

МЕЖДИСЦИПЛИНАРНЫЕ ИССЛЕДОВАНИЯ

Е. В. Коссова¹

НИУ «Высшая школа экономики» (Москва, Россия)

И. С. Слаболицкий²

НИУ «Высшая школа экономики» (Москва, Россия)

Б. С. Потанин³

НИУ «Высшая школа экономики» (Москва, Россия)

УДК: 330.062

doi: 10.55959/MSU0130-0105-6-58-4-2

ОЦЕНИВАНИЕ ВЕРОЯТНОСТЕЙ, ПРЕДЕЛЬНЫХ ЭФФЕКТОВ И ЭФФЕКТОВ ВОЗДЕЙСТВИЯ В ИЕРАРХИЧЕСКИХ СИСТЕМАХ БИНАРНЫХ УРАВНЕНИЙ⁴

Настоящая статья посвящена вопросу интерпретации результатов оценивания иерархических (рекурсивных) систем бинарных уравнений в том случае, когда уравнение, задающее эндогенную переменную, не содержит уникальных регрессоров, т.е., когда система не удовлетворяет условиям ограничений исключения. Работа дополняет существующие исследования, посвященные идентифицируемости параметров иерархических бинарных систем, анализом условий идентифицируемости вероятностей, предельных эффектов и эффектов воздействия. Обосновано теоретически и показано на симулированных данных, что даже если параметры системы идентифицировать не удастся, то можно получить состоятельные оценки вероятностей и предельных эффектов как по экзогенному, так и по эндогенному регрессору. Эффекты воздействия в этом случае так же как и параметры, не идентифицируются. Обсуждаемая в работе проблема идентифицируемости также рассматривается на реальных

¹ Коссова Елена Владимировна — к.ф.-м.н., доцент департамента прикладной экономики факультета экономических наук, НИУ «Высшая школа экономики»; e-mail: ekossova@hse.ru, ORCID: 0000-0002-4600-5929.

² Слаболицкий Илья Сергеевич — ассистент департамента прикладной экономики факультета экономических наук, стажер-исследователь международной лаборатории стохастического анализа и его приложений, НИУ «Высшая школа экономики»; e-mail: islabolitskiy@hse.ru, ORCID: 0000-0001-6706-0874.

³ Потанин Богдан Станиславович — к.э.н., старший преподаватель департамента прикладной экономики факультета экономических наук, НИУ «Высшая школа экономики»; e-mail: bpotantin@hse.ru, ORCID: 0000-0002-5862-9202.

⁴ Работа выполнена при поддержке гранта Российского научного фонда, РНФ 20-18-00427.

данных, на примере оценивания вероятности покупки лекарственных средств в зависимости от характеристик индивида и факта посещения врача. Важным практическим вкладом работы является рекомендация для исследователей интерпретировать результаты оценивания рекурсивных бинарных систем с помощью предельных эффектов в том случае, когда нет возможности включить хотя бы одну уникальную переменную в уравнение для эндогенного бинарного регрессора.

Ключевые слова: иерархическая система бинарных уравнений, бинарный эндогенный регрессор, ограничения исключения, предельные эффекты, эффекты воздействий, двумерная пробит модель.

Цитировать статью: Коссова, Е. В., Слаболицкий, И. С., & Потанин, Б. С. (2023). Оценивание вероятностей, предельных эффектов и эффектов воздействия в иерархических системах бинарных уравнений. *Вестник Московского университета. Серия 6. Экономика*, 58(4), 23–49. <https://doi.org/10.55959/MSU0130-0105-6-58-4-2>.

E. V. Kossova

HSE University (Moscow, Russia)

I. S. Slabolitskiy

HSE University (Moscow, Russia)

B. S. Potanin

HSE University (Moscow, Russia)

JEL: C18, C31, C35

doi: 10.55959/MSU0130-0105-6-58-4-2

ESTIMATING PROBABILITIES, MARGINAL EFFECTS AND TREATMENT EFFECTS IN RECURSIVE BIVARIATE PROBIT MODELS¹

This article addresses the interpreting results of recursive system of binary equations estimation in the case when the system does not satisfy exclusion restriction conditions. It means that the equation defining the endogenous variable does not contain a unique covariate. This article extends the analysis of previous studies on the identifiability of parameters of recursive binary systems by analyzing the conditions for the identifiability of probabilities, marginal effects and treatment effects. We provide a reasonable consideration suggesting that even if parameters of the model are unidentifiable, it is still possible to estimate accurately the conditional probabilities and marginal effects, but not the treatment effects. The problem of identifiability discussed in the paper is also considered on real data. We estimate the probability of purchasing medicine depending on individuals' characteristics and the fact of visiting a doctor. An important practical contribution of the work is the recommendation for researchers to interpret the result recursive binary system estimation via marginal effects in the case when it is not possible to include at least one unique variable in the equation for the binary covariate.

¹ This work is supported by the Russian Science Foundation, RSF 20-18-00427.

Keywords: recursive bivariate probit model, binary endogenous covariate, exclusion restrictions, marginal effects, treatment effects.

To cite this document: Kossova, E. V., Slabolitskiy, I. S., & Potanin, B. S. (2023). Estimating probabilities, marginal effects and treatment effects in recursive bivariate probit models. *Lomonosov Economics Journal*, 58(4), 23–49. <https://doi.org/10.55959/MSU0130-0105-6-58-4-2>.

Введение

Системы бинарных уравнений активно применяются исследователями при моделировании поведения и состояний индивидов, домохозяйств, фирм и других экономических агентов, когда речь идет о двойном выборе, который задается двумя зависимыми бинарными переменными.

Наиболее популярными являются двумерная пробит-модель внешне не связанных бинарных уравнений (bivariate probit, BP) и иерархическая двумерная пробит-модель (recursive bivariate probit, RBP). Например, ниже, в примере, рассматривается вероятностная модель, в которой в качестве зависимых переменных выступают индикатор обращения к врачу и индикатор покупки лекарств индивидом в течение последнего месяца. В предположении, что обращение к врачу непосредственно влияет на вероятность покупки лекарств, следует рассматривать модель RBP.

В модели BP зависимость между объясняемыми бинарными переменными задается корреляцией случайных ошибок, входящих в уравнения для латентных переменных (Filippini et al., 2018). Модель RBP задается системой бинарных уравнений с коррелированными случайными ошибками, в которой одна из зависимых бинарных переменных является эндогенным регрессором для второй. Эта модель впервые была предложена в работе (Heckman, 1978) как частный случай модели, задаваемой системой латентных переменных, содержащей эндогенную бинарную переменную. Модель BP является вложенной по отношению к модели RBP.

В статье (Filippini et al., 2018) было также показано, что при выборе между моделями BP и RBP следует отдавать предпочтение модели RBP, так как незначимость корреляции в модели BP может быть вызвана не включением эндогенного бинарного регрессора. Модель RBP особенно удобна при оценивании эффекта воздействия, когда наблюдаемый результат задается бинарной переменной, для которой индикатор воздействия является эндогенным регрессором.

Привлекательность модели RBP обусловлена возможностью оценивания модели с эндогенной объясняющей переменной без использования инструментальных переменных, относительно простой вычислительной процедурой и содержательной интерпретацией ее параметров. Большое количество прикладных исследований, относящихся к разным сферам экономики, основывают свои выводы на результатах оценивания модели RBP. Это исследования в области экономики труда (Morris, 2007; Castello,

2012), экономики здоровья (Sirven, Debrand, 2012; Dogbe, 2020), экономики сельского хозяйства (Ma et al., 2017), экономики экологии (Cohen et al., 2019; Blasch et al., 2019), экономики инновации (Frondel et al., 2008; Ayalew, Xianzhi, 2019) и многих других областях.

Несмотря на популярность иерархических бинарных систем, вопрос об идентифицируемости их параметров до сих пор остается открытым. В том случае, если параметры системы не идентифицируются, получить их состоятельные оценки невозможно (Martin, Quintata, 2022). Таким образом, многие исследователи применяют модель RBP «на свой страх и риск». Лишь небольшое количество работ посвящено поиску необходимых и достаточных условий идентифицируемости параметров и функций от этих параметров, и наиболее острая дискуссия развернулась вокруг так называемых условий ограничений исключения (exclusion restriction conditions). Суть этих условий состоит в том, что уравнение системы, задающее эндогенную переменную, должно включать по крайней мере одну уникальную для системы объясняющую переменную. Эти уникальные переменные в некоторых работах называют инструментами (excluded instruments).

Для прикладных исследований нередко ситуации, когда соблюдение ограничений исключения трудно выполнимо, а принудительное исключение регрессоров может привести к потере состоятельности оценок. Поэтому вопрос о том, насколько необходимы эти условия, является крайне важным.

В пионерной работе (Heckman, 1978) сформулированы достаточные условия идентифицируемости параметров для широкого класса моделей, описываемых системой уравнений и содержащих эндогенные бинарные регрессоры. Одним из этих достаточных условий является требование полного ранга матрицы регрессоров. Однако в статье (Maddala, 1983) отмечается, что это условие не является достаточным, и для идентификации параметров в модели RBP необходимо наличие ограничений исключения. При этом в обеих работах были высказаны лишь содержательные соображения и отсутствовали строгие математические доказательства. Следующей работой, посвященной проблеме идентифицируемости параметров, является статья (Wilde, 2000). В этой статье аргументы Дж. Маддалы подвергаются критике, а достаточность условия, предложенного Дж. Хекманом, еще раз обосновывается. Дж. Уайлд утверждает, что идентифицируемость параметров модели обусловлена ее функциональной формой, а именно тем фактом, что для нахождения оценок параметров решается система независимых уравнений, в которой число неизвестных параметров равняется числу уравнений, и ограничения исключения являются избыточными требованиями. Статья (Wilde, 2000) широко цитируется исследователями для обоснования отказа от ограничений исключения и оценивания модели RBP с общим набором экзогенных регрессоров. В статье

(Mourifie, Meango, 2014) опровергаются соображения Дж. Уайлда. В ней высказывается соображение, что функциональная форма не может давать гарантий единственности решения, и на симулированных данных демонстрируется, что при отсутствии регрессоров или наличии единственного общего бинарного регрессора, параметры модели RBP не могут быть оценены однозначно. Наконец, в работе (Han, Vitlacil, 2017) формулируются достаточные условия идентифицируемости параметров обобщенной модели RBP и доказываемается, что ограничения исключения являются достаточными условиями независимо от типа используемых регрессоров и необходимыми при отсутствии регрессоров или наличии единственного общего бинарного экзогенного регрессора, а условия полного ранга являются необходимыми, но не достаточными.

Общий вывод работ заключается в том, что если модель RBP не содержит непрерывных регрессоров, то ограничения исключения необходимы; если же в модели есть общий непрерывный регрессор, то однозначного ответа на вопрос о необходимости инструментов нет.

Во всех исследованиях, посвященных идентифицируемости параметров модели RBP, рассматриваются только оценки самих параметров системы. Исключения составляют работа (Shaikh, Vitlacil, 2011), в которой при выполнении ряда условий, включающих ограничения исключения, находятся границы для среднего эффекта воздействия, и работа (Han, Lee, 2019), в которой затрагивается вопрос об устойчивости среднего эффекта воздействия при нарушении предпосылок о совместном распределении случайных ошибок.

Однако в прикладных исследованиях коэффициенты при регрессорах в модели RBP малоинформативны. Как правило, для интерпретации результатов и прогнозирования исследователям требуется оценивать функции от параметров системы:

- совместные, условные и маргинальные вероятности;
- предельные эффекты;
- эффекты воздействия.

Для определения влияния объясняющих факторов на вероятности вычисляются предельные эффекты. Если же модель RBP используется для ответа на вопрос о воздействии некоторого события на бинарный показатель, вычисляются средние эффекты воздействия.

Из самих параметров системы наибольший интерес представляют коэффициент перед эндогенной переменной и корреляция между случайными ошибками, но, как было отмечено в (Filippini et al., 2018), последний должен быть интерпретирован с осторожностью. Таким образом, для прикладных исследований важна возможность вычисления предельных эффектов или эффектов воздействия.

Целью настоящей работы является ответ на вопрос, возможно ли идентифицировать и оценить вероятности, предельные эффекты и эффекты

воздействия в модели RBP в условиях отсутствия ограничений исключения, когда параметры модели неидентифицируемы.

Модель

Рассмотрим рекурсивную систему бинарных уравнений (модель RBP), состоящую из двух уравнений и не содержащую условий ограничений исключения:

$$d_i = \mathbb{I}\{\alpha_0 + \alpha_1 x_i + \varepsilon_i^d \geq 0\}, \quad (1)$$

$$y_i = \mathbb{I}\{\beta_0 + \beta_1 x_i + \delta d_i + \varepsilon_i^y \geq 0\}, \quad (2)$$

$$\begin{bmatrix} \varepsilon_i^d \\ \varepsilon_i^y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right),$$

где $\mathbb{I}\{\cdot\}$ — функция-индикатор; x_i — общая для двух уравнений независимая переменная; d_i и y_i — зависимые бинарные переменные уравнений (1) и (2) соответственно.

Если коэффициент корреляции ρ между случайными ошибками отличен от нуля, переменная d_i является эндогенным регрессором.

Отметим, что данная система должна оцениваться и в том случае, когда исследователя интересует лишь уравнение (2) системы, но переменная d_i является эндогенной.

При оценивании эффекта воздействия некоторого события на бинарный показатель y_i переменная d_i отвечает за наличие воздействия в i -м наблюдении. В примере, рассматриваемом ниже, d_i — факт обращения к врачу, y_i — факт покупки индивидом лекарства.

Мы отдельно рассматриваем случай с единственным бинарным и единственным непрерывным экзогенным регрессором. Однако обобщить полученные результаты на случай произвольного числа регрессоров и уравнений не составляет труда.

Распределение вектора случайных ошибок предполагается двумерным нормальным с нулевым вектором математических ожиданий и недиагональной ковариационной матрицей с единичными дисперсиями.

Таким образом, система содержит шесть неизвестных параметров: α_0 , α_1 , β_0 , β_1 , δ и ρ . Такая иерархическая система оценивается методом максимального правдоподобия. Логарифмическая функция правдоподобия имеет следующий вид:

$$\begin{aligned} & \log \mathcal{L}(\alpha_0, \alpha_1, \beta_0, \beta_1, \delta, \rho; x, d, y) = \\ & = \sum_{i=1}^n \left(\mathbb{I}\{d_i = 0, y_i = 0 | x_i\} \cdot \log \Phi_p \left(-(\alpha_0 + \alpha_1 x_i), -(\beta_0 + \beta_1 x_i) \right) + \right. \end{aligned}$$

$$\begin{aligned}
& + \mathbb{I}\{d_i = 0, y_i = 1 | x_i\} \cdot \log \Phi_{-\rho}(-(\alpha_0 + \alpha_1 x_i), \beta_0 + \beta_1 x_i) + \\
& + \mathbb{I}\{d_i = 1, y_i = 0 | x_i\} \cdot \log \Phi_{-\rho}(\alpha_0 + \alpha_1 x_i, -(\beta_0 + \beta_1 x_i + \delta)) + \\
& + \mathbb{I}\{d_i = 1, y_i = 1 | x_i\} \cdot \log \Phi_{\rho}(\alpha_0 + \alpha_1 x_i, \beta_0 + \beta_1 x_i + \delta),
\end{aligned}$$

где $\Phi_{\rho}(\cdot, \cdot)$ — функция распределения двумерного стандартного нормального распределения с коэффициентом корреляции ρ . Для краткости индекс наблюдения i мы далее опускаем.

Доказательство идентифицируемости параметров системы (1)–(2) и функций от них основывается на возможности однозначного их определения по вероятностям наблюдений, имеющимся в распоряжении исследователя (см. (Han, Vytlačil, 2017; Han, Lee, 2019; Freedman, Sekhon, 2020)).

Чтобы оценить вероятности любых комбинаций значений d и y , достаточно найти оценки любых трех совместных вероятностей, например, $\mathbb{P}\{d = 1, y = 0 | x\}$, $\mathbb{P}\{d = 1, y = 0 | x\}$ и $\mathbb{P}\{d = 1, y = 1 | x\}$. Очевидно, что предельные эффекты по экзогенному и эндогенному регрессорам определяются совместным распределением наблюдаемых переменных. Так, по бинарному экзогенному регрессору выражение предельного эффекта на условную вероятность $y = 1$ при условии $d = 0$ имеет следующий вид:

$$\begin{aligned}
ME_x(\mathbb{P}\{y = 1 | d = 0, x\}) &= \mathbb{P}\{y = 1 | d = 0, x = 1\} - \mathbb{P}\{y = 1 | d = 0, x = 0\} = \\
&= \mathbb{P}\{\beta_0 + \beta_1 + \varepsilon^y \geq 0 | \alpha_0 + \alpha_1 + \varepsilon^d < 0\} - \mathbb{P}\{\beta_0 + \varepsilon^y \geq 0 | \alpha_0 + \varepsilon^d < 0\} = \\
&= \frac{\mathbb{P}\{\alpha_0 + \alpha_1 + \varepsilon^d < 0, \beta_0 + \beta_1 + \varepsilon^y \geq 0\}}{\mathbb{P}\{\alpha_0 + \alpha_1 + \varepsilon^d < 0\}} - \frac{\mathbb{P}\{\alpha_0 + \varepsilon^d < 0, \beta_0 + \varepsilon^y \geq 0\}}{\mathbb{P}\{\alpha_0 + \varepsilon^d < 0\}} = \\
&= \frac{\Phi_{-\rho}(-(\alpha_0 + \alpha_1), \beta_0 + \beta_1)}{\Phi(-(\alpha_0 + \alpha_1))} - \frac{\Phi_{-\rho}(-\alpha_0, \beta_0)}{\Phi(-\alpha_0)},
\end{aligned}$$

где $\Phi(\cdot)$ — функция распределения стандартного нормального распределения.

Предельный эффект по непрерывному экзогенному регрессору на условную вероятность $y = 1$ при условии $d = 0$ определяется частной производной:

$$\begin{aligned}
ME_x(\mathbb{P}\{y = 1 | d = 0, x\}) &= \frac{\partial}{\partial x} \mathbb{P}\{y = 1 | d = 0, x\} = \frac{\partial}{\partial x} \left(\frac{\mathbb{P}\{d = 0, y = 1 | x\}}{\mathbb{P}\{d = 0 | x\}} \right) = \\
&= \frac{\partial}{\partial x} \left(\frac{\mathbb{P}\{\alpha_0 + \alpha_1 x + \varepsilon^d < 0, \beta_0 + \beta_1 x + \varepsilon^y \geq 0\}}{\mathbb{P}\{\alpha_0 + \alpha_1 x + \varepsilon^d < 0\}} \right) = \\
&= \frac{\partial}{\partial x} \left(\frac{\Phi_{-\rho}(-(\alpha_0 + \alpha_1 x), \beta_0 + \beta_1 x)}{\Phi(-(\alpha_0 + \alpha_1 x))} \right).
\end{aligned}$$

Предельный эффект по эндогенному бинарному регрессору равен разности условных вероятностей:

$$\begin{aligned}
 ME_d(\mathbb{P}\{y=1|x\}) &= \mathbb{P}\{y=1|d=1, x\} - \mathbb{P}\{y=1|d=0, x\} = \\
 &= \mathbb{P}\{\beta_0 + \beta_1 x + \delta + \varepsilon^y \geq 0 | \alpha_0 + \alpha_1 x + \varepsilon^d \geq 0\} - \\
 &\quad - \mathbb{P}\{\beta_0 + \beta_1 x + \varepsilon^y \geq 0 | \alpha_0 + \alpha_1 x + \varepsilon^d < 0\} = \\
 &= \frac{\mathbb{P}\{\alpha_0 + \alpha_1 x + \varepsilon^d \geq 0, \beta_0 + \beta_1 x + \delta + \varepsilon^y \geq 0\}}{\mathbb{P}\{\alpha_0 + \alpha_1 x + \varepsilon^d \geq 0\}} - \\
 &\quad - \frac{\mathbb{P}\{\alpha_0 + \alpha_1 x + \varepsilon^d < 0, \beta_0 + \beta_1 x + \varepsilon^y \geq 0\}}{\mathbb{P}\{\alpha_0 + \alpha_1 x + \varepsilon^d < 0\}} = \\
 &= \frac{\Phi_p(\alpha_0 + \alpha_1 x, \beta_0 + \beta_1 x + \delta)}{\Phi(\alpha_0 + \alpha_1 x)} - \frac{\Phi_{-p}(-(\alpha_0 + \alpha_1 x), \beta_0 + \beta_1 x)}{\Phi(-(\alpha_0 + \alpha_1 x))}.
 \end{aligned}$$

Таким образом, все вероятности и предельные эффекты определяются тремя совместными вероятностями наблюдаемых исследователем событий и идентифицируются всегда, независимо от того, идентифицируются ли параметры модели RBP.

Напротив, средний эффект воздействия (average treatment effect, ATE) определяется как разность вероятностей событий, одно из которых не является наблюдаемым и выступает функцией от параметров системы:

$$\begin{aligned}
 ATE &= \mathbb{E}[y_1 - y_0 | x] = \mathbb{E}[y_1 | x] - \mathbb{E}[y_0 | x] = \mathbb{P}_{d=1}\{y=1|x\} - \mathbb{P}_{d=0}\{y=1|x\} = \\
 &= \mathbb{P}\{\beta_0 + \beta_1 x + \delta + \varepsilon^y \geq 0\} - \mathbb{P}\{\beta_0 + \beta_1 x + \varepsilon^y \geq 0\} = \\
 &= \Phi(\beta_0 + \beta_1 x + \delta) - \Phi(\beta_0 + \beta_1 x).
 \end{aligned}$$

Аналогично вычисляется и средний эффект воздействия, подвергшихся воздействию (average treatment effect on treated, ATET):

$$\begin{aligned}
 ATET &= \mathbb{E}[y_1 - y_0 | d=1, x] = \mathbb{E}[y_1 | d=1, x] - \mathbb{E}[y_0 | d=1, x] = \\
 &= \mathbb{P}_{d=1}\{y=1|d=1, x\} - \mathbb{P}_{d=0}\{y=1|d=1, x\} = \\
 &= \mathbb{P}\{\beta_0 + \beta_1 x + \delta + \varepsilon^y \geq 0 | \alpha_0 + \alpha_1 x + \varepsilon^d \geq 0\} - \\
 &\quad - \mathbb{P}\{\beta_0 + \beta_1 x + \varepsilon^y \geq 0 | \alpha_0 + \alpha_1 x + \varepsilon^d \geq 0\} = \\
 &= \frac{\mathbb{P}\{\alpha_0 + \alpha_1 x + \varepsilon^d \geq 0, \beta_0 + \beta_1 x + \delta + \varepsilon^y \geq 0\}}{\mathbb{P}\{\alpha_0 + \alpha_1 x + \varepsilon^d \geq 0\}} - \\
 &\quad - \frac{\mathbb{P}\{\alpha_0 + \alpha_1 x + \varepsilon^d \geq 0, \beta_0 + \beta_1 x + \varepsilon^y \geq 0\}}{\mathbb{P}\{\alpha_0 + \alpha_1 x + \varepsilon^d \geq 0\}} =
 \end{aligned}$$

$$= \frac{\Phi_p(\alpha_0 + \alpha_1 x, \beta_2 + \beta_2 x + \delta)}{\Phi(\alpha_0 + \alpha_1 x)} - \frac{\Phi_p(\alpha_0 + \alpha_1 x, \beta_2 + \beta_2 x)}{\Phi(\alpha_0 + \alpha_1 x)}.$$

Проблемы с идентификацией параметров начинаются тогда, когда существует по крайней мере два набора значений параметров $(\alpha_0, \alpha_1, \beta_0, \beta_1, \delta, \rho)$ и $(\tilde{\alpha}_0, \tilde{\alpha}_1, \tilde{\beta}_0, \tilde{\beta}_1, \tilde{\delta}, \tilde{\rho})$, для которых $\Phi_p(\alpha_0 + \alpha_1 x, \beta_0 + \beta_1 x + \delta d) = \Phi_p(\tilde{\alpha}_0 + \tilde{\alpha}_1 x, \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\delta} d)$. При фиксировании одного из параметров δ или ρ остальные параметры определяются однозначно. Например, так как АТЕ строго возрастает по δ , разным значениям параметра δ соответствуют разные значения АТЕ.

Итак, средние эффекты воздействия могут быть вычислены только как функции от оценок параметров модели RBP, и по этой причине проблемы с идентифицируемостью самих параметров могут повлечь за собой неидентифицируемость эффектов воздействия.

Резюмируя вышеизложенное, можно утверждать, что при отсутствии ограничений исключения результаты оценивания модели RBP всегда могут быть интерпретированы с помощью предельных эффектов. Если же модель оценивается для вычисления средних эффектов воздействия, то их идентифицируемость зависит от идентифицируемости параметров.

Ниже, при помощи численного эксперимента мы рассматриваем, насколько остро стоит проблема идентифицируемости в модели RBP. В частности, мы анализируем поведение вероятностей, предельных эффектов и эффектов воздействия, варьируя объем выборки, чтобы выявить или опровергнуть наличие статистического свидетельства в пользу следующих фактов:

- 1) вероятности и предельные эффекты идентифицируемы, даже если неидентифицируемы параметры системы;
- 2) эффекты воздействия неидентифицируемы одновременно с параметрами модели
- 3) точность оценивания вероятностей и предельных эффектов растет с ростом объема выборки.

Дизайн эксперимента

Мы рассматриваем пять разных объемов выборки: 500, 1000, 5000, 1000 и 5000 наблюдений. Бинарный регрессор сгенерирован из распределения Бернулли с параметром $p = 0.85$ (эта переменная равна 1 с вероятностью 0.85 и 0 иначе). Отсутствие ограничений исключения особенно важно, когда регрессор является бинарным. Даже если условия ограничений исключения соблюдаются, но вариация регрессора мала, точность оценивания может быть низкой (Freedman, Sekhon, 2020). По этой причине мы си-

мулируем бинарный регрессор с маленькой вариацией. Непрерывный регрессор сгенерирован из хи-квадрат распределения с пятью степенями свободы¹. Случайные ошибки генерируются из двумерного стандартного нормального распределения с коэффициентом корреляции ρ .

Истинные значения параметров следующие: $\alpha_0 = -0.4$, $\alpha_1 = 0.3$, $\beta_0 = -0.3$, $\beta_1 = -0.8$, $\delta = 1.9$ и $\rho = -0.8$ для случая с бинарным экзогенным регрессором и $\alpha_0 = -0.9$, $\alpha_1 = 0.2$, $\beta_0 = -0.7$, $\beta_1 = 0.2$, $\delta = 0.2$ и $\rho = 0.2$ для случая с непрерывным экзогенным регрессором. Выбор значений параметров обусловлен желанием сгенерировать сбалансированные данные, чтобы вклад каждой из пар $(d = 0, y = 0)$, $(d = 1, y = 0)$, $(d = 0, y = 1)$ и $(d = 1, y = 1)$ составил 10–15%. Еще одним условием выбора параметров является необходимость добиться их неидентифицируемости в процессе оценивания (что происходит не всегда). Отметим, что в случае непрерывного экзогенного регрессора подобрать такие значения параметров оказалось непросто.

Таким образом, модель (1)–(2) оценивается по данным разного объема и с разными типами экзогенных регрессоров. Число симуляций для каждого случая равно 1000.

Результаты симуляций

Параметры

Таблицы 1 и 2 демонстрируют результаты оценивания параметров модели (1)–(2) для разных объемов выборки как для бинарного, так и для непрерывного регрессоров в условиях неидентифицируемости параметров.

Таблица 1

Оценки параметров модели (1)–(2) для случая бинарного экзогенного регрессора и разных объемов выборки					
$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\delta}$	$\hat{\rho}$
<i>n</i> = 500					
−0.469	0.381	0.764	−0.569	−0.818	0.912
<i>n</i> = 1000					
−0.478	0.358	0.431	−0.587	−0.200	0.612

¹ Стоит отметить, что некоторые работы выдвигают предположение, что нарушение идентифицируемости параметров может быть связано с носителем непрерывного регрессора. Для этого мы рассматривали процесс генерации непрерывного регрессора как из распределения с ограниченным носителем (равномерного распределения), так и с неограниченным (нормального распределения). Полученные результаты концептуально не отличаются от результатов генерации регрессора из хи-квадрат распределения, поэтому в данной работе они не приводятся.

Окончание табл. 1

$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\delta}$	$\hat{\rho}$
$n = 5000$					
-0.397	0.297	0.506	-0.394	-0.688	0.871
$n = 10000$					
-0.387	0.327	0.487	-0.505	-0.429	0.731
$n = 50000$					
-0.392	0.291	0.248	-0.828	-0.641	0.146

Примечание: истинные значения параметров $\alpha_0 = -0.4$, $\alpha_1 = 0.3$, $\beta_0 = -0.3$, $\beta_1 = -0.8$, $\delta = 1.9$ и $\rho = 0.8$.

Источник: расчеты авторов.

Таблица 2

**Оценки параметров модели (1)–(2)
для случая непрерывного экзогенного регрессора
и разных объемов выборки**

$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\delta}$	$\hat{\rho}$
$n = 500$					
-0.852	0.197	-0.342	0.257	-1.125	0.950
$n = 1000$					
-0.864	0.198	-0.659	0.159	0.596	-0.034
$\alpha_1 = 0.2, n = 5000$					
-0.909	0.204	-0.867	0.173	0.799	-0.219
$n = 10000$					
-0.864	0.195	-0.806	0.162	0.767	-0.151
$n = 50000$					
-0.891	0.199	-0.641	0.226	-0.186	0.415

Примечание: истинные значения параметров $\alpha_0 = -0.9$, $\alpha_1 = 0.2$, $\beta_0 = -0.7$, $\beta_1 = 0.2$, $\delta = 0.2$ и $\rho = 0.2$.

Источник: расчеты авторов.

Таблицы 1 и 2 содержат результаты оценивания на основе одной симуляции, чтобы продемонстрировать, что представляет собой проблема идентифицируемости.

Результаты симуляционного анализа для случая бинарного регрессора (см. табл. 1) говорят о том, что полученные оценки далеки от истинных значений соответствующих параметров модели. Например, даже

при $n = 50000$ оценка коэффициента корреляции ρ равна 0.146, в то время как истинное значение равно -0.8 . Такие отличия связаны с проблемой идентифицируемости, вызывающей несостоятельность получаемых оценок.

Результаты симуляций для случая непрерывного регрессора (см. табл. 2) обнаруживают проблему идентификации лишь для параметров δ и ρ , что дает основания полагать существование частичной идентификации в том случае, когда общий для двух уравнений регрессор непрерывен.

Вероятности, предельные эффекты и эффекты воздействия

Теперь перейдем к оцениванию вероятностей, предельных эффектов и эффектов воздействия. В соответствии с предположениями, выдвинутыми в предыдущем разделе, мы не ожидаем никаких проблем с идентифицируемостью вероятностей и предельных эффектов, однако, рассчитываем наблюдать неидентифицируемость эффектов воздействия.

Для измерения точности оценивания используются следующие метрики: корень из среднеквадратического отклонения (root mean squared error, RMSE), среднее абсолютное отклонение (mean absolute error, MAE) и среднее процентное отклонение (mean absolute percentage error, MAPE). Рассчитываются они следующим образом:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_i - \theta_i)^2}, \text{ MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{\theta}_i - \theta_i|$$

$$\text{и MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{\theta}_i - \theta_i}{\theta_i} \right| \cdot 100\%,$$

где $\hat{\theta}_i$ — оценка вероятности, предельного эффекта или эффекта воздействия для i -го наблюдения; θ_i — истинное значение вероятности, предельного эффекта или эффекта воздействия для i -го наблюдения; n — число наблюдений (500, 1000, 5000, 10000 и 50000 наблюдений). Значения метрик RMSE, MAE и MAPE рассчитываются для каждой симуляции и затем усредняются:

$$\text{RMSE} = \frac{1}{1000} \sum_{j=1}^{1000} \text{RMSE}_j, \text{ MAE} = \frac{1}{1000} \sum_{j=1}^{1000} \text{MAE}_j$$

$$\text{и MAPE} = \frac{1}{1000} \sum_{j=1}^{1000} \text{MAPE}_j,$$

где RMSE_j , MAE_j и MAPE_j — рассчитанные значения метрик RMSE, MAE и MAPE для j -й симуляции соответственно. Отметим, что MAPE является наиболее удобной для нас с точки зрения интерпретации метрикой.

Таблицы 3–5 демонстрируют рассчитанные значения метрик для трех совместных вероятностей, с помощью которых вычисляются любые другие вероятности, и для наиболее часто используемых в прикладных исследо-

ваниях типов предельных эффектов. Полученные результаты для других возможных типов предельных эффектов концептуально ничем не отличаются от приведенных в табл. 3–5.

Далее будут использоваться следующие обозначения вероятностей и предельных эффектов:

$$\begin{aligned} \mathbb{P}\{d=0, y=1|x\} &= P_{d=0, y=1}, \mathbb{P}\{d=1, y=0|x\} = P_{d=1, y=0}, \\ \mathbb{P}\{d=1, y=1|x\} &= P_{d=1, y=1}, ME_x(P\{y=1|d=0, x\}) = ME_x(P_{y=1|d=0}), \\ ME_x(P\{y=1|d=1, x\}) &= ME_x(P_{y=1|d=1}) \text{ и } ME_d(P\{y=1|x\}) = ME_d(P_{y=1}). \end{aligned}$$

В табл. 3 приведены значения метрик для вероятностей и предельных эффектов в случае бинарного экзогенного регрессора.

Таблица 3

**Значения метрик для оценок вероятностей и предельных эффектов
для модели (1)–(2) в случае бинарного экзогенного регрессора
и разных объемов выборки**

Метрика	$P_{d=0, y=1}$	$P_{d=1, y=0}$	$P_{d=1, y=1}$	$ME_x(P_{y=1 d=0})$	$ME_x(P_{y=1 d=1})$	$ME_d(P_{y=1})$
<i>n</i> = 500						
RMSE	0.021	0.016	0.024	0.051	0.042	0.038
MAE	0.017	0.015	0.020	0.051	0.042	0.035
MAPE, %	9.7	10.4	7.4	16.4	15.3	11.0
<i>n</i> = 1000						
RMSE	0.015	0.012	0.017	0.038	0.032	0.028
MAE	0.012	0.011	0.014	0.038	0.032	0.026
MAPE, %	7.0	8.3	5.1	11.9	11.8	8.2
<i>n</i> = 5000						
RMSE	0.007	0.005	0.008	0.017	0.015	0.013
MAE	0.005	0.005	0.006	0.017	0.015	0.012
MAPE, %	3.1	3.6	2.3	5.5	5.6	3.7
<i>n</i> = 10000						
RMSE	0.005	0.004	0.005	0.013	0.011	0.010
MAE	0.004	0.003	0.004	0.013	0.011	0.008
MAPE, %	2.2	2.6	1.6	4.1	4.1	2.7
<i>n</i> = 50000						
RMSE	0.002	0.002	0.003	0.005	0.005	0.004
MAE	0.002	0.002	0.002	0.005	0.005	0.004
MAPE, %	1.0	1.9	0.8	1.8	1.9	1.2

Источник: расчеты авторов.

Результаты симуляционного анализа соответствуют ожидаемым, оценки как вероятностей, так и предельных эффектов действительно оказываются близкими к истинным значениям. Более того, их точность растет с ростом объема выборки. Оценки вероятностей оказались немного точнее, чем оценки предельных эффектов. Наименее точными оказываются оценки коэффициента при экзогенном бинарном регрессоре (вероятно, из-за его низкой вариации), однако с ростом числа наблюдений точность оценок возрастает. Так, для $n = 500$ наибольшее среднее значение MAPE составляет 16%, но уже для 5000 наблюдений значение метрики опускается ниже 6%. Оценки предельного эффекта эндогенного регрессора близки к истинным значениям даже при малом объеме выборки. Для 500 наблюдений среднее значение MAPE немного больше 10%, но уже начиная с 1000 наблюдений становится меньше 10%. Более того, при наибольшем рассматриваемом объеме выборки — 50000 наблюдений — среднее значение MAPE снижается до 1–2%. По всей видимости, точность оценок предельных эффектов по эндогенному регрессору выше, чем по экзогенному из-за меньшей вариации последнего.

Рассмотрим результаты оценивания вероятностей и предельных эффектов для случая непрерывного регрессора (см. табл. 4).

Таблица 4

Значения метрик для оценок вероятностей и предельных эффектов для модели (1)–(2) в случае непрерывного экзогенного регрессора и разных объемов выборки

Метрика	$P_{d=0,y=1}$	$P_{d=1,y=0}$	$P_{d=1,y=1}$	$ME_x(P_{y=1 d=0})$	$ME_x(P_{y=1 d=1})$	$ME_d(P_{y=1})$
<i>n</i> = 500						
RMSE	0.022	0.017	0.023	0.009	0.011	0.043
MAE	0.019	0.015	0.021	0.008	0.009	0.040
MAPE, %	10.1	18.6	6.8	13.7	17.3	24.8
<i>n</i> = 1000						
RMSE	0.015	0.012	0.016	0.007	0.007	0.031
MAE	0.014	0.011	0.014	0.006	0.006	0.028
MAPE, %	7.6	12.7	4.7	10.0	12.6	18.3
<i>n</i> = 5000						
RMSE	0.007	0.006	0.007	0.003	0.004	0.014
MAE	0.006	0.005	0.006	0.003	0.003	0.013
MAPE, %	3.4	5.9	2.1	4.7	6.1	8.6

Метрика	$P_{d=0,y=1}$	$P_{d=1,y=0}$	$P_{d=1,y=1}$	$ME_x(P_{y=1 d=0})$	$ME_x(P_{y=1 d=1})$	$ME_d(P_{y=1})$
$n = 10000$						
RMSE	0.005	0.004	0.005	0.002	0.003	0.010
MAE	0.004	0.003	0.005	0.002	0.002	0.009
MAPE, %	2.4	4.5	1.5	3.5	4.6	6.0
$n = 50000$						
RMSE	0.002	0.002	0.002	0.001	0.001	0.005
MAE	0.002	0.002	0.002	0.001	0.001	0.004
MAPE, %	1.1	2.0	0.7	1.8	2.3	3.0

Источник: расчеты авторов.

В случае непрерывного регрессора результаты аналогичны тем, которые были получены для бинарного регрессора. Оценки предельного эффекта по эндогенной переменной $ME_d(P\{y=1|x\})$ наименее точные. Для самого маленького объема выборки среднее значение MAPE равняется 25. Тем не менее для средних и больших объемов выборки, т.е., когда $n \geq 5000$, среднее значение MAPE опускается ниже 9%, доходя до 3% для наибольшего рассматриваемого объема выборки. Одной из причин потери точности является тот факт, что среднее значение MAPE значительно выше для предельных эффектов, истинные значения которых близки к нулю, а доля таких эффектов довольно велика.

Проводить непосредственное сравнение между результатами оценивания для случая модели с бинарным и непрерывным регрессором затруднительно, поскольку модели различаются истинными значениями параметров. Как было отмечено выше, значения параметров подбирались таким образом, чтобы добиться их неидентифицируемости, однако при общем непрерывном экзогенном регрессоре отсутствие идентифицируемости параметров скорее исключение, чем правило.

Независимо от типа экзогенного регрессора значения рассчитываемых метрик уменьшается с ростом числа наблюдений. Отметим, что метрика RMSE более чувствительна к выбросам, чем MAE, однако по результатам симуляций для любого типа экзогенного регрессора и любого объема выборки среднее значение RMSE близко к среднему значению MAE.

Эффекты воздействия демонстрируют иное поведение. В табл. 5 приведены значения используемых метрик для ATE и ATET.

Таблица 5

**Значения метрик для оценок эффектов воздействия для модели (1)–(2)
в случае бинарного и непрерывного экзогенного регрессора
и разных объемов выборки**

Метрики	Бинарный экзогенный регрессор	Непрерывный экзогенный регрессор		
	<i>ATE</i>	<i>ATET</i>	<i>ATE</i>	<i>ATET</i>
<i>n</i> = 500				
RMSE	0.492	0.487	0.314	0.319
MAE	0.492	0.479	0.308	0.306
MAPE, %	76.9	80.2	535.0	546.1
<i>n</i> = 1000				
RMSE	0.500	0.495	0.284	0.286
MAE	0.500	0.486	0.277	0.272
MAPE, %	78.4	81.0	465.2	467.9
<i>n</i> = 5000				
RMSE	0.447	0.436	0.235	0.232
MAE	0.447	0.426	0.228	0.221
MAPE, %	70.0	70.7	364.5	357.7
<i>n</i> = 10000				
RMSE	0.425	0.410	0.207	0.202
MAE	0.424	0.400	0.200	0.191
MAPE, %	66.4	66.3	318.9	308.9
<i>n</i> = 50000				
RMSE	0.374	0.356	0.135	0.132
MAE	0.373	0.344	0.130	0.124
MAPE, %	58.4	60.2	199.3	192.1

Источник: расчеты авторов.

Рассчитанные метрики, относящиеся к эффектам воздействия, остаются довольно большими для выборки объемом 50 000 наблюдений. Например, в случае непрерывного регрессора для предельных эффектов средний MAPE не превышал 3%, а для эффектов воздействия он уже более 190%. Таким образом, есть повод полагать, что неидентифицируемость параметров модели влечет за собой невозможность точно оценить

эффекты воздействия, более того, увеличение объемов выборки не влечет никаких улучшений.

Пример

Рассмотрим проблему идентифицируемости параметров, предельных эффектов и эффектов воздействия на примере оценивания вероятности покупки лекарства индивидом.

Вероятность приобретения лекарства зависит от возраста респондента, пола, семейного положения, его состояния здоровья, дохода и ряда других факторов (Засимова, Коссова, 2016). Логично также предположить, что покупка лекарства тесно связана с посещением индивидом врача. Однако бинарная переменная, отвечающая за факт посещение врача, будет являться эндогенным регрессором, поскольку существуют ненаблюдаемые исследователем факторы, например, плохое самочувствие, тревожность, которые влияют как на обращение к врачу, так и на приобретение лекарства. Необходимость оценивать уравнение бинарного выбора с эндогенным бинарным регрессором определяет выбор рекурсивной системы бинарных уравнений для моделирования покупки лекарств.

Для того чтобы продемонстрировать возможную проблему идентифицируемости параметров и функций от них, оцениваются две иерархические системы: при отсутствии условий ограничений исключения и при их наличии. В качестве уникальной переменной выбирается возраст индивида. Сложность соблюдения ограничений исключения, как уже было сказано выше, заключается в трудности выбора уникальной переменной, и часто приводит к тому, что необходимый инструмент просто исключается из основного уравнения несмотря на то, что это может привести к неверной спецификации модели, а именно, к пропуску существенной переменной. В рассматриваемом примере выбранный инструмент — возраст, исходя из теоретических соображений, может являться существенной переменной для вероятности покупки лекарства.

Гипотеза исследования заключается в том, что отсутствие ограничений исключения не мешает однозначному определению предельных эффектов ни для экзогенных, ни для эндогенных регрессоров, но не дает возможности корректно интерпретировать оценки эффектов воздействия.

Анализ проводится по данным специального опроса выполняющего функции иностранного агента АНО «Левада-Центр», проведенного в 2014 г. Опрос проводился по репрезентативной всероссийской выборке городского и сельского населения. В исследовании принимали участие 1602 человека в возрасте от 18 лет. Опрос содержит перечень специальных вопросов, посвященных здоровью индивида, практике обращений за медицинской помощью и покупке лекарственных средств.

Основной зависимой переменной является переменная *buydrugs* — бинарная переменная, принимающая значение 1, если индивид покупал лекарства за последний месяц, 0 — иначе.

Переменная *docvis*, принимающая значение 1, если индивид посещал врача за последний месяц, и 0 — иначе, является эндогенным регрессором по отношению к основной зависимой переменной *buydrugs*.

Перечень объясняющих экзогенных и контрольных переменных следующий:

- *sex* — бинарная переменная, принимающая значение 1, если индивид является мужчиной, 0 — иначе;
- *marital_status* — бинарная переменная, принимающая значение 1, если индивид состоит в браке, 0 — иначе;
- *chronic_disease* — бинарная переменная, принимающая значение 1, если индивид обладает хроническими заболеваниями, 0 — иначе;
- *retirement* — бинарная переменная, принимающая значение 1, если индивид является (неработающим) пенсионером, 0 — иначе;
- *log(income)* — непрерывная переменная, представляющая собой логарифм среднемесячного дохода индивида;
- *age* — переменная, принимающая целочисленные значения и отражающая возраст индивида.

Таблицы А и В приложений демонстрируют основные дескриптивные статистики для используемых переменных (как зависимых, так и независимых).

Первая модель (модель М1), не удовлетворяющая условиям ограничений исключения, может быть описана следующими уравнениями:

$$docvis = \mathbb{I}\{\alpha_0 + \alpha_1 \cdot sex + \alpha_2 \cdot marital_status + \alpha_3 \cdot chronic_disease + \alpha_4 \cdot retirement + \alpha_5 \cdot \log(income) + \alpha_6 \cdot age + \epsilon^{docvis} \geq 0\}, \quad (3)$$

$$buydrugs = \mathbb{I}\{\beta_0 + \beta_1 \cdot sex + \beta_2 \cdot marital_status + \beta_3 \cdot chronic_disease + \beta_4 \cdot retirement + \beta_5 \cdot \log(income) + \beta_6 \cdot age + \delta \cdot docvis + \epsilon^{buydrugs} \geq 0\}. \quad (4)$$

В качестве эндогенной переменной выступает переменная *docvis*. Уравнение (3) задает данную переменную, а уравнение (4) содержит ее в качестве регрессора. Набор экзогенных регрессоров одинаков для уравнений (3) и (4).

Вторая модель (модель М2) уже удовлетворяет условиям ограничений исключения. Уравнение, задающее эндогенную переменную *docvis*, совпадает с уравнением (3). В качестве уникальной переменной выбрана переменная *age*. Она принудительно исключена из второго уравнения модели М2. Второе уравнение модели М2 представимо в следующем виде:

$$buydrugs = \mathbb{I}\{\beta_0 + \beta_1 \cdot sex + \beta_2 \cdot marital_status + \beta_3 \cdot chronic_disease + \\ + \beta_4 \cdot retirement + \beta_5 \cdot \log(income) + \delta \cdot docvis + \epsilon^{buydrugs} \geq 0\}. (5)$$

Таким образом, модель М1 не удовлетворяет ограничениям исключений, описывается уравнениями (3) и (4) и имеет следующий набор из 16 оцениваемых параметров: $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \delta$ и ρ . Модель М2 удовлетворяет условиям ограничений исключения, описывается уравнениями (3) и (5) и имеет следующий набор из 15 оцениваемых параметров: $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \delta$ и ρ . Параметр β_6 отсутствует в модели М2, поскольку уравнение (5) модели М2 не содержит переменную *age*. Параметр ρ — это коэффициент корреляции между случайными ошибками. Для обеих моделей предполагается, что случайные ошибки имеют двумерное стандартное нормальное распределение. Обе модели построены на основе 1074 наблюдений (количество наблюдений сократилось с 1602 до 1074 после удаления наблюдений, содержащих пропущенные по интересующим нас переменным значения).

В табл. 6 приведены результаты оценивания параметров обеих моделей М1 и М2.

Таблица 6

Результаты оценивания параметров моделей М1 и М2

Параметр	Оценки коэффициентов модели М1	Оценки коэффициентов модели М2
Уравнение посещения врача (зависимая переменная <i>docvis</i>)		
<i>константа</i>	−0.42511 (0.84269)	−0.54746 (0.85230)
<i>sex</i>	−0.43069*** (0.08599)	−0.43380*** (0.08257)
<i>marital_status</i>	0.22422** (0.09045)	0.21228** (0.08883)
<i>chronic_disease</i>	1.02349*** (0.10568)	1.00761*** (0.10428)
<i>retirement</i>	0.23454 (0.14785)	0.16842 (0.14966)
<i>log(income)</i>	−0.02216 (0.07955)	−0.01683 (0.07994)

Параметр	Оценки коэффициентов модели M1	Оценки коэффициентов модели M2
<i>age</i>	0.00829 ** (0.00371)	0.01047 *** (0.003673)
Уравнение покупки лекарств (зависимая переменная <i>buydrugs</i>)		
<i>константа</i>	−3.52122 *** (1.06941)	−3.04430 *** (0.88692)
<i>sex</i>	−0.32450 ** (0.15010)	−0.21265 ** (0.10889)
<i>marital_status</i>	0.09402 (0.11373)	0.07224 (0.09810)
<i>chronic_disease</i>	0.80392 ** (0.33227)	0.57074 * (0.21665)
<i>retirement</i>	0.36907 ** (0.16469)	0.44783 *** (0.14493)
<i>log(income)</i>	0.20756 ** (0.09201)	0.17861 ** (0.08442)
<i>age</i>	0.00968 ** (0.00461)	
<i>docvis</i> , $\hat{\delta}$	1.19772 (0.86632)	1.69942 *** (0.36853)
<i>корреляция</i> , $\hat{\rho}$	−0.31700 (0.52856)	−0.65300 ** (0.23761)

Примечание: *** — p-value ≤ 0.01 , ** — p-value ≤ 0.05 и * — p-value ≤ 0.1 .

Источник: расчеты авторов.

Как видно из табл. 6, оценки параметров обеих моделей довольно близки по своим значениям. Однако есть существенная разница в интерпретации результатов. Согласно модели M1 и коэффициент при посещении врача, и коэффициент корреляции являются незначимыми, хотя гипотеза о совместной значимости этих параметров не отвергается (см. табл. С приложений). Остается неясным, следует ли считать переменную, отвечающую за посещение врача, эндогенной или нет, и следует ли ее включать в модель. Согласно модели M2 переменная *docvis* оказывает значимое влияние на вероятность покупки лекарства и является эн-

догенной. Положительный коэффициент при данной переменной говорит о том, что посещение врача, при прочих равных, увеличивает вероятность покупки лекарства. В то же время коэффициент корреляции случайных ошибок значим и отрицателен, что свидетельствует о самоотборе наблюдений при посещении врача, т.е., что существуют ненаблюдаемые характеристики, влияющие на вероятность покупки и вероятность обращения к врачу в противоположном направлении. Например, дефицит общения и наличие свободного времени у пожилых людей могут увеличивать вероятность похода в поликлинику, но не вероятность покупки лекарства. В свою очередь, тот факт, что пенсионер регулярно посещает поликлинику, может свидетельствовать о его активности и относительно неплохом состоянии здоровья для его возраста, что уменьшает вероятность приобретения препаратов в условиях ограниченного пенсией бюджета.

Различие в оценках может быть обусловлено как сложностями с идентификацией параметров при отсутствии ограничений исключения модели M1, так и пропуском существенной переменной возраст и ошибкой спецификации модели M2.

Рассмотрим предельные эффекты и эффекты воздействия. В соответствии с симуляционным анализом ожидается, что если отличия в моделях M1 и M2 были обусловлены проблемами с идентификацией параметров, то оценки их предельных эффектов должны оказаться близкими, а оценки эффектов воздействия могут существенно отличаться.

В табл. 7 демонстрируются значения оценок среднего предельного эффекта на вероятность $\mathbb{P}\{buydrugs = 1 | docvis = 1\}$ по каждой экзогенной переменной (*sex*, *marital_status*, *chronic_disease*, *retirement*, *log(income)* и *age*) и на вероятность $\mathbb{P}\{buydrugs = 1\}$ по эндогенной переменной *docvis*¹.

Таблица 7

Результаты оценивания средних предельных эффектов в моделях M1 и M2

Переменная	Оценки средних предельных эффектов модели M1	Оценки средних предельных эффектов модели M2
<i>sex</i>	−0.09815*** (0.02904)	−0.08006*** (0.01921)
<i>marital_status</i>	0.03228 (0.02333)	0.03168*** (0.01874)
<i>chronic_disease</i>	0.29041*** (0.06216)	0.23784*** (0.05205)

¹ Строго говоря, оцененный средний предельный эффект рассчитывался для условных вероятностей при заданных значениях переменных *sex*, *marital_status*, *chronic_disease*, *retirement*, *log(income)*, *age* и *docvis*, однако условие было опущено для краткости записи.

Переменная	Оценки средних предельных эффектов модели M1	Оценки средних предельных эффектов модели M2
<i>retirement</i>	0.10457 ** (0.04389)	0.11254 *** (0.03799)
<i>log(income)</i>	0.00266 ** (0.00107)	0.00081 ** (0.00035)
<i>age</i>	0.04815 ** (0.02432)	0.03599 *** (0.01721)
<i>docvis</i>	0.20910 *** (0.03098)	0.20622 *** (0.03044)

Примечание: *** — $p\text{-value} \leq 0.01$, ** — $p\text{-value} \leq 0.05$ и * — $p\text{-value} \leq 0.1$.

Источник: расчеты авторов.

Полученные результаты свидетельствуют в пользу того, что идентификация предельных эффектов возможна даже в условиях отсутствия ограничений исключения. Оценки средних предельных эффектов не только близки между собой, но и являются значимыми, а следовательно, могут быть корректно интерпретированы как для модели M1, так и для модели M2. Исключение составляет средний предельный эффект по переменной *marital_status* для модели M1, немного уступающий в значимости аналогичному эффекту в модели M2.

Оценки средних предельных эффектов согласуются со здравым смыслом. Например, в модели M2 оценка среднего предельного эффекта по переменной *sex* равна -0.08 . Это говорит о том, что вероятность покупки лекарственных средств для мужчины, посетившего врача, в среднем на 8 п. п. ниже, чем для женщины. Таким образом, мужчины реже посещают врачей и реже покупают лекарства, чем женщины. Остальные переменные (возраст, доход, наличие хронических заболеваний, статус пенсионера) оказывают положительное влияние на условную вероятность покупки лекарств после посещения врача. Наиболее весомым фактором является наличие хронических заболеваний, которое увеличивает вероятность более, чем на 25 п. п.

Оценка среднего предельного эффекта по эндогенной переменной представляет отдельный интерес. Как видно из табл. 7, значения средних предельных эффектов в моделях M1 и M2 значимы и практически совпадают, при том, что оценки соответствующих коэффициентов существенно отличались. В среднем, для индивида, посетившего врача, вероятность покупки лекарства на 21 п. п. выше, чем для индивида, который врача не посещал. Полученный результат свидетельствует в пользу

того, что различия в оценках параметров модели были вызваны именно проблемами с идентификацией, а не нарушением спецификации модели.

Теперь рассмотрим эффекты воздействия. В табл. 8 приведены результаты оценивания средних эффектов воздействия (АТЕ и АТЕТ) для каждой из моделей.

Таблица 8

**Результаты оценивания средних эффектов воздействия
в моделях М1 и М2**

Эффект воздействия	Оценки средних эффектов воздействия модели М1	Оценки средних эффектов воздействия модели М1
<i>АТЕ</i>	0.38978 (0.27779)	0.56923 *** (0.12167)
<i>АТЕТ</i>	0.35670 (0.26590)	0.48300 *** (0.10451)

Примечание: *** — $p\text{-value} \leq 0.01$, ** — $p\text{-value} \leq 0.05$ и * — $p\text{-value} \leq 0.1$.

Источник: расчеты авторов.

Выдвинутая гипотеза о наличии проблем идентифицируемости параметров подтверждается полученными оценками средних эффектов воздействия. Помимо визуального различия в оценках для двух моделей, в модели М1 наблюдается та же ситуация, что и для коэффициентов при эндогенной переменной и корреляции: ни АТЕ, ни АТЕТ не являются значимыми, и не могут быть интерпретированы. При соблюдении условий ограничений исключений оба эффекта значимы и равны соответственно 0.57 и 0.48, что говорит о том, что общий условный на контрольные переменные средний эффект воздействия посещения врача на вероятность покупки лекарства равен 0.57, а аналогичный средний эффект воздействия для подвергшихся воздействию равен 0.48.

Заключение

В настоящей работе рассмотрены проблемы идентифицируемости параметров, предельных эффектов и эффектов воздействия популярной среди прикладных исследователей рекурсивной двумерной пробит-модели.

Наличие ограничений исключения — уникальных экзогенных регрессоров или инструментов, входящих в уравнение для эндогенной переменной, — гарантирует идентификацию параметров модели и любых функций от них. Однако возможность соблюдения этих ограничений в прикладных исследованиях есть далеко не всегда.

На основании теоретических соображений и симуляционного анализа, проведенного в настоящем исследовании, можно сформулировать два основных вывода, важных для прикладных работ:

1. Если у исследователя нет возможности соблюсти ограничения исключения, он может интерпретировать результаты оценивания модели с помощью предельных эффектов. Даже если значения параметров не определяются однозначно, значения предельных эффектов идентифицируются всегда.
2. Наличие единственного общего непрерывного экзогенного регрессора может приводить к потере идентификации независимо от того, является ли его носитель ограниченным или неограниченным. В этом случае оценки параметров и АТЕ могут существенно отличаться от истинных и даже иметь другой знак.

Численный эксперимент показал, что в RBP модели с непрерывным регрессором отсутствие условий ограничений исключений создает проблему идентифицируемости менее, чем в четверти случаев.

На примере оценивания вероятности покупки индивидом лекарств в зависимости от посещения врача показано, что отсутствие ограничений исключения может проявляться и как ложная незначимость параметров модели, однако предельные эффекты в этом случае будут значимы, что является дополнительным аргументом в пользу интерпретации результатов оценивания с помощью предельных эффектов.

В работе рассмотрена рекурсивная система с двумя уравнениями, однако, полученный результат можно легко распространить на случай произвольного числа уравнений. Идея оценивания не самих параметров модели, а интерпретируемых функций от них может быть полезна для более широкого класса моделей.

Список литературы

- Засимова, Л., & Коссова, Е. (2016). Расходы населения России на лекарственные средства: эмпирический анализ. *Прикладная эконометрика*, 42, 75–99.
- Ayalew, M., & Xianzhi, Z. (2019). The effect of financial constraints on innovation in developing countries: Evidence from 11 African countries. *Asian Review of Accounting*, 28 (3), 273–308. <https://doi.org/10.1108/ara-02-2019-0036>
- Blasch, J., Filippini, M., & Kumar, N. (2019). Boundedly rational consumers, energy and investment literacy, and the display of information on household appliances. *Recourse and Energy Economics*, 59, 39–58. <https://doi.org/10.2139/ssrn.2789350>
- Castello, J. (2012). Promoting employment of disabled women in Spain; Evaluating a policy. *Labour Economics*, 19 (1), 82–91. <https://doi.org/10.1016/j.labeco.2011.08.003>
- Cohen, J., Azarova, V., Kollmann, A., & Reichl, J. (2019). Q-complementarity in household adoption of photovoltaics and electricity-intensive goods: The case of electric vehicles. *Energy Economics*, 83, 567–577. <https://doi.org/10.1016/j.eneco.2019.08.004>
- Dogbe, W. (2020). Can poverty status explain obesity in developing countries? Evidence from Ghana. *Agribusiness*, 37 (2), 409–421. <https://doi.org/10.1002/agr.21663>

Filippini, M., Greene, W., Kumar, N., & Martinez-Cruz, A. (2018). A note on the different interpretation of the correlation parameters in the bivariate and the recursive bivariate probit. *Economics Letters*, 167, 104–107. <https://doi.org/10.2139/ssrn.3038804>

Freedman, D., & Sekhon, J. (2020). Endogeneity in probit response model. *Cambridge University Press*, 18 (2), 138–150. <https://doi.org/10.2139/ssrn.1138489>

Frondel, M., Horbach, J., & Rennings, K. (2008). What triggers environmental management and innovation? Empirical evidence for Germany. *Ecological Economics*, 66 (1), 153–160. <https://doi.org/10.2139/ssrn.556945>

Han, S., & Lee, S. (2019). Estimation in a generalization of bivariate probit models with dummy endogenous regressors. *Journal of Applied Econometrics*, 34 (6), 994–1015. <https://doi.org/10.2139/ssrn.3233422>

Han, S., & Vytlačil, E. (2017). Identification in a generalization of bivariate probit models with dummy endogenous regressors. *Journal of Econometrics*, 199 (1), 63–73. <https://doi.org/10.1016/j.jeconom.2017.04.001>

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46 (4), 931–959. <https://doi.org/10.3386/w0177>

Ma, W., Abdulai, A., & Goetz, R. (2017). Agricultural cooperative and investment in organic soil amendments and chemical fertilizer in China. *American Journal of Agricultural Economics*, 100 (2), 502–520. <https://doi.org/10.1093/ajae/aax079>

Maddala, G. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press. <https://doi.org/10.1017/ccol052124143x.005>

Martin, E., & Quintata, F. (2022). Consistency and identifiability revisited. *Brazilian Journal of Probability and Statistics*, 16, 99–106.

Morris, S. (2007). The impact of obesity on employment. *Labour Economics*, 14 (3), 413–433. <https://doi.org/10.1016/j.labeco.2006.02.008>

Mourifie, I., & Meango, R. (2014). A note on the identification in two equations probit model with dummy endogenous regressor. *Economics Letters*, 125 (3), 360–363. <https://doi.org/10.2139/ssrn.2340187>

Shaikh, A., & Vytlačil, E. (2011). Partial identification in triangular systems of equations with binary dependent variables. *Econometrica*, 79 (3), 949–955. <https://doi.org/10.3982/ecta9082>

Sirven, N., & Debrand, T. (2012). Social capital and health of older Europeans: Casual pathways and health inequalities. *Social Science & Medicine*, 75 (7), 1288–1295. <https://doi.org/10.1016/j.socscimed.2012.05.009>

Wilde, J. (2000). Identification of multiple equation probit models with endogenous dummy regressor. *Economics Letters*, 69 (3), 309–312. [https://doi.org/10.1016/s0165-1765\(00\)00320-7](https://doi.org/10.1016/s0165-1765(00)00320-7)

References

Zasimova, L., & Kossova, E. (2016). Empirical analysis of out-of-pocket expenditures on medicine in Russia. *Applied Econometrics*, 42, 75–99.

Приложения

Таблица А

Основные дескриптивные статистики бинарных переменных (всего 1074 наблюдения)

Переменная	Описание	Число наблюдений
<i>buydrugs</i> (покупка лекарств)	1 — покупал лекарства 0 — не покупал лекарства	504 570
<i>docvis</i> (посещение врача)	1 — посещал врача 0 — не посещал врача	552 522
<i>sex</i> (пол индивида)	1 — мужчина 0 — женщина	461 613
<i>marital_status</i> (семейное положение)	1 — состоит в браке 0 — не состоит в браке	659 415
<i>chronic_disease</i> (хронические заболевания)	1 — имеет хронические заболевания 0 — не имеет хронические заболевания	358 716
<i>retirement</i> (статус пенсионера)	1 — является (неработающим) пенсионером 0 — не является пенсионером	249 825

Источник: расчеты авторов.

Таблица В

Основные дескриптивные статистики непрерывных переменных (всего 1074 наблюдения)

Переменная	Среднее	Стандартное отклонение	Медиана	Минимум	Максимум
<i>income</i> (доход)	36 004.58	25 299.94	30 000.00	1500.00	210 000.00
<i>log(income)</i> (логарифм дохода)	10.29	0.64	10.31	7.31	12.25
<i>age</i> (возраст)	44.65	16.34	44	18	87

Источник: расчеты авторов.

Таблица С

**Результаты тестирования гипотезы о значимости модели
в целом и гипотезы об одновременной значимости коэффициентов δ и ρ
для моделей М1 и М2 при помощи теста отношения правдоподобия**

Гипотеза	Значения тестовой статистики и p-value для модели М1	Значения тестовой статистики и p-value для модели М2
Гипотеза о значимости модели в целом	LR = 475.09, df = 13 p-value = 0.0000	LR = 469.85, df = 12 p-value = 0.0000
Гипотеза об одновременной значимости коэффициентов δ и ρ	LR = 53.27, df = 2 p-value = 0.0000	LR = 48.04, df = 2 p-value = 0.0000

Источник: расчеты авторов.