

Прикладная эконометрика, 2022, т. 67, с. 121–143.

Applied Econometrics, 2022, v. 67, pp. 121–143.

DOI: 10.22394/1993-7601-2022-67-121-143

Е. В. Коссова, Б. С. Потанин¹

Оценивание модели с гауссовским мультиномиальным эндогенным переключением и неслучайным отбором

Предлагается двухшаговая процедура оценивания модели с мультиномиальным переключением при предположении о совместном нормальном распределении случайных ошибок. На симулированных данных демонстрируется преимущество предлагаемого подхода над популярными альтернативами при нарушении допущения о независимости от посторонних альтернатив. С помощью предложенного метода оценивается премия за брак для мужчин. Получены статистические свидетельства в пользу того, что для состоящих в зарегистрированном браке данная премия является отрицательной, а для состоящих в незарегистрированном — положительной.

Ключевые слова: модели с переключением; неслучайный отбор; эндогенность.

JEL classification: C34; J12.

Введение

В микроэконометрике широко используются модели с переключением, в которых вид основного уравнения определяется в соответствии с некоторым правилом. Правило, «переключающее режимы» основного уравнения, может задаваться значением переменной, эндогенной по отношению к данному уравнению. Начиная с основополагающей работы Дж. Хекмана (Heckman, 1979), экономическая литература изобилует прикладными исследованиями, использующими предложенный им метод оценивания регрессий с эндогенным бинарным переключением. Теоретическое развитие подхода Хекмана шло в двух направлениях: отказ от предпосылки о совместном нормальном распределении за счет использования более гибкой спецификации модели (Pigini, 2015) и переход к более сложным правилам переключения. Например, в работе (Коссова, Потанин, 2018) рассматривается случай, когда переключение задается несколькими бинарными правилами. Однако проблема оценивания моделей с механизмами переключения, альтернативными бинарному, остается слабо изученной. В частности, в некоторых исследованиях возникает необходимость учета

¹ **Коссова Елена Владимировна** — Национальный исследовательский университет «Высшая школа экономики», Москва; ekossova@hse.ru.

Потанин Богдан Станиславович — Национальный исследовательский университет «Высшая школа экономики», Москва; bogdanpotanin@gmail.com.

переключения, определяемого мультиномиальным (множественным) правилом отбора, при котором наблюдается лишь одна из нескольких возможных альтернатив. Так, в исследовании (Лукьянова, 2021) учитывается, что форма уравнения зарплаты может различаться в зависимости от сектора занятости индивида: частного или государственного. Кроме того, уравнение зарплаты не наблюдается для безработных. В результате возникают три «режима» уравнения зарплаты, из которых для каждого индивида исследователь наблюдает лишь тот, который был выбран самим индивидом.

Для оценивания моделей с мультиномиальным переключением, как правило, используется процедура, аналогичная методу Хекмана (Heckman, 1979). На первом шаге оцениваются параметры мультиномиального уравнения отбора (сектор занятости), а на втором — полученные оценки используются для оценивания параметров целевого уравнения (зарплата). Для упрощения процедуры оценивания на первом шаге исследователи используют мультиномиальную логит-модель, существенный недостаток которой заключается в допущении независимости от посторонних альтернатив (Bourguignon et al., 2007). Это предположение ограничивает применение модели в тех случаях, когда существуют близкие альтернативы, например, при выборе вида транспорта (McFadden, 1973).

В настоящем исследовании предлагается модель, ослабляющая предположение о независимости от посторонних альтернатив за счет использования на первом шаге мультиномиальной пробит-модели. Также на основе метода (Newey, 2009) вводятся полупараметрические модификации предложенного и классического подходов, позволяющие оценивать более гибкую спецификацию модели, отказавшись от предпосылок о совместном нормальном распределении ошибок.

В данной работе для реализации процедуры оценивания регрессии с переключением, определяемым мультиномиальной пробит-моделью, выводится формула для условного математического ожидания случайной ошибки целевого уравнения и обосновывается асимптотическая нормальность оценок предлагаемого метода. Сравнение предложенных в работе подходов с существующими проводится на симулированных данных, а также на примере оценивания отдачи от брака для мужчин. Цель исследования — показать важность учета взаимосвязи альтернатив при оценивании регрессии с мультиномиальным эндогенным переключением и предложить удобную для реализации процедуру оценивания, учитывающую зависимость всех случайных составляющих и легко адаптируемую к более гибкой спецификации модели.

1. Модель с мультиномиальным эндогенным переключением

1.1. Метод Дубина–Макфаддена

В регрессии с переключением, задаваемым номинальной переменной z_i , принимающей J значений, коэффициенты регрессии при объясняющих переменных и распределение случайной ошибки определяются значением $z_i \in \{1, 2, \dots, J\}$:

$$y_i = y_{z_i}^*, \quad \text{где } y_{z_i}^* = x_i' \beta_{z_i} + \varepsilon_{z_i i}, \quad (1)$$

β_j и x_i — вектор-столбцы коэффициентов и объясняющих переменных, а ε_{ij} — случайная ошибка. При этом сама переменная z_i часто интерпретируется как выбор индивидом i альтернативы j , имеющей наибольшую полезность z_{ji}^* :

$$z_i = j, \text{ если } z_{ji}^* > \max_{k \neq j} z_{ki}^*, \quad z_{ji}^* = w_i' \gamma_j + u_{ji}, \quad i \in \{1, \dots, n\}, \quad j \in \{1, \dots, J\}. \quad (2)$$

Латентная переменная z_{ji}^* определяет полезность выбора j -й из J взаимоисключающих альтернатив. Эта полезность зависит от случайной ошибки u_{ji} и линейного индекса $w_i' \gamma_j$, где w_i и γ_j являются вектор-столбцами регрессоров и регрессионных коэффициентов соответственно. Значение переменной z_i определяется в соответствии с правилом множественного выбора: индивид выбирает альтернативу с наибольшей полезностью. Каждому возможному значению z_i соответствует свой механизм (режим) формирования латентной целевой переменной y_{ji}^* . Число режимов совпадает с числом альтернатив J . Однако на практике исследователь наблюдает значение y_i , полученное в соответствии лишь с одним из возможных режимов, в зависимости от наблюдаемого значения z_i . Соответствующие режимы различаются величинами регрессионных коэффициентов β_j и дисперсиями случайных ошибок ε_{ji} . При этом в модели явным образом не вводится предположение о виде совместного распределения случайных ошибок уравнений латентных целевых переменных $\varepsilon_{1i}, \dots, \varepsilon_{ji}$. Отдельного внимания заслуживает случай, когда при выборе некоторых альтернатив значения объясняемой переменной y_i не наблюдаются. В классической работе Дж. Хекмана (Hecman, 1979) значения объясняемой переменной предполагались наблюдаемыми при выборе одной из двух альтернатив и ненаблюдаемыми при выборе другой. Если значения объясняемой переменной наблюдаются только при выполнении некоторого правила, то говорят о неслучайном отборе наблюдений.

В данной статье рассматривается наиболее общий случай, когда случайные составляющие уравнений полезности и основного уравнения могут быть зависимыми, и их ковариационная матрица имеет недиагональный вид:

$$\Sigma^{(j)} = \text{Cov} \begin{bmatrix} u_{1i} \\ u_{2i} \\ \dots \\ u_{ji} \\ \varepsilon_{ji} \end{bmatrix} = \begin{bmatrix} \sigma_{u_1}^2 & \rho_{12} \sigma_{u_1} \sigma_{u_2} & \rho_{13} \sigma_{u_1} \sigma_{u_3} & \dots & \rho_1^{(j)} \sigma_{u_1} \sigma_{\varepsilon_j} \\ \rho_{12} \sigma_{u_1} \sigma_{u_2} & \sigma_{u_2}^2 & \rho_{23} \sigma_{u_2} \sigma_{u_3} & \dots & \rho_2^{(j)} \sigma_{u_2} \sigma_{\varepsilon_j} \\ \rho_{13} \sigma_{u_1} \sigma_{u_3} & \rho_{23} \sigma_{u_2} \sigma_{u_3} & \sigma_{u_3}^2 & \dots & \rho_3^{(j)} \sigma_{u_3} \sigma_{\varepsilon_j} \\ \dots & \dots & \dots & \dots & \dots \\ \rho_{1j} \sigma_{u_1} \sigma_{u_j} & \rho_2^{(j)} \sigma_{u_2} \sigma_{\varepsilon_j} & \rho_3^{(j)} \sigma_{u_3} \sigma_{\varepsilon_j} & \dots & \sigma_{\varepsilon_j}^2 \end{bmatrix}. \quad (3)$$

Оценивание коэффициентов β_j при помощи метода наименьших квадратов (МНК) может дать несостоятельные оценки, поскольку условное математическое ожидание целевой переменной имеет вид

$$E(y_i | z_i = j, x_i) = x_i' \beta_j + E(\varepsilon_{ji} | z_i = j, x_i),$$

где условное математическое ожидание случайной ошибки в общем случае может не равняться нулю, поскольку зависит от результата множественного выбора.

Процедура оценивания параметров модели зависит от предположения о совместном распределении случайных ошибок. Среди различных подходов, подробно описанных в (Schmertmann, 1994; Bourguignon et al., 2007), наибольшую популярность приобрел метод Дубина–Макфаддена (Dubin, McFadden, 1984). Данный подход опирается на предположение

о том, что случайные ошибки u_{ji} независимы и имеют распределение Гумбеля, а условное математическое ожидание случайной ошибки (при условии $z_i = j$) может быть записано как

$$E(\varepsilon_{ji} | z_i = j, x_i) = \sum_{t=1}^J \frac{\sqrt{6}}{\pi} \sigma_{\varepsilon_j} \rho_t^{(j)} \lambda_{ti}, \quad (4)$$

где

$$\lambda_{ti} = \begin{cases} -\ln P(z_i = t), & \text{если } t = j, \\ \frac{P(z_i = t) \ln P(z_i = t)}{1 - P(z_i = t)}, & \text{в противном случае.} \end{cases} \quad (5)$$

В этом случае состоятельные оценки коэффициентов β_j можно получить при помощи двухшаговой процедуры, аналогичной (Hesckman, 1979). На первом шаге оцениваются параметры уравнения множественного отбора при помощи мультиномиальной логит-модели. С помощью полученных оценок нетрудно оценить вероятности $P(z_i = t)$, а затем и λ_{ti} . На втором шаге полученные оценки подставляются в регрессионное уравнение для y_{ji} , где

$\frac{\sqrt{6}}{\pi} \sigma_{\varepsilon_j} \rho_t^{(j)}$ выступают в качестве регрессионных коэффициентов при $\hat{\lambda}_{ti}$:

$$y_{ji}^* = x_i' \beta_j + \left(\frac{\sqrt{6}}{\pi} \sigma_{\varepsilon_j} \rho_1^{(j)} \right) \hat{\lambda}_{1i} + \dots + \left(\frac{\sqrt{6}}{\pi} \sigma_{\varepsilon_j} \rho_J^{(j)} \right) \hat{\lambda}_{Ji} + v_{ji}. \quad (6)$$

В уравнении (6) случайная ошибка v_{ji} является гетероскедастичной, поэтому возникает необходимость в коррекции оценки асимптотической ковариационной матрицы регрессионных коэффициентов, для чего на практике, как правило, применяется бутстрап.

Основной недостаток описанного метода связан с одной из наиболее серьезных проблем мультиномиальной логит-модели — предположением о независимости от посторонних альтернатив. При несоблюдении этого условия форма условного математического ожидания случайной ошибки целевого уравнения может существенно отличаться от предполагаемой, что может привести к серьезному падению точности оценок метода.

В исследованиях (Schmertmann, 1994; Bourguignon et al., 2007) точность оценок метода Дубина–Макфаддена, а также ряда альтернатив, была изучена на симулированных данных. Кратко опишем основные выводы данных работ. Во-первых, наличие множественного неслучайного отбора может привести к серьезному смещению МНК оценок, что мотивирует применение подходов, осуществляющих соответствующую коррекцию. Во-вторых, метод Дубина–Макфаддена позволяет получить достаточно точные оценки даже в случае нарушения предположения о независимости от посторонних альтернатив, когда совместное распределение случайных ошибок является многомерным нормальным, в том числе в случае, когда все корреляции отличны от нуля. В-третьих, данный метод практически не уступает в точности другим методам коррекции множественного неслучайного отбора в случаях, когда процесс генерации данных осуществляется в соответствии с предпосылками альтернативных подходов, накладывающих жесткие ограничения на структуру ковариационной матрицы случайных ошибок (Bourguignon et al., 2007).

Недостаток существующих на сегодняшний день исследований заключается в отсутствии работ, изучающих работоспособность метода Дубина–Макфаддена при наличии высокой корреляции между альтернативами (0.5 и более). Остается неясным, насколько правомерно

применять этот подход при наличии близких альтернатив, и какая процедура может быть использована в этом случае. Ниже будут предложены подходы, позволяющие получать состоятельные оценки коэффициентов регрессии при произвольной структуре невырожденной ковариационной матрицы случайных ошибок $\Sigma^{(j)}$, и проведено сопоставление точности их оценок с оценками Дубина–Макфаддена.

1.2. Регрессия с мультиномиальным гауссовским переключением и двухшаговая процедура ее оценивания

Регрессия с гауссовским переключением — это регрессия с эндогенным мультиномиальным переключением в предположении о том, что совместное распределение случайных ошибок уравнений множественного выбора и случайной ошибки каждого из регрессионных уравнений является многомерным нормальным с нулевым вектором математических ожиданий и ковариационной матрицей $\Sigma^{(j)}$.

Без потери общности, из соображений идентифицируемости параметров системы предполагается, что $\Sigma_{11}^{(j)} = 1$ и $z_{ji}^* = 0$. Последнее равенство позволяет исключить случайную ошибку u_{ji} . В результате совместное распределение оставшихся случайных составляющих остается многомерным нормальным с ковариационной матрицей Σ^j , получаемой из исходной $\Sigma^{(j)}$ удалением J -й строки и J -го столбца.

Для удобства введем вспомогательные обозначения для разностей случайных ошибок и линейных индексов уравнений множественного выбора. При $j < J$ обозначим:

$$\tilde{u}^{(ji)} = [u_{1i} - u_{ji}, u_{2i} - u_{ji}, \dots, u_{(j-1)i} - u_{ji}, u_{(j+1)i} - u_{ji}, \dots, u_{(J-1)i} - u_{ji}, -u_{ji}]',$$

$$z^{(ji)} = [w'_i(\gamma_j - \gamma_1), w'_i(\gamma_j - \gamma_2), \dots, w'_i(\gamma_j - \gamma_{j-1}), w'_i(\gamma_j - \gamma_{j+1}), \dots, w'_i(\gamma_j - \gamma_{J-1}), w'_i\gamma_j]'$$

Для случая $j = J$:

$$\tilde{u}^{(Ji)} = [u_{1i}, u_{2i}, \dots, u_{(J-1)i}]', \quad z^{(Ji)} = [-w'_i\gamma_1, -w'_i\gamma_2, \dots, -w'_i\gamma_{J-1}]'$$

Используя свойства усеченного нормального распределения, описанные в (Косова, Потанин, 2018), можно показать, что

$$E(\varepsilon_{ji} | z_i) = E(\varepsilon_{ji} | \tilde{u}_1^{(ji)} < \tilde{z}_1^{(ji)}, \dots, \tilde{u}_{J-1}^{(ji)} < \tilde{z}_{J-1}^{(ji)}) = -\sum_{t=1}^{J-1} \tilde{\Sigma}_{tj}^j \lambda_{ti}^{(j)}, \quad (7)$$

где

$$\tilde{\Sigma}^j = \text{Cov} \begin{bmatrix} \tilde{u}^{(ji)} \\ \varepsilon_{ji} \end{bmatrix},$$

$$\lambda_i^{(j)} = \lambda_i^{(j)}(\tilde{z}_1^{(ji)}, \dots, \tilde{z}_{J-1}^{(ji)}) = [\lambda_{1i}^{(j)}, \dots, \lambda_{(J-1)i}^{(j)}] = \nabla \left(\ln F_{ji}(\tilde{z}_1^{(ji)}, \dots, \tilde{z}_{J-1}^{(ji)}) \right) =$$

$$= \left[\frac{\partial F_{ji}(\tilde{z}_1^{(ji)}, \dots, \tilde{z}_{J-1}^{(ji)})}{\partial \tilde{z}_1^{(ji)}}, \dots, \frac{\partial F_{ji}(\tilde{z}_1^{(ji)}, \dots, \tilde{z}_{J-1}^{(ji)})}{\partial \tilde{z}_{J-1}^{(ji)}} \right] / F_{ji}(\tilde{z}_1^{(ji)}, \dots, \tilde{z}_{J-1}^{(ji)}), \quad (8)$$

а F_{ji} — функция распределения $(J-1)$ -мерного случайного вектора $\tilde{u}^{(ji)}$ с нулевым вектором математических ожиданий и ковариационной матрицей, получаемой из $\tilde{\Sigma}^j$ удалением последней строки и последнего столбца. Выражение для частных производных многомерного нормального распределения можно найти, например, в (Коссова, Потанин, 2018).

Таким образом, условное математическое ожидание случайной ошибки основного уравнения является линейной комбинацией обобщенных обратных отношений Миллса (8), оценки которых можно получить как функцию от ММП-оценок параметров мультиномиальной пробит-модели (первый шаг). Следовательно, в соответствии с теоремой о двухшаговой процедуре оценивания (Murphy, Topel, 1985), состоятельные и асимптотически нормальные оценки параметров β_j можно получить за счет оценивания с помощью МНК следующего регрессионного уравнения:

$$y_{ji}^* = x_i \beta_j - \sum_{t=1}^{J-1} \tilde{\Sigma}_{tj}^j \hat{\lambda}_{ti}^{(j)} + v_{ji} = x_i \beta_j + (-\tilde{\Sigma}_{1j}^j) \hat{\lambda}_{1i}^{(j)} + \dots + (-\tilde{\Sigma}_{(J-1)j}^j) \hat{\lambda}_{(J-1)i}^{(j)} + v_{ji}, \quad (9)$$

где взятые с обратным знаком ковариации $(-\tilde{\Sigma}_{tj}^j) = \text{Cov}(-\tilde{u}_t^{(ji)}, \varepsilon_{ji}) = \beta_{\lambda_{ti}^{(j)}}$ выступают в качестве дополнительных оцениваемых параметров (второй шаг). При этом оценивание осуществляется по индивидам, выбравшим альтернативу j .

По аналогии с (Коссова, Потанин, 2018) можно показать, что состоятельная оценка дисперсии случайной ошибки целевого уравнения имеет вид

$$\hat{\sigma}_j^2 = \sum_{i: z_i=j} (y_i - \hat{y}_i)^2 + \sum_{t_1=1}^{J-1} \hat{\beta}_{\lambda_{t_1}}^2 \tilde{z}_{t_1}^{(ji)} \hat{\lambda}_{t_1 i}^{(j)} / \hat{\Sigma}_{t_1 t_1}^j - \sum_{t_1=1}^{J-1} \hat{\beta}_{\lambda_{t_1}} \sum_{t_2 \neq t_1}^{J-1} \left(\hat{\beta}_{\lambda_{t_2}} - \hat{\Sigma}_{t_1 t_2}^j \hat{\beta}_{\lambda_{t_1}} / \hat{\Sigma}_{t_1 t_1}^j \right) \hat{\lambda}_{t_1 t_2 i}^{(j)} + \left(\sum_{t_1=1}^{J-1} \hat{\beta}_{\lambda_{t_1}} \tilde{z}_{t_1}^{(ji)} \hat{\lambda}_{t_1 i}^{(j)} \right)^2, \quad (10)$$

$$\text{где } \hat{\Lambda}_{t_1 t_2 i}^{(j)} = \frac{\partial^2 F_{ji}(\tilde{z}_1^{(ji)}, \dots, \tilde{z}_{J-1}^{(ji)})}{\partial \tilde{z}_{t_1}^{(ji)} \partial \tilde{z}_{t_2}^{(ji)}} / F_{ji}(\tilde{z}_1^{(ji)}, \dots, \tilde{z}_{J-1}^{(ji)}).$$

Для вывода состоятельной оценки асимптотической ковариационной матрицы оценок регрессионных коэффициентов необходимо учесть гетероскедастичность случайной ошибки v_{ji} и тот факт, что вместо истинных функций $\lambda_{ti}^{(j)}$ используются их оценки, полученные на первом шаге. Состоятельная оценка ковариационной матрицы оценок регрессионных коэффициентов может быть получена по аналогии с (Коссова, Потанин, 2018).

1.3. Частный случай оценивания эффекта воздействия, определяемого эндогенной номинальной переменной

Рассмотрим частный случай модели с мультиномиальным переключением, когда регрессионные уравнения, соответствующие различным альтернативам, отличаются лишь константой. Такая модель может быть использована для вычисления эффекта воздействия фактора,

задаваемого номинальной переменной. Например, при оценивании уравнения Минцера образование, задаваемое номинальной переменной (среднее, среднее специальное, высшее), может входить в уравнение как набор соответствующих фиктивных переменных, являющихся эндогенными по отношению к заработной плате (Potanin, 2019). В этом случае уравнения заработной платы будут отличаться лишь на константу, равную коэффициенту при соответствующей фиктивной переменной.

Итак, предположим, что

$$y_{ji}^* = \beta_j^0 + x_i \beta + \varepsilon_i, \quad \beta_j^0 = 0, \quad (11)$$

где через $\beta_j^0 \in R$ обозначается индивидуальная (для каждого уравнения) константа. Тогда коэффициенты β_j^0 могут быть интерпретированы как условные (на регрессоры x_i) средние эффекты воздействия номинальной переменной z_i при выборе альтернативы j по отношению к контрольной группе индивидов, выбравших альтернативу J .

Процедура оценивания может быть модифицирована таким образом, что на втором шаге вместо оценивания J регрессионных уравнений достаточно оценить лишь одно. Для этого необходимо включить в регрессионные уравнения фиктивные переменные на $J - 1$ альтернативу (кроме J -й), а также переменные, моделирующие условное математическое ожидание случайной ошибки:

$$z_i^{(j)} = \begin{cases} 1, & \text{если } z_i = j, \\ 0, & \text{в противном случае,} \end{cases} \quad \hat{\lambda}_{ti}^{(j)} = \begin{cases} \hat{\lambda}_{ti}^{(j)}, & \text{если } z_i = j, \\ 0, & \text{в противном случае.} \end{cases}$$

В результате регрессионное уравнение принимает вид

$$y_i = \sum_{j=1}^{J-1} z_i^{(j)} \beta_j^0 + x_i' \beta + \sum_{t=1}^{J-1} \sum_{j=1}^J \left(-\tilde{\Sigma}_{tj}^j \right) \hat{\lambda}_{ti}^{(j)} + v_i. \quad (12)$$

Отметим, что в последнем уравнении некоторые коэффициенты при переменных смещения совпадают с точностью до знака. Действительно, обратим внимание, что при $1 \leq j \leq t < J - 1$

$$\begin{aligned} \tilde{\Sigma}_{tj}^j &= \text{Cov}(\tilde{u}_t^{(ji)}, \varepsilon_i) = \text{Cov}(u_{(t+1)i} - u_{ji}, \varepsilon_i) = \\ &= -\text{Cov}(u_{ji} - u_{(t+1)i}, \varepsilon_i) = -\text{Cov}(\tilde{u}_j^{((t+1)i)}, \varepsilon_i) = -\tilde{\Sigma}_{jj}^{(t+1)}. \end{aligned}$$

Аналогично, учитывая, что $u_{ji} = 0$ при $t = J - 1$ и $j \leq J - 1$, получаем:

$$\begin{aligned} \tilde{\Sigma}_{(J-1)J}^j &= \text{Cov}(\tilde{u}_{(J-1)}^{(ji)}, \varepsilon_i) = \text{Cov}(-u_{ji}, \varepsilon_i) = \\ &= -\text{Cov}(u_{ji}, \varepsilon_i) = -\text{Cov}(\tilde{u}_j^{ji}, \varepsilon_i) = -\tilde{\Sigma}_{jj}^J. \end{aligned}$$

Следовательно, если $j \leq t \leq J - 1$, то коэффициенты при оценках обобщенных обратных отношений Миллса $\hat{\lambda}_{ti}^{(j)}$ и $\hat{\lambda}_{ji}^{(t+1)}$ совпадают по абсолютной величине и имеют противоположные знаки. Таким образом, число оцениваемых коэффициентов при переменных смещения равно не $J(J - 1)$, а $J(J - 1)/2$. В результате регрессионное уравнение может быть записано в виде

$$y_i = \sum_{j=1}^{J-1} z_i^{(j)} \beta_j^0 + x_i' \beta + \sum_{t=1}^{J-1} \sum_{j=1}^t (-\tilde{\Sigma}_{tj}^j) \psi_i^{(jt)} + v_i, \quad (13)$$

где

$$\psi_i^{(jt)} = \begin{cases} \hat{\lambda}_{ji}^{(j)}, & \text{если } z_i = j, \\ -\hat{\lambda}_{ji}^{(t+1)}, & \text{если } z_i = t+1, \\ 0, & \text{в противном случае,} \end{cases} \quad (1 \leq j \leq t \leq J-1).$$

1.4. Полупараметрическая модификация регрессии с мультиномиальным гауссовским переключением

Главными достоинствами регрессии с гауссовским переключением являются возможность рассматривать близкие альтернативы, точность получаемых оценок и простота интерпретации оцениваемых параметров. Однако, как и при оценивании любой параметрической модели, возникает вопрос об устойчивости результатов в случае нарушения ее предпосылок.

При нарушении предположения о совместном нормальном распределении случайных ошибок оценки предложенного метода могут оказаться несостоятельными или, по крайней мере, недостаточно точными. В работе (Newey, 2009) для моделей с бинарным механизмом неслучайного отбора была предложена полупараметрическая процедура коррекции смещения отбора, позволяющая получать состоятельные оценки коэффициентов регрессии в том случае, когда распределение ошибок неизвестно. Опираясь на данный подход, авторы предлагают полупараметрическую модификацию описанной в предыдущем разделе двухшаговой процедуры оценивания, позволяющую получить состоятельные оценки коэффициентов основного уравнения в том случае, когда совместное распределение ошибок уравнений альтернатив не является нормальным.

На **первом шаге** используется мультиномиальная пробит-модель. Параметрическая модель вместо непараметрической применяется исследователями на первом шаге тогда, когда непараметрическое или полунепараметрическое оценивание ресурсозатратно или невозможно. На практике, даже обычные бинарные уравнения непараметрически оцениваются чрезвычайно долго. В то же время, в предыдущих исследованиях было обнаружено, что коррекция смещения отбора, основанная на непараметрическом оценивании первого шага, редко дает преимущество по отношению к корректирующим полиномам, полученным с помощью параметрической модели, даже при отклонении совместного распределения от нормального (Pigini, 2015; Коссова и др., 2020).

На **втором шаге** неизвестное условное математическое ожидание случайной ошибки основного уравнения аппроксимируется при помощи полинома от обобщенных обратных отношений Миллса с перекрестными произведениями:

$$E(\varepsilon_{ji} | z_i) = \sum_{k_1=0}^k \sum_{k_2=k_1}^k \sum_{k_3=k_2}^k \dots \sum_{k_{J-1}=k_{J-2}}^k \alpha_{i_1, \dots, i_{J-1}} \lambda_{1i}^{k_1} \times \dots \times \lambda_{(J-1)i}^{k_{J-1}},$$

где $\alpha_{i_1, \dots, i_{J-1}}$ — дополнительные параметры, оцениваемые МНК вместе с коэффициентами регрессии.

Подбор оптимальной степени k можно осуществлять, например, с помощью поэлементной кросс-валидации (LOOCV). Отметим, что во избежание проблемы мультиколлинеарности при большом числе альтернатив J может быть целесообразен отказ от использования перекрестных членов. Также, при наличии достаточных вычислительных ресурсов исследователь может перебрать все возможные комбинации (для всех k из некоторого диапазона) учитываемых перекрестных коэффициентов и степеней. Отказ от учета перекрестных произведений может быть мотивирован тем, что информация о связи между уравнениями отбора частично отражена в оцениваемых на первом шаге корреляциях между случайными ошибками.

Аналогичную полупараметрическую модификацию можно осуществить и для метода Дубина–Макфаддена. В следующем разделе на симулированных данных будут проанализированы свойства оценок предложенной двухшаговой процедуры и ее полупараметрической модификации.

2. Симуляции

При проведении анализа симулированных данных рассматривались следующие вопросы.

1. Сохранится ли работоспособность подхода Дубина–Макфаддена при нарушении предпосылки о независимости от посторонних альтернатив, когда альтернативы тесно связаны? Можно ли получить заметный прирост в точности оценок при использовании предложенного метода в качестве альтернативы классическому при нарушении предположения о независимости от посторонних альтернатив?

2. Являются ли оценки предлагаемого и классического методов устойчивыми к нарушениям предположения о совместном нормальном распределении случайных ошибок? Можно ли в случае нарушения предпосылок о распределении повысить точность оценок за счет использования полупараметрических модификаций?

3. Можно ли достаточно точно скорректировать смещение, возникающее вследствие множественного отбора, при помощи классического метода Хекмана²?

При проведении симуляционного анализа сравнивались точности оценок метода наименьших квадратов (LS), метода Дубина–Макфаддена (DMF) и предлагаемого метода (MNGS, MultiNomial Gaussian Switch). Кроме того, в сравнении участвовал метод Хекмана, основанный на методе максимального правдоподобия (ML) и на двухшаговой процедуре (TS). Наконец, рассматривались предложенные полупараметрические модификации MNGS и DMF методов с перекрестными коэффициентами (MNGS2 и DMF2) и без них (MNGS1 и DMF1).

С целью обеспечения сопоставимости результатов за основу был взят дизайн симуляций, предложенный в (Bourguignon et al., 2007). Предполагается, что $J = 3$, причем $\rho_{12} = 0.1$, $\rho_{13} = 0.45$, $\rho_{23} = -0.35$, $\rho_1^{(1)} = 0.64$, $\rho_2^{(1)} = -0.24$ и $\rho_3^{(1)} = 0.14$, $\sigma_1^2 = 32$, и $\sigma_{u_1} = \sigma_{u_2} = \sigma_{u_3} = 1$. Без потери общности будем оценивать коэффициенты лишь первого уравнения, т. е. β_1 :

$$y_{1i}^* = \beta_{11} + \beta_{12}x_{1i} + \beta_{13}x_{2i} + \varepsilon_{1i}, \quad (14)$$

$$z_{1i}^* = \gamma_{11} + u_{1i}, \quad z_{2i}^* = \gamma_{21} + \gamma_{22}w_{1i} + \gamma_{23}w_{2i} + u_{2i}, \quad z_{3i}^* = \gamma_{31} + \gamma_{32}w_{1i} + \gamma_{33}w_{2i}, \quad (15)$$

² Когда для каждого из J уравнений вместо номинальной переменной z_i неслучайный отбор задает лишь одна бинарная переменная, принимающая значение 1 при $z_i = j$ и 0 в противном случае.

$$\beta_{11} = \beta_{12} = \beta_{13} = 1, \gamma_{11} = 3.3, \gamma_{21} = \gamma_{22} = \gamma_{31} = \gamma_{33} = 1, \gamma_{23} = \gamma_{32} = 0,$$

$$w_{1i}, w_{2i} \sim \mathcal{N}(0, 16), r_{xw} = 0.9,$$

$$x_{1i} = r_{xw} w_{1i} + v_{1i}, \quad x_{2i} = r_{xw} w_{2i} + v_{2i}, \quad v_{1i}, v_{2i} \sim \mathcal{N}(0, (1 - r_{xw}^2)16).$$

В исследовании (Bourguignon et al., 2007) рассматривались малая и большая выборки, включавшие 500 и 5000 наблюдений y_{1i} соответственно. Однако при таком дизайне общий объем выборки варьируется из-за наличия наблюдений, для которых $z_i \neq 1$, что делает процесс симуляции менее удобным с технической точки зрения. В качестве альтернативы рассматриваются выборки объемом $n = 1000$ и $n = 10000$ наблюдений. При этом значение γ_{11} было заменено с 0.9 на 3.3, вследствие чего $z_i = 1$ примерно в половине случаев. Все прочие параметры совпадают с теми, что использовались в (Bourguignon et al., 2007) для спецификации NORMAL(c).

По результатам $m = 1000$ симуляций для каждого метода и каждого коэффициента β_{1l} $l \in \{1, 2, 3\}$, были рассчитаны среднее (Mean), стандартное отклонение (Sd) и среднее абсолютное процентное отклонение (MAPE):

$$\text{Mean}(\beta_{1l}) = \frac{1}{m} \sum_{t=1}^m \hat{\beta}_{1l,t}, \quad \text{Sd}(\beta_{1l}) = \sqrt{\frac{1}{m-1} \sum_{t=1}^m (\hat{\beta}_{1l,t} - \text{Mean}(\beta_{1l}))^2},$$

$$\text{MAPE}(\beta_{1l}) = \frac{1}{m} \sum_{t=1}^m |\hat{\beta}_{1l,t} - \beta_{1l}| / \beta_{1l},$$

где $\hat{\beta}_{1l,t}$ является оценкой коэффициента β_{1l} в t -й симуляции. Также были рассчитаны среднеквадратическое отклонение RMSE и среднее отклонение MAE оценок, однако эти метрики давали результаты, аналогичные MAPE, поэтому для краткости приводим лишь процентные отклонения из-за простоты содержательной интерпретации данной метрики.

Результаты симуляционного анализа, представленные в табл. 1, согласуются с выводом (Bourguignon et al., 2007) об устойчивости оценок DMF метода к нарушению предположения о независимости от посторонних альтернатив. Даже при 1000 наблюдениях DMF-оценки коэффициентов β_{12} и β_{13} в среднем близки к истинным значениям, что свидетельствует в пользу малой величины смещения. Однако, согласно MAPE, оценки DMF метода уступают в точности MNGS-оценкам. Так, например, при 1000 наблюдениях MNGS-оценки коэффициента β_{13} в среднем отклоняются от истинных значений приблизительно на 6.9%, а DMF-оценки — на 7.5%. Заметное различие в точности между оценками этих методов можно наблюдать в отношении константы β_{11} : процентное отклонение DMF-оценок оказалось в 1.3 и 1.7 раз больше, чем у MNGS метода, для 1000 и 10000 наблюдений соответственно³. Оценки остальных методов оказались существенно менее точными, что говорит о нежелательности использования классического метода наименьших квадратов или метода Хекмана при наличии мультиномиального механизма неслучайного отбора. Тем не менее,

³ Заметное смещение в DMF-оценках константы может на практике привести к серьезному смещению при оценивании эффекта воздействия и элементов декомпозиции разности в средних целевой переменной при различных альтернативах (построение аналогов декомпозиции Оахаса–Блиндера).

метод Хекмана продемонстрировал преимущество над МНК, а значит, он может быть использован для уменьшения смещения в оценках в случаях, когда оценивание модели с учетом множественного неслучайного отбора является вычислительно сложным.

Таблица 1. Спецификация с гауссовскими случайными ошибками с невысокими корреляциями

	$\beta_{11} = 1$			$\beta_{12} = 1$			$\beta_{13} = 1$		
	Mean	Sd	MAPE	Mean	Sd	MAPE	Mean	Sd	MAPE
<i>1000 наблюдений</i>									
LS	2.422	0.308	142.2	1.200	0.080	20.04	1.118	0.079	12.35
DMF	0.883	0.560	45.60	0.997	0.100	7.892	0.986	0.094	7.525
DMF1	0.865	0.629	49.94	0.996	0.103	8.147	0.984	0.099	7.896
DMF2	0.942	0.727	57.49	1.001	0.104	8.237	0.988	0.101	8.030
MNGS	1.024	0.422	33.71	1.005	0.092	7.362	1.002	0.087	6.933
MNGS1	1.008	0.474	37.65	1.003	0.096	7.613	1.001	0.089	7.132
MNGS2	1.014	0.436	34.75	1.004	0.093	7.434	1.001	0.087	6.947
ML	0.529	0.632	64.09	1.009	0.096	7.762	0.927	0.096	9.712
TS	0.545	0.701	67.08	1.008	0.102	8.158	0.928	0.100	9.849
<i>10000 наблюдений</i>									
LS	2.426	0.095	142.6	1.198	0.024	19.814	1.118	0.025	11.84
DMF	0.873	0.178	17.77	0.994	0.032	2.579	0.985	0.031	2.752
DMF1	0.896	0.222	19.39	0.995	0.034	2.695	0.987	0.033	2.846
DMF2	0.936	0.223	18.25	0.998	0.034	2.660	0.988	0.033	2.806
MNGS	1.004	0.132	10.40	1.001	0.029	2.332	1.000	0.028	2.262
MNGS1	0.999	0.149	11.70	1.000	0.031	2.427	1.000	0.029	2.336
MNGS2	1.002	0.137	10.83	1.001	0.030	2.355	1.000	0.029	2.280
ML	0.513	0.191	48.78	1.005	0.029	2.333	0.925	0.030	7.495
TS	0.533	0.225	47.11	1.004	0.033	2.600	0.927	0.032	7.295

Далее рассматривался ряд спецификаций с более сильными корреляциями между случайными ошибками уравнений множественного отбора, в предположении, что при более сильном нарушении гипотезы о независимости от посторонних альтернатив преимущество MNGS над DMF может усилиться. Однако во всех случаях полученные результаты оказались схожими с теми, что были описаны ранее. В качестве примера в табл. 2 представлены результаты для $\rho_{12} = 0.5$, $\rho_{13} = 0.6$ и $\rho_{23} = 0.7$.

Далее были проведены эксперименты с иными комбинациями параметров симуляционного анализа, в том числе при варьировании r_{xy} в диапазоне от нуля до единицы, добавлении дополнительных, в том числе бинарных регрессоров, переменных взаимодействия и регрессоров из равномерного распределения, изменении величины коэффициентов γ_j и ряда других преобразований. Однако во всех случаях относительная точность оценок различных методов оказывалась схожей с продемонстрированной на предыдущих симуляциях, из-за чего, для краткости, таблицы с соответствующими результатами опускаются.

Таблица 2. Спецификация с гауссовскими случайными ошибками с высокими корреляциями

	$\beta_{11} = 1$			$\beta_{12} = 1$			$\beta_{13} = 1$		
	Mean	Sd	MAPE	Mean	Sd	MAPE	Mean	Sd	MAPE
<i>1000 наблюдений</i>									
LS	2.419	0.295	141.9	1.207	0.080	20.75	1.114	0.077	11.95
DMF	0.965	0.456	36.29	0.997	0.092	7.357	0.999	0.088	6.948
DMF1	0.964	0.497	39.07	0.998	0.096	7.592	0.997	0.091	7.182
DMF2	0.970	0.557	44.37	1.002	0.096	7.630	0.997	0.092	7.402
MNGS	1.010	0.383	30.38	1.004	0.087	6.979	1.000	0.083	6.631
MNGS1	1.007	0.411	32.35	1.003	0.090	7.141	0.999	0.085	6.808
MNGS2	1.010	0.390	30.88	1.004	0.088	7.041	1.000	0.084	6.696
ML	0.522	0.620	62.77	1.014	0.096	7.729	0.917	0.096	10.25
TS	0.540	0.685	65.74	1.012	0.101	8.118	0.920	0.099	10.27
<i>10000 наблюдений</i>									
LS	2.422	0.095	142.2	1.206	0.024	20.59	1.115	0.026	11.46
DMF	0.961	0.146	12.01	0.994	0.030	2.397	1.000	0.029	2.316
DMF1	0.971	0.172	13.59	0.997	0.032	2.510	0.999	0.031	2.482
DMF2	0.983	0.169	13.38	1.001	0.031	2.428	0.999	0.030	2.411
MNGS	1.004	0.121	9.543	1.001	0.028	2.201	1.000	0.028	2.185
MNGS1	1.001	0.133	10.43	1.000	0.029	2.284	1.000	0.028	2.253
MNGS2	1.003	0.124	9.731	1.001	0.028	2.232	1.000	0.028	2.198
ML	0.515	0.181	48.66	1.011	0.029	2.480	0.917	0.031	8.339
TS	0.538	0.218	46.609	1.010	0.032	2.687	0.920	0.033	8.002

2.1. Устойчивость к различным спецификациям совместного распределения случайных ошибок

На следующем этапе симуляционного анализа была протестирована устойчивость рассмотренных подходов к нарушению предположения о нормальном совместном распределении случайных ошибок. Для этого были рассмотрены случайные ошибки из равномерного распределения, распределения Стьюдента с пятью степенями свободы, экспоненциального распределения и бета-распределения с параметрами $\alpha = \beta = 0.1$, при которых оно является бимодальным. Для случайных ошибок уравнения отбора все эти распределения были стандартизированы, а случайная ошибка основного уравнения была получена умножением стандартизированной ошибки на стандартное отклонение. Совместное распределение случайных ошибок задавалось при помощи копулы Стьюдента с пятью степенями свободы. Прочие параметры симуляций, включая ковариационную матрицу, подаваемую в копулу, оставались прежними.

Согласно таблице 3, при равномерном распределении случайных ошибок как абсолютная, так и относительная точность оценок DMF и MNGS методов остается приблизительно такой же, как и в случае с нормальным распределением. При этом полупараметрические подходы MNGS1, MNGS2, DMF1 и DMF2 демонстрируют несколько меньшую точность даже

Таблица 3. Спецификация с равномерным распределением случайных ошибок

	$\beta_{11} = 1$			$\beta_{12} = 1$			$\beta_{13} = 1$		
	Mean	Sd	MAPE	Mean	Sd	MAPE	Mean	Sd	MAPE
<i>1000 наблюдений</i>									
LS	2.348	0.315	134.8	1.190	0.075	19.04	1.111	0.083	11.84
DMF	0.818	0.567	47.56	0.989	0.098	7.869	0.984	0.100	8.141
DMF1	0.838	0.626	50.85	0.990	0.101	8.081	0.986	0.102	8.321
DMF2	0.856	0.742	59.32	0.991	0.103	8.281	0.985	0.106	8.595
MNGS	1.032	0.431	34.79	1.006	0.090	7.238	1.003	0.092	7.390
MNGS1	1.006	0.479	38.03	1.002	0.094	7.526	1.002	0.095	7.592
MNGS2	1.013	0.440	35.21	1.004	0.091	7.309	1.001	0.093	7.464
ML	0.615	0.611	58.26	1.013	0.092	7.448	0.936	0.098	9.425
TS	0.545	0.720	68.44	1.004	0.102	8.210	0.930	0.105	10.21
<i>10000 наблюдений</i>									
LS	2.359	0.097	135.9	1.189	0.024	18.91	1.111	0.025	11.14
DMF	0.826	0.178	20.54	0.989	0.031	2.559	0.984	0.030	2.736
DMF1	0.872	0.226	20.86	0.991	0.033	2.646	0.987	0.032	2.767
DMF2	0.883	0.218	19.95	0.991	0.032	2.605	0.987	0.032	2.753
MNGS	1.020	0.137	11.01	1.003	0.028	2.265	1.001	0.028	2.246
MNGS1	0.993	0.161	12.61	1.000	0.030	2.367	1.000	0.029	2.313
MNGS2	1.005	0.148	11.68	1.002	0.029	2.279	1.000	0.028	2.264
ML	0.640	0.183	36.27	1.015	0.028	2.581	0.938	0.029	6.211
TS	0.583	0.221	42.11	1.007	0.032	2.597	0.933	0.032	6.762

при большом объеме выборки. Однако средние значения оценок полупараметрических подходов ближе к истинным, что свидетельствует в пользу меньшего смещения в сравнении с оценками параметрических методов. Поэтому различия в точности, вероятно, обусловлены заметно большей дисперсией оценок полупараметрических методов. Так, например, среднее значение MNGS1-оценки коэффициента β_{12} при 10000 наблюдениях совпадает с истинным значением с точностью до тысячных, а средняя MNGS-оценка этого же коэффициента превышает истинное значение на 0.003. Однако разность стандартных отклонений оценок, равная приблизительно 0.02, говорит о большей точности параметрического подхода.

При экспоненциальном распределении случайных ошибок (табл. 4) наиболее точными оказались оценки DMF модели. Причем разрыв в точности MNGS- и DMF-оценок заметно возрос вместе с объемом выборки. Возможно, MNGS метод менее устойчив к асимметрии, поскольку нормальное распределение, в отличие от распределения Гумбеля, лежащего в основе DMF модели, является симметричным. При этом оценки полупараметрической MNGS1 модели также оказались достаточно точными и близкими по MAPE к DMF-оценкам.

При случайных ошибках из распределения Стьюдента (табл. 5) DMF и MNGS подходы, включая их полупараметрические модификации, продемонстрировали примерно одинаковую точность оценок. При этом у большинства методов парадоксальным образом наблюдается рост смещения в оценке константы при увеличении объема выборки. Однако это смещение компенсируется заметным снижением дисперсии, что приводит к повышению точности оценок константы на большой выборке.

Таблица 4. Спецификация с экспоненциальным распределением случайных ошибок

	$\beta_{11} = 1$			$\beta_{12} = 1$			$\beta_{13} = 1$		
	Mean	Sd	MAPE	Mean	Sd	MAPE	Mean	Sd	MAPE
<i>1000 наблюдений</i>									
LS	2.235	0.370	123.5	1.184	0.088	18.52	1.121	0.093	12.96
DMF	0.952	0.586	47.10	0.998	0.103	8.162	0.988	0.105	8.528
DMF1	0.964	0.677	51.76	0.999	0.107	8.499	0.988	0.110	8.880
DMF2	0.816	0.799	64.89	0.988	0.110	8.782	0.980	0.110	8.974
MNGS	0.384	0.518	66.67	0.943	0.098	9.048	0.950	0.102	9.089
MNGS1	0.746	0.621	54.33	0.977	0.106	8.548	0.980	0.106	8.656
MNGS2	0.484	0.574	63.70	0.952	0.103	9.051	0.958	0.103	8.896
ML	-2.381	1.100	343.6	0.714	0.139	29.88	0.656	0.142	35.35
TS	0.127	0.824	98.90	0.967	0.114	9.472	0.908	0.115	11.90
<i>10000 наблюдений</i>									
LS	2.258	0.118	125.8	1.187	0.029	18.73	1.124	0.028	12.43
DMF	0.950	0.200	16.49	0.999	0.034	2.654	0.988	0.034	2.852
DMF1	1.000	0.228	17.70	1.001	0.035	2.789	0.993	0.035	2.860
DMF2	0.815	0.252	25.46	0.989	0.035	2.905	0.981	0.035	3.223
MNGS	0.346	0.158	65.43	0.941	0.032	6.004	0.946	0.032	5.528
MNGS1	0.983	0.249	19.73	1.000	0.036	2.901	0.998	0.036	2.822
MNGS2	0.702	0.317	34.99	0.973	0.040	3.887	0.974	0.039	3.747
ML	-2.638	0.187	363.8	0.688	0.031	31.19	0.629	0.031	37.12
TS	0.126	0.258	87.37	0.968	0.036	3.861	0.908	0.036	9.173

Результаты, наблюдаемые при бета-распределении ошибок (табл. 6), схожи с теми, что были получены в случае с нормальным и равномерным распределениями, что свидетельствует в пользу устойчивости оценок рассматриваемых методов к бимодальной структуре распределения случайных ошибок.

Наконец, в качестве дополнительной проверки устойчивости была исследована спецификация, в которой ошибке каждого из четырех уравнений случайным образом назначалось одно из четырех рассмотренных распределений (без повторений). Данный эксперимент был воспроизведен с использованием как копулы Стьюдента (с прежней ковариационной матрицей), так и копулы Клейтона с параметром, равным единице, при котором корреляции между случайными ошибками варьировались в диапазоне от $1/3$ до $1/2$. При этом случайная ошибка u_{it} была умножена на -1 , поскольку в противном случае корреляции между разностями случайных ошибок уравнения отбора и случайной ошибкой основного уравнения оказывались близки к нулю, что нивелировало смещение отбора. Вследствие указанного преобразования соответствующие корреляции варьировались от $-1/2$ до $-1/3$. Полученные результаты, представленные в таблицах 7 и 8, свидетельствуют о высокой устойчивости рассматриваемых методов к нарушению предпосылки о распределении. При этом независимо от объема выборки наиболее точными, как правило, оказывались оценки MNGS метода. Исключение составляют лишь оценки ML метода при 1000 наблюдениях с копулой Клейтона, которые неожиданно оказались наиболее точными для коэффициентов β_{12} и β_{13} .

Таблица 5. Спецификация со случайными ошибками из распределения Стьюдента

	$\beta_{11} = 1$			$\beta_{12} = 1$			$\beta_{13} = 1$		
	Mean	Sd	MAPE	Mean	Sd	MAPE	Mean	Sd	MAPE
<i>1000 наблюдений</i>									
LS	2.336	0.311	133.6	1.183	0.077	18.37	1.110	0.083	11.76
DMF	1.032	0.523	42.01	1.004	0.092	7.349	0.992	0.096	7.632
DMF1	1.004	0.629	48.10	1.001	0.099	7.842	0.991	0.100	7.947
DMF2	1.023	0.692	54.76	1.005	0.099	7.908	0.990	0.101	8.004
MNGS	0.923	0.450	36.21	0.989	0.088	7.046	0.991	0.093	7.441
MNGS1	1.000	0.473	37.59	0.997	0.091	7.208	0.996	0.094	7.551
MNGS2	0.943	0.453	36.53	0.990	0.090	7.140	0.992	0.093	7.461
ML	0.240	0.773	89.86	0.970	0.108	9.000	0.898	0.113	12.37
TS	0.516	0.724	70.42	0.996	0.101	8.160	0.926	0.104	10.25
<i>10000 наблюдений</i>									
LS	2.351	0.099	135.1	1.184	0.025	18.44	1.112	0.025	11.19
DMF	1.044	0.176	14.58	1.005	0.030	2.448	0.994	0.030	2.430
DMF1	1.033	0.205	16.23	1.003	0.032	2.540	0.993	0.032	2.540
DMF2	1.064	0.224	18.30	1.007	0.032	2.606	0.995	0.031	2.504
MNGS	0.888	0.145	14.93	0.985	0.029	2.576	0.988	0.029	2.486
MNGS1	1.027	0.185	14.99	0.999	0.032	2.511	0.998	0.031	2.426
MNGS2	0.964	0.177	14.63	0.992	0.032	2.569	0.993	0.030	2.413
ML	0.263	0.260	73.68	0.973	0.036	3.525	0.901	0.035	9.924
TS	0.539	0.230	46.36	0.998	0.032	2.574	0.929	0.032	7.138

2.2. Точность оценивания эффекта воздействия эндогенной номинальной переменной

На финальном этапе анализа симулированных данных сравнивалась точность оценок эффекта эндогенного номинального регрессора z_i в уравнении (13), в предположении процесса генерации данных со значениями параметров, аналогичными тем, что использовались в спецификации (15). Таким образом, оценивалось уравнение вида:

$$y_i = 1 + x_{1i} + x_{2i} + \beta_1^0 z_{1i} + \beta_2^0 z_{2i} + \varepsilon_i, \text{ где } \beta_1^0 = -1 \text{ и } \beta_2^0 = 1.$$

Результаты, представленные в табл. 9, говорят о том, что при малом объеме выборки DMF и MNGS подходы существенно превосходят LS. Однако, согласно MAPE, при малом объеме выборки оба подхода дают весьма неточные оценки, примерно в равной степени отклоняющиеся от истинных значений. Тем не менее, при большом объеме выборки точность оценок обоих методов заметно возрастает, при этом оценки MNGS метода оказываются существенно более точными и характеризуются крайне малым смещением.

Таблица 6. Спецификация со случайными ошибками из бета-распределения

	$\beta_{11} = 1$			$\beta_{12} = 1$			$\beta_{13} = 1$		
	Mean	Sd	MAPE	Mean	Sd	MAPE	Mean	Sd	MAPE
<i>1000 наблюдений</i>									
LS	2.112	0.317	111.2	1.157	0.077	15.82	1.091	0.082	10.25
DMF	0.846	0.593	49.13	0.991	0.100	7.966	0.990	0.101	8.217
DMF1	0.863	0.673	53.81	0.991	0.103	8.251	0.991	0.104	8.454
DMF2	0.774	0.810	66.75	0.985	0.106	8.559	0.985	0.108	8.781
MNGS	1.042	0.444	35.05	1.009	0.093	7.411	1.003	0.093	7.464
MNGS1	1.003	0.513	39.76	1.004	0.097	7.632	1.002	0.096	7.761
MNGS2	1.025	0.465	36.13	1.007	0.094	7.482	1.003	0.094	7.567
ML	0.752	0.644	55.20	1.019	0.095	7.682	0.955	0.099	8.741
TS	0.661	0.739	64.79	1.008	0.104	8.366	0.946	0.107	9.573
<i>10000 наблюдений</i>									
LS	2.121	0.097	112.1	1.155	0.024	15.54	1.091	0.025	9.131
DMF	0.846	0.184	19.55	0.989	0.032	2.649	0.989	0.031	2.575
DMF1	0.910	0.279	22.66	0.993	0.036	2.867	0.992	0.034	2.783
DMF2	0.819	0.262	25.78	0.986	0.034	2.947	0.988	0.033	2.819
MNGS	1.042	0.141	11.64	1.006	0.029	2.393	1.002	0.028	2.259
MNGS1	0.983	0.193	15.28	0.999	0.034	2.691	0.999	0.031	2.469
MNGS2	1.011	0.159	12.74	1.003	0.031	2.469	1.002	0.029	2.311
ML	0.751	0.195	26.74	1.017	0.029	2.658	0.953	0.030	4.782
TS	0.689	0.229	32.77	1.009	0.032	2.647	0.947	0.033	5.380

3. Оценивание отдачи от брака среди мужчин

В качестве примера применения модели со множественным переключением рассмотрим задачу оценивания уравнения Минцера. Один из вопросов, поднимаемых в исследованиях, посвященных уравнению Минцера, посвящен изучению «премии» за брак у женатых мужчин. Проверяется гипотеза о том, что женатые мужчины получают, при прочих равных условиях, большую заработную плату, чем неженатые что часто объясняется тем, что брак является неким сигналом работодателю о наличии положительных качеств у работника (McDonald, 2020). Как правило, в этом случае оценивается уравнение заработной платы с учетом неслучайного отбора в статус занятых и фиктивной переменной статуса брака в качестве регрессора (Kossova et al., 2020). Премию за брак отражает коэффициент при соответствующей фиктивной переменной. Однако при таком подходе возникают две проблемы. Первая связана с эндогенностью статуса брака. Есть основания предполагать, что одни и те же ненаблюдаемые исследователем качества индивида влияют как на вероятность брака, так и на величину заработной платы.

Второй вопрос связан с самой переменной статуса брака, поскольку неясно, к какой категории относить мужчин, состоящих в незарегистрированном браке: к холостым или к состоящим в официальном браке. При этом исключение или объединение данной категории

Таблица 7. Спецификация со случайными ошибками из разных распределений и копулой Стьюдента

	$\beta_{11} = 1$			$\beta_{12} = 1$			$\beta_{13} = 1$		
	Mean	Sd	MAPE	Mean	Sd	MAPE	Mean	Sd	MAPE
<i>1000 наблюдений</i>									
LS	0.998	0.776	63.56	1.000	0.125	10.14	0.999	0.120	9.819
DMF	1.018	0.609	47.86	1.004	0.100	8.081	0.996	0.106	8.434
DMF1	1.023	0.662	51.67	1.003	0.102	8.196	0.997	0.109	8.589
DMF2	1.022	0.769	60.70	1.004	0.105	8.475	0.997	0.110	8.783
MNGS	0.934	0.455	36.08	0.998	0.093	7.529	0.991	0.096	7.694
MNGS1	0.972	0.524	40.86	1.001	0.096	7.740	0.994	0.101	8.031
MNGS2	0.941	0.471	37.20	0.999	0.094	7.570	0.991	0.097	7.746
ML	0.734	1.265	80.07	0.974	0.153	11.20	0.973	0.155	11.23
TS	1.016	0.755	60.14	1.002	0.114	9.221	1.001	0.115	9.257
<i>10000 наблюдений</i>									
LS	1.017	0.764	64.93	1.002	0.098	8.616	1.002	0.100	8.772
DMF	1.028	0.243	19.49	1.000	0.035	2.771	1.001	0.033	2.675
DMF1	1.027	0.241	18.73	1.000	0.034	2.719	1.001	0.034	2.755
DMF2	1.050	0.248	20.10	1.002	0.034	2.724	1.003	0.034	2.722
MNGS	0.930	0.160	13.86	0.994	0.030	2.412	0.994	0.030	2.437
MNGS1	1.021	0.190	15.11	1.001	0.031	2.505	1.002	0.031	2.527
MNGS2	0.986	0.195	15.39	0.998	0.031	2.466	0.999	0.031	2.494
ML	0.602	1.187	59.63	0.960	0.128	8.212	0.959	0.129	8.418
TS	0.994	0.315	25.32	1.000	0.057	4.725	1.000	0.060	4.923

с холостыми или состоящими в официальном браке может привести к существенному смещению оценок, о чем свидетельствуют результаты приведенного выше анализа симулированных данных, продемонстрировавшего заметное смещение оценок моделей с бинарным переключением в случае мультиномиального отбора.

Будем рассматривать статус брака с тремя альтернативами, включив в уравнение две фиктивные переменные для находящихся в официальном и незарегистрированном браке, а также учитывая смещение отбора, вызванное выбором соответствующей альтернативы:

$$\ln(y_i) = x_i' \beta + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \varepsilon_i.$$

Будут рассмотрены модель MNGS, обычная линейная регрессионная модель LS, модель Дубина–Макфаддена DMF и двухшаговая процедура оценивания MVGS (MultiVariate Gaussian Switch), предложенная Коссовой, Потаниным (2018) и позволяющая, как и в работе (Kossova et al., 2020), учесть неслучайный отбор не только в число женатых, но и в число занятых. При этом в рамках данного подхода те, кто состоят в незарегистрированном браке, сначала были отнесены к холостым (MVGS0), а затем к женатым (MVGS1). Наконец, для учета неслучайного отбора в число занятых при трех возможных формах семейного статуса используется модификация MNGS модели (2MNGS), в которой добавляется уравнение

Таблица 8. Спецификация со случайными ошибками из разных распределений и копулой Клейтона

	$\beta_{11} = 1$			$\beta_{12} = 1$			$\beta_{13} = 1$		
	Mean	Sd	MAPE	Mean	Sd	MAPE	Mean	Sd	MAPE
<i>1000 наблюдений</i>									
LS	-0.455	0.304	145.5	0.846	0.076	15.55	0.841	0.077	16.03
DMF	1.055	0.636	50.41	0.999	0.103	8.132	0.999	0.103	8.282
DMF1	1.033	0.719	56.04	0.997	0.106	8.401	0.997	0.109	8.702
DMF2	0.930	0.867	68.85	0.990	0.110	8.783	0.991	0.109	8.800
MNGS	1.106	0.491	40.25	1.004	0.095	7.457	1.004	0.096	7.742
MNGS1	1.074	0.572	44.98	1.002	0.100	7.866	1.002	0.100	8.000
MNGS2	1.102	0.510	41.44	1.004	0.097	7.596	1.003	0.097	7.814
ML	1.013	0.682	51.11	0.988	0.094	7.432	0.984	0.097	7.732
TS	1.317	0.782	65.99	1.018	0.102	8.152	1.013	0.105	8.418
<i>10000 наблюдений</i>									
LS	-0.439	0.145	143.9	0.846	0.028	15.44	0.846	0.029	15.37
DMF	1.103	0.209	18.91	1.003	0.032	2.552	1.004	0.034	2.717
DMF1	1.050	0.276	21.69	0.999	0.035	2.794	1.000	0.037	2.978
DMF2	1.011	0.300	23.61	0.996	0.034	2.738	0.997	0.036	2.894
MNGS	1.133	0.172	17.60	1.007	0.030	2.454	1.008	0.032	2.598
MNGS1	1.065	0.216	18.04	1.003	0.032	2.547	1.003	0.035	2.770
MNGS2	1.103	0.193	17.68	1.005	0.031	2.525	1.006	0.033	2.650
ML	1.049	0.276	23.09	0.990	0.034	2.879	0.991	0.036	2.966
TS	1.356	0.253	37.25	1.020	0.033	3.032	1.021	0.034	3.202

занятости⁴ с нормально распределенной случайной ошибкой. При этом для облегчения процедуры оценивания предполагается, что случайная ошибка уравнения занятости не коррелирует со случайными ошибками уравнения мультиномиального отбора⁵.

Оценивание осуществляется на данных РМЭЗ⁶ за 2020 г. (29-я волна). Следуя (Kossova et al., 2020), рассматриваются мужчины в возрасте от 25 до 60 лет, не являющиеся вдовцами. В качестве контрольных переменных в уравнениях (логарифма) зарплаты и семейного статуса (официальный брак — 1, незарегистрированный брак — 2, холостой — 3)

⁴ Его спецификация совпадает со спецификацией уравнения брака за тем исключением, что вместо региональных характеристик семейного статуса используется показатель безработицы по региону. Также учитывается факт наличия детей.

⁵ Нетрудно показать, что в таком случае для учета неслучайного отбора в число занятых в выражение для условного математического ожидания MNGS метода достаточно включить обычную лямбду Хекмана, получаемую из уравнения занятости с помощью пробит-модели.

⁶ Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ (RLMS-HSE)», проводимый Национальным исследовательским университетом «Высшая школа экономики» и ООО «Демоскоп» при участии Центра народонаселения Университета Северной Каролины в Чапел Хилле и Института социологии Федерального научно-исследовательского социологического центра РАН.

Таблица 9. Спецификация с гауссовскими случайными ошибками с невысокой корреляцией и эндогенным номинальным регрессором

	$\beta_1^0 = -1$			$\beta_2^0 = 1$		
	Mean	Sd	MAPE	Mean	Sd	MAPE
<i>1000 наблюдений</i>						
LS	1.170	0.546	217.0	-0.215	0.627	122.8
DMF	-1.101	0.709	57.41	0.640	0.821	71.93
MNGS	-0.982	0.719	57.99	0.947	0.881	70.90
<i>10000 наблюдений</i>						
LS	1.182	0.175	218.2	-0.188	0.204	118.8
DMF	-1.103	0.222	19.54	0.674	0.260	35.43
MNGS	-1.000	0.227	17.75	1.000	0.278	22.00

используются уровень образования (ниже среднего, среднее, среднее специальное, высшее), наличие инвалидности и место проживания (Москва, Санкт-Петербург, региональный центр, город, поселок городского типа или село)⁷. В уравнение зарплаты также был включен стаж индивида, а в уравнение брака — возраст и подсчитанные по регионам доли мужчин, состоящих в официальном и не официальном браке.

Результаты, представленные в табл. 10, свидетельствуют о том, что статус брака является эндогенной переменной, т. к. коэффициенты при переменных, отвечающих за смещение отбора по браку, значимы во всех моделях за исключением MVGS1, где, однако, p -значение оказалось достаточно маленьким (0.1081).

Без учета эндогенности отдача от брака оказывается положительной для мужчин, состоящих как в официальном браке, так и в незарегистрированном. При этом оценки методов, учитывающих эндогенность, оказались достаточно близкими (за исключением модели MVGS1) и показывают наличие *отрицательной* премии за брак для мужчин, состоящих в официальном браке, что согласуется с результатом, полученным ранее (Kossova et al., 2020). Также полученный результат говорит о наличии положительной премии за брак для мужчин, состоящих в незарегистрированном браке. По всей видимости, это отличие связано с разным отношением к риску данных категорий мужчин, которое отражается на процессе поиска работы и на готовности менять место работы.

Объединение категории «незарегистрированный брак» с категорией неженатых в модели MVGS0 и с категорией «официальный брак» в модели MVGS1 дает противоположные по смыслу результаты: положительную премию за брак при объединении с официальным браком и отрицательную премию при объединении с холостыми. Учитывая результаты оценивания моделей MNGS и DMF, можно сделать вывод о том, что категории «незарегистрированный брак» и «официальный брак» не являются близкими, и их объединение является существенной ошибкой спецификации модели.

⁷ Описательные статистики схожи с представленными в (Kossova et al., 2020) и опускаются для краткости.

Таблица 10. Результаты регрессионного анализа

	LS	DMF	MNGS	MVGS1	MVGS0	2MNGS
Константа	10.05*** (0.0481)	10.30*** (0.0839)	10.18*** (0.0973)	10.21*** (0.0787)	10.49*** (0.0937)	10.29*** (0.1021)
Среднее образование	-0.0178 (0.0340)	-0.0180 (0.0351)	-0.0098 (0.0349)	-0.0268 (0.0347)	-0.0361 (0.0355)	-0.0134 (0.0347)
Среднее специальное образование	0.0528 (0.0355)	0.0944** (0.0375)	0.1076** (0.0385)	0.0225 (0.0395)	0.0346 (0.0398)	0.0787** (0.0397)
Высшее образование	0.2090*** (0.0348)	0.3225*** (0.0411)	0.3386*** (0.0441)	0.1797*** (0.0397)	0.2369*** (0.0429)	0.3156*** (0.0444)
Стаж	0.0073** (0.0036)	0.0230*** (0.0047)	0.0213*** (0.0047)	0.0077** (0.0038)	0.0219*** (0.0049)	0.0181*** (0.0050)
Квадрат стажа	-0.0003*** (0.0001)	-0.0005*** (0.0001)	-0.0005*** (0.0001)	-0.0002** (0.0001)	-0.0005*** (0.0001)	-0.0004*** (0.0001)
Инвалидность	-0.1432* (0.0738)	-0.2515*** (0.0827)	-0.2029** (0.0806)	0.0403 (0.1370)	-0.0575 (0.1391)	0.0985 (0.1585)
Москва	0.5881*** (0.0401)	0.4968*** (0.0456)	0.4889*** (0.0470)	0.5737*** (0.0426)	0.5227*** (0.0437)	0.4622*** (0.0459)
Санкт-Петербург	0.3331*** (0.0485)	0.2158*** (0.0460)	0.2031*** (0.0493)	0.2713*** (0.0516)	0.2083*** (0.0543)	0.1019 (0.0634)
Областной центр	0.1823*** (0.0270)	0.1637*** (0.0277)	0.1569*** (0.0277)	0.1326*** (0.0388)	0.1329*** (0.0368)	0.0814** (0.0397)
Город	0.0989*** (0.0273)	0.0618** (0.0284)	0.0615** (0.0287)	0.0413 (0.0408)	0.0257 (0.0396)	-0.0222 (0.0426)
Поселок городского типа	0.1132*** (0.0434)	0.1567*** (0.0467)	0.1629*** (0.0470)	0.0849* (0.0468)	0.1113** (0.0470)	0.1283*** (0.0458)
Официальный брак	0.1691*** (0.0265)	-0.5566*** (0.1312)	-0.4228*** (0.1381)	0.0780*** (0.0278)	-0.5641*** (0.1341)	-0.3884*** (0.1375)
Незарегистрированный брак	0.0843** (0.0347)	0.5562** (0.1980)	0.8420*** (0.2795)			1.0737*** (0.3133)
$\hat{\lambda}_1(\psi^{(1)})$		0.5134*** (0.0927)	7.225*** (1.5121)			8.277*** (1.6270)
$\hat{\lambda}_2(\psi^{(12)})$		-0.2456*** (0.0756)	0.2906*** (0.0709)			-0.2557*** (0.0746)
$\hat{\lambda}_3(\psi^{(22)})$		-0.0531 (0.0492)	-4.462** (1.8630)	0.0247 (0.0253)	0.3989*** (0.0815)	-5.965*** (2.0793)
$\lambda_{занятость}$				-0.2249 (0.1399)	-0.1447 (0.1409)	-0.3398** (0.1516)
R^2	0.1791	0.1937	0.1920	0.1826	0.1889	0.1979
Поэлементная кросс-валидация (RMSE)	0.4396	0.4362	0.4367	0.4392	0.4376	0.4358
Число наблюдений	1958 (официальный брак), 368 (незарегистрированный брак), 532 (холостых)					

Примечание. *** — $p < 0.01$, ** — $p < 0.05$, * — $p < 0.1$, в скобках указаны оценки стандартных ошибок.

Заключение

В данном исследовании предложен новый подход к оцениванию моделей с мультиномиальным эндогенным переключением, задаваемым выбором одной из нескольких неупорядоченных альтернатив. Отличие предложенного подхода от существующих заключается в том, что рассматриваемые альтернативы могут быть близкими, в том смысле, что соответствующие им случайные полезности могут быть зависимыми. Используемая спецификация, основанная на многомерном нормальном распределении, позволяет ослабить предположение о независимости от посторонних альтернатив, которое содержится в классическом подходе Дубина–Макфаддена, однако накладывает ограничение на совместное распределение случайных составляющих в уравнениях. Модель может быть оценена как методом максимального правдоподобия, так и с помощью более простой (с вычислительной точки зрения) двухшаговой процедуры, обобщающей двухшаговую процедуру, предложенную Дж. Хекманом. В работе также предложена полупараметрическая модификация двухшаговой процедуры, позволяющая отказаться от предпосылки о нормальном распределении.

Анализ моделей на симулированных данных с различными распределениями случайных ошибок позволил сделать следующие выводы:

- при выполнении предпосылки о совместном нормальном распределении оценки предложенного подхода и метода Дубина–Макфаддена достаточно близки при любых значениях корреляций между альтернативами, однако предложенный метод более точно оценивает константу, что особенно важно при оценивании эффекта воздействия эндогенной номинальной переменной;
- оценки предложенного метода проигрывают методу Дубина–Макфаддена при несимметричном (экспоненциальном) распределении и дают чуть более точные результаты в остальных случаях;
- полупараметрические модификации методов характеризуются меньшим смещением, но заметно большим разбросом оценок.

На примере оценивания «премии за брак» по данным РМЭЗ за 2020 г. показано, что при наличии эндогенной номинальной переменной необходимо оценивать регрессию с поправкой на смещение отбора, задаваемого мультиномиальной моделью, а попытки сведения множественного выбора к бинарному приводят к серьезному искажению оценок.

Благодарности. Работа выполнена при поддержке Российского Научного Фонда (РНФ), проект 21-18-00427.

Список литературы

Косова Е. В., Потанин Б. С. (2018). Обобщение метода Хекмана и модели с переключением на случай произвольного числа уравнений отбора. *Прикладная эконометрика*, 50, 114–143.

Косова Е. В., Куприянова Л. А., Потанин Б. С. (2020). Сравнение точности оценок параметрических и полупараметрических методов коррекции многомерного смещения отбора. *Прикладная эконометрика*, 57, 119–139. DOI: 10.22394/1993-7601-2020-57-119-139.

Лукиянова А. Л. (2021). Что держит бюджетников на низкооплачиваемых рабочих местах? Роль отбора и некогнитивных факторов в объяснении межсекторных различий в оплате труда. *Прикладная эконометрика*, 62, 32–53. DOI: 10.22394/1993-7601-2021-62-32-53.

Bourguignon F., Fournier M., Gurgand M. (2007). Selection bias correction based on the multinomial logit model: Monte Carlo comparison. *Journal of Economic Surveys*, 21, 174–205. DOI: 10.1111/j.1467-6419.2007.00503.x.

Dubin J. A., McFadden D. L. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 52, 345–362. DOI: 10.2307/1911493.

Heckman J. (1979). Sample selection bias as a specification error. *Econometrica*, 47 (1), 153–161. DOI: 10.2307/1912352.

Kossova E. V., Potanin B. S., Sheluntcova M. A. (2020). Estimating effect of marriage on male wages in Russia. *Journal of Economic Studies*, 47 (7), 1649–1667. DOI: 10.1108/JES-04-2019-0184.

McDonald P. (2020). The male marriage premium: Selection, productivity, or employer preferences? *Journal of Marriage and Family*, 82 (5), 1553–1570. DOI: 10.1111/jomf.12683.

McFadden D. (1973). Conditional logit analysis of qualitative choice behavior. In: Zarembka P. (ed.). *Frontiers in Econometrics*, 105–142. Academic Press, NY.

Murphy K. M., Topel R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, 3, 370–379. DOI: 10.1198/073500102753410417.

Newey W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal*, 12, 217–229. DOI: 10.1111/j.1368-423X.2008.00263.x.

Pigini C. (2015). Bivariate non-normality in the sample selection model. *Journal of Econometric Methods*, 4 (1), 123–144. DOI: 10.1515/jem-2013-0008.

Potanin B. (2019). Estimating the effect of higher education on an employee's wage. *Studies on Russian Economic Development*, 30 (3), 319–326. DOI: 10.1134/S1075700719030146.

Schmertmann C. P. (1994). Selectivity bias correction methods in polychotomous sample selection models. *Journal of Econometrics*, 60, 101–132. DOI: 10.1016/0304-4076(94)90039-6.

Поступила в редакцию 26.06.2022;
принята в печать 22.08.2022.

Kossova E. V., Potanin B. S. Estimation of Gaussian multinomial endogenous switching model. *Applied Econometrics*, 2022, v. 67, pp. 121–143.
DOI: 10.22394/1993-7601-2022-67-121-143

Elena Kossova

National Research University Higher School of Economics (NRU HSE), Moscow, Russian Federation;
ekossova@hse.ru

Bogdan Potanin

National Research University Higher School of Economics (NRU HSE), Moscow, Russian Federation;
bogdanpotanin@gmail.com

Estimation of Gaussian multinomial endogenous switching model

We propose a two-step estimation procedure for multinomial endogenous switching model assuming joint normality of random errors. Simulated data analysis suggests that our method outperforms popular alternatives when independence of irrelevant alternatives (IIA) assumption is violated. We also apply

our method to estimate marriage male premium. We have found statistical evidence that this premium is negative for officially married men and positive for unofficially married men.

Keywords: sample selection; switching regression model; endogeneity.

JEL classification: C34; J12.

Acknowledgments. The work is done under the support of Russian Science Foundation project 21-18-00427.

References

- Bourguignon F., Fournier M., Gurgand M. (2007). Selection bias correction based on the multinomial logit model: Monte Carlo comparison. *Journal of Economic Surveys*, 21, 174–205. DOI: 10.1111/j.1467-6419.2007.00503.x.
- Dubin J. A., McFadden D. L. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 52, 345–362. DOI: 10.2307/1911493.
- Heckman J. (1979). Sample selection bias as a specification error. *Econometrica*, 47 (1), 153–161. DOI: 10.2307/1912352.
- Kossova E. V., Potanin B. S. (2018). Heckman method and switching regression model multivariate generalization. *Applied Econometrics*, 50, 114–143 (in Russian).
- Kossova E. V., Potanin B. S., Sheluntcova M. A. (2020). Estimating effect of marriage on male wages in Russia. *Journal of Economic Studies*, 47 (7), 1649–1667. DOI: 10.1108/JES-04-2019-0184.
- Kossova E. V., Kupriianova L. A., Potanin B. S. (2020). Parametric and semiparametric multivariate sample selection models estimators' accuracy: Comparative analysis on simulated data. *Applied Econometrics*, 57, 119–139 (in Russian). DOI: 10.22394/1993-7601-2020-57-119-139.
- Lukyanova A. L. (2021). What keeps public sector workers in low-paid jobs? The role of self-selection and non-cognitive skills in explaining the public-private wage gap. *Applied Econometrics*, 62, 32–53 (in Russian). DOI: 10.22394/1993-7601-2021-62-32-53.
- McDonald P. (2020). The male marriage premium: Selection, productivity, or employer preferences? *Journal of Marriage and Family*, 82 (5), 1553–1570. DOI: 10.1111/jomf.12683.
- McFadden D. (1973). Conditional logit analysis of qualitative choice behavior. In: Zarembka P. (ed.). *Frontiers in Econometrics*, 105–142. Academic Press, NY.
- Murphy K. M., Topel R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics*, 3, 370–379. DOI: 10.1198/073500102753410417.
- Newey W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal*, 12, 217–229. DOI: 10.1111/j.1368-423X.2008.00263.x.
- Pigini C. (2015). Bivariate non-normality in the sample selection model. *Journal of Econometric Methods*, 4 (1), 123–144. DOI: 10.1515/jem-2013-0008.
- Potanin B. (2019). Estimating the effect of higher education on an employee's wage. *Studies on Russian Economic Development*, 30 (3), 319–326. DOI: 10.1134/S1075700719030146.
- Schmertmann C. P. (1994). Selectivity bias correction methods in polychotomous sample selection models. *Journal of Econometrics*, 60, 101–132. DOI: 10.1016/0304-4076(94)90039-6.

Received 26.06.2022; accepted 22.08.2022.