

Вестник Череповецкого государственного университета. 2024. № 3 (120). С. 46–54.
Cherepovets State University Bulletin, 2024, no. 3 (120), pp. 46–54.

Научная статья

УДК 330.4

<https://doi.org/10.23859/1994-0637-2024-3-120-4>

EDN: YFPREW

Выбор оптимальных лагов и оценивание параметров линейной регрессионной модели с запаздывающими переменными

Сергей Иванович Носков¹, Егор Сергеевич Попов^{2✉}

^{1,2} Иркутский государственный университет путей сообщения,
Иркутск, Россия

¹sergey.noskov.57@mail.ru, <https://orcid.org/0000-0003-4097-2720>

^{2✉}eglr5732@mail.ru, <https://orcid.org/0009-0002-2307-0995>

Аннотация. В статье дан краткий обзор работ по использованию запаздывающих (лаговых) переменных при моделировании сложных объектов. В частности, рассмотрены: смещение оценки метода наименьших квадратов регрессионной модели с лаговой зависимой переменной и автокоррелированными возмущениями; способ включения запаздывающих переменных в регрессионную модель с использованием пошаговой процедуры; решение проблемы параметрической идентификации в моделях дискретного выбора панельных данных, когда набор объясняющих переменных включает экзогенные переменные, лаги зависимой переменной, а также ненаблюдаемые индивидуальные эффекты. Разработана алгоритмическая схема выбора оптимальных временных лагов для независимых переменных и оценивания неизвестных параметров линейной регрессионной модели с запаздыванием. При выборе в качестве расстояния между векторами фактических и расчетных значений зависимой переменной манхэттенской метрики, соответствующей методу наименьших модулей, данный способ сводится к решению задачи линейно-булевого программирования приемлемой для практических ситуаций размерности. Построена регрессионная модель с запаздыванием для описания динамики количества дорожно-транспортных происшествий, которая обладает меньшими ошибками аппроксимации по сравнению с обычной линейной моделью при использовании методов наименьших квадратов и модулей.

Ключевые слова: регрессионная модель, переменные с запаздыванием, лаги, дорожно-транспортные происшествия, метод наименьших модулей, задача линейно-булевого программирования

Для цитирования: Носков С. И., Попов Е. С. Выбор оптимальных лагов и оценивание параметров линейной регрессионной модели с запаздывающими переменными // Вестник Череповецкого государственного университета. 2024. № 3 (120). С. 46–54. <https://doi.org/10.23859/1994-0637-2024-3-120-4>; EDN: YFPREW

Selection of optimal lags and estimation of parameters of a linear regression model with lagging variables

Sergey I. Noskov¹, Egor S. Popov^{2✉}

^{1, 2} Irkutsk State Transport University,
Irkutsk, Russia

¹sergey.noskov.57@mail.ru, <https://orcid.org/0000-0003-4097-2720>

^{2✉}eglr5732@mail.ru, <https://orcid.org/0009-0002-2307-0995>

Abstract. The paper provides a brief overview of works on the use of lagging variables in modeling complex objects. In particular, the following are considered: bias in the least squares estimation of a regression model with a lagged dependent variable and autocorrelated disturbances; a method for including lagged variables in a regression model using a stepwise procedure; solving the problem of parametric identification in discrete choice models of panel data, when the set of explanatory variables includes exogenous variables, lags of the dependent variable, as well as unobserved individual effects. An algorithmic scheme has been developed for selecting optimal time lags for independent variables and estimating unknown parameters of a linear regression model with a lag. When choosing the Manhattan metric corresponding to the method of least moduli as the distance between the vectors of actual and calculated values of the dependent variable, this method is reduced to solving a linear-Boolean programming problem of a dimension acceptable for practical situations. A regression model with lag was developed to describe the dynamics of the number of road accidents, which has smaller approximation errors than the usual linear model when using least squares and moduli methods.

Keywords: regression model, lagged variables, lags, traffic accidents, least modulus method, linear-Boolean programming problem

For citation: Noskov S. I., Popov E. S. Selection of optimal lags and estimation of parameters of a linear regression model with lagging variables. *Cherepovets State University Bulletin*, 2024, no. 3 (120), pp. 46–54. (In Russ.) <https://doi.org/10.23859/1994-0637-2024-3-120-4>; EDN: YFPREW

Введение

При построении математических моделей сложных объектов характер их функционирования может предполагать необходимость использования запаздывающих (лаговых) переменных. Так, в работе А. Маеширо¹ утверждается, что смещение оценки метода наименьших квадратов (МНК) регрессионной модели с лаговой зависимой переменной и автокоррелированными возмущениями определяется двумя эффектами: динамическим эффектом и эффектом корреляции, которые могут быть усиливающими или компенсирующими. Последствия этих двух эффектов исследуются с помощью метода Монте-Карло. В работе Y. Yanti и S. Rahardianto² включе-

¹ Maeshiro A. Teaching Regressions with a Lagged Dependent Variable and Autocorrelated Disturbances // *The Journal of Economic Education*. 1996. Vol. 27, is. 1. P. 72–84.

² Yanti Y., Rahardianto S. Stepwise Approach in Lagged Variables Time Series Modeling: A Simple Illustration // *Materials Science and Engineering. The 7th International Conference on Global Optimization and Its Application (ICoGOIA 2018)* (August 30–31, 2018, Bali, Indonesia). 2019. Vol. 621. P. 012009.

ние запаздывающих переменных в регрессионную модель осуществляется с использованием пошагового МНК. В результате можно определить значимость переменных-предикторов в модели, а также сформировать соответствующие критерии согласия при прогнозировании. В статье L. Kelly и N. J. Kelly¹ справедливо отмечается, что запаздывающая зависимая переменная в регрессии часто используется как средство учета динамических эффектов и как способ избавления модели от автокорреляции. Но если в модели присутствует остаточная автокорреляция, лаговая зависимая переменная вызывает смещение коэффициентов для объясняющих переменных в сторону понижения. На основе соответствующего анализа разрабатывается несколько практических предложений о том, когда и как использовать зависимые переменные с запаздыванием в правой части модели. В работе В. Е. Нопоре и Е. Кюриазидзу² рассматривается решение проблемы параметрической идентификации в моделях дискретного выбора панельных данных, когда набор объясняющих переменных включает экзогенные переменные, лаги зависимой переменной, а также ненаблюдаемые индивидуальные эффекты. Показано, что в результате оценки оказываются непротиворечивыми и асимптотически нормальными, хотя скорость их сходимости медленнее, чем обратная величина квадратного корня из размера выборки. Статья J. Z. Huang и H. Shein³ посвящена описанию метода глобального сглаживания, основанного на полиномиальных сплайнах, для оценки моделей регрессии функциональных коэффициентов для нелинейных временных рядов. Обсуждаются способы автоматического выбора пороговой переменной и значимых переменных с лагами. Оценочная модель используется для построения прогнозов на несколько шагов вперед, включая интервальные прогнозы и прогнозы плотности. Методика иллюстрируется двумя примерами с реальными данными. В работе J. Anatolyev⁴ отмечается, что в условно гетероскедастической регрессии временных рядов с мартингаловыми разностными ошибками использование запаздывающих регрессоров может не повысить эффективность оценки по сравнению с обычным МНК. Приводится пример аналогичного явления в модели с серийно коррелированными ошибками, где оптимальная оценка инструментальных переменных асимптотически столь же эффективна, как и оценка инструментальных переменных, построенная при игнорировании наличия условной гетероскедастичности. О возможности построения регрес-

¹ Keele L., Kelly N. J. Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables // *Political Analysis*. 2006. Vol. 14, no. 2. P. 186–205.

² Honoré B. E., Kyriazidou E. Panel Data Discrete Choice Models with Lagged Dependent Variables // *Econometrica*. 2000. Vol. 68, no. 4. P. 839–874.

³ Huang J. Z., Shen H. Functional Coefficient Regression Models for Non-Linear Time Series: A Polynomial Spline Approach // *Scandinavian Journal of Statistics*. 2004. Vol. 31, no. 4. P. 515–534.

⁴ Anatolyev S. Redundancy of Lagged Regressors Revisited // *Econometric Theory*. 2007. Vol. 23, no. 2. P. 364–368.

сионной модели с запаздывающими переменными упоминается и в работе М. П. Базилевского и С. И. Носкова¹.

Основная часть

Рассмотрим описывающую некоторый динамический процесс линейную регрессионную модель с запаздывающими независимыми переменными общего вида:

$$y_t = \alpha_0 + \sum_{i=1}^m \alpha_i x_{(t-\tau_i),i} + \varepsilon_t, \quad (1)$$

где y – зависимая, а x_i – i -ая независимая переменные; α_i , $i = \overline{0, m}$ – неизвестные идентифицируемые параметры; ε_t – ошибки аппроксимации; t – номер наблюдения (момент времени); τ_i , $i = \overline{1, m}$ – подлежащие определению оптимальные лаги запаздывания, при этом каждый лаг τ_i может принимать любое значение из набора $T_i = \{0, 1, \dots, \bar{\tau}\}$. Будем считать заданными значения независимых переменных для всех $t = \overline{-\tau+1, n}$, а зависимой для $t = \overline{1, n}$.

Заметим, что модель (1) является одним из возможных динамических аналогов обычной линейной регрессии (см., например, работу С. И. Носкова²).

Введем обозначения: $\alpha = (\alpha_1, \dots, \alpha_m)$, $\tau = (\tau_1, \dots, \tau_m)$.

Поставим задачу определения неизвестных параметров и оптимальных лагов запаздывания путем минимизации городского (манхэттенского) расстояния между заданными и вычисленными по модели значениями зависимой переменной, соответствующего методу наименьших модулей (МНМ)³.

Другими словами, необходимо минимизировать функцию $J(\alpha, \tau)$:

$$J(\alpha, \tau) = \sum_{t=1}^n \left| y_t - \alpha_0 - \sum_{i=1}^m \alpha_i x_{(t-\tau_i),i} \right| \rightarrow \min. \quad (2)$$

Для решения задачи воспользуемся приемом сведения задачи построения так называемой аддитивной по параметрам регрессионной зависимости к задаче ЛБП⁴.

¹ Базилевский М. П., Носков С. И. Методические и инструментальные средства построения некоторых типов регрессионных моделей // Системы. Методы. Технологии. 2012. № 1 (13). С. 80–87.

² Носков С. И. Применение многокритериального метода наименьших модулей для моделирования количества дорожно-транспортных происшествий // Ученые записки Комсомольского-на-Амуре государственного технического университета. 2023. № 5 (69). С. 30–35.

³ Там же.

⁴ Носков С. И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных: монография / ответственный редактор А. И. Тятюшкин. Иркутск: Облинформпечать, 1996. 320 с.

Введем в рассмотрение новый вектор неизвестных $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_{1+m(\bar{\tau}+1)})$ и матрицу наблюдений $Z = \|z_{ty}\|$, $t = \overline{1, n}$, $i = \overline{1, 1+m(\bar{\tau}+1)}$ для новых предикторов, сформированную следующим образом:

$$z_{t1} = 1, \quad t = \overline{1, n},$$

$$z_{t, 2+j} = x_{t-j, i}, \quad i = \overline{1, m}, \quad j = \overline{0, \bar{\tau}}, \quad t = \overline{1-j, n-j}.$$

Построим линейное регрессионное уравнение

$$y_t = \sum_{i=1}^{1+m(\bar{\tau}+1)} \beta_i z_{ti} + \delta_t, \quad t = \overline{1, n}, \quad (3)$$

в котором для каждой исходной независимой переменной x_i лишь один из соответствующих ей параметров β_s , $s = 2 + (i-1)(\bar{\tau}+1), 2 + (i-1)(\bar{\tau}+1) + \bar{\tau}$ не равен нулю, а именно тот, который соответствует оптимальному для нее лагу запаздывания.

Введем в рассмотрение блок новых неотрицательных неизвестных переменных $u_t, v_t, t = \overline{1, n}$ следующим образом:

$$u_t = \begin{cases} \delta_t, & \text{если } \varepsilon_t > 0 \\ 0, & \text{в противном случае,} \end{cases}$$

$$v_t = \begin{cases} -\delta_t, & \text{если } \varepsilon_t < 0 \\ 0, & \text{в противном случае.} \end{cases}$$

Тогда задача ЛБП для определения оценок параметров модели (1) примет вид:

$$\sum_{i=1}^{1+m(\bar{\tau}+1)} \beta_i z_{ti} + u_t - v_t = y_t, \quad t = \overline{1, n}, \quad (4)$$

$$-M\sigma_i \leq \beta_i \leq M\sigma_i, \quad i = \overline{2, 1+m(\bar{\tau}+1)}, \quad (5)$$

$$\sum_{j=2+(i-1)(\bar{\tau}+1)}^{2+(i-1)(\bar{\tau}+1)+\bar{\tau}} \sigma_j = 1, \quad i = \overline{1, m}, \quad (6)$$

$$u_t \geq 0, \quad v_t \geq 0, \quad t = \overline{1, n}, \quad (7)$$

$$\sigma_i \in \{0, 1\}, \quad i = \overline{2, 1+m(\bar{\tau}+1)}, \quad (8)$$

$$\sum_{t=1}^n (u_t + v_t) \rightarrow \min, \quad (9)$$

где M – наперед заданное большое положительное число.

Пусть вектор $(\beta^*, u^*, v^*, \sigma^*)$ – решение задачи ЛБП (4) – (9). Тогда оптимальный лаг запаздывания τ_i для i -ой независимой переменной в модели (1) будет равен значению $j_i \in \{0, 1, \dots, \bar{\tau}\}$, если

$$\sigma_{2+(i-1)(\bar{\tau}+1)+j_i}^* = 1,$$

при этом, в соответствии с (6)

$$\sigma_{2+(i-1)(\bar{\tau}+1)+s}^* = 0, \quad s = \overline{2+(i-1)(\bar{\tau}+1), 2+(i-1)(\bar{\tau}+1)+\bar{\tau}}, \quad s \neq j_i.$$

Оценки параметров α_i , $i = \overline{0, m}$, рассчитанные на основе построенного уравнения (3), будут равны величинам:

$$\alpha_0 = \beta_1^*,$$

$$\alpha_i = \sum_{j=2+(i-1)(\bar{\tau}+1)}^{2+(i-1)(\bar{\tau}+1)+\bar{\tau}} \beta_j^* \sigma_j^*, \quad i = \overline{1, m}.$$

В работах С. И. Носкова и Ю. А. Бычкова¹ разработаны линейные регрессионные модели количества ДТП в Российской Федерации, включающие в себя в качестве независимых переменных количество троллейбусов, общую численность населения, численность трудоспособного сельского населения, ежегодное потребление алкоголя, протяженность автомобильных дорог общего назначения. Разработаем на основе обработки данных официальной статистики за 2004–2015 гг.² регрессионную модель вида (1) для этого важного показателя с учетом возможного вхождения в ее правую часть лаговых переменных. При этом задействуем три из пяти перечисленных выше предикторов. Введем следующие обозначения:

y – количество дорожно-транспортных происшествий с пострадавшими, чел.;

x_1 – количество троллейбусов, тыс. ед.;

x_2 – численность населения в России, млн. чел.;

¹ Носков С. И. Применение многокритериального метода наименьших модулей для моделирования количества дорожно-транспортных происшествий // Ученые записки Комсомольского-на-Амуре государственного технического университета. 2023. № 5 (69). С. 30–35; Носков С. И., Бычков Ю. А. Регрессионная модель динамики дорожно-транспортных происшествий в Российской Федерации // Южно-Сибирский научный вестник. 2023. № 1 (47). С. 37–42.

² Носков С. И., Бычков Ю. А. Регрессионная модель динамики дорожно-транспортных происшествий в Российской Федерации // Южно-Сибирский научный вестник. 2023. № 1 (47). С. 37–42.

x_3 – ежегодное потребление алкоголя, литры чистого спирта на человека в возрасте более 15 лет.

Примем в качестве максимально возможного лага число $\bar{\tau} = 3$. Тогда значение $n = 1$ будет соответствовать 2007 г., а общее число наблюдений – 9. В результате решения задачи ЛБП (4) – (9) получим следующую модель с запаздывающими переменными:

$$y_t = -739404,67 + 42998,31x_{t-2,1} + 3164,55x_{t-3,2} + 1338,81x_{t3} + \varepsilon_t, \quad t = \overline{1,9}, \quad (10)$$
$$H = 23772,7.$$

Здесь H – сумма абсолютных ошибок аппроксимации:

$$H = \sum_{t=1}^9 (u_t^* + v_t^*).$$

Таким образом, только объем потребления алкоголя отражается на количестве ДТП в том же году, остальные две независимые переменные влияют на него с запаздыванием в два (количество троллейбусов) и три (численность населения в России) года соответственно. Эти обстоятельства требуют дополнительного изучения.

Построим регрессионную модель для того же состава переменных, но без запаздываний, с помощью МНМ и МНК. В результате получим:

а) МНМ

$$y_t = 1418140 - 47024,1x_{t1} - 5982,92x_{t2} + 14989x_{t3} + \varepsilon_t, \quad t = \overline{1,9}, \quad (11)$$
$$H = 38174,4.$$

б) МНК

$$y_t = 496869 - 10970,2x_{t1} - 1814,99x_{t2} + 8292,89x_{t3} + \varepsilon_t, \quad t = \overline{1,9}, \quad (12)$$
$$H = 43874,4.$$

Анализ моделей (10) – (12) показывает, что лишь переменная уровня потребления алкоголя во все три входит с положительным коэффициентом, что, впрочем, вполне естественно. Знаки трех других параметров различаются. Обращает на себя внимание, что модель (10) лучше модели (11) по критерию H на 60,5 %, а модели (12) – на 84,5 %.

Выводы

В работе представлен способ выбора оптимальных временных лагов для независимых переменных и оценивания неизвестных параметров линейной регрессионной модели с запаздыванием. При выборе в качестве расстояния между векторами фактических и расчетных значений зависимой переменной манхэттенской метрики, со-

ответствующей методу наименьших модулей, данный способ сводится к решению задачи ЛБП. Построена регрессионная модель с запаздыванием для описания динамики количества ДТП, которая обладает меньшими ошибками аппроксимации по сравнению с обычной линейной моделью при использовании методов наименьших квадратов и модулей.

Список источников

Базилевский М. П., Носков С. И. Методические и инструментальные средства построения некоторых типов регрессионных моделей // Системы. Методы. Технологии. 2012. № 1 (13). С. 80–87.

Носков С. И. Применение многокритериального метода наименьших модулей для моделирования количества дорожно-транспортных происшествий // Ученые записки Комсомольского-на-Амуре государственного технического университета. 2023. № 5 (69). С. 30–35.

Носков С. И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных: монография / ответственный редактор А. И. Тятюшкин. Иркутск: Облиформпечать, 1996. 320 с.

Носков С. И., Бычков Ю. А. Регрессионная модель динамики дорожно-транспортных происшествий в Российской Федерации // Южно-Сибирский научный вестник. 2023. № 1 (47). С. 37–42.

Anatolyev S. Redundancy of Lagged Regressors Revisited // Econometric Theory. 2007. Vol. 23, no. 2. P. 364–368.

Honoré B. E., Kyriazidou E. Panel Data Discrete Choice Models with Lagged Dependent Variables // Econometrica. 2000. Vol. 68, no. 4. P. 839–874.

Huang J. Z., Shen H. Functional Coefficient Regression Models for Non-Linear Time Series: A Polynomial Spline Approach // Scandinavian Journal of Statistics. 2004. Vol. 31, no. 4. P. 515–534.

Keele L., Kelly N. J. Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables // Political Analysis. 2006. Vol. 14, no. 2. P. 186–205.

Maeshiro A. Teaching Regressions with a Lagged Dependent Variable and Autocorrelated Disturbances // The Journal of Economic Education. 1996. Vol. 27, is. 1. P. 72–84.

Yanti Y., Rahardianto S. Stepwise Approach in Lagged Variables Time Series Modeling: A Simple Illustration // Materials Science and Engineering. The 7th International Conference on Global Optimization and Its Application (ICoGOIA 2018) (August 30–31, 2018, Bali, Indonesia). 2019. Vol. 621. P. 012009.

References

Bazilevskii M. P., Noskov S. I. Metodicheskie i instrumental'nye sredstva postroeniia nekotorykh tipov regressionnykh modelei [Methodology and instrumental tools for construction some types of regression models]. *Sistemy. Metody. Tekhnologii* [Systems. Methods. Technologies], 2012, no. 1 (13), pp. 80–87.

Noskov S. I. Primenenie mnogokriterial'nogo metoda naimen'shikh modulei dlia modelirovaniia kolichestva dorozhno-transportnykh proisshествii [Application of the multicriteria method of least modules for modeling road accident number]. *Uchenye zapiski Komsomol'skogo-na-Amure gosudarstvennogo tekhnicheskogo universiteta* [Scholarly Notes of Komsomolsk-na-Amure State Technical University], 2023, no. 5 (69), pp. 30–35.

Noskov S. I. *Tekhnologiya modelirovaniia ob"ektov s nestabil'nykh funktsionirovaniem i neopredelennost'iu v dannykh: monografiia* [Technology of modelling objects with unstable functioning and uncertainty in data; ed. by A. I. Tyatyushkin]. Irkutsk: Oblinformpechat', 1996. 320 p.

Noskov S. I., Bychkov Iu. A. Regressionnaia model' dinamiki dorozhno-transportnykh proisshествii v Rossiiskoi Federatsii [Regression model of traffic accident dynamics in the Russian Federation]. *Iuzhno-Sibirskii nauchnyi vestnik* [South-Siberian Scientific Bulletin], 2023, no. 1 (47), pp. 37–42.

Anatolyev S. Redundancy of lagged regressors revisited. *Econometric Theory*, 2007, vol. 23, no. 2, pp. 364–368.

Honoré B. E., Kyriazidou E. Panel data discrete choice models with lagged dependent variables. *Econometrica*, 2000, vol. 68, no. 4, pp. 839–874.

Huang J. Z., Shen H. Functional coefficient regression models for non-linear time Series: A polynomial spline approach. *Scandinavian Journal of Statistics*, 2004, vol. 31, no. 4, pp. 515–534.

Keele L., Kelly N. J. Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Political Analysis*, 2006, vol. 14, no. 2, pp. 186–205.

Maeshiro A. Teaching regressions with a lagged dependent variable and autocorrelated disturbances. *The Journal of Economic Education*, 1996, vol. 27, is. 1, pp. 72–84.

Yanti Y., Rahardiantoro S. Stepwise approach in lagged variables time series modeling: A simple illustration. *Materials Science and Engineering. The 7th International Conference on Global Optimization and Its Application (ICoGOIA 2018) (August 30–31, 2018, Bali, Indonesia)*, 2019, vol. 621, p. 012009.

Сведения об авторах

Сергей Иванович Носков – доктор технических наук, профессор; <https://orcid.org/0000-0003-4097-2720>, sergey.noskov.57@mail.ru, Иркутский государственный университет путей сообщения (д. 15, ул. Чернышевского, 664074 Иркутск, Россия); **Sergey I. Noskov** – Doctor of Technical Sciences, Professor, <https://orcid.org/0000-0003-4097-2720>, sergey.noskov.57@mail.ru, Irkutsk State Transport University (15, ul. Chernyshevskogo, 664074 Irkutsk, Russia).

Егор Сергеевич Попов – магистрант; <https://orcid.org/0009-0002-2307-0995>, eglr5732@mail.ru, Иркутский государственный университет путей сообщения (д. 15, ул. Чернышевского, 664074 Иркутск, Россия) **Egor S. Popov** – Student in the master's programme, <https://orcid.org/0009-0002-2307-0995>, eglr5732@mail.ru, Irkutsk State Transport University (15, ul. Chernyshevskogo, 664074 Irkutsk, Russia).

Заявленный вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Статья поступила в редакцию 22.11.2023; одобрена после рецензирования 12.03.2024; принята к публикации 26.03.2024.

The article was submitted 22.11.2023; Approved after reviewing 12.03.2024; Accepted for publication 26.03.2024.