МОДЕЛЬ КРЕДИТНОГО РИСКА НА ОСНОВЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ С ИЗМЕНЯЮЩИМИСЯ ВО ВРЕМЕНИ ПАРАМЕТРАМИ

А. С. Сорокин

Российский экономический университет им. Г. В. Плеханова, Москва, Россия
E-mail: alsorokin@mail.ru

Работа посвящена разработке математического подхода для решения проблемы изменения со временем распределения предикторов в модели логистической регрессии, используемой для оценки кредитного риска. Такие изменения приводят к нарушение стабильности скоринговой модели при оценке вероятности дефолта. Для решения данной проблемы необходимо периодически перестраивать модель логистической регрессии для оценки кредитного риска по новым накопленным историческим данным. Автором предложен новый метод на основе изменяющихся во времени параметров скоринговых моделей. Такой подход обладает рядом преимуществ по сравнению с классическим подходом перестроения скоринговой модели с течение времени. В основе предлагаемого метода – прогнозирование коэффициентов модели логистической регрессии с использованием комбинации моделей класса ARIMA, DCC-CARCH и пространственно-временных моделей.

CREDIT RISK MODEL BASED ON LOGISTIC REGRESSION WITH TIME-VARYING PARAMETERS

A. S. Sorokin

The paper is devoted to the development of a mathematical approach to solve the problem of changes in the distribution of predictors over time in a logistic regression model used to assess credit risk. Such changes lead to a violation of the stability of the scoring model when assessing the probability of default. To solve this problem, it is necessary to periodically rebuild the logistic regression model to assess credit risk based on new accumulated historical data. The author proposes a new method based on time-varying parameters of scoring models. This approach has a number of advantages over the classical approach of rebuilding the scoring model over time. The proposed method is based on predicting the coefficients of a logistic regression model using a combination of ARIMA, DCC-CARCH class models and a state-space model.

Кредитному скорингу посвящено множество научных работ, как отечественных, так и зарубежных авторов. В качестве скоринговой модели может быть выбрана статистическая модель или модель машинного обучения: линейный дискриминантный анализ [1], деревья решений [2], анализ Марковских цепей [3,4], [5] и логистическая регрессия [5, 6]. Логистическая регрессия является классическим и наиболее распространенными методом при прогнозировании дефолта заемщика в кредитном скоринге.

Тематикой данного исследования является моделирование вероятности дефолта заемщика на базе логистической регрессии с динамически изменяю-

щимися параметрами. Данная тематика является новой и глубоко не проработанной в научной литературе. Среди отечественной литературы есть статьи, которые содержат обзоры применяемых в скоринге методов, как, например, статья [7], в которой показан широкий набор современных методов в задаче кредитного скоринга. Однако, тема динамически меняющихся коэффициентов в ней не освещается, методы представлены отдельно друг от друга, и не рассматривается их комбинация, например, [8-10]. Динамические модели распространены в прогнозировании макроэкономических процессов. Динамическое моделирование риска получило широкое распространение в управлении финансовыми рисками [11]. Динамические модели, популярные в академической среде и редко используемые в бизнесе ввиду своей сложности, дают более точные оценки, чем более простые статические модели, основанные на исторических данных.

В литературе по кредитному скорингу данная тематика практически не освещается, хотя она очень популярна в прогнозировании временных рядов [12-14]. В кредитном скоринге этот подход может повысить эффективность применяемой модели логистической регрессии или альтернативных моделей [15, 16].

Таким образом, в отечественной научной среде поднималась проблема учета динамических факторов в задаче кредитного скоринга, рассматривались методы машинного обучения для решения этой задачи. Однако, редко встречаются попытки эконометрического моделирования и разработки соответствующего математического подхода для учета динамических параметров. Не рассматривался вопрос меняющихся во времени коэффициентов регрессии. Также и в иностранной литературе данная тематика набирает популярность, но тем не менее не разработаны соответствующие модели, которые можно было бы применить в практике кредитных организаций.

Рассмотрим более подробную идею предлагаемого метода, более подробное изложение которого с эмпирическим доказательством и имитационным моделированием представлено автором в работе [17].

Рассмотрим выборку $\{y_t, X_t : t = 1,\ldots,T\}$, где $y_t = (y_{1t}, y_{2t}, \ldots, y_{nt})^T$ – бинарная целевая переменная, которая равна единице в случае дефолта по кредиту и нулю в противном случае, а $X_t = (1, x_{1t}, x_{2t}, \ldots, x_{mt})$ – это вектор объясняющих переменных в разные моменты времени, где $x_{it} = (x_{1it}, x_{2it}, \ldots, x_{nit})^T$. Предположим, что вероятность дефолта может быть смоделирована с помощью логистической функции следующим образом:

$$p(1 \mid X_t) = \frac{1}{1 + e^{-X_t \beta_t}} \tag{1}$$

где β_t – вектор изменяющихся во времени параметров.

Динамику каждого параметра β_{it} можно смоделировать либо с помощью модели сезонного тренда, либо с помощью некоторой модели класса ARIMA вида:

$$\hat{\beta}_{it} = \alpha_0 + \alpha_1 \hat{\beta}_{i(t-1)} + \ldots + \alpha_p \hat{\beta}_{i(t-p)} + \gamma_1 \hat{\varepsilon}_{i(t-1)} + \ldots + \gamma_q \hat{\varepsilon}_{i(t-q)} + \hat{\varepsilon}_{it}. \tag{2}$$

Теоретическая гипотеза состоит в том, что если вместо полученных параметров в логистической регрессии использовать прогнозируемые по уравнению (2) значения модель будет более эффективной:

$$p(1 \mid X_t) = \frac{1}{1+e^{-X_t \hat{\beta}_t}}. \quad (3)$$

Более того, в каждый период времени получаются две независимые оценки вектора истинных параметров β_t: одна – из уравнения прогнозирования (3), а другая – из подгонки модели в соответствии с уравнением (1). Таким образом, эти две оценки с помощью фильтра Калмана могут быть объединены для получения более точного вектора параметров. Функция плотности вероятности для таких параметров может быть вычислена следующим образом:

$$f(\beta_t \mid \hat{\beta}_{t-1}, ..., \hat{\beta}_{t-p}, X_t) = \frac{f_1(\beta_t \mid \hat{\beta}_{t-1}, ..., \hat{\beta}_{t-p}) f_2(\beta_t \mid X_t)}{\int_{\mathbf{R}} f_1(\beta_t \mid \hat{\beta}_{t-1}, ..., \hat{\beta}_{t-p}) f_2(\beta_t \mid X_t) d\beta_t}, \quad (4)$$

где:

$$f_1(\beta_t \mid \hat{\beta}_{t-1}, ..., \hat{\beta}_{t-p}) = (2\pi)^{-\frac{m+1}{2}} \det(\sum_t)^{-\frac{1}{2}} e^{-\frac{1}{2}(\beta_t - \hat{\beta}_t)^T \sum_t^{-1} (\beta_t - \hat{\beta}_t)}, \quad (5)$$

$$f_2(\beta_t \mid X_t) = (2\pi)^{-\frac{m+1}{2}} \det(\Omega_t)^{-\frac{1}{2}} e^{-\frac{1}{2}(\beta_t - \tilde{\beta}_t)^T \Omega_t^{-1} (\beta_t - \tilde{\beta}_t)}. \quad (6)$$

Ковариационная матрица вектора параметров Ω_t из уравнения (6) обычно оценивается информационной матрицей Фишера следующим образом:

$$\Omega_t = X_t^T \tilde{W} X_t, \quad (7)$$

где:

$$\tilde{W} = diag\left(\frac{e^{\sum_{j=0}^{m} \tilde{\beta}_{jt} x_{1jt}}}{\left(1+e^{\sum_{j=0}^{m} \tilde{\beta}_{jt} x_{1jt}}\right)^2}, ..., \frac{e^{\sum_{j=0}^{m} \tilde{\beta}_{jt} x_{njt}}}{\left(1+e^{\sum_{j=0}^{m} \tilde{\beta}_{jt} x_{njt}}\right)^2} \right). \quad (8)$$

Ковариационная матрица вектора параметров Σ_t из уравнения (5) может быть получена с помощью слегка модифицированной модели DCC-GARCH.

$$\sum_t = D_t R_t D_t, \quad (9)$$

$$D_t = \begin{bmatrix} \sqrt{h_{1t}} & 0 & \cdots & 0 \\ 0 & \sqrt{h_{2t}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{h_{mt}} \end{bmatrix}, \quad (10)$$

где каждый h_{it} представляет собой дисперсию численно сгенерированной функции плотности вероятности $(pdf_{it}(\beta_{it}))$, которая представляет собой среднее значение из *N* функций плотности вероятности для прогнозируемого значения β_{it}, полученное из модели (2). Причиной численной генерации, в данном случае, является тот факт, что у нас нет истинных значений β_{it}, мы имеем их оценки с некоторой долей неопределенности. Вот почему мы генерируем результаты $\beta_{t-1}, \ldots, \beta_{t-p}$ на основе их pdf, полученных на предыдущих шагах, чтобы получить

соответствие $pd(\beta_{it})$ следующим образом:

$$pdf_{it}(\beta_{it}) = \sum_{j=1}^{N} pdf_{ijt}(\beta_{ijt}) / N \quad . \tag{11}$$

Поскольку R_t представляет собой условную корреляционную матрицу для β_t и выглядит следующим образом:

$$R_t = \begin{bmatrix} 1 & \rho_{12t} & \cdots & \rho_{1nt} \\ \rho_{21t} & 1 & \ddots & \rho_{2nt} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{n1t} & \rho_{n2t} & \cdots & 1 \end{bmatrix}. \tag{12}$$

тогда из уравнений (10) и (12) каждый элемент Σ_t может быть представлен следующим образом:

$$\left[\sum_t\right]_{ij} = \sqrt{h_{it}h_{jt}}\rho_{ijt}\,. \tag{13}$$

Корреляционную матрицу R_t можно смоделировать следующим образом:

$$R_t = Q_t^{*-1} Q_t Q_t^{*-1}, \tag{14}$$

$$Q_t = (1-a)\bar{\Omega} + a\Omega_{t-1}, \tag{15}$$

где:

$$\bar{\Omega} = \frac{1}{T}\sum_{t=1}^{T}\Omega_t\,, \tag{16}$$

$$Q_t^* = \begin{bmatrix} \sqrt{q_{11t}} & 0 & \cdots & 0 \\ 0 & \sqrt{q_{22t}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{q_{mmt}} \end{bmatrix}. \tag{17}$$

Здесь R_t разлагается на Q_t^{*-1} и Q_t для обеспечения того, чтобы абсолютные значения всех записей были меньше или равны единице. Для оценки параметров модели DCC-GARCH может быть использован метод оценки максимального правдоподобия.

Получив pdf для β_t, скорректированные оценки вычисляются путем максимизации правдоподобия уравнения (4):

$$\hat{\beta}_t = \underset{\beta_t}{\operatorname{argmax}}\, f(\beta_t \mid \hat{\beta}_{t-1}, ..., \hat{\beta}_{t-m}, X_t)\,. \tag{18}$$

Эти скорректированные оценки и pdf затем используются в уравнении прогнозирования (2) для получения прогнозов для следующих значений истинных параметров β_{t+1}. Стоит отметить, что, если каждый β_{it} подчиняется или может быть аппроксимирован нормальным распределением и Σ_t, Ω_t – диагональные, скорректированная плотность вероятности также нормальна, и уравнение (4) может быть записано явно. Чтобы показать, что в исследовании вводятся следующие обозначения для упрощения дальнейших вычислений: $E(\beta_{it} \mid \hat{\beta}_{it-1}, ..., \hat{\beta}_{it-m}) = \mu_1$, $E(\beta_{it} \mid X_t) = \mu_2$.

$\operatorname{var}(\beta_{it} \mid \hat{\beta}_{i,t-1}, ..., \hat{\beta}_{i,t-m}) = \sigma_1^2$, $\operatorname{var}(\beta_{it} \mid X_t) = \sigma_2^2$. Затем pdf для β_{it} из уравнения (4)

можно переписать, как показано ниже:

$$pdf(\beta_{it}) = \frac{\frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(\beta_{it}-\mu_1)^2}{2\sigma_1^2}-\frac{(\beta_{it}-\mu_2)^2}{2\sigma_2^2}}}{\int_{\mathbf{R}} \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(\beta_{it}-\mu_1)^2}{2\sigma_1^2}-\frac{(\beta_{it}-\mu_2)^2}{2\sigma_2^2}} d\beta_{it}}. \quad (19)$$

Чтобы показать, что этот pdf является нормальным, важно сначала рассмотреть только числитель этой дроби. Упрощение этого выражения даст результат, показанный ниже:

$$\frac{1}{2\pi\delta_1\delta_2} e^{-\frac{(\beta_{it}-\mu_1)^2}{2\delta_1^2}-\frac{(\beta_{it}-\mu_2)^2}{2\delta_2^2}} = \frac{1}{2\pi\delta_1\delta_2} e^{-\frac{\beta_{it}^2(\delta_1^2+\delta_2^2)-2\beta_{it}(\mu_1\delta_2^2+\mu_2\delta_1^2)+\mu_1^2\delta_2^2+\mu_2^2\delta_1^2}{2\delta_1^2\delta_2^2}} = \frac{1}{2\pi\delta_1\delta_2} e^{-\frac{\left(\beta_{it}-\frac{\mu_1\delta_2^2+\mu_2\delta_1^2}{\delta_1^2+\delta_2^2}\right)^2+\frac{\mu_1\delta_2^2+\mu_2\delta_1^2}{\delta_1^2+\delta_2^2}-\left(\frac{\mu_1\delta_2^2+\mu_2\delta_1^2}{\delta_1^2+\delta_2^2}\right)^2}{2\frac{\delta_1\delta_2}{\delta_1^2+\delta_2^2}}}. \quad (20)$$

После интегрирования знаменателя дроби в уравнении (19) и сокращения $\frac{\mu_1\delta_2^2+\mu_2\delta_1^2}{\delta_1^2+\delta_2^2} - \left(\frac{\mu_1\delta_2^2+\mu_2\delta_1^2}{\delta_1^2+\delta_2^2}\right)^2$, получено следующее выражение для функции плотности вероятности:

$$pdf(\beta_{it}) = \frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sqrt{2\pi}\sigma_1\sigma_2} e^{-\frac{\left(\beta_{it}-\frac{\mu_1\sigma_2^2+\mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2}{2\frac{\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}}. \quad (21)$$

Что доказывает, что скорректированный pdf для β_{it} также является нормальным со средним значением, показанным в уравнении (22), и дисперсией, показанной в уравнении (23)

$$\mu = \frac{\mu_1\sigma_2^2+\mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}, \quad (22)$$

$$\sigma^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}. \quad (23)$$

Из уравнения (23) ясно, что $\sigma^2 < \min(\sigma_1^2, \sigma_2^2)$, из чего следует, что, применяя этот метод, получается лучшая оценка β_{it}. Поэтому уравнение выражения (4) может быть заменено прямой и более простой в вычислении формой, учитывая справедливость вышеупомянутых предположений.

Таким образом, автором предлагается новый метод работы с изменяющимися во времени параметрами модели логистической регрессии, используемой для оценки вероятности дефолта заемщика. В данной статье изложено теоретическое обоснование метода. В работе [17] также эмпирически было показано, что в постоянно меняющейся экономической среде влияние факторов на целевую переменную также меняется. Таким образом, прогнозирование коэффициентов

дает лучший финансовый результат, чем простое применение параметров, полученных на основе накопленной статистики за прошлые периоды времени

СПИСОК ЛИТЕРАТУРЫ

1. *Bansal G., Sinha A. P., Zhao H.* Tuning data mining methods for cost-sensitive regression: A study in loan charge-off forecasting // J. Manag. Inf. Syst. 2008. Vol. 25. P. 315-336.
2. *Zhang H., Legro R. S., Zhang J., Zhang L., Chen X., Huang H., Casso P. R., Schlaff W. D., Diamond M. P.; Krawetz S. A., et al.* Decision trees for identifying predictors of treatment effectiveness in clinical trials and its application to ovulation in a study of women with polycystic ovary syndrome // Hum. Reprod. 2010. Vol. 25. P. 2612-2621.
3. *Smith L. D., Lawrence E. C.* Forecasting losses on a liquidating long-term loan portfolio // J. Bank. Financ. 1995. Vol. 19. P. 959-985.
4. *Greenidge K., Grosvenor T.* Forecasting non-performing loans in Barbados. // Financ. Econ. Emerg. Econ. 2010. Vol. 5. P. 80-108.
5. *Сорокин А. С.* Практика построения скоринговых карт с использованием модели логистической регрессии // Интернет-журнал «Науковедение». 2014. № 2 (21). С. 82.
6. *Darroch J. N., Ratcliff D.* Generalized iterative scaling for log-linear models // Ann. Math. Stat. 1972. Vol. 43. P. 1470-1480.
7. *Волкова Е. С., Гисин В. Б., Соловьев В. И.* Современные подходы к применению методов интеллектуального анализа данных в задаче кредитного скоринга // Финансы и кредит. 2017. № 34 (754). С. 2044-2060.
8. *Исаев Д. В.* Динамическое ансамблевое обучение для оценки кредитоспособности // Инновации и инвестиции. 2022. № 3. С. 74-79.
9. *Широбокова М. А.* Модель оценки риска дефолта на всем протяжении жизни кредита // Вестник Удмуртского университета. Серия «Экономика и право». 2018. С. 228-233.
10. *Гришин А. А., Строев С. П.* Разработка модели поведенческого скоринга с использованием методов градиентного бустинга // Научно-технический вестник Поволжья. 2018. № 9. С. 93-98.
11. *Carol A., Yang H., Xiaochun M.* Static and dynamic models for multivariate distribution forecasts: Proper scoring rule tests of factor-quantile versus multivariate GARCH models // International Journal of Forecasting. 2022.
12. *Bitto A., Frühwirth-Schnatter S.* Achieving shrinkage in a time-varying parameter model framework. // Journal of Econometrics. 2019. Vol. 210. P. 75-97.
13. *Chan J. C., Eisenstat E.* Bayesian model comparison for time-varying parameter VARs with stochastic volatility. // Journal of Econometrics. 2018. Vol. 33. P. 509-532.
14. *Kalli M., Griffin J. E.* Time-varying sparsity in dynamic regression models // Journal of Econometrics. 2014. 178 (2). PP. 779-793.
15. *Orlando G., Pelosi R.* Non-performing loans for Italian companies: When time matters: An empirical research on estimating probability to default and loss given default // International Journal of Financial Studies. 2020. Vol. 8. P. 68.
16. *Aslan A., Poppe L., Posch P.* Are sustainable companies more likely to default? Evidence from the dynamics between credit and ESG ratings // Sustainability. 2021. Vol. 13. P. 8568.
17. *Moiseev N., Sorokin A., Zvezdina N., Mikhaylov A., Khomyakova L., Mir Sayed Sh. D.* Credit risk theoretical model on the base of DCC-GARCH in time-varying parameters framework // Mathematics. 2021. Т. 9. № 19. С. 2423.