看作是总体信息和样本信息对先验分布$P(\beta)$所作出调整的结果。

贝叶斯回归主要步骤为⑦：

(1)选择先验分布

即选择回归系数β和回归误差ε的方差σ^2的先验分布。

通常假设β的先验分布为多元正态分布：

$\beta \sim N(\mu_0, \Sigma_0)$

σ^2的先验分布为逆伽马分布：

$\sigma^2 \sim G^{-1}(a,b)$

其中，a为形状参数；b为尺度参数。

(2)计算似然函数

$$P(D|\beta)=\prod N(D|\beta) \tag{3}$$

其中，$N(\cdot)$为正态分布的概率密度函数。

(3)计算后验分布

由于$P(D)$为常数，对所以β都相同，因此，只需要比较分子：

$$P(\beta|D) \propto P(D|\beta)\cdot P(\beta) \tag{4}$$

(4)后验分布推断

在先验分布为多元正态分前提下，后验分布亦为多元正态分布：

$\beta \sim N(\mu_n, \Sigma_n)$，

其中，μ_n为后验分布的均值向量；Σ_n为后验分布协方差。为了推动的灵活便捷、准确可靠，后验分布通常采用采用马尔科夫链蒙特卡罗法(MCMC)优化迭代推断，具体如下⑧：

(a)给定β和σ^2初值：$\beta^{(0)}$和$\sigma^{2(0)}$；

(b)对β和σ^2进行采用迭代，即根据它们的分布特性，随机生成候选值$\beta^{(t)}$和$\sigma_{2(t)}$，以及相应的候选样本；

(c)计算新生成数据的接受概率，若接受则保留$\beta^{(t)}$和$\sigma^{2(t)}$及样本，否则拒接舍弃新值，仍保留旧值；

(d)重复步骤(b)–(c)，直到满足收敛条件或达到最大迭代步数；

(e)估计后验分布参数均值：

$$\hat{\beta}=\frac{1}{T-T'}\sum_{t=T-T'}^{T}\beta^{(t)} \tag{5}$$

$$\hat{\sigma}^2=\frac{1}{T-T'}\sum_{t=T-T'}^{T}\sigma^{2(t)} \tag{6}$$

其中，T为总迭代次数；T为拒接新值的迭代次数。

(5)数据点预测

回归系数β求出后，可建立对应的回归方程对数据点进行预测。

③ 胡玉梅,赵明涛,朱家明.基于贝叶斯分位数回归的中国绿色产业发展影响因素研究[J].哈尔滨师范大学学报(自然科学版),2023,39(4):43–57.

④ 孙德红,周亿迎,蒋佳伶.基于贝叶斯岭回归的长江航运服务业集聚动力机制研究[J].中国航海,2024,47(2):48–55.

⑤ 胡玉梅.基于 Knockoff–贝叶斯分位数回归的我国绿色产业发展影响因素研究[D].安徽:安徽财经大学,2024:18–21.

⑥ 俞翰君,赵碗迪,于力超.基于广义非对称拉普拉斯分布的贝叶斯线性混合效应分位数回归模型[J].数理统计与管理,2024,43(5):830–846.

⑦ 袁剑波,郭平,曾恬宁.基于贝叶斯线性回归的造价预测模型[J].武汉理工大学学报(信息与管理工程版),2024,46(1):90–94.

⑧ 朱朋辉,赵全忠,廖志文,等.基于贝叶斯线性回归的鸟害故障分析[J].南京信息工程大学学报(自然科学版),2022,14(2):227–232.

可见，贝叶斯回归模型的基本思想是通过贝叶斯定理，将先验分布与样本数据结合，推导出参数的后验分布，以此来对实现回归系数进行概率估计。后验分布是先验证分布和样本点的折中，后验均值 μ_n 是样本数据和先验信息的加权平均。后验协方差 Σ_n 代表参数估计的精度。与传统的线性回归不同，贝叶斯线性回归提供回归系数的概率分布，而不仅仅是一个确定的数值，包括其参数值、置信区间等，可以对估计结果进行更加全面的评估。

2 浙江省贷款余额预测

贷款的用途多种多样，但最主要的还是用于投资和消费，广义上的投资包括购买生产资料再生产或扩大再生产、投资新项目、原有项目技术改造等等，消费主要是贷款购买大件商品，当然也有普通商品，尤其在当前金融和消费深度融合的背景下，大部分商品可以通过贷款按揭方式购买，实现超前消费，如大到住房、汽车，小到家电、手机、电脑，甚至旅游、教育等服务产品也可以贷款方式满足消费。因此，直接影响贷款数额的是投资和消费。当然，房贷也是重要的消费范畴，是个人贷款的重头戏。居民对住房的刚性需求或者改善性需求，贷款购房贷款买房习以为常。这样会使他们的居住条件得到改善，如居住面积增加，居住环境变好，而城镇居民是房贷的主角。从可操作、可量化的角度出发，这两类类驱动因子选取固定资产投资、社会消费品零售额和城镇居民人均居住面积来代表。贷款用于投资时，带来的结果一方面是推动经济进一步发展，为社会创造更多的财富，从而带动 GDP 增长，另一方面增加了就业岗位，促进居民充分、高质量就业，进而使居民的收入增加。从全面考察的方式来看，因此，驱动因子还增加了国内生产总值(GDP)、全体居民收入、城镇新增就业人数 3 项，那么自变量总个数为 6。

图 1 为 2009–2019 年浙江省本外币各项贷款余额统计数据[1]。2020–2022 年由于全球遭遇受新冠病毒疫情，中国也不例外，而浙江省是疫情重灾区之一。浙江的经济和社会活动受到严重的影响，数据出现异常。2023 年后虽然疫情结束，但仍然未恢复到疫情前的正常状况，导致 2019 年之后的数值或多或少存在失真现象，因此未被采用。从图 1 可见，在 2009–2019 年期间，浙江省各项贷款余额逐年增长态势，从 2009 年的 39234 亿元增长到 2019 年的 121754 亿元，年均复合增长率 15.4%，为经济社会持续健康发展打下了坚实的基础。

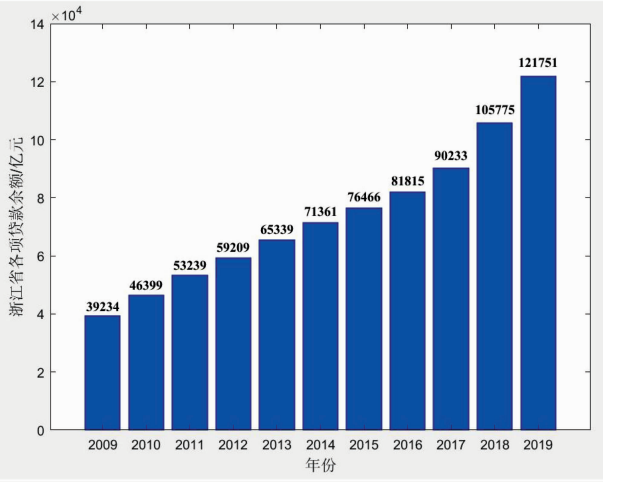

图 1 浙江省各项贷款余额统计数据

设浙江省各项贷款余额为因变量 Y，固定资产投资、社会消费品零售额、国内生产总值、全体居民收入、城镇新增就业人数、城镇居民人均居住面积分别为自变量 x_1、x_2、x_3、x_4、x_5、x_5、x_6，它们的具体值如下所示(数据来源于浙江省统计年鉴)，并令 $X=[x_1,x_2,x_3,x_4,x_5,x_6]$，则 Y 对 X 的回归可表示为式(1)所示的回归方程。

其中：

Y= [39234, 46399, 53239, 59209, 65339, 71361, 76466, 81815, 90233, 105775, 121751];

x_1= [8523, 10123, 12346, 14523, 17346, 20123, 22346, 24523, 26457, 28346, 31166];

x_2= [10742, 12488, 14290, 16665, 17096, 20194, 23555, 29571, 31126, 33336, 36670];

x_3= [22990, 27748, 32363, 34734, 37757, 40173, 42886, 47251, 51768, 58003, 62352];

x_4= [18463, 21193, 24235, 27133, 30010, 32658, 35537, 38529, 42043, 45840, 49899];

x_5= [83.2, 86.5, 90.1, 94.2,98.7, 104.2, 109.8, 116.5, 125.6, 127.8, 129.5]。

[1] (数据来源于 2009–2019 年浙江省国民经济和社会发展统计公报)

x_6= [34.8, 36.9, 37.2, 37.7, 38.1, 39.6, 40.9, 42.3, 43.6, 44.9, 46.0]。

利用贝叶斯原理求各自变量回归系数 β_i (i=1,2,3,4,5,6)。

借助 matlab2021 软件工具，调用其自带 bayeslm 函数，即系统默认采用 MCMC 法中吉布斯采样确定后验分布。参数选择主要为 β 的先验分布函数，之前已假设它为正态分布，其均值为 0，则主要参数为协方差 Σ_0。由于样本数据较大，方差 σ_2 预计比较大，那么先验分布对参数的约束较弱，后验分布主要由数据主导，Σ_0 的取值对求后验分布影响有限，因此，Σ_0 选取不必过于拘泥，经过尝试取 Σ_0=3*I6 效果稍好。同样，样本似然函数很强，导致 σ_2 的形状参数 a 和尺度参数 b 对后验分布估计影响不敏感，故参数 a 和 b 大胆选用，取 a=1，b=1。运行程序得到了回归系数的求解结果，见图 2 所示。

```
Method: Analytic posterior distributions
Number of observations: 11
Number of predictors:   7
Log marginal likelihood: -151.948

             |      Mean          Std              CI95           Positive  Distribution
-------------------------------------------------------------------------------------------
 Intercept   | -18181.7318   29823.7376   [-77448.958, 41085.495]   0.260   t (-18181.73, 27433.85^2, 13)
 Beta(1)     |     -4.5700       1.0457     [-6.648, -2.492]        0.000   t (-4.57, 0.96^2, 13)
 Beta(2)     |     -0.0318       0.3877     [-0.802,  0.739]        0.465   t (-0.03, 0.36^2, 13)
 Beta(3)     |     -1.6734       0.7825     [-3.228, -0.118]        0.018   t (-1.67, 0.72^2, 13)
 Beta(4)     |      8.9281       1.6659     [ 5.618, 12.239]        1.000   t (8.93, 1.53^2, 13)
 Beta(5)     |   -879.7631     217.2648   [-1311.523, -448.004]     0.000   t (-879.76, 199.85^2, 13)
 Beta(6)     |   1207.5332     972.0025     [-724.079, 3139.145]    0.900   t (1207.53, 894.11^2, 13)
 Sigma2      |      1.3238e+06   6.2403e+05  [588680.251, 2907184.549]  1.000   IG(6.50, 1.4e-07)
```

图 2　回归系数求解结果

从图 2 可见，x_1、x_2、x_3、x_4、x_5、x_6 的回归系数为后验分布的均值，其中，β_1=−4.5700，β_2=−0.0318，β_3=−1.6734，β_4=8.9281，β_5=−869.7631，β_6=1297.5332，截距(常数项) b=−18181.7318，回归误差 ε 的方差 σ^2=1.3238×10^6。

图 2 中，std、CI95、Positive、Distribution 分别表示标准差、95%置信区间、回归系数为正的概率、概率分布参数。由于 β 比较多，不方便显示，这里仅展示运用 MCMC 法采样迭代估计 σ^2 的过程及分布图，如图 3 所示。

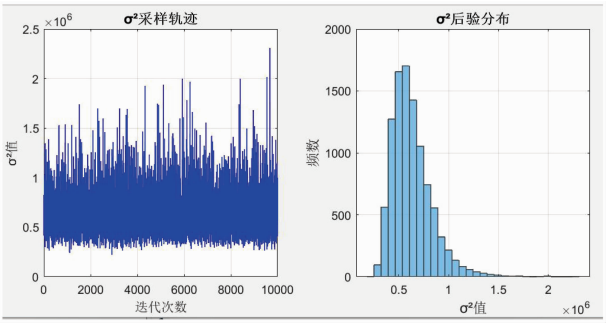

图 3　σ^2 采样迭代过程及分布

于是，可以得到相应的贝叶斯回归方程：

$$Y=-4.5700x_1-0.0318x_2-1.36734x_3+8.9281x_4-869.7631x_5+1296.5332x_6-18181.7318 \quad (7)$$

由方程(7)得到 2009–2019 年浙江省各项贷款余额的预测值，结果如表 1 所示。

回归的均方根误差为 1119.42，相关系数为 0.9979，平均预测误差为 1.4088%。反映模型有较高拟合精度。

需要提醒的是：贝叶斯回归属于概率统计模型，回归的系数主要具有统计学上的意义，重点是从拟合精度上来考量。由于指标数据都是随着时间增长而单调增长，因此，数据之间存在多重共线性，导致回一些回归出现负数，失去了解释意义的，这不在贝叶斯回归的研究范畴。要想得到解释意义，就要消除共线性影响，则可用岭回归、偏最

表 1　贝叶斯回归与 OLS 回归结果

年份	实际值	贝叶斯回归值	误差/%	OLS 回归值	误差/%
2009	39234	37582.72139	−4.20879	37719.72081	3.85961
2010	46399	46719.76902	0.69133	46396.33139	0.00575
2011	53239	52774.83969	−0.87184	52811.45363	0.80307
2012	59209	61576.87202	3.99918	61689.71743	−4.18976
2013	65339	65745.46913	0.62209	65926.43297	−0.89905
2014	71361	69762.17794	−2.24047	69708.24883	2.31604
2015	76466	77428.69394	1.25898	77249.34849	−1.02444
2016	81815	82220.75068	0.49594	82313.54724	−0.60936
2017	90233	90761.21985	0.58540	90804.14132	−0.63296
2018	105775	105183.36917	−0.55933	105201.59768	0.54230
2019	121751	121065.11717	−0.56335	121002.27838	0.61496
平均误差/%	–		1.4633	–	1.4088

小二乘法回归、主成分回归等模型来解决。

为了检验贝叶斯回归的性能，以上述相同的数据，建立 Y 和 X 的回归方程，运用最小二乘法(OLS)求自变量的回归系数，结果为：

γ_1=–4.6001，γ_2=–0.1180，γ_3=–1.7941，γ_4=9.0540，γ_5=–908.9622，γ_6=–1676.8043，c=–30590.24422。

回归的均方根误差为 1110.46，相关系数为 0.9978，平均预测误差为 1.4633%。于是得到相应的 OLS 回归方程：

$$Y=-4.6001x_1-0.1180x_2-1.7941x_3+9.0540x_4-908.9622x_5-1676.8043x_6-30590.24422 \quad (8)$$

由方程(8)得到 2009–2019 年浙江省各项贷款余额的预测值，结果如表 1 所示。

可见，尽管 OLS 回归的精度也比较高，但仍比贝叶斯回归还是要稍逊一筹，贝叶斯回归的平均预测比 OLS 的减小了 3.7245%，且贝叶斯回归获得回归信息更为丰富。之所以贝叶斯回归精度提高不大，主要是样本数量太小，如果样本数量足够大，贝叶斯回归优势就会更加充分显现出来。

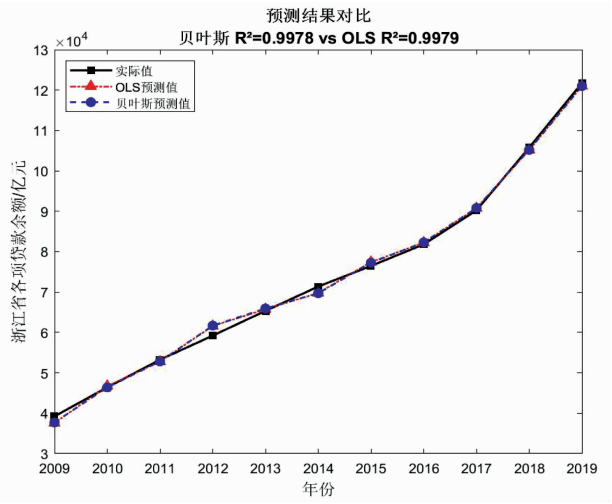

图 4　浙江省各项贷款余额预测曲线

预测曲线如图 4 所示，由于两个模型的预测精度都很高，因此，三条曲线几乎重合。

至于对样本外的数据预测，只要知道自变量的值，将其代入方程(7)就可以直接求出因变量的值(异常数据年份慎用)。

3　结语

金融机构各项贷款余额是优化信贷资源配置、提升金融服务实体经济能力、提高人民生活质量等的重要基础。科学预测对地区各项贷款余额进行预测，能够使金融机构更加精准地把握信贷需求变化趋势，制定针对性的信贷政策和措施，提高信贷的有效性，更好发挥金融服务于经济社会的发展功能。信贷工作的核心是提高资金使用效率，优化信贷结构，支持实体经济提质增效，为经济高质量发展创造了良好的金融环境。金融机构认真履行职责，强化使命担当，不断提升金融服务专业性、主动性、创造性，为经济社会发展提供优质服务的需要。要积极应用金融科技，加快产业数字化转型，利用大数据、人工智能等技术手段，精准预测信贷需求变化趋势，制定差异化的信贷政策，引导资金流向实体经济重点领域和薄弱环节，特别是支持小微企业、科技创新和绿色发展等，优化信贷结构，提升资金使用效率。同时，提高服务效率，降低运营成本，增强风险识别和防控能力。要建立完善的风险预警和处置机制，及时有效防范和化解金融风险，为金融发展营造安全稳定的环境。要加快构建高效、安全、普惠，具有中国特色的现代金融体系，充分发挥金融对实体经济的支撑作用，推动我国经济高质量发展。贝叶斯回归是回归分析的重要方法之一。其核心是将模型的参数视为随机变量，通过贝叶斯定理，将先验知识和观测数据相结合，推断参数的后验概率分布，继而对目标值作出预测。在不确定量化、先验知识整合、避免过拟合方面等具有独特的优点，适用于多种复杂场景。估计的参数全面系统，信息量丰富，稳定可靠。运用贝叶斯回归方法对浙江省各项贷款余额进行了预测，取得了良好的效果，在样本数量较小的情况下，模型的平均预测误差仅为 1.4088%，比最小二乘法的减小了 3.7245%，而且获取的回归信息量更为丰富。如果样本再大一些，模型的优势会表现得更加显著。

（责任编辑：元小佩）