DOI 10.37882/2223-2966.2023.12.06

МОДЕЛИРОВАНИЕ И ИССЛЕДОВАНИЕ СПУТЫВАЮЩЕЙ ПЕРЕМЕННОЙ С ИСПОЛЬЗОВАНИЕМ МОДЕЛЕЙ ЛИНЕЙНОЙ РЕГРЕССИИ

MODELING AND INVESTIGATION OF A CONFUSING VARIABLE USING LINEAR REGRESSION MODELS

O. Buchnev
A. Atalyan

Summary. Modeling of the impact of a confounding variable was carried out when the exposure is binary, and the confounding variable and the response variable are continuous. Experiments were carried out with the model. Linear regression models were used to examine the change in the value of the regression coefficient and the standard error of the regression coefficient before and after including the confounding variable in the regression model.

Keywords: confounding, regression models, modelling, observational data analysis.

Бучнев Олег Сергеевич
к.т.н., доцент, Иркутский национальный исследовательский технический университет
buchnevo81@mail.ru
Аталян Алина Валерьевна
к.б.н., старший научный сотрудник, руководитель функциональной группы информационных систем и биостатистики, Научный центр проблем здоровья семьи и репродукции человека (г. Иркутск)

Аннотация. Выполнено моделирование спутывающей переменной, когда признак-фактор бинарный, а сама спутывающая переменная и переменная отклика непрерывные. Проведены эксперименты с моделью. С помощью моделей линейной регрессии исследовано как изменяются значения коэффициента регрессии и стандартной ошибки коэффициента регрессии до и после включения спутывающей переменной в регрессионную модель при различных значениях сбалансированности выборки.

Ключевые слова: спутывающая переменная, регрессионные модели, моделирование, анализ обсервационных данных.

В современном анализе данных, будь то статистическая обработка данных, полученных в результате подготовленного и выполненного исследования, или данных, для сбора которых не проводилось планирование, всегда были актуальными вопросы анализа причинно-следственных связей. В наборе данных могут быть переменные, которые оказывают влияние как на зависимую переменную, так и на оказывающий воздействие фактор. Такие переменные называются конфаундер, спутывающая переменная, спутывающий фактор и т.д. Они вносят смещение в результирующие оценки, из-за которого можно не увидеть причинно-следственную связь там, где она есть, или обнаружить там, где в действительности ее нет. Для того, чтобы переменная была конфаундером, или спутывающей переменной, она должна удовлетворять трем условиям: 1) эта переменная оказывает влияние на отклик; 2) имеет различное распределение в группах, образованных различными значениями признака-фактора *X*; 3) не подвержена воздействию признака-фактора *X*. [6]

В отечественной и зарубежной литературе предложены множество способов устранения влияния спутывающих переменных [1, 3, 4, 5, 7, 9] от построения регрессионных моделей до использования различных методов машинного обучения: Байесовская поправка на спутывающие переменные (Bayesian adjustment for confounding, BAC), обобщенная Байесовская оценка причинности (generalized Bayesian causal effect estimation, GBCEE), групповая оценка Лассо регрессии и дважды надежные оценки (Group Lasso and doubly robust estimation (GLiDeR), оценка, основанная на максимальном правдоподобии (the scalable collaborative targeted maximum likelihood estimation (SC-TMLE)), многомерные оценки меры склонности (High-dimensional propensity scores (hdPS)). Эти методы имеют множество ограничений и требовательны к исходным данным, поэтому область их применения ограничена [5].

В статистических исследованиях, которые выполняются на выборках относительно небольшого объема, для идентификации спутывающих переменных актуально применение регрессионных моделей. Идентификация спутывающих переменных и устранение влияния, которые они оказывают, является важным этапом прикладных исследований в эпидемиологии, маркетинге, социологии, экономике и других областях. [7].

Применение статистических моделей является наиболее общим способом контроля спутывающих переменных, когда отсутствует возможность постановки активного эксперимента, или при планировании эксперимента не было учтено влияние спутывающих переменных. Этот способ основан на построении регрессионной модели, в которой кроме зависимой переменной

Y и действующего признака-фактора X участвует возможная спутывающая переменная C. На первом этапе строится регрессионная модель зависимой переменной и действующего фактора:

$$Y = b_0 + b_1 X. \quad (1)$$

Статистическая связь признака-фактора и зависимой переменной доказывается значимостью коэффициента регрессии b_1 в модели. Далее выполняют корректировку, для чего в регрессионную модель включают переменную, которая, предположительно, является спутывающей переменной:

$$Y = b_0 + b_1 X + b_2 C. \quad (2)$$

Существуют различные подходы к оценке различия параметров и их статистических свойств в регрессионных моделях до включения в них спутывающей переменной и после. Они описаны, например, в работе [3]. В данной работе изучено изменение значения *p-value* для коэффициентов и изменение значения коэффициентов в процентах. Первый подход предполагает, что есть основания предполагать, что переменная C является спутывающей переменной, если коэффициент в регрессионной модели b_1, после включения в нее спутывающей переменной, перестает быть статистически значимым, но при этом статистически значимым становится коэффициент b_2. Второй подход предполагает оценку изменения коэффициента при b_1: если он, после включения в модель признака C, изменяется более чем на 10%, то считают, что переменная C спутывающая переменная [2,8]. Именно такой метод контроля рекомендован для использования в прикладных исследованиях [3]. Однако она обладает следующими недостатками: оценки коэффициентов регрессии и *p-value* могут быть искажены неверным предположением о виде зависимости между переменными. Кроме того, необходимо учитывать направление причинной связи между фактором и зависимой переменной.

В данной работе выполнено моделирование влияния спутывающей переменной и исследовано как изменяются значения коэффициентов и стандартные ошибки коэффициентов до включения и после включения спутывающей переменной в модель. При этом исследовано, как влияет сбалансированность выборки на значения коэффициентов и стандартные ошибки коэффициентов.

Целью исследования является моделирование спутывающей переменной и оценка зависимости отношения коэффициентов модели регрессии, а также стандартного отклонения оценки в зависимости от меры влияния спутывающей переменной при различных значениях сбалансированности выборки до и после включения спутывающей переменной в модель.

Для исследования построена и реализована на языке R модель, основанная на примере [7]. Программный модуль с текстом программы размещен в GitHub и доступен по ссылке [https://github.com/BuchnevOS/confounder_CIE/blob/main/confounders.R].

Этапы алгоритма моделирования спутывающей переменной:

1. Используя коэффициент k, создать значения бинарного признака X — признак-фактор. Коэффициент k отвечает за сбалансированность выборки: $k = 0{,}1$ — плохо сбалансированная выборка, $k = 0{,}5$ — хорошо сбалансированная выборка.
2. Создать значения нового признака с нормальным распределением C, используя коэффициент T, определяющий значение математического ожидания в каждой группе: $M[C \mid X = 0] = 0$, $M[C \mid X = 1] = T$.
3. Создать значения зависимой переменной

$$Y = 1 + b_1 * C + 20 * X + \varepsilon, \quad (3)$$

где ε — нормально распределенная случайная величина с нулевым математическим ожиданием и стандартным отклонением, равным 10.

Признак C в этой модели является спутывающей переменной, так как удовлетворяет всем трем условиям: 1) формирует значение отклика с помощью коэффициента b_1 в (3); 2) величина коэффициента T обуславливает различное распределение в группах, образованных различными значениями признака-фактора X; 3) не является эффектом воздействия признака-фактора X [6]. Последнее требование актуально для тех задач, в которых исследуется определенная предметная область.

Эксперименты с моделью проводились многократно. При каждом значении набора параметров генерировались сто выборок объемом одна тысяча элементов каждая. При проведении экспериментов изменялись значение коэффициента k, указывающего на сбалансированность выборки. Изменялось смещение (разность средних) $T = M[C \mid X = 0] - M[C \mid X = 1]$, обусловленное значениями признака-фактора: оно принимало значения 0; 0,1; 0,5; 1; 5; 10. Также изменялась мера влияния спутывающей переменной на зависимую переменную (коэффициент b_1 в (3)): принимались значения 0,5; 1,5; 3. В каждом эксперименте с моделью оценивалась регрессия по формулам (1) и (2). При этом фиксировались значения коэффициентов b_1 в моделях до и после корректировки на спутывающую переменную, значения *p-value*, а также значения стандартных ошибок коэффициентов регрессии. Далее было вычислено процентное изменение коэффициентов до корректировки и после и процентное изменение стандартных ошибок коэффициентов регрессии до корректировки и после.

По данным, полученным в результате экспериментов, построены графики, на которых видно, как зависят значения регрессионных коэффициентов при переменной *X* в (2) и их стандартные ошибки от того, насколько выборка сбалансирована, от величины смещения *T*, обусловленного значениями признака-фактора, и от того, какой вклад вносит спутывающая переменная в значение зависимой переменной. Поскольку значения коэффициентов регрессии распределены нормально, значения коэффициентов, полученные при выполнении 100 экспериментов при каждом наборе значений параметров были усреднены для всех оценок, за исключением изменения оценки коэффициента b_1 до и после поправки. Для этого показателя бралось медианное значение.

На рисунках ниже приведены графики, построенные по результатам компьютерного моделирования, которые отражают изменение коэффициента b_1 в (1) и в (2) (на графиках *b* и *b_*) и их изменения в процентах, стандартной ошибки коэффициентов регрессии (*sd* и *sd_*) для моделей (1) и (2) и их изменения в процентах при различных значениях сбалансированности выборки, величины смещения *T*, обусловленного значениями признака-фактора, и мерой влияния спутывающей переменной на значение зависимой переменной (b_1 в (3)).

Основные выводы, которые можно сделать по результатам проведенных с моделью экспериментов.

1) Изменение значения коэффициента регрессии в процентах связано с показателем сбалансированности выборки незначительно. При этом от сбалансированности выборки зависит изменение в процентах значения ошибки для коэффициента: для плохо сбалансированных выборок он изменяется сильнее.

2) Уменьшение в процентах от исходного значения коэффициента b_1 в модели (2) зависит не только от величины смещения, обусловленного значениями признака-фактора, но и от вклада спутывающей переменной в значение зависимой переменной: чем этот вклад больше, тем сильнее изменяется значение коэффициента.

3) На хорошо сбалансированных выборках при незначительном смещении *T*, обусловленном значениями признака-фактора, ошибка оценки коэффициента изменяется существенно. Чем больше это смещение, тем слабее в процентах уменьшается ошибка оценки коэффициента, причем, с ростом влияния спутывающей переменной на значение зависимой переменной, изменение стандартной ошибки коэффициента в процентах становится сильнее.

Последний вывод говорит о том, что имеет смысл оценивать действие спутывающей переменной не только по изменению значения коэффициента регрессии до и после введения в модель спутывающей перемен-

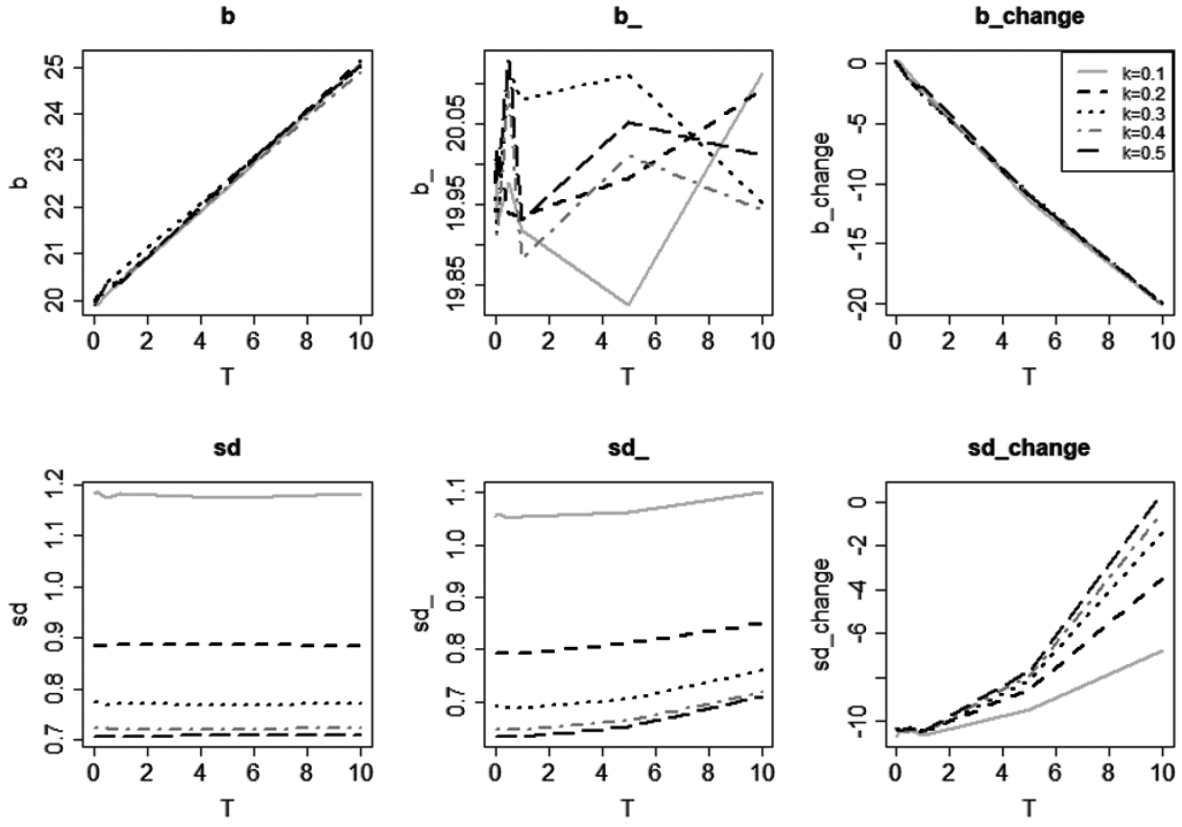

Рис. 1. Изменение оценок в регрессионных моделях при b_1=0,5 в модели (3)

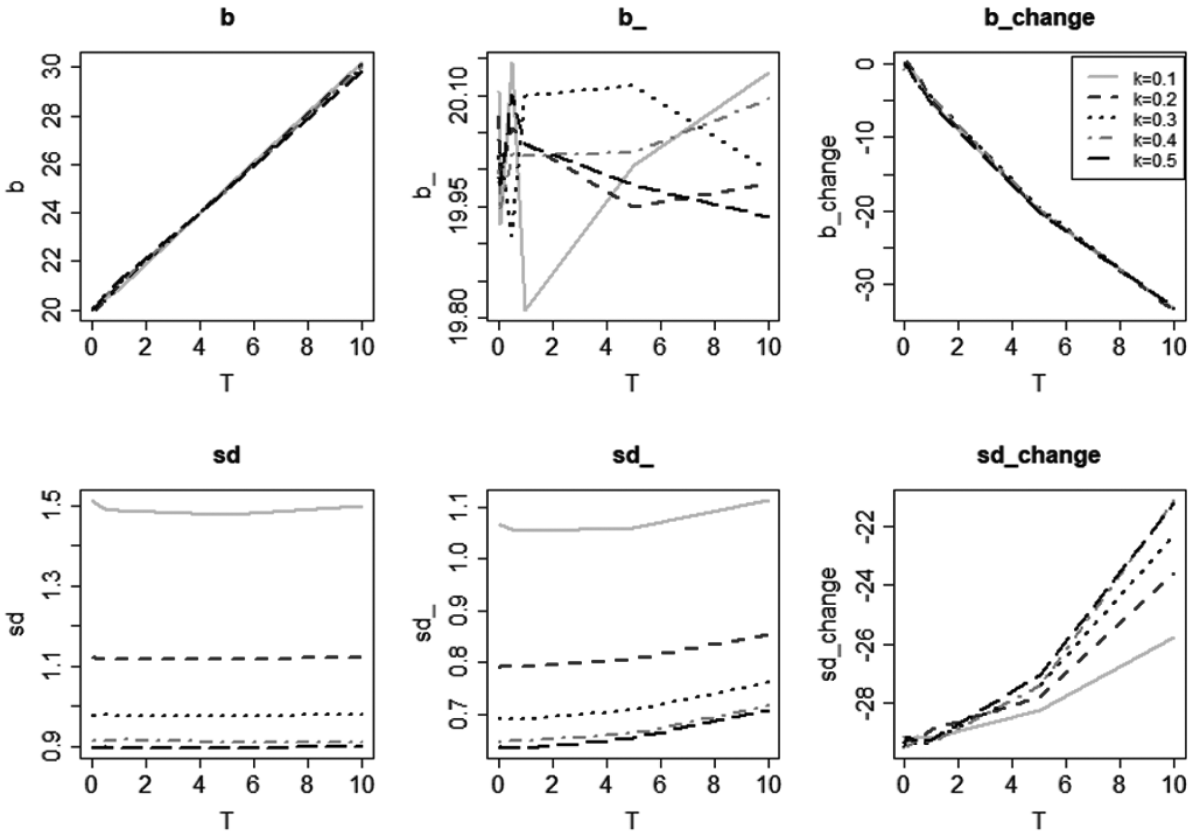

Рис. 2. Изменение оценок в регрессионных моделях при b_1=1 в модели (3)

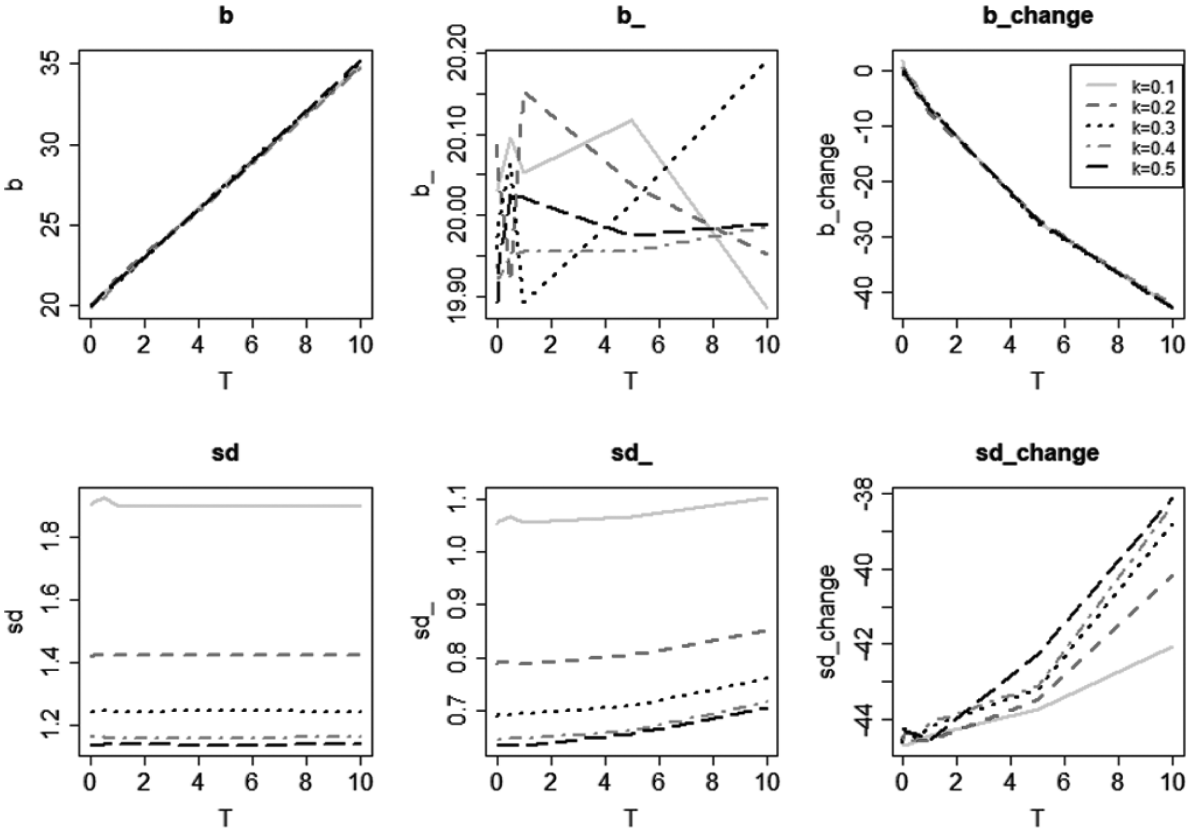

Рис. 3. Изменение оценок в регрессионных моделях при b_1=1,5 в модели (3)

ной, но и по изменению ошибки оценки этого коэффициента.

Заключение

В работе реализована и описана модель, с помощью которой проведены компьютерные эксперименты с влиянием спутывающей переменной. Эксперименты с моделью показали, что правило о 10 % изменении коэффициента регрессии после введения спутывающей переменной в модель работает при сильном влиянии спутывающей переменной на зависимую переменную. В случае слабого влияния информативным может быть изменение стандартной ошибки коэффициента регрессии. При этом изменение значения коэффициента регрессии практически не зависит от сбалансированности выборки, в то время как изменение значения стандартной ошибки регрессии зависит от сбалансированности выборки.

ЛИТЕРАТУРА

1. Вараксин А.Н., Шалаумова Ю.В., Панов В.Г. Принципы контроля конфаундеров в сравнительных исследованиях в экологии: стандартизация и регрессионные модели // Принципы экологии. 2014. Т. 3. № 1. С. 4–14.
2. Budtz-Jørgensen E, Keiding N, Grandjean P, Weihe P. Confounder selection in environmental epidemiology: Assessment of health effects of prenatal mercury exposure. Ann Epidemiol. 2007;17:27–35.
3. Denis Talbot, Awa Diop, Mathilde Lavigne-Robichaud and Chantal Brisson. The change in estimate method for selecting confounders: A simulation study//Statistical Methods in Medical Research 2021, Vol. 30(9) 2032–2044.
4. G Maldonado, S Greenland. Simulation study of confounder-selection strategies//Am J Epidemiol. 1993 Dec 1;138(11):923–36. doi: 10.1093/oxfordjournals.aje.a116813.
5. Imane Benasseur, Denis Talbot, Madeleine Durand, Anne Holbrook, Alexis Matteau, Brian J Potter, Christel Renoux , Mireille E Schnitzer, Jean-Éric Tarride, Jason R Guertin. A comparison of confounder selection and adjustment methods for estimating causal effects using large healthcare databases//Pharmacoepidemiol Drug Saf. 2022 Apr;31(4):424–433. doi: 10.1002/pds.5403. Epub 2022 Jan 7.
6. K J Jager, C Zoccali, A Macleod, F W Dekker. Confounding: what it is and how to deal with it//Kidney Int. 2008 Feb;73(3):256–60. doi: 10.1038/sj.ki.5002650.
7. McNamee R. Regression modelling and other methods to control confounding. Occup Environ Med. 2005;62:500–6.
8. Paul H. Lee. Is a Cutoff of 10 % Appropriate for the Change-in-Estimate Criterion of Confounder Identification?// J Epidemiol. 2014; 24(2): 161–167. doi: 10.2188/jea.JE20130062.
9. Tyler J VanderWeele. Principles of confounder selection//Eur J Epidemiol. 2019 Mar;34(3):211–219. doi: 10.1007/s10654-019-00494-6. Epub 2019 Mar 6.

© Бучнев Олег Сергеевич (buchnevo81@mail.ru); Аталян Алина Валерьевна

Журнал «Современная наука: актуальные проблемы теории и практики»